Universitat Oberta de Catalunya (UOC)

Master Degree - Data Science

Master Thesis

Area 5: Healthcare and Environment

COVID-19: Prediction of infections based on the mobility trends reported by mobile phone networks in Spain

PEC3 - Design and Implementation

Autor: Álvaro Rodríguez Sans

Tutor: Carlos Luis Sánchez Bocanegra

Tutor: Rafael Pastor Vargas

Professor: Albert Solé Ribalta

Delivery PEC3 - 23/05/2021

Copyright

WORK SHEET

Tittle of the master thesis:	COVID-19: Infection prediction based on the mobility reported by mobile phone networks in Spain
Autor's name:	Álvaro Rodríguez Sans
Tutor's name:	Carlos Luis Sánchez Bocanegra
Tutor's name:	Rafael Pastor Vargas
PRA's name:	Albert Solé Ribalta
Delivery date (mm/yyyy):	06/2021
Degree:	Master Degree – Data Science
Area of work:	Area 5 – Healtcare and Environment
Language:	English
Keywords:	COVID-19, Data-Mining, Deep-Learning, Mobility

Abstract

The pandemic caused by the virus called COVID-19, has lead our society to face new paradigms and challenges from healthcare and social life perspectives. Healthcare systems around the world face extreme situations and general population is impacted in their life-style / behaviours (e.g. mobility) since this new virus appears.

Bearing in mind that we live in a world completely connected on a physical and virtual levels, it is important to understand how people's mobility may or not affect a greater spread of this virus. Our society (individuals and public / private organizations) request tools and best practices that help in predicting a potential number or percentage of the population susceptible to being infected by this and other types of virus with similar behaviour. Planning human and material resources of health-care systems has become a priority.

There are a large number of data sources available that can be consumed, crossed and analysed to see the spread of the disease considering different dimensions. In Spain, the National Statistics Institute (INE - by its acronym in Spanish), as well as the different regional governments and other organizations, publish this type of data.

The present study / work seeks to understand if mobility is a relevant element in the spread of this virus in Spain using mobility data in combination with the virus evolution information. Data mining techniques, traditional statistical methods and neural networks are used to confirm this hypothesis and establish a predictive model of COVID-19 based on mobility.

Keywords: COVID-19, Data-Mining, Deep-Learning, Mobility

Resumen

La pandemia provocada por el virus COVID-19 ha llevado a nuestra sociedad a tener que afrontar nuevos paradigmas y retos desde la perspectiva de la salud y en normal desarrollo de la vida cotideanea. Los sistemas de salud de todo el mundo se enfrentan situaciones extremas y la población en general se ve afectada en su estilo de vida / comportamientos (por ejemplo, movilidad) desde que este nuevo virus irrumpio en nuestras vidas.

Teniendo en cuenta que vivimos en un mundo que está completamente conectado tanto a nivel físico como virtual, es importante comprender cómo la movilidad de las personas puede o no afectar a una mayor propagación de este virus. Nuestra sociedad (individuos y organizaciones tanto públicas como privadas) requieren de herramientas y buenas prácticas que ayuden a predecir el número o porcentaje potencial de población susceptible de ser infectada por este y otros tipos de virus con comportamiento similar. La planificación de los recursos humanos y materiales de los sistemas de salud se ha convertido en una prioridad para todos los entes sociales.

Existen una gran cantidad de fuentes de datos disponibles que se pueden consumir, cruzar y analizar para ver la propagación de la enfermedad considerando diferentes dimensiones. En España, el Instituto Nacional de Estadística (INE), así como los diferentes gobiernos autonómicos y otros organismos, publican este tipo de datos.

El presente estudio / trabajo busca comprender si la movilidad es un elemento relevante en la propagación de este virus en España. Se utilizan datos de movilidad en combinación con la evolución del virus. Se utilizan técnicas de minería de datos, métodos estadísticos tradicionales y redes neuronales para confirmar esta hipótesis y establecer un modelo predictivo de COVID-19 basado en la movilidad.

Palabras clave: COVID-19, Minería de Datos, Aprendizaje Profundo, Movilidad

Contents

List of Figures

List of Tables

Chapter 1

Definition and planning

1.1 Description, interest and relevance of the proposal

COVID-19 is an infectious disease caused by a newly discovered coronavirus. In one hand, the majority of people affected by this virus will experience from zero to slight-moderate respiratory symptoms / illness (similar to the seasonal flu) with, apparently, no side effects. On the other hand, people +65 years and those with previous / chronic medical diseases (e.g. diabetes, respiratory disease and / or cancer) it has been revealed as the risk group that can develop a serious illness and lead to death or important sequelae like altered cognition, pulmonary function abnormalities, etc. [8, 51].

The virus spreads primarily through two main ways. First, **droplets** (both large and small) of saliva or discharge from the nose when a person coughs or sneezes. Transmission through airborne of smaller droplets / particles is included into this category due to those small particles have the ability to remain suspended more time and go further compared with droplet transmission (Figure 1.1). Second, contact transmission, when direct contact with an infected person or contaminated surface takes place [34, 35, 36].

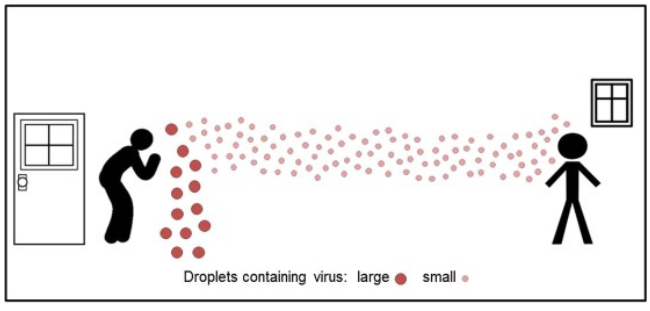


Figure 1.1: Droplets spread. Source: Morawska and Cao [35]

There is a real need to predict the virus spread based on new dimensions. Mobility has been considered as one of the most important due to restrictions adopted by governments around the world. Several data sources offers up-to-date information related to the number of infections, deceased and recovery population in Spain [10, 25], mobility data collected from mobile telephony companies [22] and big-tech providers that publish mobility trends thanks to the massive usage of their mobile applications and devices [3, 19].

The main objective of this work is to offer a series of best practises on how to analyse the importance of population mobility spreading the virus and how this patterns can be predicted thanks to the usage of data-mining and deep-learning techniques [2, 5, 13, 30, 31, 40].

1.2 Objectives

The primary objectives are:

- Understand the infective behaviour of the COVID-19.
- Identify relevant data, provided by mobile telephony providers, to predict the spread of COVID-19 in Spain.
- Use **data-mining and machine learning** techniques to assess the impact / influence of population mobility patterns spreading COVID-19 in Spain.
- Articulate the general best practices to be used in order to afford such scenarios.

The secondary objectives are to apply:

- The necessary **data-mining** techniques to find, create join and / or discard the raw data needed and provided by different organizations (public and private), performing the necessary **exploratory data analysis and transformation** to extract initial insights.
- Different **machine learning** techniques to assess which one produce the better performance / accuracy predictions.

1.3 Methodology to be used

Research into the appropriated academic and scientific databases, journals and books to understand COVID-19 behaviour and how it can be combined with Data-Science techniques in order to predict its spread based on mobility by:

- Looking for the relevant data related to mobility and COVID-19 in Spain (public and private sources of data), to understand its behaviour, how to combine it / use it, etc.
- Analysing / exploring the art state of the data-mining and machine-learning techniques and tools to extract the insights from data. The selection of tools / program languages to elaborate the necessary exploratory data analysis during development phase.
- The usage of **CRISP-DM** [17, 18, 37, 56], **PDSA** [15] and **Agile** [57] methodologies, among others, to meet previous stages from a project management perspective.

Based on the methods stated and thanks to the analysis / review of the models applied, we will confirm the initial hypothesis: "Mobility patterns affect / impact virus spread".

1.4 Planning

The following tables (Table 1.1, 1.2, 1.3, 1.4, 1.5 and 1.6) and Figure 1.2 shows a summary of the initial list of the different activities to be delivered in each PEC. This list will be enriched with further additions or modifications.

Name	Duration	Start	Finish
Pec1 - Definition and planning	12 days	Wed 17/02/21	Sun 28/02/21
- Topic selection and initial research	3 days	Wed 17/02/21	Fri 19/02/21
- Description, interest and relevance of the proposal	2 days	Sat 20/02/21	Sun 21/02/21
- Objectives and personal motivation	2 days	Mon 22/02/21	Tue 23/02/21
- Methodology to be used and planning	2 days	Wed 24/02/21	Thu 25/02/21
- Abstract preparation + Pec1 delivery	3 days	Fri 26/02/21	Sun 28/02/21

Table 1.1: Pec1 plan

Name	Duration	Start	Finish
Pec2 - State of the art / Market analysis of the project	21 days	Mon 01/03/21	Sun 21/03/21
- Research - COVID-19 spread	4 days	Mon 01/03/21	Thu 04/03/21
- Research - Data-Mining techniques (Clean-up, EDA, etc.)	4 days	Fri 05/03/21	Mon 08/03/21
- Research - Machine Learning (forecast time series, neural network, etc.)	3 days	Tue 09/03/21	Thu 11/03/21
- Research / review existing jobs that address the objective	3 days	Fri 12/03/21	Sun 14/03/21
- Incorporate new ideas / refine current approach - Pec2 delivery	7 days	Mon 15/03/21	Sun 21/03/21

Table 1.2: Pec2 plan

Name	Duration	Start	Finish
Pec3 - Design and implementation	63 days	Mon 22/03/21	Sun 23/05/21
- Program tool + lenguage selection (R / Python)	5 days	Tue 23/03/21	Sat 27/03/21
- Data gathering + Clean-up + EDA + Transformations	27 days	Sun 28/03/21	Fri 23/04/21
- Models review	5 days	Sat 24/04/21	Wed 28/04/21
- Models selection	17 days	Thu 29/04/21	Sat 15/05/21
- Prediction analysis	5 days	Sun 16/05/21	Thu 20/05/21
- Review + Pec3 delivery	3 days	Fri 21/05/21	Sun 23/05/21

Table 1.3: Pec3 plan

Name	Duration	Start	Finish
Pec4 - Report writing	14 days	Mon 24/05/21	Sun 06/06/21
- Review documentation generated	4 days	Mon 24/05/21	Thu 27/05/21
- Conclusions from of the results	5 days	Fri 28/05/21	Tue 01/06/21
- Write master thesis (official format)	5 days	Wed 02/06/21	Sun 06/06/21

Table 1.4: Pec4 plan

Name	Duration	Start	Finish
Pec5 - Presentation, exposure and defence of the project	8 days	Mon 07/06/21	Mon 14/06/21
- Generate video + presentation	6 days	Mon 07/06/21	Sat 12/06/21
- Exposure of the project	2 days	Sun 13/06/21	Mon 14/06/21

Table 1.5: Pec5 plan

Name	Duration	Start	Finish
Pec6 - Public exposure	1 day	Sat 19/06/21	Sat 19/06/21
- Public exposure of the project	1 day	Sat 19/06/21	Sat 19/06/21

Table 1.6: Pec6 plan

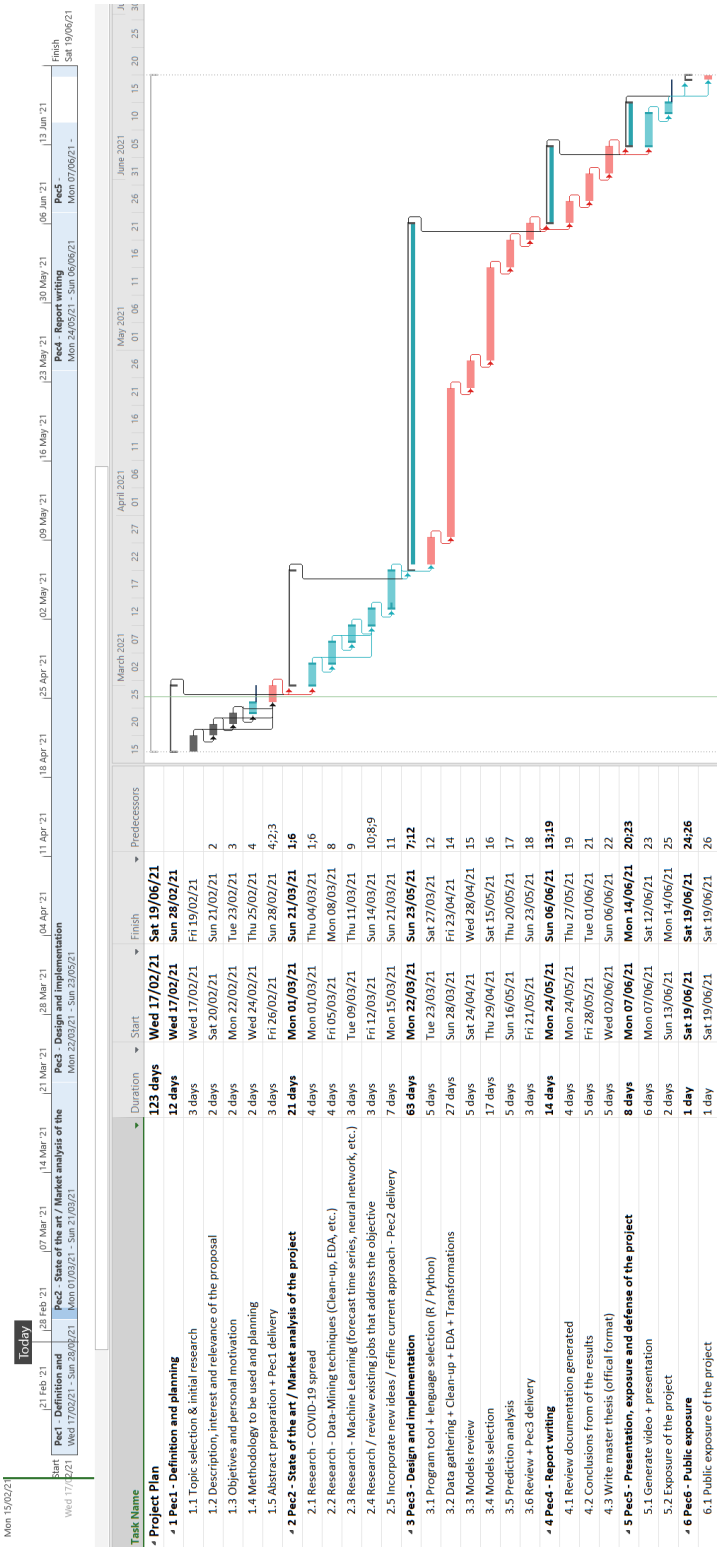


Figure 1.2: Gantt diagram

Chapter 2

State of the art

2.1 Machine learning process - Action plan

When using **Machine Learning**, it is convenient to follow a process to obtain good results, make better use of our time and have guidance on what to do in case our results are not as good as expected. Figure 2.1 and following points describes the different phases of the machine learning process and how they interact each other based on the **Cross Industry Standard Process for Data Mining** (CRISP-DM) standard [17, 18, 37, 56].

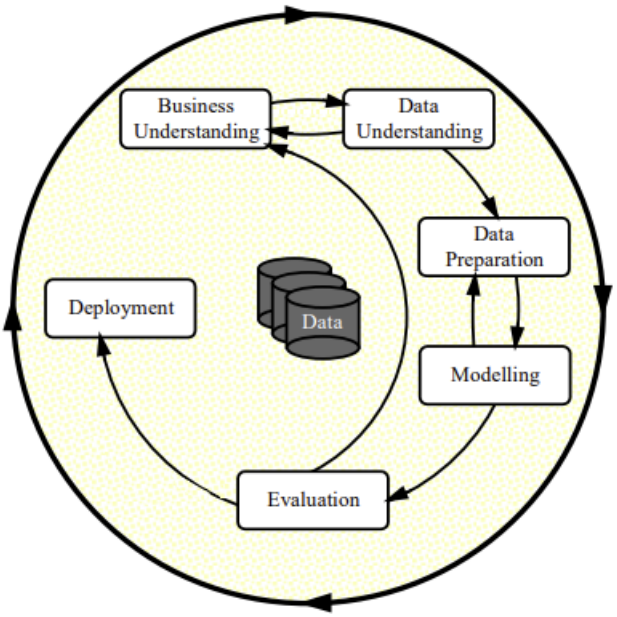


Figure 2.1: CRIPS-DM. Source: Wirth, [56]

2.1.1 Problem understanding

It is very important to understand the problem to solve. Understanding the problem usually takes a long time, especially if the problem comes from an industry or field where the analyst have little knowledge [2, 11, 18, 30, 31].

This task has a medium relative effort due to in this phase the idea is to collaborate with people who know about the problem (SME - Subject Matter Expert).

To get a better understanding of the problem is to ask why? several times until the answer satisfies the question. Knowing the why of things helps to understand the way of thinking / behaviour in that industry or field.

2.1.2 Data understanding

As important as understanding the problem is understanding the data we have available. It is common to do an **exploratory analysis of data** (EDA) to familiarize ourselves with them [2, 11, 18, 30, 31].

In exploratory analysis graphs, correlations and descriptive statistics are usually made to better understand what story the data is telling us. It also helps to estimate if the data we have is sufficient, and relevant, to build a model.

2.1.3 Data preparation

Data preparation is one of the phases in machine learning that requires a lot attention and effort. Main challenges to be addressed are the following [2, 11, 18, 30, 31]:

- **Incomplete data** - It is quite normal not having all the data needed (e.g. if is necessary to predict which customers are more likely to buy a product and the data comes from online surveys, there will be many of the fields needed not filled). In order to deal with incomplete data these are some of the strategies to be adopted:
  - **Eliminate** - Easy option it is just keep the complete data. This may be an option dealing with few incomplete data.
  - **Impute** - Reasonable value will be added / imputed if makes sense (e.g. if someone didn't add his / her age in a survey, the average age of the other participants can be used).

  - **Do nothing** and use some machine learning technique that can handle incomplete data.

- **Combine data from multiple / different sources** - Some data can come from databases, spreadsheets, flat files, etc. It is necessary to combine data to let machine learning algorithms consider / manage all the information.

- **Format data properly** - The usage of machine learning libraries implies to pass the data in a format the can understand. In general, these libraries expect the data to be in the form of a matrix or a tensor. A tensor is a generalization of a matrix. If the matrix has 2 dimensions, the tensor has a number "n" of dimensions.

- **Calculate relevant characteristics (features)** - Machine learning algorithms work much better if they can work only with relevant features instead of the raw data. This phase requires a lot of effort. It is necessary identify which features are going to be relevant to solve the problem and test it.

- **Data normalization** - It is useful to normalize the data to make learning easier for machine learning. Normalizing is the act of putting all the data on a similar / same scale. There are several ways to normalize the data.

2.1.4 Modelling

The phase of building a machine learning model, once we have the data ready, requires less effort. This is due to there are already several machine learning libraries available and a lot of them are free and open source.

During this phase, it is necessary choose the machine learning technique to be used. The machine learning algorithm will automatically learn to obtain the appropriate results with the historical data that has been prepared in previous phases. Of course, it will have an error [2, 11, 18, 30, 31].

2.1.5 Evaluation

The error analysis requires a medium relative effort. Analyzing errors (evaluation) is important to understand what to do to improve machine learning results. In particular the options will be:

- Use another more **complex** model.

- Use a **simpler** one.
- Review if more data and / or more **features** are needed.
- Review if a **better understanding** of the problem and / or the next steps needed during the whole process.

In the evaluation phase we will try to ensure that our model is capable of generalization. Generalization is the ability of machine learning models to produce good results when using new data.

In general, it is not difficult to achieve acceptable results using this process. However, to get really good results over the time, it is mandatory to iterate over the previous phases several times. With each iteration, the understanding of the problem and the data will increase. This allows to design better relevant features and reduce generalization error. A greater understanding will also offers the possibility to choose more precisely the machine learning technique / model that best suits the problem to be addressed [2, 11, 18, 30, 31].

The majority of the situations states that having more data helps to increase the performance of the model. In practice, more data and a simple model tend to perform better than a complex model with less data.

2.1.6 Deployment

Once the evaluation phase is completed and the model error is acceptable, it is a good practice to compare it with the error in the current solution (if is available). If it is better enough, the machine learning model will be deployed / integrated into the current system [49].

The integration phase of a machine learning model into a live system requires a relatively greater effort due to:

- It is needed to **automatically repeat** the data preparation stages. This requires the machine learning model have to communicate with other parts of the system and the results of the model have to be used by the system.
- In addition, **monitoring** the errors of the model and warn if model errors grow over time to rebuild the machine learning model with new data.

A considerable part of the effort it is consumed building the data interfaces necessary that provides data to the model in an automated way and then the system can use its prediction

automatically (Figure 2.2). For machine learning and artificial intelligence to be useful, in most cases they must be integrated into a larger system (e.g. an online translation system would not be very useful if had only built a machine learning model. This would be fine for academic proposes but will not have no commercial value unless it will be implemented. The real value is that the model is integrated into the system and allows multiple end users to use its machine learning model capabilities to translate sentences).

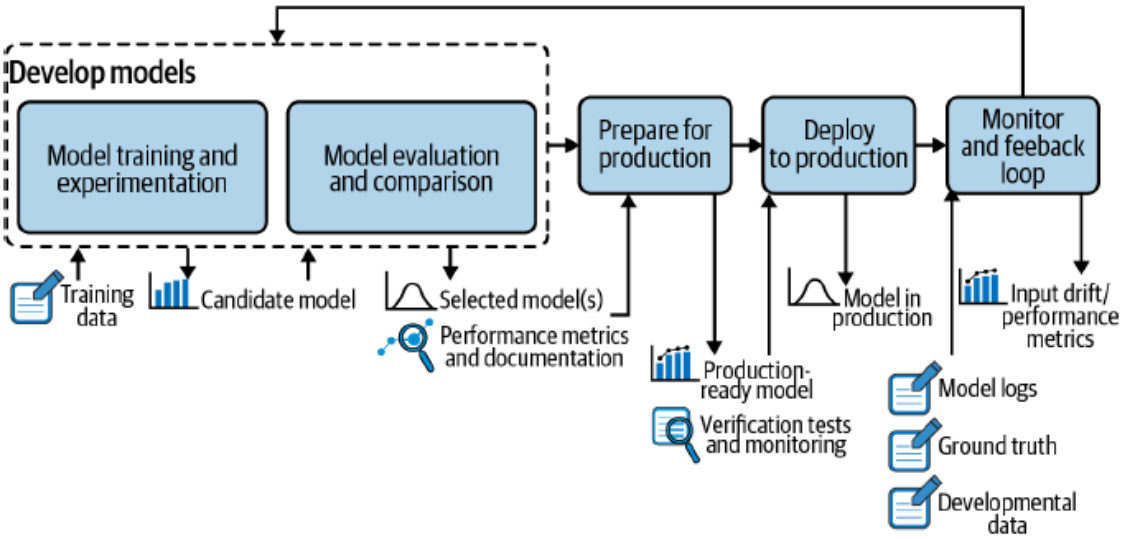


Figure 2.2: End to end process. Source: Treveil, [49]

2.2 COVID-19

2.2.1 Global spread

In early December 2019, local healthcare authorities in the city of **Wuhan** (China), were surprised by a new pneumonia disease of unknown origin. This new pneumonia had a great facility for its spread and consequently was easy to find parallelism with the previous epidemics, like the **Severe Acute Respiratory Syndrome Coronavirus** (SARS-CoV - 2003) [6] and the **Middle East respiratory syndrome** (MERS - 2012) [7]. It is important to remark that the new pneumonia is / was responsible to cause more deaths, although it is a virus with lower lethality rates [46] compare to other similar virus.

It is highly suspected that bats coronaviruses may have led to the evolution of COVID-19 and its introduction into humans [14]. Due to the fast spread of this new virus the **World Health Organization** (WHO) started to activate the different protocols available with the

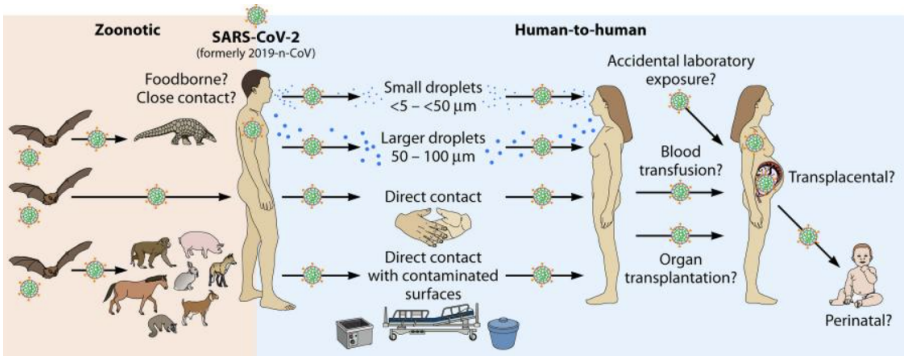


Figure 2.3: Covid-19 transmission. Source: Dhama et al., [14]

aim to contain its expansion in January 2020. During these days the number of new cases reported increases alarmingly [54]. **In March 2020 WHO** declare the global pandemic, reported a grand total of **132,000 cases** of COVID-19, from **123 countries** and territories. **In April 2020, 1.000.000 million of new cases** were reported and more than **50.000 deaths**. Unfortunately these figures experienced a grow until a grand total in **March 2021 of confirmed 115,653,459 cases and 2,571,823 of deaths** [55] around the world.

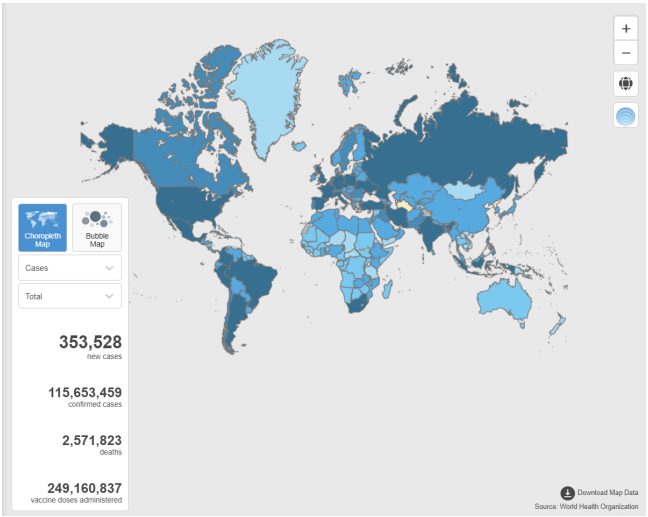


Figure 2.4: Covid-19 transmission march 2021. Source: WHO [55]

Along with the rapid spread of the virus worldwide, new variants have emerged (UK, South African, etc.). Even it has not been demonstrated these new variants present relevant higher levels of contagion, greater impact on the health and recovery of infected patients or more mortality rates, healthcare authorities consider that more investigation should be done [32].

To avoid these new variants adds more uncertainty to the current global / local situations, healthcare authorities recommends to all countries and parties involved in control and promote

local and international travels, to provide the necessary information related to the current status and the measures to follow in order to kept population safe [53].

2.2.2 Spain - Spread and lessons learned

The first case of COVID-19 in Spain was detected on the island of La Gomera - Canarias (January 2020), and it was considered as an imported case from Germany. In February 2020, the first cases were reported in mainland Spain. In May 2020, a grand total of 237,906 confirmed cases were reported and the virus was responsible of 27,119 deaths [46]. Figures in Spain experienced a grow until a grand total in March 2021 of confirmed 3,142,358 cases and 70,501 of deaths [55].

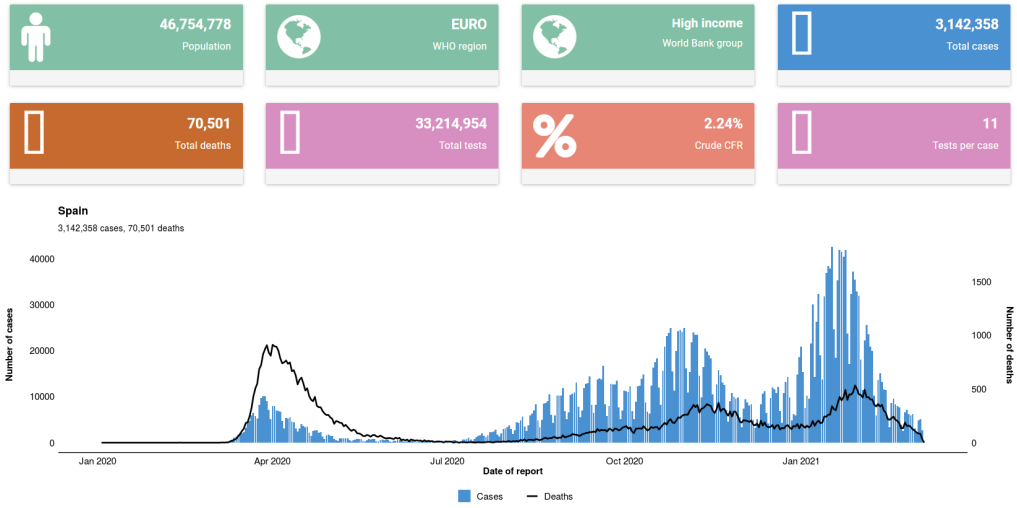


Figure 2.5: Covid-19 evolution until march 2021 Spain. Source: WHO [55]

It is necessary to take into consideration that until all the vaccines [52] will not be available for distribution worldwide, measures that helps to find a balance between daily behaviour and the maintenance of adequate self-protection measures that offers a rapid response to any new outbreak should be kept (usage of masks, social distance, etc.) [46].

A recent study [33], **focused on mobility trends in Spain, reveals that a proper assessment of mobility is crucial in order to understand the effects of measures** (quarantines, selective re-openings, etc.) containing the virus. High levels of mobility contribute to virus spread, where "multi-seeding" (consider as independent infected individuals arriving at a new region or city) can be consider a potential to boost new local infections, and could be impact negatively over the effectiveness of tracing measures. In Spain, multi-seeding, could be considered as a key player spreading COVID-19.

The study indicates that peaks in number of infections and mortality rates are highly correlated with mobility (from and to) occurred in Madrid, city consider the "hub" in Spain due to, as capital, attracts workers, students, visitors all around the country and daily commuters from neighbour provinces / cities. [33].

Figures 2.6 and 2.7 offers a graphical representation of the conclusions extracted by this study.

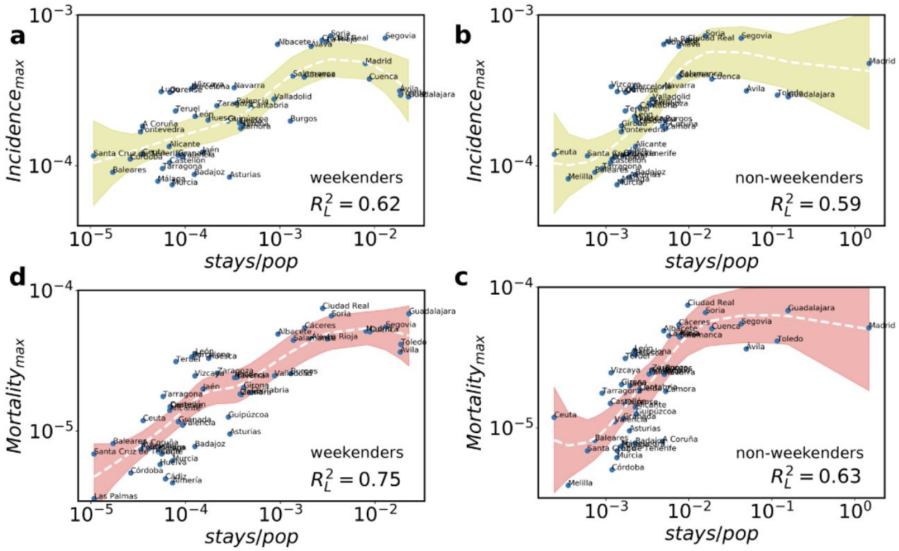


Figure 2.6: Mobility madrid (from - to). Source: Mazzoli et al., [33]

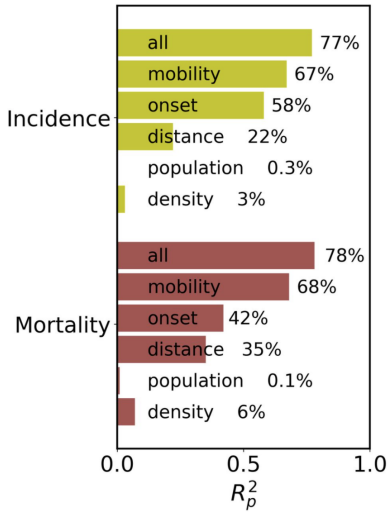


Figure 2.7: Multivariate analysis. Source: Mazzoli et al., [33]

2.2.3 Spread study based on mobility

In USA a study performed in 2020, based on daily mobility data (aggregated and anonymised) from January 2020 to April 2020, was used to retrieve trends in movement patterns for each US state under observation. Social distancing metrics were generated in combination with epidemiological data in order to compute COVID-19 growth rate, for a given state on a given day. The study was able to evaluate how social distancing (measured by relative change in mobility patterns), impacted the rate of new infections in all USA states under evaluation [4].

The **analysis concluded that mobility patterns have a strong correlation with the decreased of COVID-19 infections for the states with more incidence**. The effects of changes in mobility, are not visible until 9 to 12 days, which is consider the time needed to incubate and report the virus infection to the healthcare system [4].

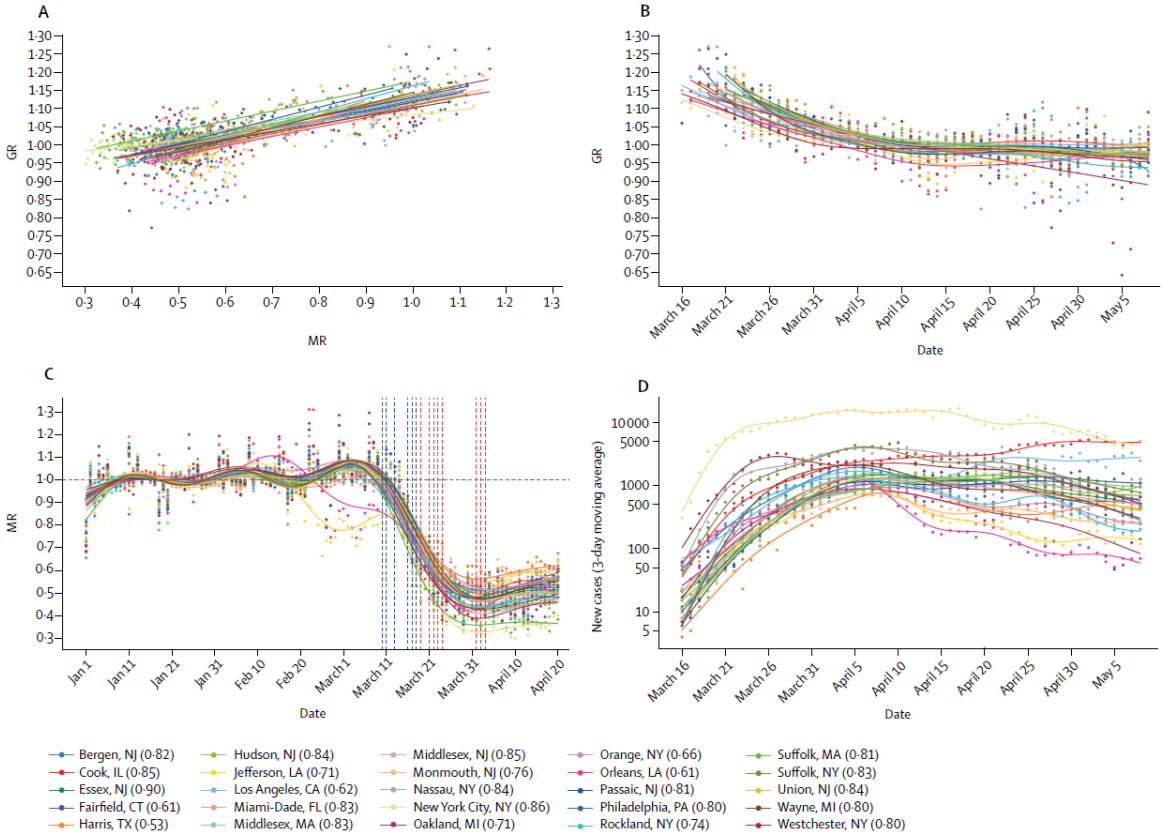


Figure 2.8: Correlation (mobility - virus spread). Source: The Lancet [4]

One of the key elements this study offer, is the inclusion of a quantitative dimension for analysis adding new metrics as MR (mobility ratio) and GR (growth ratio of the COVID-19). For MR (formula below), it was considered the difference between trips / movements (from - to and into the state) compared with a base reference date (before the virus outbreak) by a

given date. Here, i and j are the states, V is the number of commutes / trips, t is the given date and $t0$ is the reference date [4].

$$MR_j^t = \frac{\sum_{i \neq j} V_{ij}^t + \sum_{i \neq j} V_{ji}^t + V_{jj}^t}{\sum_{i \neq j} V_{ij}^{t0} + \sum_{i \neq j} V_{ji}^{t0} + V_{jj}^{t0}}$$

For the GR (formula below), it was considered the difference of cases reported in a period of 3 days compared with the cases reported in a period of 7 days. Here, j is the state, t is the given date, i is the number of cases reported, V is the number of trips / movements and $t0$ is the reference date [4].

$$GR_j^t = \frac{\log\left(\sum_{i-3}^{t} \frac{C_j^t}{3}\right)}{\log\left(\sum_{i-7}^{t} \frac{C_j^t}{7}\right)}$$

Thanks to the introduction of this approach and formulas, it was possible to demonstrate that there is a strong correlation between mobility and virus spread (Figure 2.10). Other studies in USA have adopted this approach trying to find out new indexes, like SDI (Social Distance Index), where mobility is also the foundation to discover correlations between mobility and virus spread. Here, five dimensions (*X1* - Percentage of residents staying home, *X2* - Daily work trips per person, *X3* - Daily non-work trips per person, *X4* - Distances travelled per person and *X5* - Out-of-county trips -in thousands-) were combined in a formula that states the level of social distance achieve by population in percentage [38].

$$SDI = [\beta 1 * X1 + 0.01 * (100 - X1) * (\beta 2 * X2 + \beta 3 * X3 + \beta 4 * X4)] * (1 - \beta 5) + \beta 5 * X5$$

In summary, according the authors, the formula above, was divided into two main blocks, the first ones it is focused on residential and inside trips and the second one on out of county trips. For the first one, it has been used the percentage of residents staying home (residents - trips less than 1.61km from home) so the weight is one ($\beta 1 = 1$). For the ones not staying at home, the percentage is 100 minus X1. For the individuals with more work and non-work trips and longer distances, the assumption is that the are practising less social distancing. Then the weights for each variable should sum up to one ($\beta 2 + \beta 3 + \beta 4 = 1$). That it is meant the resident travellers are comparable to residents staying at home [38].

The assignment of the appropriate weights to each variable, were based on actual observations and conceptual guidelines offered by federal agencies. The relative ratio between resident trips and out-of-county trips was four to one, it was assign a weight of 0.2 to $\beta 5$ [38].

2.3 Mobility trends - Spain (INE / Google)

In 2019, the Statistical National Institute (INE - by its acronym in Spanish), carried out an initial mobility study based on mobile telephony and due to the outbreak of COVID-19 (March 2020 in Spain), the study was extended to measure mobility during the alarm state and onwards [22, 23, 25].

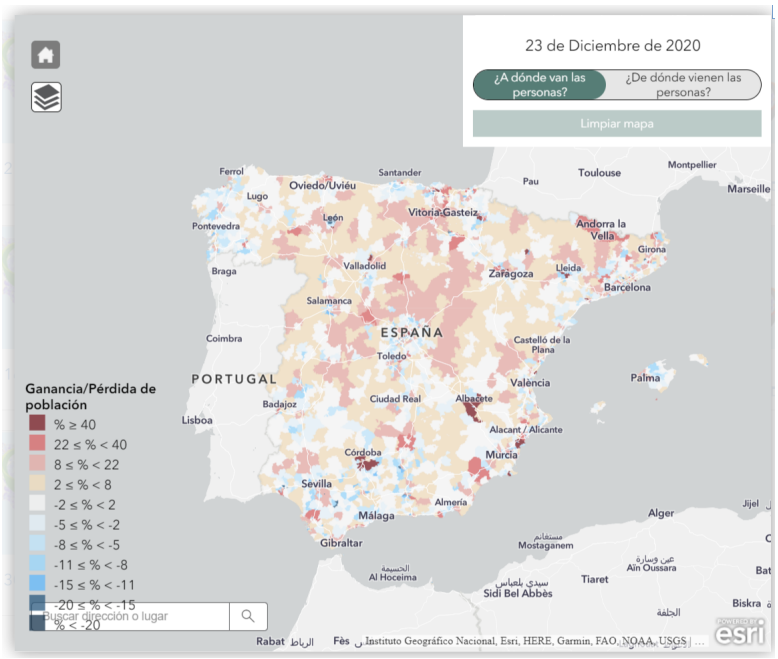


Figure 2.9: Spain mobility maps. Source: INE [22]

The study is based on mobile phones part of the national telephone numbering plan in Sapin (foreign phone lines are excluded). Data obtained is aggregated to population totals. The geographical scope is the entire national territory which is divided into 3,214 specific "mobility areas", each one consisting of a minimum of 5,000 inhabitants and an average of almost 15,000 inhabitants (Figure 2.8). The "mobility area" is a more homogeneous unit than the municipality, but less detailed than the coverage area of each antenna. For daily mobility data, the most frequent position of mobile phones at night (from 10 p.m. to 6 a.m.) is analysed for a given day, to determine the area of residence, and compared with the position in the schedule from 10:00 to 16:00, which determines the destination area [22, 23, 25].

The study observed that in 2019, weekdays (monday to friday) 30% of the population left their area of residence during the central hours (probably due to the need to go to work or study). The areas that received the most population on a daily basis in November 2019 were Madrid and Barcelona due to they are consider as city hubs in Spain [22, 23, 25]. The figures

observed in 2019 can be used as a reference for further comparisons due to the absence of mobility restrictions.

The analysis of 2020 figures, confirms the percentage of population that left their regular area of residence on weekdays during central hours, experienced a high reduction during the second half of the year compared with 2019 figures. The percentage observed was between 15% and 20%, compared to levels close to 30% in a "normal" week of 2019 (18 to 21 November 2019 was the reference week for this study)[22, 23, 25].

A similar approach was taken by Google, where local mobility reports are broken down by location and shows the number of visits to supermarkets, parks, etc. (Figure 2.10). The information in these reports is generated from aggregated and anonymized data sets, where users accept to have activated its location history. Anonymization technology to protect privacy and security is implemented, allowing Google and INE co-create valuable information while preserving the anonymity of any specific person [19, 23].

Figure 2.10: Google mobility trends. Source: Google [19]

The datasets provided by INE and Google are used in this work [19, 22, 23, 25].

2.4 COVID-19 - Machine and deep learning techniques

Multiple researchers and scientific teams related to statistics and machine learning are working in the development of solutions that helps to understand COVID-19 behaviour, patterns, char-

acteristics, future spread, and how the data available can help to implement the right decisions. There are several machine and deep learning techniques that can used in order to estimate the number of new possible inflections based on epidemic aspects (mortality, recovers, etc.) and social aspects (mobility in our case) [1].

The following subsections are going to define general concepts and techniques for better understanding.

2.4.1 Autoregressive Integrated Moving Average (ARIMA)

ARIMA (Autoregressive Moving Integrated Average) models are widely used to predict future values of a time-series. This is a statistical model for time series that take into account the dependence between the data, which is considered each observation at a given moment, and its modelling based on the previous values [11, 21, 28]. The relevant elements in ARIMA are:

- **Autoregressive Processes - AR (p)** - Autoregressive models are based on the idea the current value of the series, Xt, can be explained or predicted based on p past values $Xt-1$, ... $Xt-p$ plus an error term, where p determines the number of past values needed to forecast a current value.

- **Moving Average Processes - MA (q)** - A moving average model is one that explains the value of a certain variable in a period t as a function of an independent term and a succession of errors corresponding to preceding periods, appropriately weighted. These models are usually denoted by the initials MA, followed by the order in parentheses. All moving average processes are stationary processes but not all moving average processes are invertible.

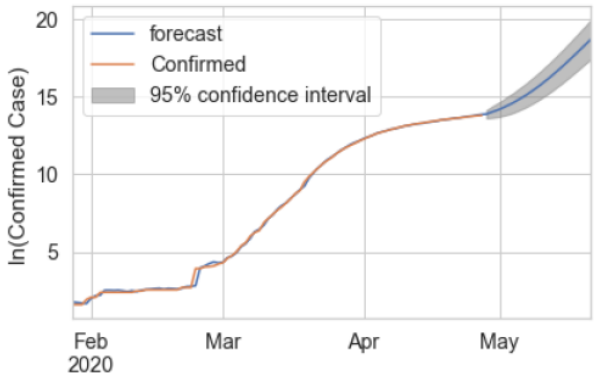
- **Autoregressive Process of Moving Averages - ARMA (p,q)** - A natural extension of the AR (p) and MA (q) models is a type of model that includes both autoregressive and moving average terms. The autoregressive models of moving averages, ARMA (p, q), are the sum of an autoregressive process of order p and one of moving averages of order q. It is very likely that a time series has characteristics of AR and MA at the same time and, therefore, is ARMA. An ARMA (p, q) process is stationary if its autoregressive component is stationary, and it is invertible if its moving average component is.

- **Integrated Process - I (d)** - Not all time series are stationary, some of them shows changes over time or the variance is not constant, so the series is differentiated d times to make it stationary. These types of processes are considered integrated processes, and an

ARMA (p, q) model can be applied to this differentiated series to give rise to an ARIMA (p, d, q) model.

In summary, ARIMA is an integrated autoregressive time series of moving average, where p denotes the number of autoregressive terms, d the number of times the series must be differentiated to make it stationary and q the number of terms of the invertible moving average [11, 21, 28].

A study performed to forecast COVID-19, based on time series in several countries [29], used ARIMA and its results stated that, compared with other similar models (like Facebook Prophet [48]), ARIMA was performing better when analysing the **Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Root Relative Squared Error (RRSE), and Mean Absolute Percentage Error (MAPE) error levels**. The dimensions forecasted were recoveries, deaths, confirmed and active cases of COVID-19 (Figure 2.11 - Confirmed cases).



(a) ARIMA Forecasting for US confirmed cases

Figure 2.11: ARIMA forecast. Source: Taylor and Letham [48]

In order to deal with multiple variables, ARIMA should be adopt the form of a VAR (Vector Autoregressive Model) for multivariate time series data.

2.4.2 Long-Short Term Memory (LSTM)

The **Artificial Neuron** (AN) is considered the minimum unit of calculation able to simulate the behaviour of a biological neuron and it is considered the basics of constitutes artificial neural networks (Figures 2.12, 2.13). The **Artificial Neural Networks** (ANN) are multiple AN connected each other with the aim to transmit signals / information that crosses over the

ANN. During these movements a series of mathematical operations happens and produces a series of output values.

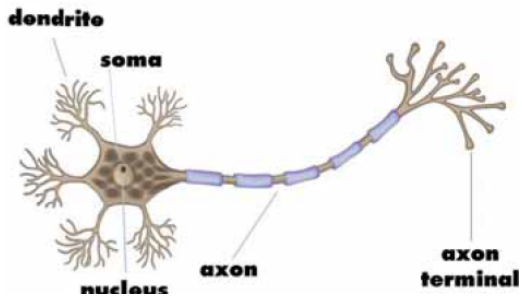


Figure 2.12: Neuron. Source: Ciaburro and Venkateswaran [9]

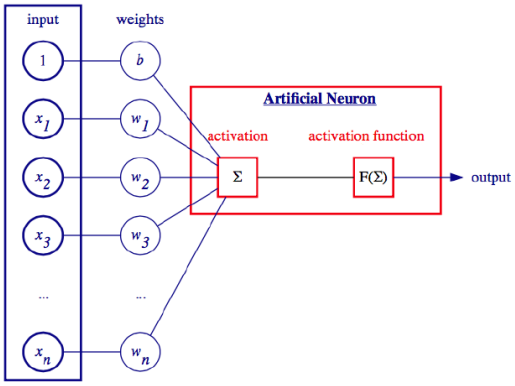


Figure 2.13: Artifical neuron. Source: Vasilev et al., [50]

As an example (Figure 2.14), we can observe the different connection possibilities. Here, we have one input layer, two hidden layers and one output Layer). The input layers receive the initial data, the hidden layers perform the mathematical calculations, and the output layer returns the prediction.

An **LSTM** it is a type of Recurrent Neural Network (RNN), with the ability to reuse valuable information (recent or not form other neurons) as input. RNNs are not able to learn from long-term memory dependencies, here LTSM solves this issue due to its ability to avoid the vanishing gradient problem from RNNs. The LSTM approach to face the vanishing gradients is via the introduction of gating mechanism able to remove or add information to cell state that control the information moved through them [20, 39, 50].

The gate system provided by LSTM (Figure 2.15) it is composed by the input state, output, and forget gates [50] as follows:

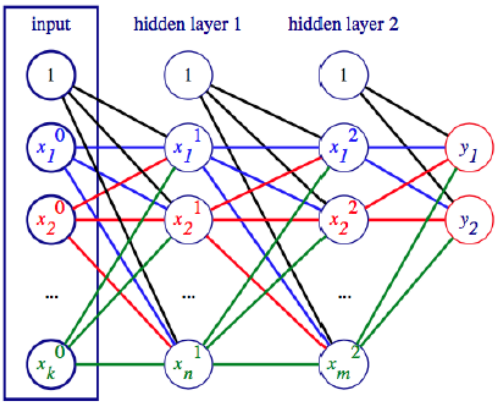


Figure 2.14: Artifical neuron multilayer. Source: Vasilev et al., [50]

- **Input state (write)** - Define the amount of information of the newly computed state that will be moved to the succeeding states for the current input, x_t.
- **Forget gate (reset)** - Control the amount of information from the previous state that will be moved to the next cell, c_t.
- **Output gate (read)** - Define the amount of internal state information that will be moved to the next state, h_t.

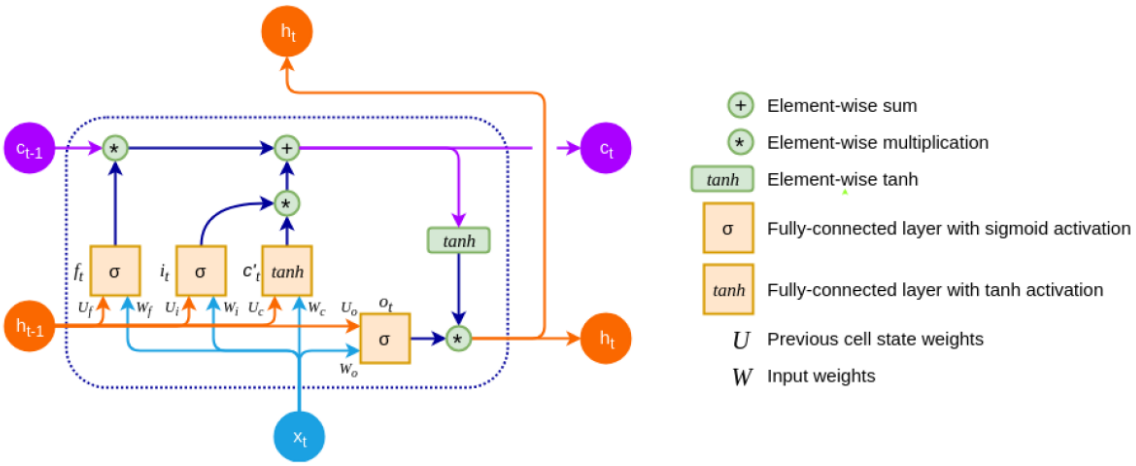


Figure 2.15: LSTM. Source: Vasilev et al., [50]

The figure above represents x_t as the LSTM input, c_t as the cell memory state and h_t as the output / hidden state in a given moment t. Here, x_t and previous h_{t-1} are connected to each gate. Candidate cell vector sets weights W and U. c_t is the cell state at moment t. f_t, i_t, and o_t are the forget, input, and output gates of the LSTM cell [50].

The first step in building an LSTM network is to identify the data / information that is not required and will be omitted / removed, forget gate. This identification is decided by the sigmoid function σ, which takes the output from the last LSTM unit h_{t-1} at time $t-1$ and the current input x_t at time t. The sigmoid function determines how much of the above output should be removed. This door is a vector with values ranging from 0 to 1, corresponding to each number in the cell state c_{t-1}. 0 value remove c_{t-1} the cell block and a value of 1 move on the information.

$$f_t = \sigma(W_f x_t + U_f h_{t-1})$$

The second step, input gate, decides the new information will be added to the memory cell based on h_{t-1} and x_t. Like in forget gate, this door is a vector with values from 0 or 1 for each cell. 0 means no information will be added to the cell block's memory.

$$i_t = \sigma(W_i x_t + U_i h_{t-1})$$

Then the a candidate input is added with a tangential, tanh, function that gives weight to the passed values, deciding their level of importance (-1 to 1).

$$c_t^{'} = \tanh(W_c x_t + U_c h_{t-1})$$

And now the forget and input gates the new cell state by selecting the old and new parts.

$$c_t = f_t * c_{t-1} + i_t * c_t^{'}$$

The third step, output gate, the output values (h_t) are based on the state of the output cell (o_t) but it is a filtered version. Here, a sigmoid layer decides which parts of the cell state get to the output.

$$o_t = \sigma(W_o x_t + U_o h_{t-1})$$

Finally, the output of the sigmoid gate is then multiplied by the new values created by the tanh layer from the cell state (c_t), with a value ranging from -1 to 1.

$$h_t = o_t * \tanh(c_t)$$

2.4.3 Studies carried-out

Several studies have been adopted the LSTM to forecast COVID-19 comparing with other models (ARIMA, etc.). The outcome from these studies is that LSTM can be consider a robust model forecasting COVID-19 due to the MAE, RSME, etc. offered [26, 27, 42, 43, 47].

One of these studies was focused to test the performance LSTM, RNN, and Gated Recurrent Units (GRU) to predict number of COVID-19 infections during a period of 10 days based on data published by WHO (World Health Organization). The outcome states that these models were able to obtain precisions rates close or greater than 90%. Countries under study were **Pakistan, India, Afghanistan, and Bangladesh** (Figure 2.16) [43].

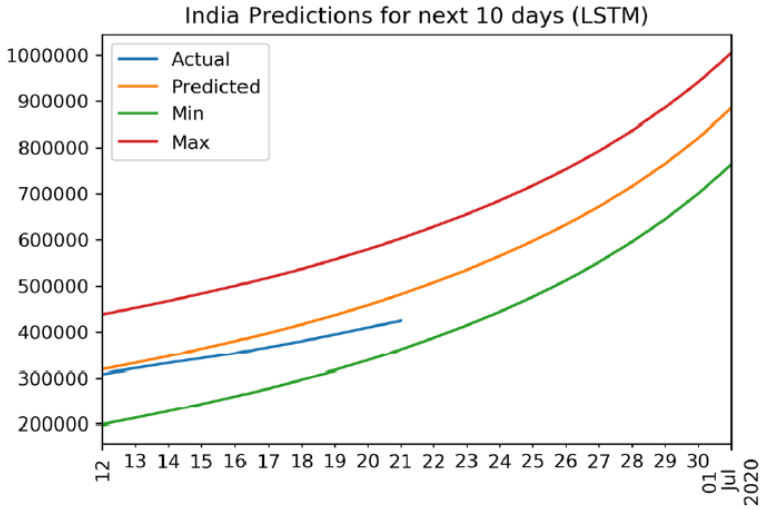


Figure 2.16: LSTM India prediction. Source: Rauf et al., [43]

A study carried-out in the city of Chennai (India) was able to forecast COVID-19 death and recovered cases using a Stacked LSTM model [12]. Figure 2.17 shows the performance of the Stacked LSTM model compared with others in different based on MAPE.

Other of these studies, with a global geographical scope [47], where ten countries were under observation. Again it was observed that the LSTM models were performing better compared with Support Vector Regression (SVR) and ARIMA. The outcomes of this study states clearly that ARIMA and SVR were not able to predict (Figure 2.18).

The information stated in these studies and the outcomes offered, leads us to adopt a similar approach in order to create a model that predict COVID-19 based on mobility dimension. We are going to compare ARIMA and LSTM.

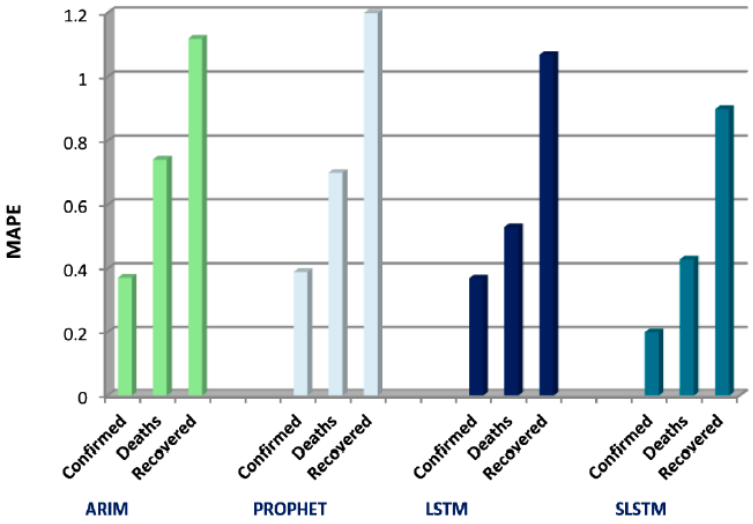


Figure 2.17: SLSTM India MAPE. Source: Devaraj et al., [12]

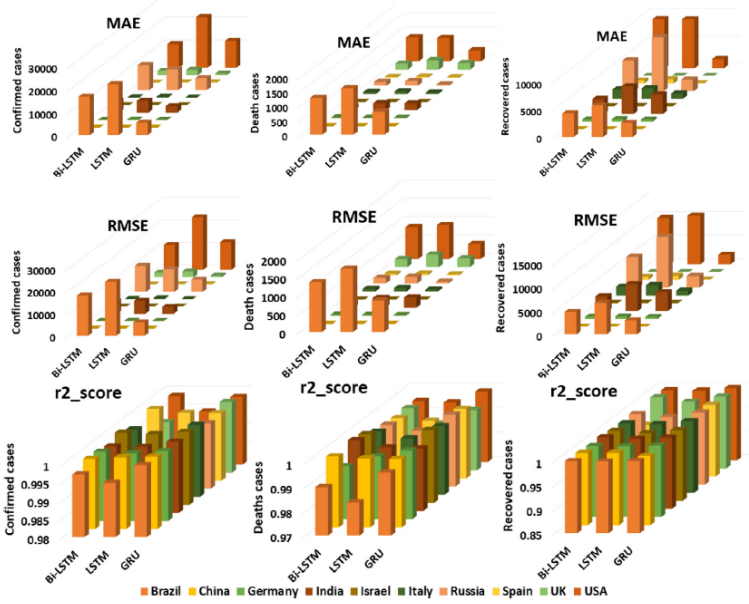


Figure 2.18: LSTM comparative - 2. Source: Shahid, Zameer and Muneeb, [47]

2.5 Datasets to be used

As it has been stated in previous sections, this study will use the data offered by CNE [10] Google [19] and INE [22]. CRISP-DM will be the foundation for data preparation, analysis and model selection.

- **CNE** - This dataset offers information related to infections, recoveries and deaths reported by local and regional governments in Spain.

- **Google** - This dataset offers information related to mobility using Google application services.
- **INE** - These datasets (there are several) offers the mobility of the population captured by the mobile telephony infrastructure system in Spain.

Due to all these datasets are compressed and offers a huge quantity of records, it will be necessary to perform a lot of previous data preparation activities as was mention when CRISP-DM was explained.

2.6 Data-science IDE and language to be used

R [41] is a high-level program language and environment for data analysis and graphics to perform statistical tasks. It is free and can be downloaded from the project site (CRAN - Comprehensive R Archive Network).

RStudio [45] is an integrated development environment (IDE) for the R programming (statistical computing language). It origins are based to statistical computing and graphics. It is open-source, includes a console syntax editor that supports code execution, as well as tools for plotting, debugging, and managing the workspace. It is available for **Windows, Mac, and Linux.**

This is the software to be used for this study due to is widely used in industry, is open-source and there are available a lot of predictive, explore and analysis libraries ready for use like **Tensorflow, Tidymodels, Tidyverse, ggplot2, dplyr, tidyr, etc.**

Chapter 3

Methodology

3.1 Steps followed

As has been stated at **"State of the art"** section and to establish a proper order in the elaboration of the work, **KDD** methodology (Knowledge Discovery in Databases) and **CRISP-DM** (Cross Industry Standard Process for Data Mining) were used, covering all the tasks and phases for the project [16, 17, 18, 37, 56] (Figure 3.1).

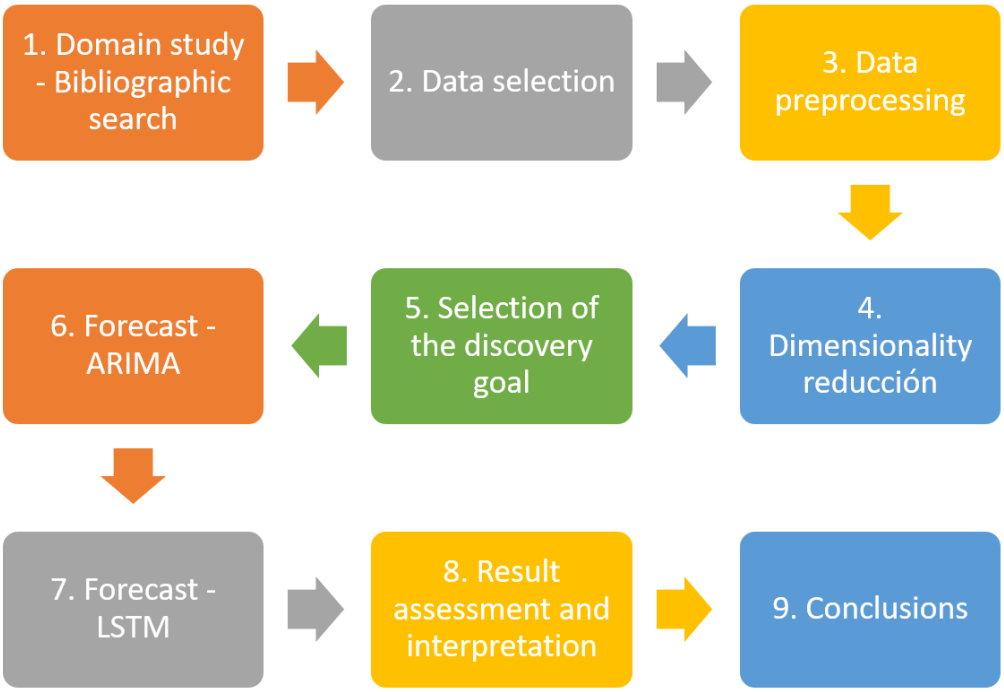


Figure 3.1: Methodology used

3.1.1 Domain Study - Bibliographic research

A bibliographic research was carried out in **"The Lancet", "Nature", "PUBMED",** etc. seeking for a better understanding of COVID-19 pandemic, measures taken by governments, open databases available to monitor COVID-19 and studies performed to predict its impact based on time series and neural networks (this research is contained at "State of the art" section).

3.1.2 Data selection

Due to the geographical focus of this project is Spain, as it has been stated in previous sections, this study will use the data offered by **CNE** [10] and **INE** [22] due to they are official institutions in Spain and offers open data. **Google** [19], is not and official institution, but the data offered it is widely used and as well as the CNE and INE, ensures anonymity and its data privacy policy, data gathering and methods applied to the data can be accessed and reviewed.

3.1.3 Data preprocessing

At this section we summarize only the most relevant data preparations accomplished. A separate report it is attached with all the different transformations, modifications, executions (pre and post with its results), etc. with specified and documented comments at the appropriated code block section [44].

The software used was **RStudio** [45] and the following libraries for manipulation, forecasting, visualization and document preparation were downloaded and installed (plus others): **corrplot, DescTools, fable, forecast, fpp3, ggplot2, imputeTS, keras, knitr, latex2exp, latexpdf, lubridate, psych, stats, tensorflow and tidyverse**.

- **CNE** - These datasets offers information related to infections, recoveries and deaths reported by local and regional governments in Spain. Two datasets were used:

  1. **cases_technic_province.csv** - Number of cases by diagnostic technique and province of residence, with **23426** obs. of **8** variables, start-date **01-01-2020**, end-date **17-03-2021**, **442** records per province and **0.01886792%** of missing values for column **"provincia-iso"** (those rows were omitted from the dataset).
  2. **cases_hosp_uci_def_sexo_edad_provres.csv** - Number of hospitalizations, number of ICU admissions and number of deaths by sex, age and province of residence, with **702780** obs. of **8** variables, start-date **01-01-2020**, end-date **17-03-2021, 13260** records per province (we have sub-age groups per province) and

0.01886792% of missing values for column **"provincia-iso"** (those rows were omitted from the dataset).

3. In both datasets we transformed column **"Fecha"** from **"character"** to **"date"**. Columns **"Grupo_edad"** and **"Sexo"** were eliminated from the dataset "CNE_casos" due to they are not adding value (mobility datasets do not include this variables). We changed **"NC"** values at iso code level to **"NA" (Navarra)** in both dataframes.

- **Google** - This dataset offers information related to mobility using Google application services. We downloaded the regional compressed dataset **"Google_Region_Mobility_Report_CSVs.zip"**, and we looked for the Spain one named as **"Google-2020_ES_Region_Mobility_Report.csv".**

1. This dataset has **24242** obs. of **15** variables, start-date **15-02-2020**, end-date **05-03-2021** and **385** records per province (we have groups per autonomous-communities). Several checks and transformations related to the appropriated **"province name"** and **"ISO code"** were carried-out.
2. Rows with **"na"** and **""** in **"sub_region_1"** and **"sub_region_2"** columns were eliminated. **"Date"** was transformed from **"character"** to **"date"**. Some columns were eliminated due to they are not adding value or they contain only blanks (**country_region_code, country_region, metro_area, census_fips_code,pace_id**). **"ES-"** characters were eliminated from **"iso_3166_2_code"** column.
3. For the missing values at the different areas of interest explored, as **"retail_and_recreation_percent_change_from_baseline"** we have used **"na_seadec()"** function from **"imputeTS"** package due to it takes into account seasonality for the time series (for this particular case we have generated one time series dataset per dimension of interest and finally we merged again into a cleaned data-frame).

- **INE** - This dataset offers the mobility population captured by the mobile telephony infrastructure system in Spain. We downloaded the dataset offered by the **INE web application** related only to the provinces in Spain [24].

1. **EM3-Movimiento de personas por provincias.csv** - This dataset has **9198** obs. of **3** variables, start-date **16-03-2020**, end-date **31-12-2020** and **146** records per province.
2. **"Total"** column was changed from **"character"** to **"numerical"** and **"Periodo"** column from **"character"** to **"date"**.

3. Due to the nature of this dataset, we have had to transpose it in order to analyse the missing values by province and impute them. As in the case of Google, we have used **"na_seadec()"** function from **"imputeTS"** package due to it takes into account seasonality for the time series (for this particular case we have generated a time series from the data-frame and finally we converted back to a data-frame).

- **Total** - This dataset is the result of a merge process carried-out for the previous ones (**CNE+INE+Google**) and it is the one used to perform our study, **ARIMA vs LSTM** [21]. This dataset has **15080** obs. of **20** variables, start-date **16-03-2020**, end-date **31-12-2020** and 290 **records per province**.

  The dataset has been also converted to a time-series and a CSV file version (Total.csv) it is provided for its review and usage, if necessary, at the Github [44] repository generated for this project. This dataset was converted to time-series with a daily frequency (Total_ts and Total_ts_b -this one with 5 provinces and 1450 rows-). The following are some figures extracted from Total dataset (Figures 3.2).

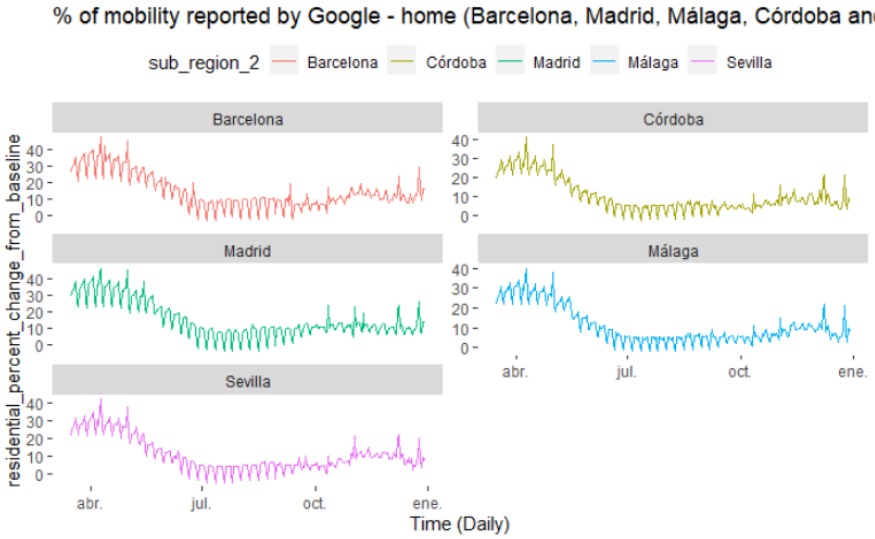


Figure 3.2: Google - Change residential

3.1.4 Dimensionality reduction

Correlation and **PCA** (Principal Component Analysis) were carried-out to address which are the mos important variables for our analysis (Figures 3.3, 3.4).

For our case we consider that up to **PC3, which explain the 84%** of the accumulated variance is enough and we have eliminated **"num_casos_prueba_test_ac", "num_casos_prueba**

_ag", "num_casos_prueba_elisa" and "num_casos.y" columns. The case of "num_casos.y" is due to we can consider it as duplicated against "num_casos.x". This means we count with 15 columns that it is suppose will offer valuable information for further stages (15080 total rows - 290 per province). This approach was based on **Barcelona** and was extrapolated to the rest of provinces.

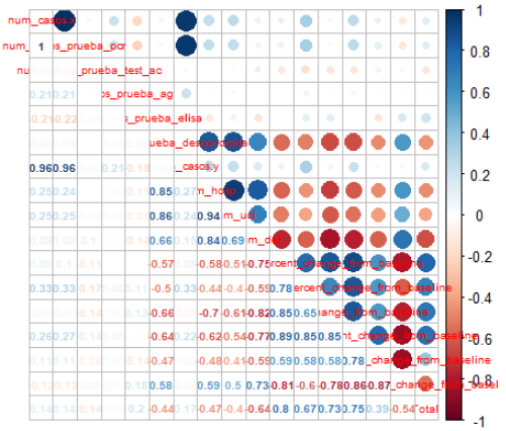


Figure 3.3: Correlation observed - Barcelona

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9     PC10    PC11    PC12    PC13    PC14    PC15    PC16      PC17
Standard deviation     3.0590 1.8914 1.12142 1.02409 0.83504 0.57825 0.46266 0.41006 0.33620 0.28445 0.21986 0.17734 0.16911 0.15372 0.1369 0.01274 1.826e-17
Proportion of Variance 0.5504 0.2104 0.07398 0.06169 0.04102 0.01967 0.01259 0.00989 0.00665 0.00476 0.00284 0.00185 0.00168 0.00139 0.0011 0.00001 0.000e+00
Cumulative Proportion  0.5504 0.7609 0.83485 0.89655 0.93756 0.95723 0.96982 0.97971 0.98636 0.99112 0.99397 0.99582 0.99750 0.99889 1.0000 1.00000 1.000e+00
                                                                PC1          PC2          PC3          PC4          PC5          PC6          PC7          PC8          PC9
num_casos.x                                          -0.1354826901 -0.469400471 -0.002875288 -0.028445983 -0.15653475  0.16549669  0.11538168  0.07515692 -0.204798939
num_casos_prueba_pcr                                 -0.1356873829 -0.469237915 -0.012292151 -0.026795983 -0.15256567  0.17028866  0.11048811  0.07713501 -0.206074966
num_casos_prueba_test_ac                             -0.0782958432  0.237502756 -0.392712472 -0.268237834 -0.81534671 -0.10365823 -0.08510797 -0.12117026  0.031408747
num_casos_prueba_ag                                   0.0005323686  0.008790618 -0.034460138  0.950253303 -0.25969423 -0.01360833 -0.03926649 -0.14832473 -0.027397924
num_casos_prueba_elisa                                0.0495912937  0.004385553  0.811470962 -0.060011278 -0.34805371 -0.37185910  0.24859208  0.01374512  0.085159377
num_casos_prueba_desconocida                         -0.2907430443  0.113686282  0.019674417 -0.078105637  0.20580692 -0.11429059  0.19845835 -0.75630827 -0.349092565
num_casos.y                                          -0.1546876637 -0.450080493  0.041668217 -0.036946432 -0.15827091  0.15297764  0.04117476 -0.02860659 -0.130984317
num_hosp                                             -0.2956764064 -0.199498016 -0.051419434 -0.010413059  0.03855580 -0.11688342 -0.11367668 -0.07611704  0.283065654
num_uci                                              -0.2788863801 -0.230508957 -0.018126448  0.006323566  0.06172358 -0.06173530 -0.02723992 -0.20583241  0.717067401
num_def                                              -0.3153112111 -0.030904089 -0.031133681  0.009747098  0.08791553 -0.26888200 -0.20890547  0.15436921 -0.009119052
retail_and_recreation_percent_change_from_baseline    0.3077977291 -0.116093786  0.033200491 -0.075398568 -0.02731712  0.18700422 -0.10130686 -0.31700087  0.099531350
grocery_and_pharmacy_percent_change_from_baseline     0.2553316609 -0.261732823  0.027494995 -0.030995413  0.01445198 -0.34534708 -0.63305922 -0.04093499 -0.110976209
parks_percent_change_from_baseline                    0.3128404870  0.012776948  0.029299658 -0.011018427 -0.01705528  0.39173313  0.15076652  0.04293438  0.283874181
transit_stations_percent_change_from_baseline         0.2855716515 -0.218931952 -0.099387689 -0.018021579  0.03520761 -0.26003250 -0.06852096 -0.19335324  0.094114251
workplaces_percent_change_from_baseline               0.2622358364 -0.170838668 -0.303553292  0.028981871  0.10943614 -0.48832683  0.37588850  0.19906444 -0.125028631
residential_percent_change_from_baseline             -0.2895648649  0.166332144  0.211975965  0.003369575  0.03104587  0.13040337 -0.39494991  0.23079899 -0.178700468
Total                                                 0.2993175108 -0.081611685  0.173800838 -0.065471155 -0.05468582  0.19337305 -0.26415699 -0.26912855 -0.118124812
```

Figure 3.4: PCA - Variance explained - Barcelona

3.1.5 Selection of the discovery goal

Data were processed using the following regression and neural network forecasting methods: **"ARIMA" and "LSTM"** [21].

The goal for this study is to predict infections caused by COVID-19 based on mobility data (quantitative time-series forecast). We have selected "num_casos.x" as target / dependent

variable and the rest of mobility variables are considered as the independent ones. The **forecast horizon will be 7, 14 and 21 days** (our frequency is daily - 365 based on the data obtained).

3.1.6 ARIMA

As stated by Hyndman [21], "...

- **ARIMA** models aim to describe the autocorrelations in the data...
- **Trend** exists when there is a long-term increase or decrease in the data...
- **Seasonal** pattern occurs when a time series is affected by seasonal factors such as the time of the year...
- **Cyclic** occurs when the data exhibit rises and falls that are not of a fixed frequency...
- **A stationary** time series is one whose statistical properties do not depend on the time at which the series is observed...
- When a decomposition of a time series happens, trend and cycle are combined into a single trend-cycle component (called "the trend"). So, a time series is composed by three components: trend-cycle, seasonal component, and remainder (this one contains anything else in the time series)."

In our case and for simplicity reasons we have analysed 5 provinces (Barcelona, Madrid, Málaga, Cádiz and Sevilla) from a univariate and multivariate perspective (in this report we have focused on Barcelona results due to the other provinces offers a similar behaviour). We have followed some of the general steps stated by Hyndman [21] for the analysis of our time series.

- Plot the data.
- If necessary, transform the data (stabilise the variance).
- If the data are non-stationary, take first differences till its became stationary.
- Examine the ACF/PACF.
- Try chosen model(s) / Use the AICc to search for a better model.
- Check residuals from chosen model by plotting the ACF of the residuals, and doing a portmanteau test of the residuals (if no white noise is observed try a modified model).

- If residuals can be consider white noise, calculate forecasts.

We have observed a week seasonality thanks to the Seasonal and Trend decomposition using Loess (STL) test performed and thanks to difference method applied we were able to convert our time series to a stationary one (Figures 3.5 and 3.6)

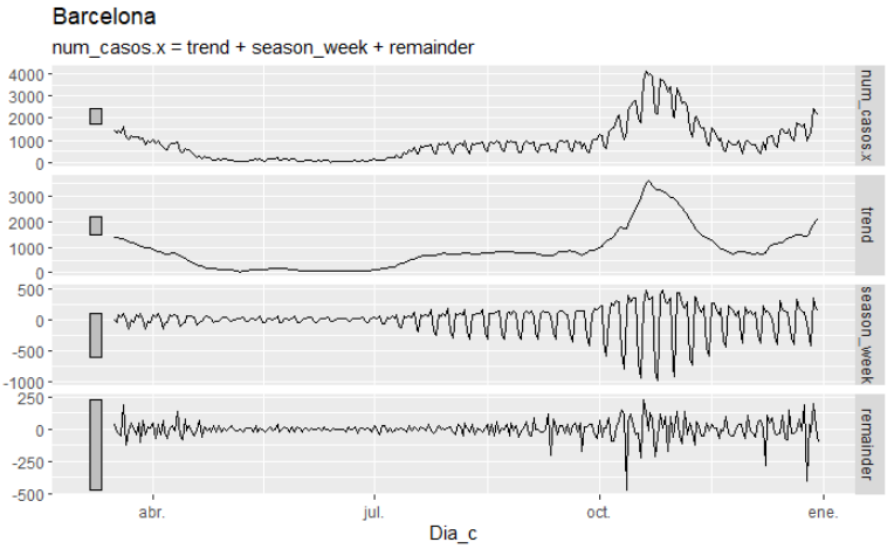


Figure 3.5: Barcelona - Seasonality / Trend - STL

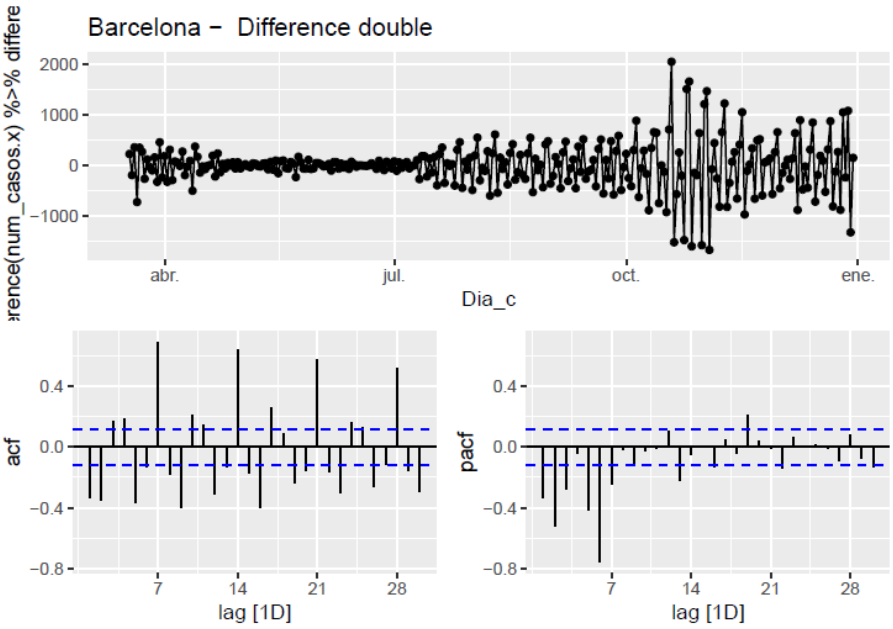


Figure 3.6: Barcelona - Residuals and difference

Univariate and multivariate (this one using the dynamic regression using errors approach, using fable() function from fpp3 library). The two approaches were carried out for

Barcelona with similar results and the univariate model was performed over Madrid, Málaga, Cádiz and Sevilla.

ARIMA models performs better compare with SNaive in both cases. But ARIMA automated option with extensive search of parameters performs better under the multivariate option (**arima_at2**). For the manual (**arima_mn**) and easy automate (**arima_at1**) approaches, results were similar or even worst for multivariate. We can conclude that under the ARIMA perspective, the inclusion of externals variables increases the performance of the forecast only for the automate hard way option (Figures 3.7, 3.8 and 3.9).

.model <chr>	sub_region_2 <chr>	.type <chr>	ME <dbl>	RMSE <dbl>	MAE <dbl>	MPE <dbl>	MAPE <dbl>
arima_at1	Barcelona	Test	539.9930	659.0784	546.1993	34.00993	34.64712
arima_at2	Barcelona	Test	580.3954	697.0175	580.3954	36.61504	36.61504
arima_man	Barcelona	Test	379.2112	507.4258	409.4153	23.75412	26.85516
SNaive	Barcelona	Test	700.3810	805.6091	700.3810	46.00523	46.00523

Figure 3.7: Barcelona - Accuracy univariate based on errors

.model <chr>	sub_region_2 <chr>	.type <chr>	ME <dbl>	RMSE <dbl>	MAE <dbl>	MPE <dbl>	MAPE <dbl>
arima_at1	Barcelona	Test	557.7329	670.1695	557.7329	36.07949	36.07949
arima_at2	Barcelona	Test	439.3400	551.6422	454.9878	27.47750	29.08404
arima_man	Barcelona	Test	557.2254	662.7884	574.6976	39.42849	41.22235
SNaive	Barcelona	Test	700.3810	805.6091	700.3810	46.00523	46.00523

Figure 3.8: Barcelona - Accuracy multivariate based on errors

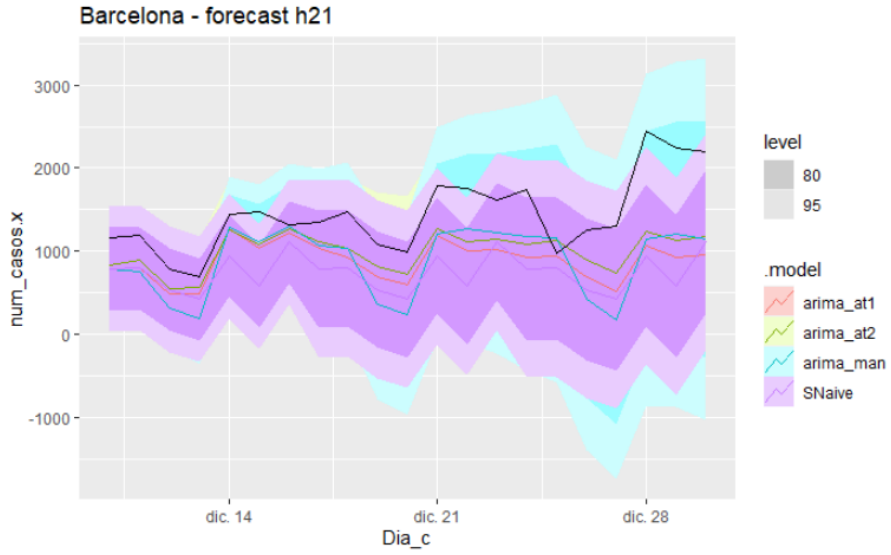


Figure 3.9: Barcelona - Multivariate forecast plot

In this exercise we have notice that for Málaga, Cádiz and Sevilla provinces, SNaive model performed better for the univariate analysis (Figures 3.10 and 3.11).

.model <chr>	sub_region_2 <chr>	.type <chr>	ME <dbl>	RMSE <dbl>	MAE <dbl>	MPE <dbl>	MAPE <dbl>
arima_at1	Málaga	Test	143.31269	165.2677	143.31269	66.67529	66.67529
arima_at2	Málaga	Test	172.87829	200.1143	172.87829	81.65825	81.65825
SNaive	Málaga	Test	80.80952	118.1792	83.38095	33.03723	34.63352

Figure 3.10: Málaga - Accuracy univariate based on errors

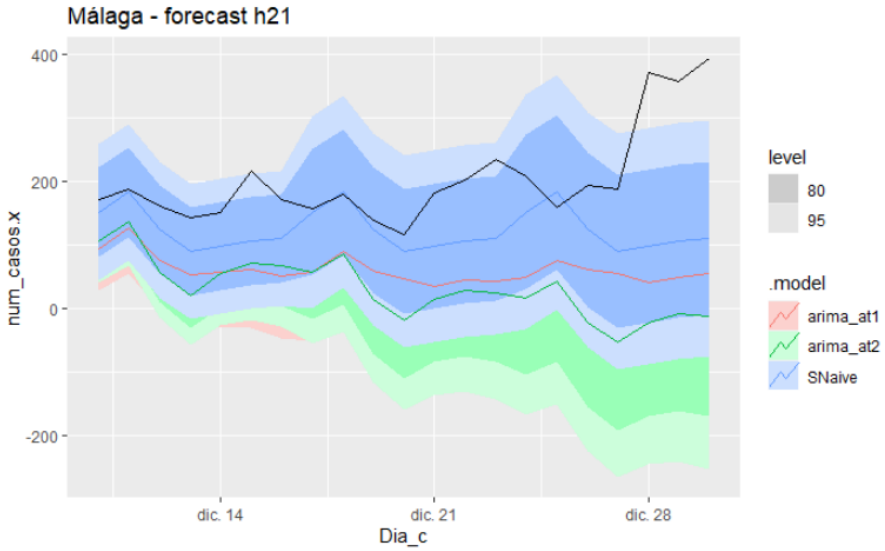


Figure 3.11: Málaga - Univariate forecast plot

This situation leads us to consider that, for ARIMA models, it is necessary to manage each time series as a unique study case (by province) even if the area of study and variables used are the same (Covid spread based on mobility).

3.1.7 LSTM

3.1.8 Results assessment

3.1.9 Conclusions

Bibliography

[1] Mobility-based prediction of sars-cov-2 spreading. https://arxiv.org/abs/2102.08253. [Online; accessed 25-Feb-2021].

[2] Charu C Aggarwal. *Data Mining*. Springer, 2015.

[3] Apple. Covid-19 - mobility trends repors. https://covid19.apple.com/mobility, 2021. [Online; accessed 22-Feb-2021].

[4] Hamada S Badr, Hongru Du, Maximilian Marshall, Ensheng Dong, Marietta M Squire, and Lauren M Gardner. Association between mobility patterns and covid-19 transmission in the usa: a mathematical modelling study. *The Lancet Infectious Diseases*, 20(11):1247–1254, 2020.

[5] Taweh Beysolow II. *Introduction to Deep Learning Using R*. Apress, 2020.

[6] CDC. Severe acute respiratory syndrome (sars). https://www.cdc.gov/sars/about/index.html, 2013. [Online; accessed 6-Mar-2021].

[7] CDC. Middle east respiratory syndrome (mers). https://www.cdc.gov/coronavirus/mers/about/index.html, 2019. [Online; accessed 6-Mar-2021].

[8] CDC. Late sequelae of covid-19. https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-care/late-sequelae.html, 2020. [Online; accessed 21-Feb-2021].

[9] Giuseppe Ciaburro and Balaji Venkateswaran. *Neural networks with R*. Packt Publishing, 2017.

[10] CNE. Covid-19. https://cnecovid.isciii.es/covid19/#documentaci%C3%B3n-y-datos, 2021. [Online; accessed 21-Feb-2021].

[11] Gergely Daróczi. *Mastering data analysis with R*. Packt Publishing, 2015.

[12] Jayanthi Devaraj, Rajvikram Madurai Elavarasan, Rishi Pugazhendhi, G.M. Shafiullah, Sumathi Ganesan, Ajay Kaarthic Jeysree, Irfan Ahmad Khan, and Eklas Hossain. Forecasting of covid-19 cases using deep learning models: Is it reliable and practically significant? *Results in Physics*, 21:103817, 2021.

[13] Roger Devine and Michael Pawlus. *Hands-On Deep Learning with R*. Packt Publishing, 2020.

[14] Kuldeep Dhama, Sharun Khan, Ruchi Tiwari, Shubhankar Sircar, Sudipta Bhat, Yashpal Singh Malik, Karam Pal Singh, Wanpen Chaicumpa, D. Katterine Bonilla-Aldana, and Alfonso J. Rodriguez-Morales. Coronavirus disease 2019–covid-19. *Clinical Microbiology Reviews*, 33(4), 2020.

[15] Deming Edwards. Pdsa cycle - the w. edwards deming institute. https://deming.org/explore/pdsa/, 2021. [Online; accessed 22-Feb-2021].

[16] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37, Mar. 1996.

[17] Jordi Giroes. *Metodologías y estándares*. Editorial UOC, nd.

[18] Jordi Gironés, Jordi Casas, and Julià Minguillón. *Minería de datos modelos y algoritmos*. Editorial UOC, 2017.

[19] Google. Covid-19 community mobility report. https://www.google.com/covid19/mobility/, 2021. [Online; accessed 22-Feb-2021].

[20] Swarna Gupta, Rehan Ali Ansari, and Dipayan Sarkar. *Deep learning with R cookbook*. Packt Publishing, 2020.

[21] R.J. Hyndman and G. Athanasopoulos. Forecasting: Principles and practice (3rd ed). https://otexts.com/fpp3/, 2021. [Online; accessed 5-Mar-2021].

[22] INE. Estadística experimental. https://www.ine.es/experimental/movilidad/experimental_em.htm, 2020. [Online; accessed 21-Feb-2021].

[23] INE. Estudio de movilidad a partir de la telefonía móvil durante el periodo julio-diciembre 2020 (em-3). https://www.ine.es/experimental/movilidad/exp_em3_proyecto.pdf, 2020. [Online; accessed 21-Feb-2021].

[24] INE. Estudio de movilidad a partir de la telefonía móvil durante el periodo julio-diciembre 2020 (em-3) - provincias. https://www.ine.es/jaxiT3/Tabla.htm?t=37812, 2020. [Online; accessed 21-Feb-2021].

[25] INE. Información estadística para el análisis del impacto de la crisis covid-19. https://www.ine.es/covid/covid_inicio.htm, 2020. [Online; accessed 21-Feb-2021].

[26] Shwet Ketu and Pramod Kumar Mishra. A hybrid deep learning model for covid-19 prediction and current status of clinical trials worldwide. *Computers, Materials and Continua*, 66(2):1896–1919, 2021.

[27] Erdinç Koç and Muammer Türkoğlu. Forecasting of medical equipment demand and outbreak spreading based on deep long short-term memory network: the covid-19 pandemic in turkey. *Signal, Image and Video Processing*, 2021.

[28] Rami Krispin. *Hands-On time series analysis with R.* Packt Publishing, 2019.

[29] Naresh Kumar and Seba Susan. Covid-19 pandemic prediction using time series forecasting models, 2020.

[30] Daniel T Larose. *Data Mining Methods and Models.* John Wiley and Sons, 2 edition, 2015.

[31] Johannes Ledolter. *Business analytics and data mining with R.* John Wiley and Sons, 2013.

[32] Smriti Mallapaty. What's the risk of dying from a fast-spreading covid-19 variant? *Nature*, 590(7845):191–192, 2021.

[33] Mattia Mazzoli, David Mateo, Alberto Hernando, Sandro Meloni, and José J. Ramasco. Effects of mobility and multi-seeding on the propagation of the covid-19 in spain. *medRxiv*, 2020. [https://doi.org/10.1101/2020.05.09.20096339].

[34] The Lancet Respiratory Medicine. Covid-19 transmission—up in the air. *The Lancet Respiratory Medicine*, 8(12):1159, 2020.

[35] Lidia Morawska and Junji Cao. Airborne transmission of sars-cov-2: The world should face the reality. *Environment International*, 139:105730, 2020.

[36] Lidia Morawska, Julian W. Tang, William Bahnfleth, Philomena M. Bluyssen, Atze Boerstra, Giorgio Buonanno, Junji Cao, Stephanie Dancer, Andres Floto, Francesco Franchimon, and et al. How can airborne transmission of covid-19 indoors be minimised? *Environment International*, 142:105832, 2020.

[37] Braulio Nuria and Josep Curto. *Customer analytics.* Editorial UOC, nd.

[38] Yixuan Pan, Aref Darzi, Aliakbar Kabiri, Guangchen Zhao, Weiyu Luo, Chenfeng Xiong, and Lei Zhang. Quantifying human mobility behaviour changes during the covid-19 outbreak in the united states. *Scientific Reports*, 10(1), 2020.

[39] Michael Pawlus and Rodger Devine. *Hands-On Deep Learning with R*. Packt Publishing, 2020.

[40] PKS Prakash and Achyutuni Sri Krishna Rao. *R deep learning cookbook*. Packt Publishing, 2017.

[41] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[42] Firda Rahmadani and Hyunsoo Lee. Hybrid deep learning-based epidemic prediction framework of covid-19: South korea case. *Applied Sciences*, 10(23):8539, 2020.

[43] Hafiz Tayyab Rauf, M. Ikram Ullah Lali, Muhammad Attique Khan, Seifedine Kadry, Hanan Alolaiyan, Abdul Razaq, and Rizwana Irfan. Time series forecasting of covid-19 transmission in asia pacific countries using deep neural networks. *Personal and Ubiquitous Computing*, 2021.

[44] Alvaro Rodriguez S. Github repository. https://github.com/arodriguezsans/TFM_PEC3, 2021. [Online; accessed 21-Feb-2021].

[45] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA, 2020.

[46] A. Serrano-Cumplido, P.B. Antón-Eguía Ortega, A. Ruiz García, V. Olmo Quintana, A. Segura Fragoso, A. Barquilla Garcia, and Á. Morán Bayón. Covid-19. la historia se repite y seguimos tropezando con la misma piedra. *Medicina de Familia. SEMERGEN*, 46:48–54, 2020.

[47] Farah Shahid, Aneela Zameer, and Muhammad Muneeb. Predictions for covid-19 with deep learning models of lstm, gru and bi-lstm. *Chaos, Solitons and Fractals*, 140:110212, 2020.

[48] Sean J Taylor and Benjamin Letham. Forecasting at scale. 2017. [Online; accessed 5-Mar-2021].

[49] Mark Treveil and the Dataiku team. *Introducing MLOps*. OReilly Media, Inc., 2020.

[50] Ivan Vasilev, Daniel Slater, Gianmario Spacagna, Peter Roelants, and Valentino Zocca. *Python Deep Learning - Second Edition*. Packt Publishing, 2 edition, 2019.

[51] WHO. Coronavirus. https://www.who.int/health-topics/coronavirus#tab=tab_1, 2021. [Online; accessed 21-Feb-2021].

[52] WHO. Covid-19 - vaccines. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/covid-19-vaccines, 2021. [Online; accessed 6-Mar-2021].

[53] WHO. Sars-cov-2 variants. https://www.who.int/csr/don/31-december-2020-sars-cov2-variants/en/, 2021. [Online; accessed 6-Mar-2021].

[54] WHO. Timeline: Who's covid-19 response. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline#!, 2021. [Online; accessed 6-Mar-2021].

[55] WHO. Who coronavirus (covid-19) dashboard. https://covid19.who.int/, 2021. [Online; accessed 6-Mar-2021].

[56] Rüdiger Wirth. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, pages 29–39, 2000.

[57] Wysocki and K. Sybex Brown. *Effective Project Management*. John Wiley and Sons, 8 edition, 2019.

Appendix A

Code used

PEC 3: Desing and Implementation

UOC - Alumno: Álvaro Rodríguez Sans

May 2020 - Delivery 23/05/2021

Índex

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code. Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*. When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file). The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.

The bibliographic references used for this practice have been: (Baayen 2008; Hothorn and Everitt 2014; Hyndman and Athanasopoulos 2021; Liviano Solas and Pujol Jover, n.d.; Teetor 2011; Vegas Lozano, n.d.).

```
# At section - Data types and modifications
if(!require(knitr)){
    install.packages('knitr', repos='http://cran.us.r-project.org')
    library(knitr)}
```

```
## Loading required package: knitr
```

```
if(!require(latexpdf)){
    install.packages('latexpdf', repos='http://cran.us.r-project.org')
    library(latexpdf)}
```

```
## Loading required package: latexpdf
```

```
if(!require(latex2exp)){
    install.packages('latex2exp', repos='http://cran.us.r-project.org')
    library(latex2exp)}
```

```
## Loading required package: latex2exp
```

```
if(!require(data.table)){
    install.packages('data.table', repos='http://cran.us.r-project.org')
    library(data.table)}
```

```
## Loading required package: data.table
```

```
if(!require(tidyverse)){
    install.packages("tidyverse", repos='http://cran.us.r-project.org')
    library(tidyverse)}
```

```
## Loading required package: tidyverse
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.3.3     v purrr   0.3.3
## v tibble  3.0.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::between()   masks data.table::between()
## x dplyr::filter()    masks stats::filter()
## x dplyr::first()     masks data.table::first()
## x dplyr::lag()       masks stats::lag()
## x dplyr::last()      masks data.table::last()
## x purrr::transpose() masks data.table::transpose()
```

```r
if(!require(VIM)){
    install.packages('VIM', repos='http://cran.us.r-project.org')
    library(VIM)}
```

```
## Loading required package: VIM
## Loading required package: colorspace
## Loading required package: grid
## VIM is ready to use.
##  Since version 4.0.0 the GUI is in its own package VIMGUI.
##
##           Please use the package to use the new (and old) GUI.
## Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM/issues
##
## Attaching package: 'VIM'
## The following object is masked from 'package:datasets':
##
##     sleep
```

```r
if(!require(imputeTS)){
    install.packages("imputeTS", repos='http://cran.us.r-project.org')
    library(imputeTS)}
```

```
## Loading required package: imputeTS
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```r
if(!require(xts)){
    install.packages("xts", repos='http://cran.us.r-project.org')
    library(xts)}
```

```
## Loading required package: xts
## Loading required package: zoo
##
## Attaching package: 'zoo'
## The following object is masked from 'package:imputeTS':
##
##     na.locf
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
##
## Attaching package: 'xts'
## The following objects are masked from 'package:dplyr':
##
##     first, last
## The following objects are masked from 'package:data.table':
##
```

```
##     first, last
```

```
if(!require(tsbox)){
    install.packages("tsbox", repos='http://cran.us.r-project.org')
    library(tsbox)}
```

```
## Loading required package: tsbox
```

```
# At section - Visual analysis
if(!require(fpp3)){
    install.packages("fpp3", repos='http://cran.us.r-project.org')
    library(fpp3)}
```

```
## Loading required package: fpp3
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v lubridate   1.7.8     v feasts      0.2.1
## v tsibble     1.0.0     v fable       0.3.0
## v tsibbledata 0.3.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::between()     masks data.table::between()
## x lubridate::date()    masks base::date()
## x dplyr::filter()      masks stats::filter()
## x xts::first()         masks dplyr::first(), data.table::first()
## x lubridate::hour()    masks data.table::hour()
## x tsibble::index()     masks zoo::index()
## x tsibble::intersect() masks base::intersect()
## x tsibble::interval()  masks lubridate::interval()
## x lubridate::isoweek() masks data.table::isoweek()
## x tsibble::key()       masks data.table::key()
## x dplyr::lag()         masks stats::lag()
## x xts::last()          masks dplyr::last(), data.table::last()
## x lubridate::mday()    masks data.table::mday()
## x lubridate::minute()  masks data.table::minute()
## x lubridate::month()   masks data.table::month()
## x lubridate::quarter() masks data.table::quarter()
## x lubridate::second()  masks data.table::second()
## x tsibble::setdiff()   masks base::setdiff()
## x purrr::transpose()   masks data.table::transpose()
## x tsibble::union()     masks base::union()
## x lubridate::wday()    masks data.table::wday()
## x lubridate::week()    masks data.table::week()
## x lubridate::yday()    masks data.table::yday()
## x lubridate::year()    masks data.table::year()
```

```
if(!require(corrplot)){
    install.packages('corrplot', repos='http://cran.us.r-project.org')
    library(corrplot)}
```

```
## Loading required package: corrplot
```

```
## corrplot 0.84 loaded
```

```
if(!require(DescTools)){
    install.packages("DescTools", repos='http://cran.us.r-project.org')
    library(DescTools)}
```

```
## Loading required package: DescTools
##
## Attaching package: 'DescTools'
## The following objects are masked from 'package:fabletools':
##
##     MAE, MAPE, MSE, RMSE
## The following object is masked from 'package:data.table':
##
##     %like%
```

```
# At sections - ARIMA
#if(!require(forecast)){
#    install.packages('forecast', repos='http://cran.us.r-project.org')
#    library(forecast)}
#if(!require(tseries)){
#    install.packages('tseries', repos='http://cran.us.r-project.org')
#    library(tseries)}
#if(!require(astsa)){
#    install.packages('astsa', repos='http://cran.us.r-project.org')
#    library(astsa)}

#if(!require(psych)){
#    install.packages("psych", repos='http://cran.us.r-project.org')
#    library(psych)}
#if(!require(stats)){
#    install.packages("stats", repos='http://cran.us.r-project.org')
#    library(stats)}
#if(!require(keras)){
#    install.packages('keras', repos='http://cran.us.r-project.org')
#    library(keras)}
#if(!require(tensorflow)){
#    install.packages('tensorflow', repos='http://cran.us.r-project.org')
#    library(tensorflow)}

#if(!require(DataExplorer)){
#    install.packages('DataExplorer', repos='http://cran.us.r-project.org')
#    library(DataExplorer)}

knitr::opts_chunk$set(echo = TRUE)
```

1 Data load

Data is loaded from the sources stated at PEC1 and PEC2 (CNE, INE and Google).

- CNE-Covid-19
- INE-Covid-19
- Google-Covid-19

```
#library(dplyr)
# Source INE
EM3 <- read.csv('EM3-Movimiento de personas por provincias.csv',
                header=TRUE,
                sep = ";",
```

```
                stringsAsFactors = FALSE)

# Source Google
Google <- read.csv('Google-2020_ES_Region_Mobility_Report.csv',
                   header=TRUE,
                   sep = ";",
                   stringsAsFactors = FALSE)

# Source CNE
CNE_tecnica <- read.csv('CNE-casos_tecnica_provincia.csv',
                        header=TRUE,
                        sep = ",",
                        stringsAsFactors = FALSE)
CNE_casos <- read.csv('CNE-casos_hosp_uci_def_sexo_edad_provres.csv',
                      header=TRUE,
                      sep = ",",
                      stringsAsFactors = FALSE)
```

2 Initial descriptive statistics and visualization (na and impute)

2.1 Data types and modifications

We are gonig to check the **type of variable** that corresponds to each of the variables (numerical, factor, etc.) and **missing data / values or other anomalies** in each dataset.

2.1.1 EM3 review

We have the movement of people by provinces (we can see 146 rows by province, that correspond to days). In order to facilitate the comparison and have a valid reference on to what extent the mobility of the population should be considered to have varied, the data of a day of a week that can be considered "normal" are taken as a reference. For this study, the "normal" day that has been considered is the one that results from the average of the days 18 (Monday) to 21 (Thursday) of November 2019. It is indicated in the tables as the reference date 18/11/2019.

```
# Source INE
summary(EM3)
```

```
##  Zonas.de.movilidad   Periodo             Total
##  Length:9198        Length:9198        Length:9198
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
```

```
head(str(EM3,vec.len=2))
```

```
## 'data.frame':    9198 obs. of  3 variables:
##  $ Zonas.de.movilidad: chr  "Almería" "Almería" ...
##  $ Periodo           : chr  "30/12/2020" "27/12/2020" ...
##  $ Total             : chr  "17,17" "11,53" ...
```

```
## NULL
```

```
table(EM3$Zonas.de.movilidad)
```

```
##
##               Albacete       Alicante/Alacant                Almería
##                    146                    146                    146
```

```
##            Araba/Álava               Asturias                  Ávila
##                    146                    146                    146
##                Badajoz         Balears, Illes              Barcelona
##                    146                    146                    146
##                Bizkaia                 Burgos                Cáceres
##                    146                    146                    146
##                  Cádiz              Cantabria     Castellón/Castelló
##                    146                    146                    146
##                  Ceuta            Ciudad Real                Córdoba
##                    146                    146                    146
##             Coruña, A                 Cuenca             Formentera
##                    146                    146                    146
##          Fuerteventura                Gipuzkoa                Girona
##                    146                    146                    146
##             Gomera, La           Gran Canaria                Granada
##                    146                    146                    146
##            Guadalajara             Hierro, El                 Huelva
##                    146                    146                    146
##                 Huesca                  Ibiza                   Jaén
##                    146                    146                    146
##              Lanzarote                  León                  Lleida
##                    146                    146                    146
##                   Lugo                 Madrid                 Málaga
##                    146                    146                    146
##               Mallorca                Melilla                Menorca
##                    146                    146                    146
##                 Murcia                Navarra                Ourense
##                    146                    146                    146
##               Palencia             Palma, La            Palmas, Las
##                    146                    146                    146
##             Pontevedra             Rioja, La              Salamanca
##                    146                    146                    146
## Santa Cruz de Tenerife               Segovia                Sevilla
##                    146                    146                    146
##                  Soria              Tarragona               Tenerife
##                    146                    146                    146
##                 Teruel                 Toledo      Valencia/València
##                    146                    146                    146
##             Valladolid                Zamora               Zaragoza
##                    146                    146                    146
```

2.1.2 EM3 data transformation

We are going to **transform**:

- "Total" from "character" to "numerical"
- "Periodo" from "character" to "date"

```
EM3$Total <- sub(",", ".", EM3$Total)
EM3$Total <- as.numeric(EM3$Total)
EM3$Periodo <- as.Date(EM3$Periodo,format="%d/%m/%Y")
head(EM3)
```

```
##   Zonas.de.movilidad    Periodo Total
## 1            Almería 2020-12-30 17.17
## 2            Almería 2020-12-27 11.53
```

```
## 3            Almería 2020-12-23 17.81
## 4            Almería 2020-12-20 12.13
## 5            Almería 2020-12-16 18.28
## 6            Almería 2020-12-13 11.97
```

2.1.3 EM3 transpose and dates missing generation

Due to the nature of this dataset we have to transpose it in order to analyse the missing values by province and impute them. There are some dates that are not provided by EM3 study.

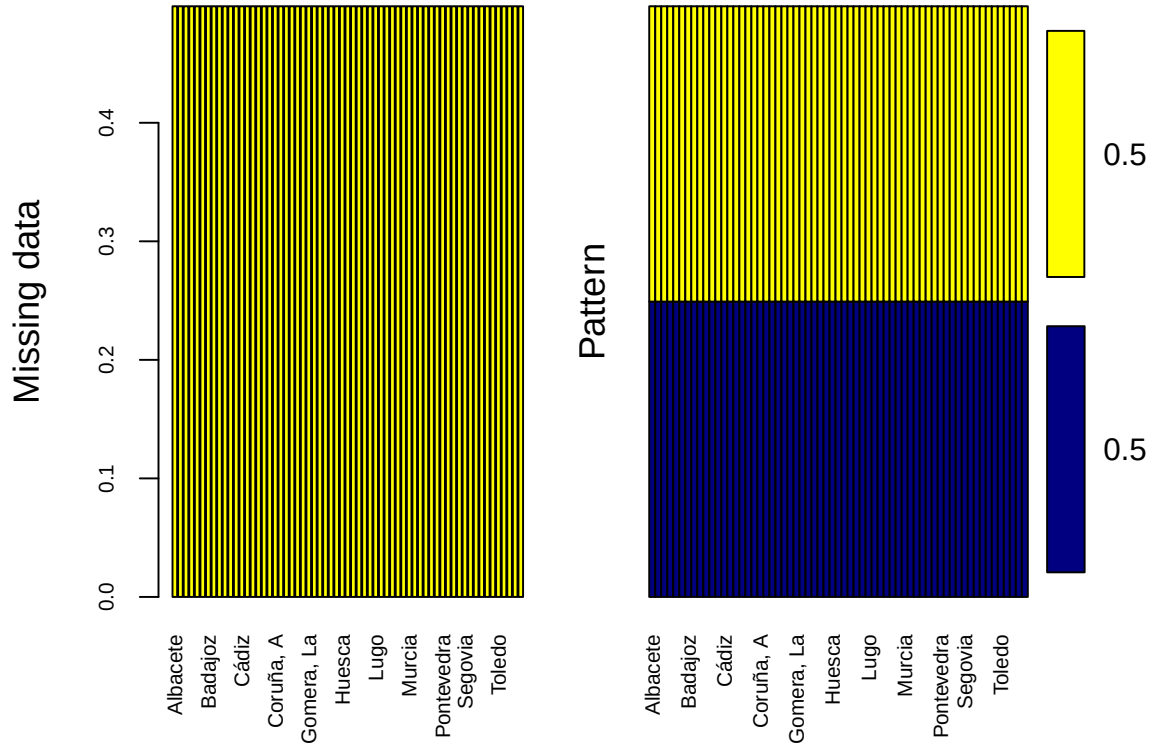
```
#library(data.table)
# Transpose dataframe
EM3_t<-dcast(EM3, Periodo~Zonas.de.movilidad)#, fill=NA)

#library(tidyverse)
# Create dates missing (for time series).
# Note: Acccording INE some "dates" are not provided.
EM3_t<-EM3_t %>%
  complete(Periodo = seq.Date(min(Periodo), max(Periodo), by="day"))

# Filter the interest period according INE EM3 study
# "2019-11-18" is the reference date EM3 study (for us it is excluded)
EM3_t<- EM3_t %>%
  filter(Periodo <= "2019-11-18" | Periodo >= "2020-03-16")

EM3_t
```

```
## # A tibble: 291 x 64
##    Periodo    Albacete `Alicante/Alaca~ Almería `Araba/Álava` Asturias Ávila
##    <date>        <dbl>            <dbl>   <dbl>         <dbl>    <dbl> <dbl>
##  1 2019-11-18    25.2             28.1    24.4          31.9     29.9  26.6
##  2 2020-03-16     9.9             14.4    11.0          15.9     13.1   9.44
##  3 2020-03-17    NA               NA      NA            NA       NA    NA
##  4 2020-03-18     9.51            13.4     7.28         14.5     12.0   9.17
##  5 2020-03-19    NA               NA      NA            NA       NA    NA
##  6 2020-03-20     8.75            12.0     6.87         11.9     11.3   8.69
##  7 2020-03-21    NA               NA      NA            NA       NA    NA
##  8 2020-03-22     4.5              6.14    4.19          6.46     5.64  4.53
##  9 2020-03-23    NA               NA      NA            NA       NA    NA
## 10 2020-03-24     9.02            10.9     8.98         13.3     11.2   8.26
## # ... with 281 more rows, and 57 more variables: Badajoz <dbl>, `Balears,
## #   Illes` <dbl>, Barcelona <dbl>, Bizkaia <dbl>, Burgos <dbl>, Cáceres <dbl>,
## #   Cádiz <dbl>, Cantabria <dbl>, `Castellón/Castelló` <dbl>, Ceuta <dbl>,
## #   `Ciudad Real` <dbl>, Córdoba <dbl>, `Coruña, A` <dbl>, Cuenca <dbl>,
## #   Formentera <dbl>, Fuerteventura <dbl>, Gipuzkoa <dbl>, Girona <dbl>,
## #   `Gomera, La` <dbl>, `Gran Canaria` <dbl>, Granada <dbl>, Guadalajara <dbl>,
## #   `Hierro, El` <dbl>, Huelva <dbl>, Huesca <dbl>, Ibiza <dbl>, Jaén <dbl>,
## #   Lanzarote <dbl>, León <dbl>, Lleida <dbl>, Lugo <dbl>, Madrid <dbl>,
## #   Málaga <dbl>, Mallorca <dbl>, Melilla <dbl>, Menorca <dbl>, Murcia <dbl>,
## #   Navarra <dbl>, Ourense <dbl>, Palencia <dbl>, `Palma, La` <dbl>, `Palmas,
## #   Las` <dbl>, Pontevedra <dbl>, `Rioja, La` <dbl>, Salamanca <dbl>, `Santa
## #   Cruz de Tenerife` <dbl>, Segovia <dbl>, Sevilla <dbl>, Soria <dbl>,
## #   Tarragona <dbl>, Tenerife <dbl>, Teruel <dbl>, Toledo <dbl>,
## #   `Valencia/València` <dbl>, Valladolid <dbl>, Zamora <dbl>, Zaragoza <dbl>
```

2.1.4 EM3 review missing values & impute

We check the missing values by province (we are close to 150 by province).

```
#library(VIM)
aggr(EM3_t[,-1], col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(EM3_t[,-1]), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```

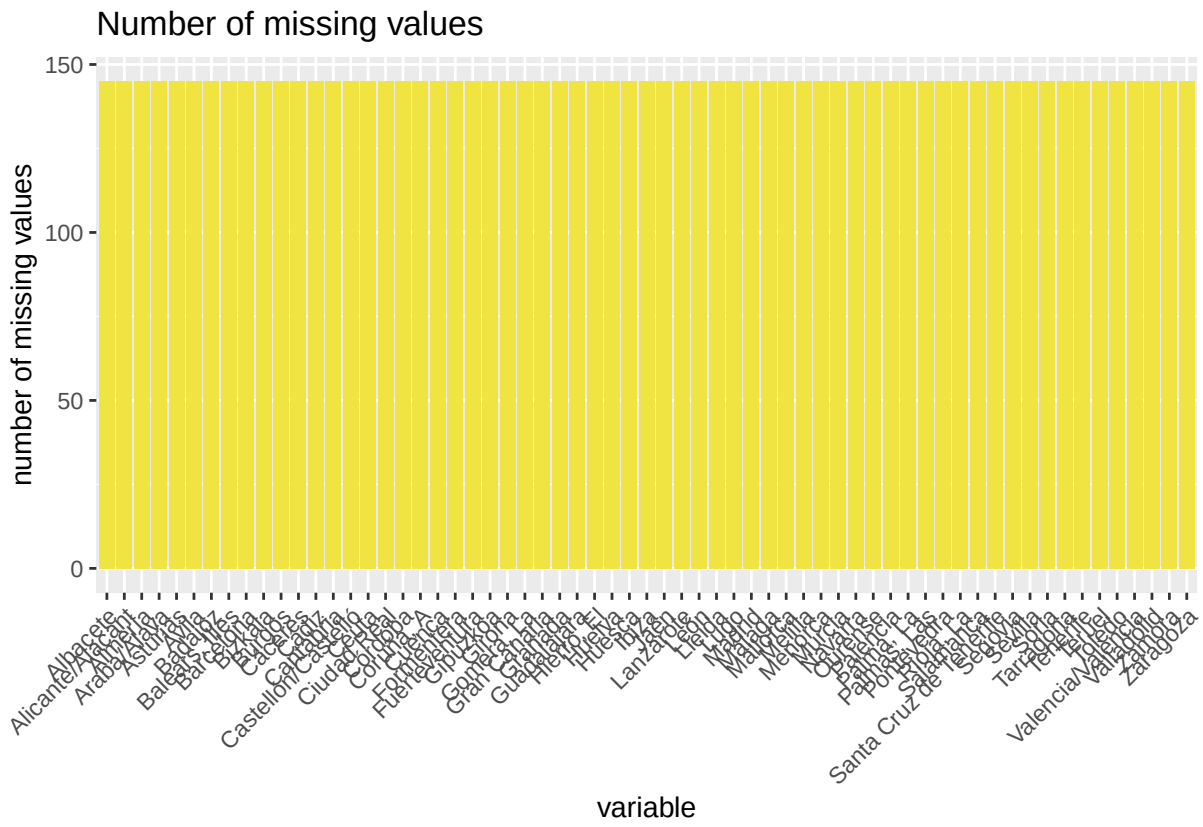



```
##
##  Variables sorted by number of missings:
##                Variable     Count
##                Albacete 0.4982818
##       Alicante/Alacant 0.4982818
##                Almería 0.4982818
##            Araba/Álava 0.4982818
##               Asturias 0.4982818
##                  Ávila 0.4982818
##                Badajoz 0.4982818
##         Balears, Illes 0.4982818
##              Barcelona 0.4982818
##                Bizkaia 0.4982818
##                 Burgos 0.4982818
##                Cáceres 0.4982818
##                  Cádiz 0.4982818
##              Cantabria 0.4982818
```

```
##      Castellón/Castelló 0.4982818
##                   Ceuta 0.4982818
##             Ciudad Real 0.4982818
##                 Córdoba 0.4982818
##               Coruña, A 0.4982818
##                 Cuenca 0.4982818
##              Formentera 0.4982818
##           Fuerteventura 0.4982818
##                Gipuzkoa 0.4982818
##                  Girona 0.4982818
##              Gomera, La 0.4982818
##            Gran Canaria 0.4982818
##                 Granada 0.4982818
##             Guadalajara 0.4982818
##              Hierro, El 0.4982818
##                  Huelva 0.4982818
##                  Huesca 0.4982818
##                   Ibiza 0.4982818
##                    Jaén 0.4982818
##                Lanzarote 0.4982818
##                   León 0.4982818
##                  Lleida 0.4982818
##                    Lugo 0.4982818
##                  Madrid 0.4982818
##                  Málaga 0.4982818
##                 Mallorca 0.4982818
##                  Melilla 0.4982818
##                  Menorca 0.4982818
##                   Murcia 0.4982818
##                  Navarra 0.4982818
##                  Ourense 0.4982818
##                 Palencia 0.4982818
##               Palma, La 0.4982818
##             Palmas, Las 0.4982818
##              Pontevedra 0.4982818
##                Rioja, La 0.4982818
##                Salamanca 0.4982818
##  Santa Cruz de Tenerife 0.4982818
##                  Segovia 0.4982818
##                  Sevilla 0.4982818
##                    Soria 0.4982818
##                Tarragona 0.4982818
##                 Tenerife 0.4982818
##                   Teruel 0.4982818
##                  Toledo 0.4982818
##       Valencia/València 0.4982818
##                Valladolid 0.4982818
##                   Zamora 0.4982818
##                 Zaragoza 0.4982818
```

```
EM3_t %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
```

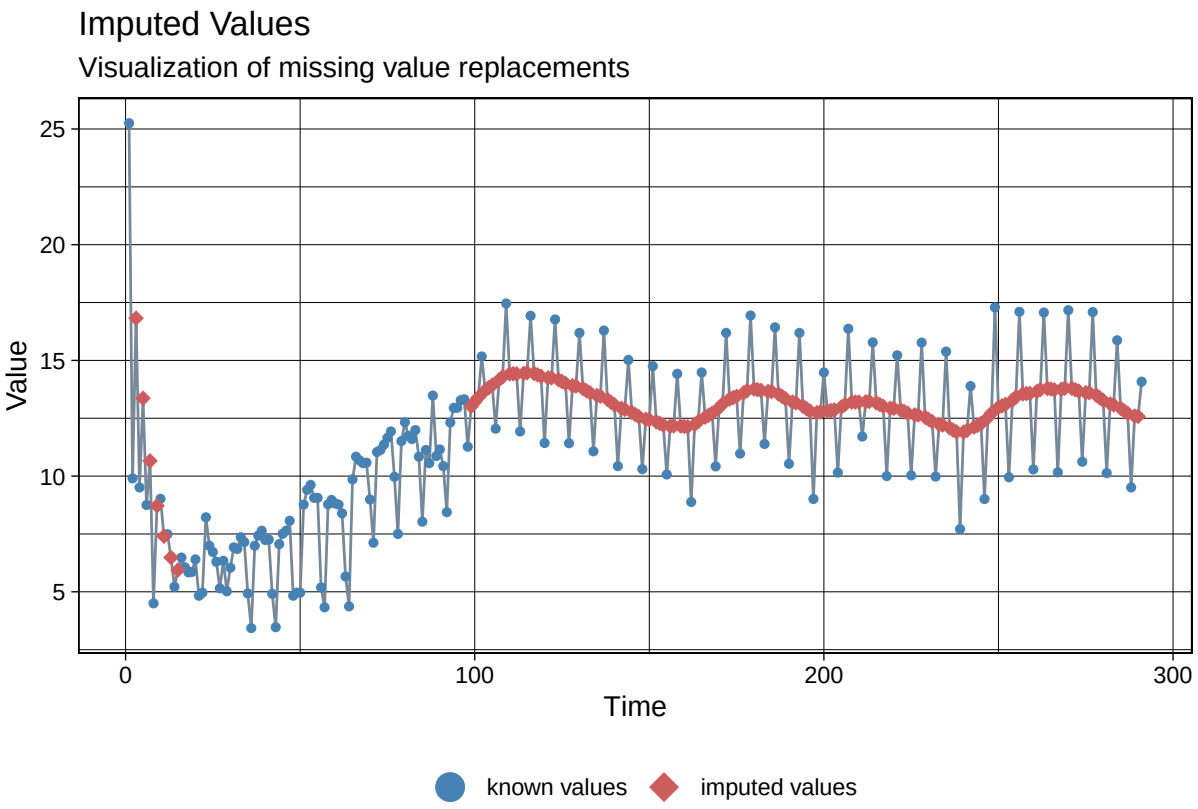
```
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



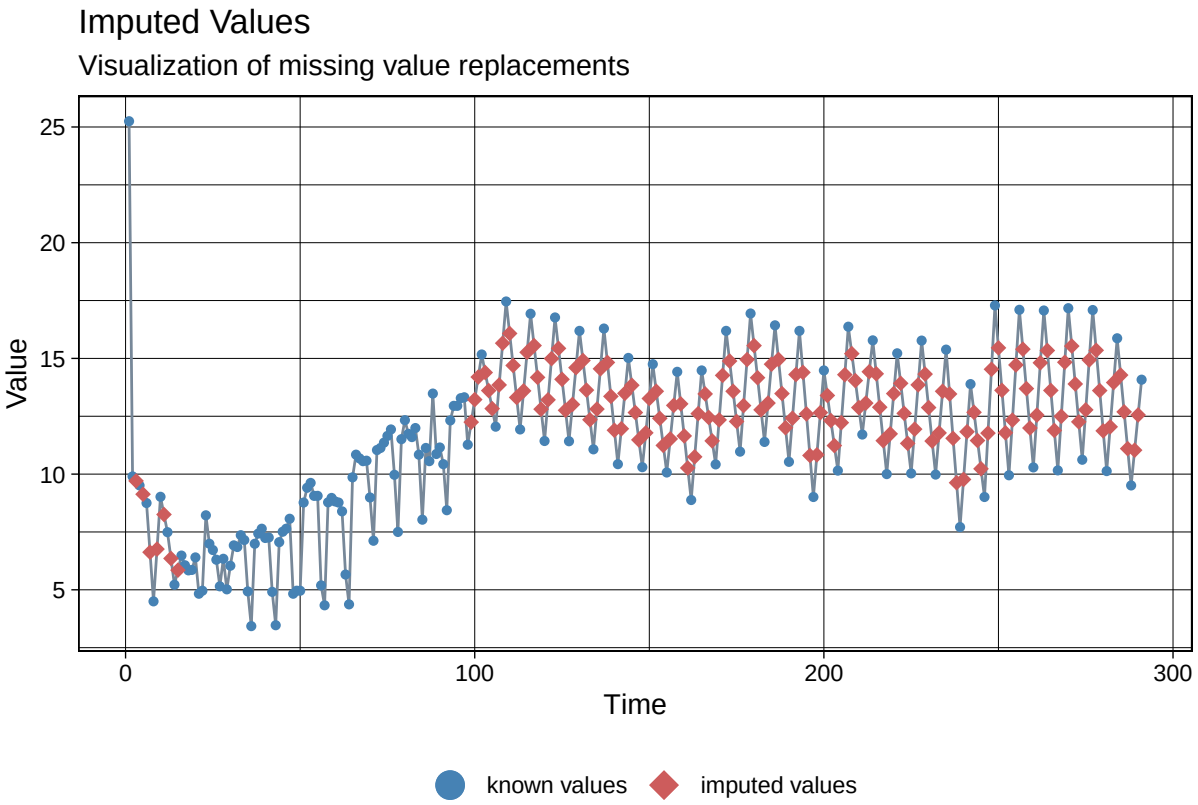
We impute the missing values following the principales stated for imputeTS. Thanks to this approach we almost double the amount of data for analysis by province (It was selected "na_seadec" due to it covers seasonality aspects -weekdays/weekends in our case-).

It is needed to transform the dataframe to a time series object.

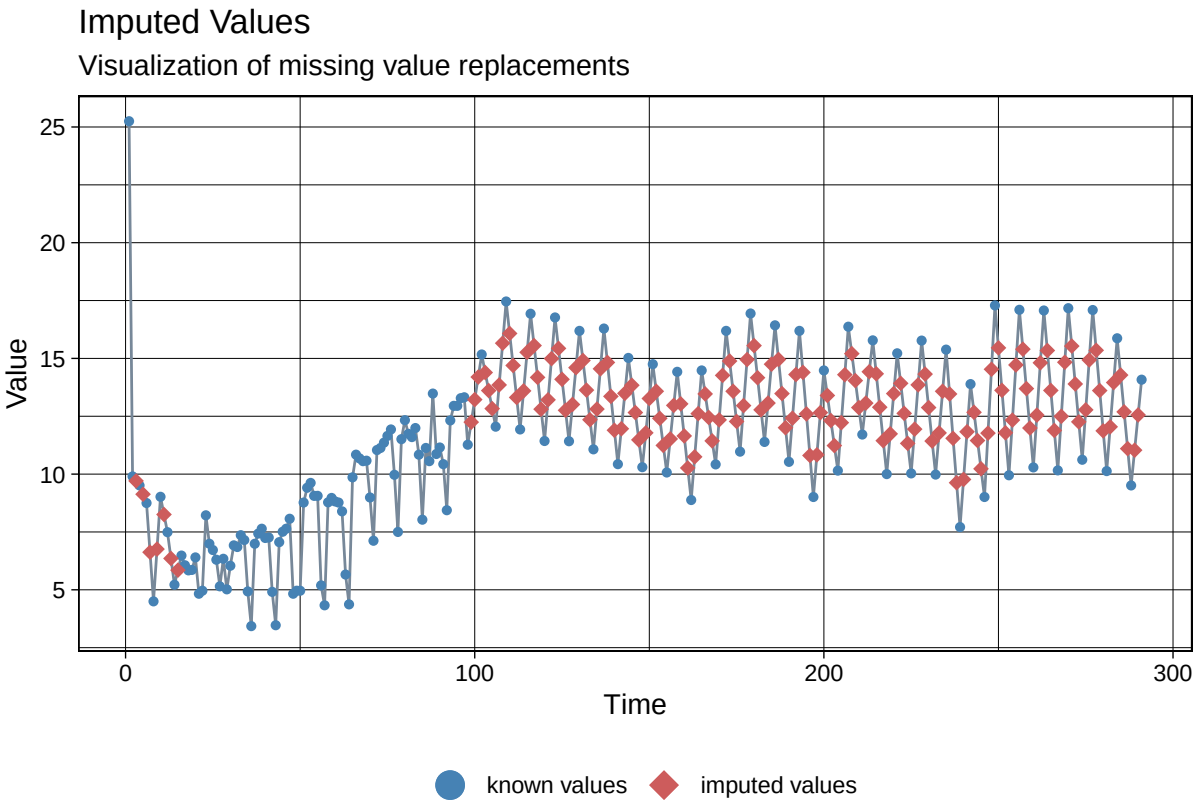
```
# Used to convert dataframe to ts object
#library(xts)
EM3_t_ts<-xts(EM3_t[-1],EM3_t$Periodo)

# Impute the missing values with na_kalman, na_seadec, na_interpolation & na_seasplit
#library(imputeTS)
imp <- na_kalman(EM3_t_ts[,1])
ggplot_na_imputations(EM3_t_ts[,1], imp)
```



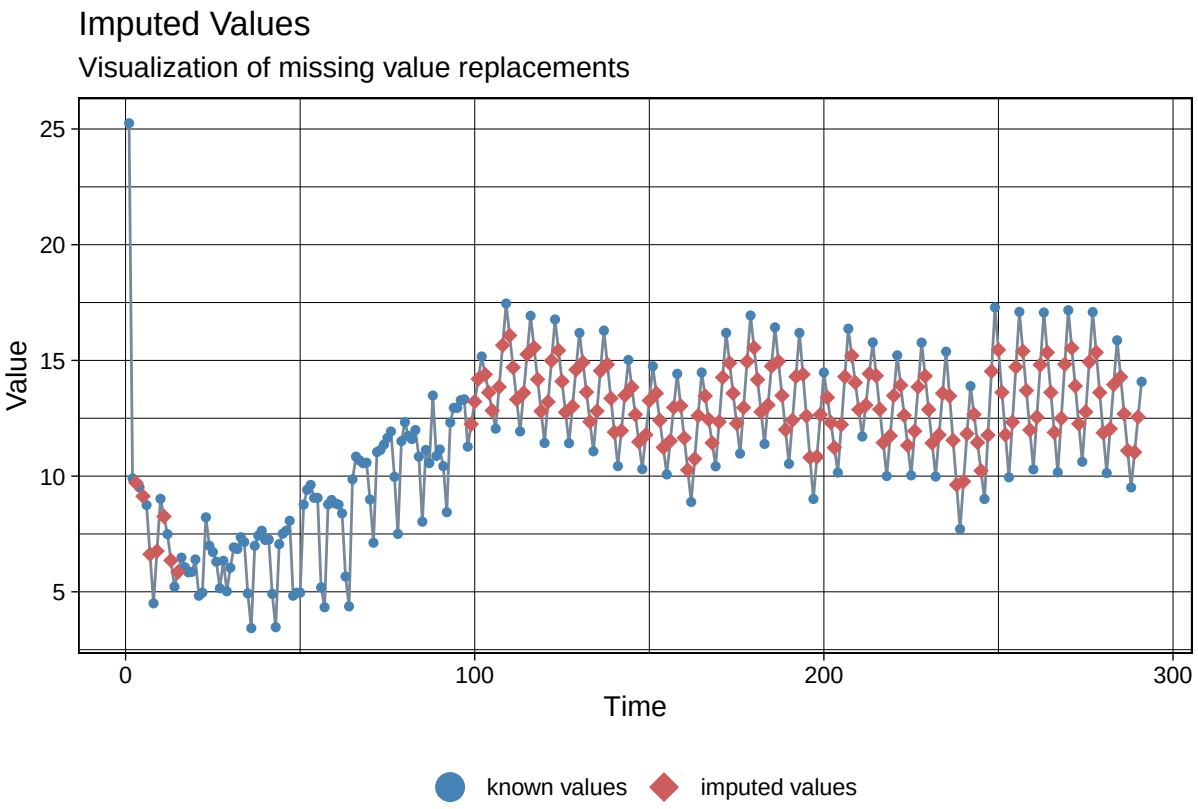

```
imp2 <- na_seadec(EM3_t_ts[,1])
ggplot_na_imputations(EM3_t_ts[,1], imp2)
```



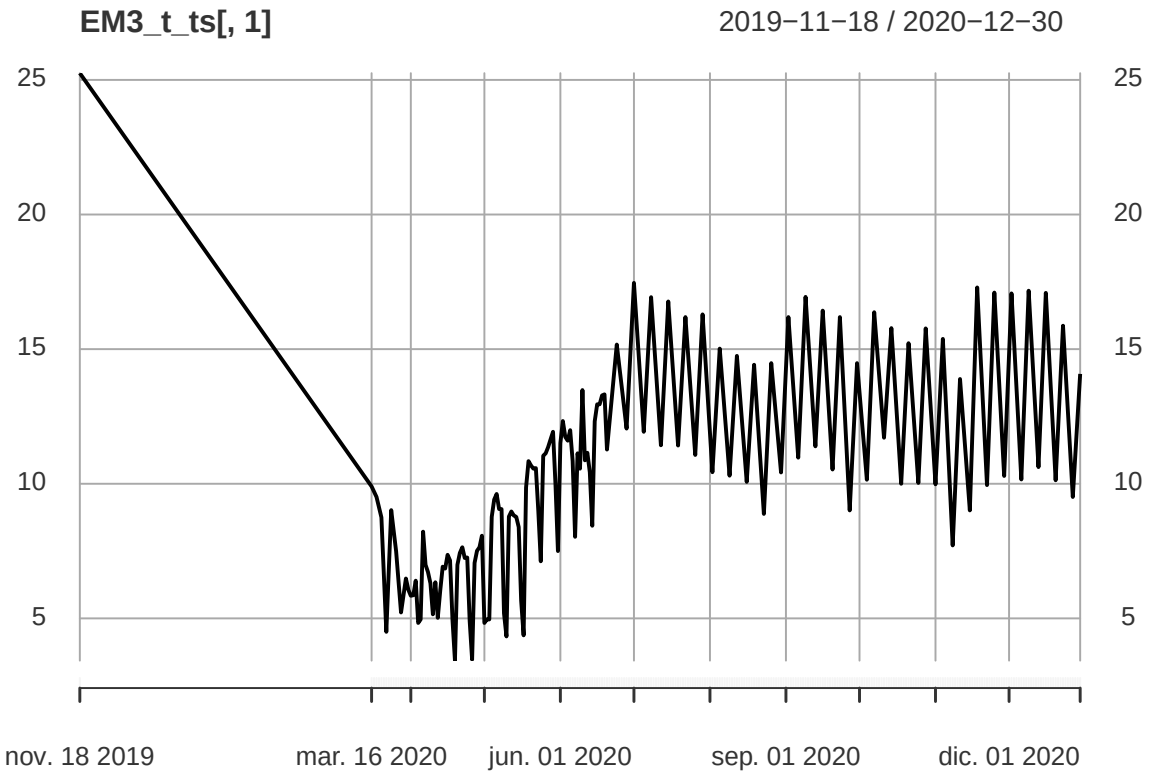

```
imp3 <- na_seasplit(EM3_t_ts[,1])
ggplot_na_imputations(EM3_t_ts[,1], imp3)
```



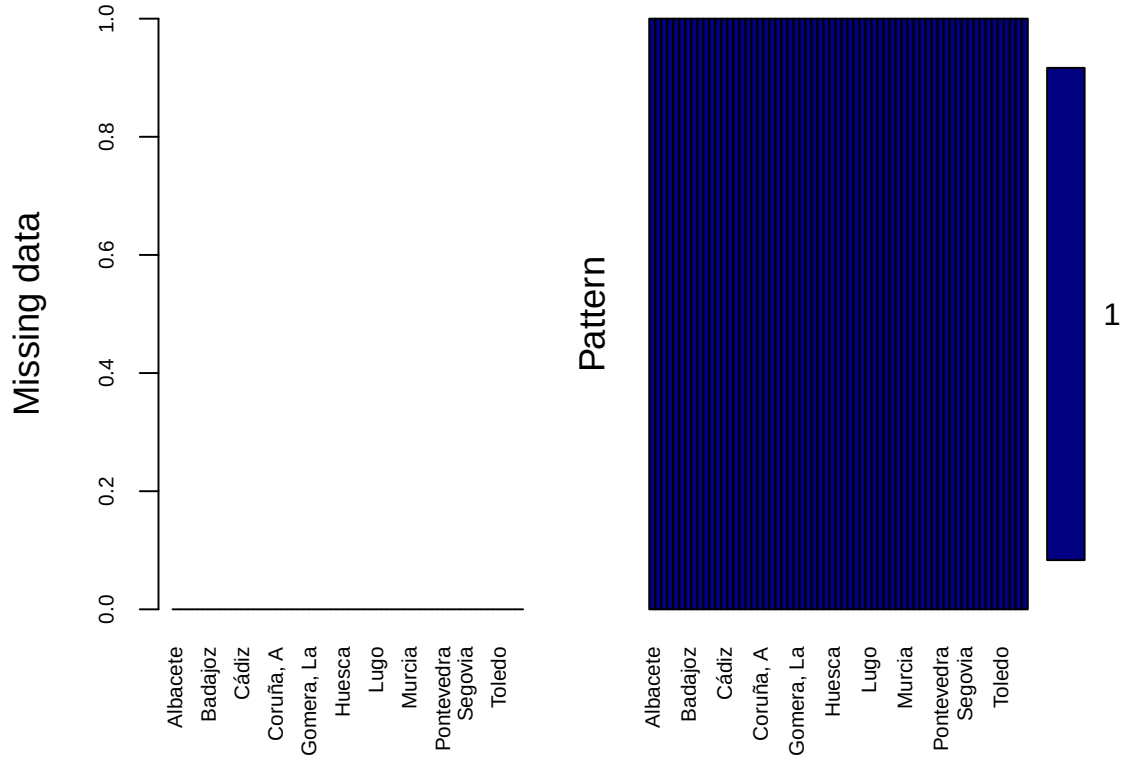

```
imp4 <- na_interpolation(EM3_t_ts[,1])
ggplot_na_imputations(EM3_t_ts[,1], imp4)
```




```
# We select na_seadec for the dataset
EM3_t_ts <- na_seadec(EM3_t_ts)
plot(EM3_t_ts[,1])
```




```
# We convert the time series object to a dataframe
#library(tsbox)
EM3 <- ts_df(EM3_t_ts)

names(EM3)[names(EM3) == "id"] <- "Zonas.de.movilidad"
names(EM3)[names(EM3) == "time"] <- "Periodo"
names(EM3)[names(EM3) == "value"] <- "Total"

# Transpose dataframe
EM3_t<-dcast(EM3, Periodo~Zonas.de.movilidad, fill=NA)

# We check again missing values (result should be zero)
aggr(EM3_t[,-1], col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(EM3_t[,-1]), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```




```
##
##  Variables sorted by number of missings:
##                Variable Count
##                Albacete     0
##       Alicante/Alacant     0
##                Almería     0
##            Araba/Álava     0
##                Asturias     0
##                   Ávila     0
##                Badajoz     0
##         Balears, Illes     0
##              Barcelona     0
##                Bizkaia     0
##                 Burgos     0
##                Cáceres     0
##                  Cádiz     0
##               Cantabria     0
##     Castellón/Castelló     0
##                   Ceuta     0
##             Ciudad Real     0
##                Córdoba     0
##              Coruña, A     0
##                 Cuenca     0
##             Formentera     0
##          Fuerteventura     0
##                Gipuzkoa     0
```

```
##                  Girona     0
##              Gomera, La     0
##            Gran Canaria     0
##                 Granada     0
##             Guadalajara     0
##              Hierro, El     0
##                  Huelva     0
##                  Huesca     0
##                   Ibiza     0
##                   Jaén     0
##               Lanzarote     0
##                   León     0
##                  Lleida     0
##                    Lugo     0
##                  Madrid     0
##                  Málaga     0
##                 Mallorca     0
##                  Melilla     0
##                  Menorca     0
##                   Murcia     0
##                  Navarra     0
##                  Ourense     0
##                 Palencia     0
##               Palma, La     0
##             Palmas, Las     0
##               Pontevedra     0
##                Rioja, La     0
##                Salamanca     0
##  Santa Cruz de Tenerife     0
##                  Segovia     0
##                  Sevilla     0
##                    Soria     0
##                Tarragona     0
##                 Tenerife     0
##                   Teruel     0
##                   Toledo     0
##       Valencia/València     0
##              Valladolid     0
##                  Zamora     0
##                 Zaragoza     0
```

```r
EM3_t %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


```
head(str(EM3_t,vec.len=2))
```

```
## 'data.frame':    291 obs. of  64 variables:
##  $ Periodo               : Date, format: "2019-11-18" "2020-03-16" ...
##  $ Albacete              : num  25.2 9.9 ...
##  $ Alicante/Alacant      : num  28.1 14.4 ...
##  $ Almería               : num  24.4 11 ...
##  $ Araba/Álava           : num  31.9 15.9 ...
##  $ Asturias              : num  29.9 13.1 ...
##  $ Ávila                 : num  26.62 9.44 ...
##  $ Badajoz               : num  19.5 10.7 ...
##  $ Balears, Illes        : num  25.9 12.3 ...
##  $ Barcelona             : num  34.9 14.6 ...
##  $ Bizkaia               : num  33.9 17.9 ...
##  $ Burgos                : num  24.9 11.3 ...
##  $ Cáceres               : num  21.37 9.08 ...
##  $ Cádiz                 : num  24.5 10.7 ...
##  $ Cantabria             : num  31.5 14.5 ...
##  $ Castellón/Castelló    : num  26.9 11.7 ...
##  $ Ceuta                 : num  31.9 12.3 ...
##  $ Ciudad Real           : num  17.07 8.29 ...
##  $ Córdoba               : num  22.4 11.9 ...
##  $ Coruña, A             : num  26.8 15.8 ...
##  $ Cuenca                : num  20.27 8.54 ...
##  $ Formentera            : num  12.98 3.08 ...
##  $ Fuerteventura         : num  11.9 5.76 5.31 4.86 4.62 ...
```

```
##  $ Gipuzkoa              : num  31.3 14.1 ...
##  $ Girona                : num  25.6 10.9 ...
##  $ Gomera, La            : num  11.43 4.26 ...
##  $ Gran Canaria          : num  31.4 16 ...
##  $ Granada               : num  24.2 11.4 ...
##  $ Guadalajara           : num  31.3 12.8 ...
##  $ Hierro, El            : num  14.82 5.75 ...
##  $ Huelva                : num  21.8 11.4 ...
##  $ Huesca                : num  25.5 10.3 ...
##  $ Ibiza                 : num  20.3 10.1 ...
##  $ Jaén                  : num  18.71 8.81 ...
##  $ Lanzarote             : num  25.1 13.6 ...
##  $ León                  : num  29.8 13.2 ...
##  $ Lleida                : num  28.4 11.9 ...
##  $ Lugo                  : num  23.6 12.4 ...
##  $ Madrid                : num  36.7 13.9 ...
##  $ Málaga                : num  23.7 11.4 ...
##  $ Mallorca              : num  28.1 13.3 ...
##  $ Melilla               : num  35 12.5 ...
##  $ Menorca               : num  16.26 7.21 ...
##  $ Murcia                : num  24.6 12.4 ...
##  $ Navarra               : num  30.6 13.7 ...
##  $ Ourense               : num  27.5 15.6 ...
##  $ Palencia              : num  31.1 12.9 ...
##  $ Palma, La             : num  24.6 13.1 ...
##  $ Palmas, Las           : num  28.5 14.6 ...
##  $ Pontevedra            : num  26.2 17.1 ...
##  $ Rioja, La             : num  28 12.1 ...
##  $ Salamanca             : num  28.1 12.9 ...
##  $ Santa Cruz de Tenerife: num  28 13.7 ...
##  $ Segovia               : num  29 12.6 ...
##  $ Sevilla               : num  25.9 12.3 ...
##  $ Soria                 : num  17.67 7.85 ...
##  $ Tarragona             : num  28.4 11.4 ...
##  $ Tenerife              : num  28.9 14.1 ...
##  $ Teruel                : num  16.35 6.77 ...
##  $ Toledo                : num  25.7 12.3 ...
##  $ Valencia/València     : num  31 15.3 ...
##  $ Valladolid            : num  29 14.3 ...
##  $ Zamora                : num  26 11.5 ...
##  $ Zaragoza              : num  32.3 14.8 ...
```

```
## NULL
```

```
summary(EM3_t)
```

```
##     Periodo               Albacete      Alicante/Alacant    Almería
##  Min.   :2019-11-18   Min.   : 3.430   Min.   : 5.72    Min.   : 4.10
##  1st Qu.:2020-05-26   1st Qu.: 9.738   1st Qu.:14.42    1st Qu.:11.31
##  Median :2020-08-07   Median :11.960   Median :17.19    Median :13.84
##  Mean   :2020-08-06   Mean   :11.592   Mean   :16.16    Mean   :13.19
##  3rd Qu.:2020-10-18   3rd Qu.:13.873   3rd Qu.:18.70    3rd Qu.:15.56
##  Max.   :2020-12-30   Max.   :25.250   Max.   :28.12    Max.   :24.35
##   Araba/Álava       Asturias         Ávila          Badajoz
##  Min.   : 6.30   Min.   : 5.24   Min.   : 4.51   Min.   : 4.90
```

```
##  1st Qu.:14.48   1st Qu.:11.93   1st Qu.:10.10   1st Qu.:11.29
##  Median :18.09   Median :16.28   Median :12.71   Median :13.10
##  Mean   :17.52   Mean   :15.65   Mean   :11.93   Mean   :12.74
##  3rd Qu.:21.11   3rd Qu.:19.41   3rd Qu.:14.26   3rd Qu.:14.79
##  Max.   :31.92   Max.   :29.94   Max.   :26.62   Max.   :19.46
##  Balears, Illes    Barcelona        Bizkaia          Burgos
##  Min.   : 3.92   Min.   : 6.42   Min.   : 7.53   Min.   : 3.97
##  1st Qu.:13.09   1st Qu.:13.95   1st Qu.:15.41   1st Qu.:10.74
##  Median :16.14   Median :17.51   Median :19.70   Median :13.06
##  Mean   :15.29   Mean   :17.05   Mean   :19.21   Mean   :12.68
##  3rd Qu.:18.46   3rd Qu.:20.25   3rd Qu.:23.06   3rd Qu.:15.26
##  Max.   :25.95   Max.   :34.90   Max.   :33.93   Max.   :24.88
##     Cáceres          Cádiz         Cantabria     Castellón/Castelló
##  Min.   : 4.02   Min.   : 5.42   Min.   : 5.84   Min.   : 4.95
##  1st Qu.:10.34   1st Qu.:12.59   1st Qu.:13.40   1st Qu.:12.59
##  Median :12.39   Median :15.98   Median :17.57   Median :15.00
##  Mean   :11.81   Mean   :14.92   Mean   :16.93   Mean   :14.43
##  3rd Qu.:13.84   3rd Qu.:17.82   3rd Qu.:20.63   3rd Qu.:16.76
##  Max.   :21.37   Max.   :24.46   Max.   :31.50   Max.   :26.93
##      Ceuta        Ciudad Real        Córdoba        Coruña, A
##  Min.   : 4.87   Min.   : 3.080   Min.   : 5.68   Min.   : 6.02
##  1st Qu.:11.90   1st Qu.: 8.291   1st Qu.:11.49   1st Qu.:12.73
##  Median :13.93   Median :10.050   Median :13.69   Median :15.84
##  Mean   :13.34   Mean   : 9.597   Mean   :13.41   Mean   :15.63
##  3rd Qu.:15.34   3rd Qu.:11.453   3rd Qu.:15.67   3rd Qu.:18.96
##  Max.   :31.88   Max.   :17.070   Max.   :22.38   Max.   :26.84
##      Cuenca         Formentera     Fuerteventura       Gipuzkoa
##  Min.   : 3.010   Min.   : 0.830   Min.   : 1.140   Min.   : 4.71
##  1st Qu.: 9.298   1st Qu.: 2.520   1st Qu.: 5.745   1st Qu.:11.95
##  Median :11.620   Median : 3.562   Median : 7.490   Median :15.49
##  Mean   :10.843   Mean   : 4.559   Mean   : 6.954   Mean   :14.94
##  3rd Qu.:12.892   3rd Qu.: 7.320   3rd Qu.: 8.503   3rd Qu.:18.27
##  Max.   :20.270   Max.   :12.980   Max.   :11.900   Max.   :31.31
##      Girona        Gomera, La      Gran Canaria      Granada
##  Min.   : 4.16   Min.   : 1.340   Min.   : 5.70   Min.   : 5.71
##  1st Qu.:10.24   1st Qu.: 4.860   1st Qu.:15.07   1st Qu.:11.38
##  Median :14.60   Median : 6.635   Median :18.21   Median :14.72
##  Mean   :13.45   Mean   : 6.111   Mean   :17.54   Mean   :14.07
##  3rd Qu.:16.65   3rd Qu.: 7.416   3rd Qu.:20.93   3rd Qu.:16.66
##  Max.   :25.56   Max.   :11.430   Max.   :31.44   Max.   :24.25
##   Guadalajara      Hierro, El         Huelva          Huesca
##  Min.   : 4.89   Min.   : 1.560   Min.   : 5.51   Min.   : 4.24
##  1st Qu.:12.19   1st Qu.: 6.645   1st Qu.:10.94   1st Qu.:11.12
##  Median :14.93   Median : 8.470   Median :13.65   Median :13.77
##  Mean   :14.52   Mean   : 7.933   Mean   :13.23   Mean   :12.96
##  3rd Qu.:17.29   3rd Qu.: 9.710   3rd Qu.:15.62   3rd Qu.:15.14
##  Max.   :31.33   Max.   :14.820   Max.   :21.79   Max.   :25.48
##      Ibiza             Jaén         Lanzarote          León
##  Min.   : 2.960   Min.   : 4.17   Min.   : 4.73   Min.   : 5.56
##  1st Qu.: 9.411   1st Qu.: 9.43   1st Qu.:12.87   1st Qu.:12.58
##  Median :11.550   Median :11.77   Median :14.89   Median :15.65
##  Mean   :12.017   Mean   :11.24   Mean   :14.22   Mean   :14.92
##  3rd Qu.:15.325   3rd Qu.:13.22   3rd Qu.:16.58   3rd Qu.:17.50
##  Max.   :20.650   Max.   :18.71   Max.   :25.07   Max.   :29.80
```

```
##      Lleida           Lugo           Madrid          Málaga
##  Min.   : 5.16   Min.   : 4.88   Min.   : 6.03   Min.   : 4.56
##  1st Qu.:11.73   1st Qu.:10.94   1st Qu.:13.21   1st Qu.:12.25
##  Median :14.89   Median :13.64   Median :16.58   Median :15.41
##  Mean   :14.25   Mean   :13.11   Mean   :16.19   Mean   :14.51
##  3rd Qu.:16.81   3rd Qu.:15.65   3rd Qu.:19.51   3rd Qu.:17.51
##  Max.   :28.38   Max.   :23.60   Max.   :36.70   Max.   :23.71
##     Mallorca        Melilla         Menorca           Murcia
##  Min.   : 4.36   Min.   : 6.48   Min.   : 1.590   Min.   : 5.13
##  1st Qu.:14.31   1st Qu.:13.60   1st Qu.: 8.134   1st Qu.:11.81
##  Median :17.52   Median :15.85   Median :10.110   Median :14.19
##  Mean   :16.44   Mean   :15.37   Mean   :10.846   Mean   :13.97
##  3rd Qu.:19.72   3rd Qu.:17.39   3rd Qu.:15.880   3rd Qu.:16.45
##  Max.   :28.06   Max.   :35.05   Max.   :19.490   Max.   :24.65
##     Navarra         Ourense         Palencia       Palma, La
##  Min.   : 5.33   Min.   : 6.43   Min.   : 5.05   Min.   : 5.28
##  1st Qu.:13.45   1st Qu.:12.87   1st Qu.:12.18   1st Qu.:12.96
##  Median :16.51   Median :15.62   Median :14.78   Median :15.39
##  Mean   :15.89   Mean   :15.35   Mean   :14.40   Mean   :14.79
##  3rd Qu.:18.87   3rd Qu.:18.06   3rd Qu.:16.94   3rd Qu.:17.03
##  Max.   :30.61   Max.   :27.48   Max.   :31.07   Max.   :24.63
##   Palmas, Las      Pontevedra      Rioja, La       Salamanca
##  Min.   : 5.11   Min.   : 6.15   Min.   : 4.74   Min.   : 6.02
##  1st Qu.:13.98   1st Qu.:13.41   1st Qu.:12.51   1st Qu.:12.61
##  Median :16.65   Median :17.23   Median :15.78   Median :15.07
##  Mean   :15.98   Mean   :16.63   Mean   :15.01   Mean   :14.54
##  3rd Qu.:18.99   3rd Qu.:20.39   3rd Qu.:17.73   3rd Qu.:17.00
##  Max.   :28.53   Max.   :26.22   Max.   :28.00   Max.   :28.12
##  Santa Cruz de Tenerife    Segovia         Sevilla          Soria
##  Min.   : 5.20          Min.   : 4.96   Min.   : 5.44   Min.   : 3.03
##  1st Qu.:13.71          1st Qu.:11.83   1st Qu.:12.13   1st Qu.: 8.78
##  Median :16.54          Median :14.48   Median :14.96   Median :10.95
##  Mean   :15.99          Mean   :13.80   Mean   :14.73   Mean   :10.19
##  3rd Qu.:18.98          3rd Qu.:16.27   3rd Qu.:17.70   3rd Qu.:12.20
##  Max.   :28.02          Max.   :29.04   Max.   :25.95   Max.   :17.67
##    Tarragona        Tenerife         Teruel           Toledo
##  Min.   : 4.93   Min.   : 5.32   Min.   : 1.950   Min.   : 4.15
##  1st Qu.:10.64   1st Qu.:14.13   1st Qu.: 6.582   1st Qu.:11.00
##  Median :14.95   Median :17.02   Median : 8.680   Median :13.52
##  Mean   :13.90   Mean   :16.42   Mean   : 8.210   Mean   :13.01
##  3rd Qu.:16.90   3rd Qu.:19.58   3rd Qu.: 9.999   3rd Qu.:15.57
##  Max.   :28.38   Max.   :28.87   Max.   :16.350   Max.   :25.70
##  Valencia/València   Valladolid        Zamora         Zaragoza
##  Min.   : 6.17     Min.   : 5.55   Min.   : 4.82   Min.   : 5.80
##  1st Qu.:15.39     1st Qu.:12.60   1st Qu.:10.79   1st Qu.:13.44
##  Median :18.16     Median :15.95   Median :13.09   Median :16.45
##  Mean   :17.53     Mean   :15.71   Mean   :12.79   Mean   :16.37
##  3rd Qu.:20.38     3rd Qu.:19.10   3rd Qu.:14.95   3rd Qu.:19.37
##  Max.   :30.97     Max.   :28.99   Max.   :25.99   Max.   :32.26
```

```
#library(DataExplorer)
#plot_histogram(EM3_t)
table(EM3$Zonas.de.movilidad)
```

```
##
```

```
##               Albacete       Alicante/Alacant                Almería
##                    291                    291                    291
##           Araba/Álava               Asturias                  Ávila
##                    291                    291                    291
##               Badajoz         Balears, Illes              Barcelona
##                    291                    291                    291
##               Bizkaia                 Burgos                Cáceres
##                    291                    291                    291
##                 Cádiz              Cantabria     Castellón/Castelló
##                    291                    291                    291
##                 Ceuta            Ciudad Real                Córdoba
##                    291                    291                    291
##             Coruña, A                 Cuenca             Formentera
##                    291                    291                    291
##         Fuerteventura              Gipuzkoa                Girona
##                    291                    291                    291
##            Gomera, La           Gran Canaria                Granada
##                    291                    291                    291
##           Guadalajara             Hierro, El                 Huelva
##                    291                    291                    291
##                Huesca                  Ibiza                  Jaén
##                    291                    291                    291
##              Lanzarote                  León                 Lleida
##                    291                    291                    291
##                  Lugo                Madrid                 Málaga
##                    291                    291                    291
##              Mallorca                Melilla                Menorca
##                    291                    291                    291
##                Murcia                Navarra                Ourense
##                    291                    291                    291
##              Palencia             Palma, La           Palmas, Las
##                    291                    291                    291
##            Pontevedra             Rioja, La              Salamanca
##                    291                    291                    291
## Santa Cruz de Tenerife               Segovia                Sevilla
##                    291                    291                    291
##                 Soria              Tarragona               Tenerife
##                    291                    291                    291
##                Teruel                 Toledo      Valencia/València
##                    291                    291                    291
##            Valladolid                 Zamora               Zaragoza
##                    291                    291                    291
```

2.1.5 Google review

Here we have data mobility from autonomus-communities and provinces.

```
#Source Google
summary(Google)
```

```
##  country_region_code country_region     sub_region_1       sub_region_2
##  Length:24242        Length:24242       Length:24242       Length:24242
##  Class :character    Class :character   Class :character   Class :character
##  Mode  :character    Mode  :character   Mode  :character   Mode  :character
##
```

```
##
##
##
##  metro_area     iso_3166_2_code    census_fips_code   place_id
##  Mode:logical   Length:24242       Mode:logical     Length:24242
##  NA's:24242     Class :character   NA's:24242       Class :character
##                 Mode  :character                    Mode  :character
##
##
##
##
##      date           retail_and_recreation_percent_change_from_baseline
##  Length:24242       Min.   :-97.00
##  Class :character   1st Qu.:-53.00
##  Mode  :character   Median :-32.00
##                     Mean   :-36.42
##                     3rd Qu.:-17.00
##                     Max.   : 71.00
##                     NA's   :56
##  grocery_and_pharmacy_percent_change_from_baseline
##  Min.   :-96.000
##  1st Qu.:-18.000
##  Median : -4.000
##  Mean   : -9.973
##  3rd Qu.:  3.000
##  Max.   :194.000
##  NA's   :396
##  parks_percent_change_from_baseline
##  Min.   :-94.0000
##  1st Qu.:-30.0000
##  Median : -5.0000
##  Mean   : -0.0038
##  3rd Qu.: 22.0000
##  Max.   :543.0000
##  NA's   :305
##  transit_stations_percent_change_from_baseline
##  Min.   :-100.00
##  1st Qu.: -46.00
##  Median : -30.00
##  Mean   : -32.63
##  3rd Qu.: -16.00
##  Max.   : 177.00
##  NA's   :832
##  workplaces_percent_change_from_baseline
##  Min.   :-92.0
##  1st Qu.:-37.0
##  Median :-24.0
##  Mean   :-26.7
##  3rd Qu.:-13.0
##  Max.   : 55.0
##  NA's   :42
##  residential_percent_change_from_baseline
##  Min.   :-10.000
##  1st Qu.:  4.000
```

```
##  Median :  7.000
##  Mean   :  9.419
##  3rd Qu.: 13.000
##  Max.   : 48.000
##  NA's   :267
```

```
head(str(Google,vec.len=1))
```

```
## 'data.frame':    24242 obs. of  15 variables:
##  $ country_region_code                                : chr  "ES" ...
##  $ country_region                                     : chr  "Spain" ...
##  $ sub_region_1                                       : chr  "" ...
##  $ sub_region_2                                       : chr  "" ...
##  $ metro_area                                         : logi  NA ...
##  $ iso_3166_2_code                                    : chr  "" ...
##  $ census_fips_code                                   : logi  NA ...
##  $ place_id                                           : chr  "ChIJi7xhMnjjQgwR7KNoB5Qs7KY" ...
##  $ date                                               : chr  "15/02/2020" ...
##  $ retail_and_recreation_percent_change_from_baseline: int  2 2 ...
##  $ grocery_and_pharmacy_percent_change_from_baseline : int  -1 3 ...
##  $ parks_percent_change_from_baseline                 : int  26 13 ...
##  $ transit_stations_percent_change_from_baseline      : int  8 5 ...
##  $ workplaces_percent_change_from_baseline            : int  0 -1 ...
##  $ residential_percent_change_from_baseline           : int  -2 -2 ...
```

```
## NULL
```

```
table(Google$sub_region_1)
```

```
##
##                              Andalusia                 Aragon               Asturias
##                   385                3465                 1540                   385
##    Balearic Islands      Basque Country      Canary Islands             Cantabria
##                   385                1540                 1155                   385
##   Castile-La Mancha   Castile and LeÃ³n           Catalonia                  Ceuta
##                  2310                3850                 1925                   378
## Community of Madrid         Extremadura              Galicia              La Rioja
##                   385                1155                 1925                   385
##              Melilla             Navarre    Region of Murcia Valencian Community
##                   379                 385                  385                  1540
```

```
table(Google$sub_region_2)
```

```
##
##                                  A CoruÃ±a                 Ã\201lava
##                    7687                  385                    385
##                 Ã\201vila                Albacete               Alicante
##                    385                   385                    385
##               AlmerÃa                Badajoz              Barcelona
##                    385                   385                    385
##                  Biscay                 Burgos              CÃ¡ceres
##                    385                   385                    385
##                 CÃ¡diz                CÃ³rdoba            CastellÃ³n
##                    385                   385                    385
##             Ciudad Real                 Cuenca              Gipuzkoa
##                    385                   385                    385
##                 Girona                Granada           Guadalajara
```

```
##                   385                   385                   385
##                Huelva                Huesca                JaÃ©n
##                   385                   385                   385
##            Las Palmas                LeÃ³n               Lleida
##                   385                   385                   385
##                 Lugo                MÃ¡laga              Palencia
##                   385                   385                   385
##            Pontevedra    Province of Ourense             Salamanca
##                   385                   385                   385
## Santa Cruz de Tenerife               Segovia               Seville
##                   385                   385                   385
##                Soria              Tarragona                Teruel
##                   385                   385                   385
##               Toledo               Valencia            Valladolid
##                   385                   385                   385
##               Zamora               Zaragoza
##                   385                   385
```

```
table(Google$iso_3166_2_code)
```

```
##
##        ES-A ES-AB ES-AL ES-AN ES-AR ES-AS ES-AV  ES-B ES-BA ES-BI ES-BU  ES-C
##   385   385   385   385   385   385   385   385   385   385   385   385   385
## ES-CA ES-CB ES-CC ES-CE ES-CL ES-CM ES-CN ES-CO ES-CR ES-CS ES-CT ES-CU ES-EX
##   385   385   385   378   385   385   385   385   385   385   385   385   385
## ES-GA ES-GC ES-GI ES-GR ES-GU  ES-H ES-HU ES-IB  ES-J  ES-L ES-LE ES-LU ES-MA
##   385   385   385   385   385   385   385   385   385   385   385   385   385
## ES-MC ES-MD ES-ML ES-NC ES-OR  ES-P ES-PO ES-PV ES-RI ES-SA ES-SE ES-SG ES-SO
##   385   385   379   385   385   385   385   385   385   385   385   385   385
## ES-SS  ES-T ES-TE ES-TF ES-TO  ES-V ES-VA ES-VC ES-VI  ES-Z ES-ZA
##   385   385   385   385   385   385   385   385   385   385   385
```

2.1.6 Google autonomous-communities & provinces

We check data grouped by autonomous communities and provinces.

```
Google %>% group_by(sub_region_1) %>% tally()
```

```
## # A tibble: 20 x 2
##    sub_region_1              n
##    <chr>                 <int>
##  1 ""                      385
##  2 "Andalusia"            3465
##  3 "Aragon"               1540
##  4 "Asturias"              385
##  5 "Balearic Islands"      385
##  6 "Basque Country"       1540
##  7 "Canary Islands"       1155
##  8 "Cantabria"             385
##  9 "Castile-La Mancha"    2310
## 10 "Castile and LeÃ³n"    3850
## 11 "Catalonia"           1925
## 12 "Ceuta"                378
## 13 "Community of Madrid"  385
## 14 "Extremadura"         1155
## 15 "Galicia"             1925
```

```
## 16 "La Rioja"              385
## 17 "Melilla"               379
## 18 "Navarre"               385
## 19 "Region of Murcia"      385
## 20 "Valencian Community"  1540
```

```
Google %>% group_by(sub_region_1) %>% count(sub_region_2)
```

```
## # A tibble: 63 x 3
## # Groups:   sub_region_1 [20]
##    sub_region_1 sub_region_2     n
##    <chr>        <chr>        <int>
##  1 ""           ""             385
##  2 "Andalusia"  ""             385
##  3 "Andalusia"  "AlmerÃa"      385
##  4 "Andalusia"  "CÃ¡diz"       385
##  5 "Andalusia"  "CÃ³rdoba"     385
##  6 "Andalusia"  "Granada"      385
##  7 "Andalusia"  "Huelva"       385
##  8 "Andalusia"  "JaÃ©n"        385
##  9 "Andalusia"  "MÃ¡laga"      385
## 10 "Andalusia"  "Seville"      385
## # ... with 53 more rows
```

In Spain there are **autonomous communities (AC)** and **autonomous cities (C)** that are considered as **provinces (Pr)**. This is the case for:

- AC - Asturias, Principality - Pr - Asturias
- AC - Balears, Illes - Pr - Balears, Illes
- AC - Cantabria - Pr - Cantabria
- AC - Madrid, Community - Pr - Madrid
- AC - Murcia, Region - Pr- Murcia
- AC - Navarra, Foral Community - Pr - Navarra
- AC - Rioja, La - Pr - Rioja, La
- C - Ceuta - C/Pr - Ceuta
- C - Melilla - C/Pr - Melilla

In this data set, the empty values in the "sub_region_2" column, for the autonomous communities mentinoed, will be replaced by the value contained in the "sub_region_1" column (A). Also we are going to modify the names of the provinces that have special characters in order to adopt the INE standards (B). See note.

Note The following links states the provinces in Spain INE CCAA and its ISO codes are going to be used as tables of referencence.

```
# Modification provinces - A
Google$sub_region_2[Google$sub_region_1=="Balearic Islands"] <- "Balears, Illes"
Google$iso_3166_2_code[Google$sub_region_2=="Balears, Illes"] <- "PM"

Google$sub_region_2[Google$sub_region_1=="Asturias"] <- "Asturias"
Google$iso_3166_2_code[Google$sub_region_2=="Asturias"] <- "O"

Google$sub_region_2[Google$sub_region_1=="Cantabria"] <- "Cantabria"
Google$iso_3166_2_code[Google$sub_region_2=="Cantabria"] <- "S"

Google$sub_region_2[Google$sub_region_1=="Community of Madrid"] <- "Madrid"
Google$iso_3166_2_code[Google$sub_region_2=="Madrid"] <- "M"
```

```r
Google$sub_region_2[Google$sub_region_1=="Region of Murcia"] <- "Murcia"
Google$iso_3166_2_code[Google$sub_region_2=="Murcia"] <- "MU"

Google$sub_region_2[Google$sub_region_1=="Navarre"] <- "Navarra"
Google$iso_3166_2_code[Google$sub_region_2=="Navarra"] <- "NA"

Google$sub_region_2[Google$sub_region_1=="La Rioja"] <- "Rioja, La"
Google$iso_3166_2_code[Google$sub_region_2=="Rioja, La"] <- "LO"

Google$sub_region_2[Google$sub_region_1=="Ceuta"] <- "Ceuta"
Google$iso_3166_2_code[Google$sub_region_2=="Ceuta"] <- "CE"

Google$sub_region_2[Google$sub_region_1=="Melilla"] <- "Melilla"
Google$iso_3166_2_code[Google$sub_region_2=="Melilla"] <- "ML"

# Modification provinces - B
Google$sub_region_2[Google$sub_region_2=="A CoruÃ±a"]<-"Coruña, A"
Google$sub_region_2[Google$sub_region_2=="Ã\u0081lava"]<-"Araba/Álava"
Google$sub_region_2[Google$sub_region_2=="Ã\u0081vila"]<-"Ávila"
Google$sub_region_2[Google$sub_region_2=="Alicante"]<-"Alicante/Alacant"
Google$sub_region_2[Google$sub_region_2=="Biscay"]<-"Bizkaia"
Google$sub_region_2[Google$sub_region_2=="CÃ¡ceres"]<-"Cáceres"
Google$sub_region_2[Google$sub_region_2=="CÃ¡diz"]<-"Cádiz"
Google$sub_region_2[Google$sub_region_2=="CÃ³rdoba"]<-"Córdoba"
Google$sub_region_2[Google$sub_region_2=="CastellÃ³n"]<-"Castellón/Castelló"
Google$sub_region_2[Google$sub_region_2=="JaÃ©n"]<-"Jaén"
Google$sub_region_2[Google$sub_region_2=="Las Palmas"]<-"Palmas, Las"
Google$sub_region_2[Google$sub_region_2=="LeÃ³n"]<-"León"
Google$sub_region_2[Google$sub_region_2=="MÃ¡laga"]<-"Málaga"
Google$sub_region_2[Google$sub_region_2=="Province of Ourense"]<-"Ourense"
Google$sub_region_2[Google$sub_region_2=="Seville"]<-"Sevilla"
Google$sub_region_2[Google$sub_region_2=="Valencia"]<-"Valencia/València"
Google$sub_region_2 <- with(Google, ifelse(grepl("^Almer", sub_region_2),
                                           "Almería", sub_region_2))
# Table check
table(Google$sub_region_2)
```

```
##
##                                Albacete       Alicante/Alacant
##                    4235             385                    385
##                  Almería     Araba/Álava               Asturias
##                     385             385                    385
##                   Ávila          Badajoz         Balears, Illes
##                     385             385                    385
##               Barcelona         Bizkaia                 Burgos
##                     385             385                    385
##                  Cáceres          Cádiz              Cantabria
##                     385             385                    385
##     Castellón/Castelló           Ceuta            Ciudad Real
##                     385             378                    385
##                  Córdoba      Coruña, A                 Cuenca
##                     385             385                    385
##                Gipuzkoa         Girona                Granada
##                     385             385                    385
```

```
##            Guadalajara                 Huelva                 Huesca
##                    385                    385                    385
##                  Jaén                  León                 Lleida
##                    385                    385                    385
##                  Lugo                 Madrid                 Málaga
##                    385                    385                    385
##                Melilla                 Murcia                Navarra
##                    379                    385                    385
##                Ourense               Palencia           Palmas, Las
##                    385                    385                    385
##             Pontevedra             Rioja, La              Salamanca
##                    385                    385                    385
## Santa Cruz de Tenerife                Segovia                Sevilla
##                    385                    385                    385
##                  Soria              Tarragona                 Teruel
##                    385                    385                    385
##                 Toledo      Valencia/València             Valladolid
##                    385                    385                    385
##                Zamora                Zaragoza
##                    385                    385
```

```
table(Google$iso_3166_2_code)
```

```
##
##          CE  ES-A ES-AB ES-AL ES-AN ES-AR ES-AV  ES-B ES-BA ES-BI ES-BU  ES-C
##   385   378   385   385   385   385   385   385   385   385   385   385   385
## ES-CA ES-CC ES-CL ES-CM ES-CN ES-CO ES-CR ES-CS ES-CT ES-CU ES-EX ES-GA ES-GC
##   385   385   385   385   385   385   385   385   385   385   385   385   385
## ES-GI ES-GR ES-GU  ES-H ES-HU  ES-J  ES-L ES-LE ES-LU ES-MA ES-OR  ES-P ES-PO
##   385   385   385   385   385   385   385   385   385   385   385   385   385
## ES-PV ES-SA ES-SE ES-SG ES-SO ES-SS  ES-T ES-TE ES-TF ES-TO  ES-V ES-VA ES-VC
##   385   385   385   385   385   385   385   385   385   385   385   385   385
## ES-VI  ES-Z ES-ZA    LO     M    ML    MU    NA     O    PM     S
##   385   385   385   385   385   379   385   385   385   385   385
```

2.1.7 Google data transformation

We are going to **transform / eliminate**:

- A - Rows with “na” / "" in “sub_region_1” and “sub_region_2” columns are eliminated.
- B - Date column is transformed from “character” to “date”.
- C - Some columns are eliminated due to they are not adding value or they contain blanks (country_region_code, country_region, metro_area, census_fips_code, pace_id).
- D - “ES-” is elimianted from “iso_3166_2_code” column.
- E - We changed from integer to numeric, integer columns.

```
# Transform / eliminate A
Google <- filter(Google, sub_region_1 != "", sub_region_2 != "" )

# Transform / eliminate B
Google$date <- as.Date(Google$date ,format="%d/%m/%Y")

# Transform / eliminate C
Google<-within(Google, rm(country_region_code,
                  country_region,
                  metro_area,
```

```
                  census_fips_code,
                  place_id))

# Transform / eliminate D
Google$iso_3166_2_code <- gsub("ES-", "", Google$iso_3166_2_code)

# We pass from integer to numeric
Google$retail_and_recreation_percent_change_from_baseline <-
  as.numeric(Google$retail_and_recreation_percent_change_from_baseline)
Google$grocery_and_pharmacy_percent_change_from_baseline <-as.numeric(Google$grocery_and_pharmacy_perce
Google$parks_percent_change_from_baseline <-as.numeric(Google$parks_percent_change_from_baseline)
Google$transit_stations_percent_change_from_baseline <-as.numeric(Google$transit_stations_percent_chang
Google$workplaces_percent_change_from_baseline <-as.numeric(Google$workplaces_percent_change_from_basel
Google$residential_percent_change_from_baseline <-as.numeric(Google$residential_percent_change_from_bas

# Check table
head(Google,5)
```

```
##   sub_region_1 sub_region_2 iso_3166_2_code       date
## 1    Andalusia      Almería              AL 2020-02-15
## 2    Andalusia      Almería              AL 2020-02-16
## 3    Andalusia      Almería              AL 2020-02-17
## 4    Andalusia      Almería              AL 2020-02-18
## 5    Andalusia      Almería              AL 2020-02-19
##   retail_and_recreation_percent_change_from_baseline
## 1                                                  5
## 2                                                 -2
## 3                                                  0
## 4                                                 -3
## 5                                                 -1
##   grocery_and_pharmacy_percent_change_from_baseline
## 1                                                -3
## 2                                                 0
## 3                                                -2
## 4                                                -3
## 5                                                -3
##   parks_percent_change_from_baseline
## 1                                 40
## 2                                 -2
## 3                                  3
## 4                                 -2
## 5                                  3
##   transit_stations_percent_change_from_baseline
## 1                                            10
## 2                                             1
## 3                                             5
## 4                                             5
## 5                                             4
##   workplaces_percent_change_from_baseline
## 1                                       1
## 2                                       1
## 3                                       3
## 4                                       3
## 5                                       3
```

```
##   residential_percent_change_from_baseline
## 1                                       -2
## 2                                       -1
## 3                                       -1
## 4                                        0
## 5                                        0
```

```
table(Google$sub_region_2)
```

```
##
##               Albacete       Alicante/Alacant                Almería
##                    385                    385                    385
##            Araba/Álava               Asturias                  Ávila
##                    385                    385                    385
##                Badajoz         Balears, Illes              Barcelona
##                    385                    385                    385
##                Bizkaia                 Burgos                Cáceres
##                    385                    385                    385
##                 Cádiz              Cantabria     Castellón/Castelló
##                    385                    385                    385
##                 Ceuta            Ciudad Real                Córdoba
##                    378                    385                    385
##              Coruña, A                 Cuenca              Gipuzkoa
##                    385                    385                    385
##                Girona                Granada           Guadalajara
##                    385                    385                    385
##                Huelva                Huesca                  Jaén
##                    385                    385                    385
##                 León                 Lleida                  Lugo
##                    385                    385                    385
##               Madrid                 Málaga                Melilla
##                    385                    385                    379
##               Murcia                Navarra                Ourense
##                    385                    385                    385
##              Palencia            Palmas, Las             Pontevedra
##                    385                    385                    385
##             Rioja, La              Salamanca Santa Cruz de Tenerife
##                    385                    385                    385
##               Segovia                Sevilla                  Soria
##                    385                    385                    385
##              Tarragona                Teruel                 Toledo
##                    385                    385                    385
##      Valencia/València             Valladolid                 Zamora
##                    385                    385                    385
##               Zaragoza
##                    385
```

```
table(Google$iso_3166_2_code)
```

```
##
##   A  AB  AL  AV   B  BA  BI  BU   C  CA  CC  CE  CO  CR  CS  CU  GC  GI  GR  GU
## 385 385 385 385 385 385 385 385 385 385 385 378 385 385 385 385 385 385 385 385
##   H  HU   J   L  LE  LO  LU   M  MA  ML  MU  NA   O  OR   P  PM  PO   S  SA  SE
## 385 385 385 385 385 385 385 385 385 379 385 385 385 385 385 385 385 385 385 385
##  SG  SO  SS   T  TE  TF  TO   V  VA  VI   Z  ZA
```

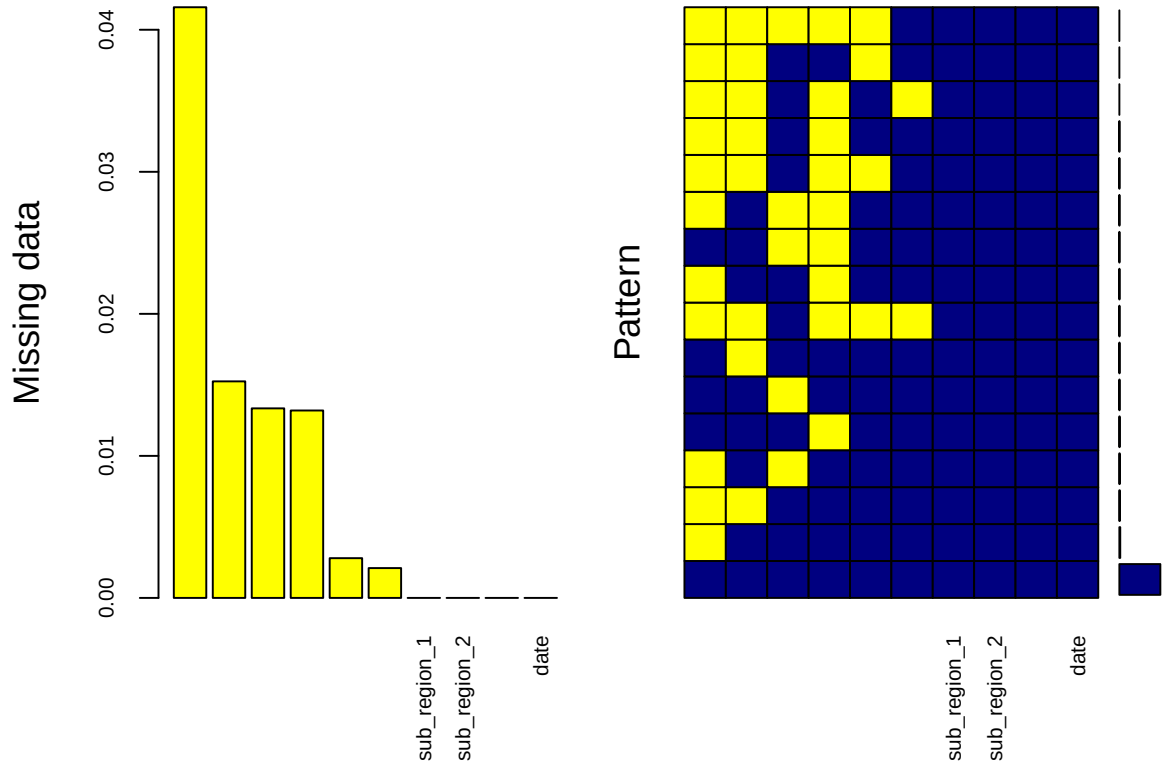
```
## 385 385 385 385 385 385 385 385 385 385 385 385
```

```
#unique(Google$sub_region_2)
#unique(EM3$Zonas.de.movilidad)
```

2.1.8 Google review missing values & impute

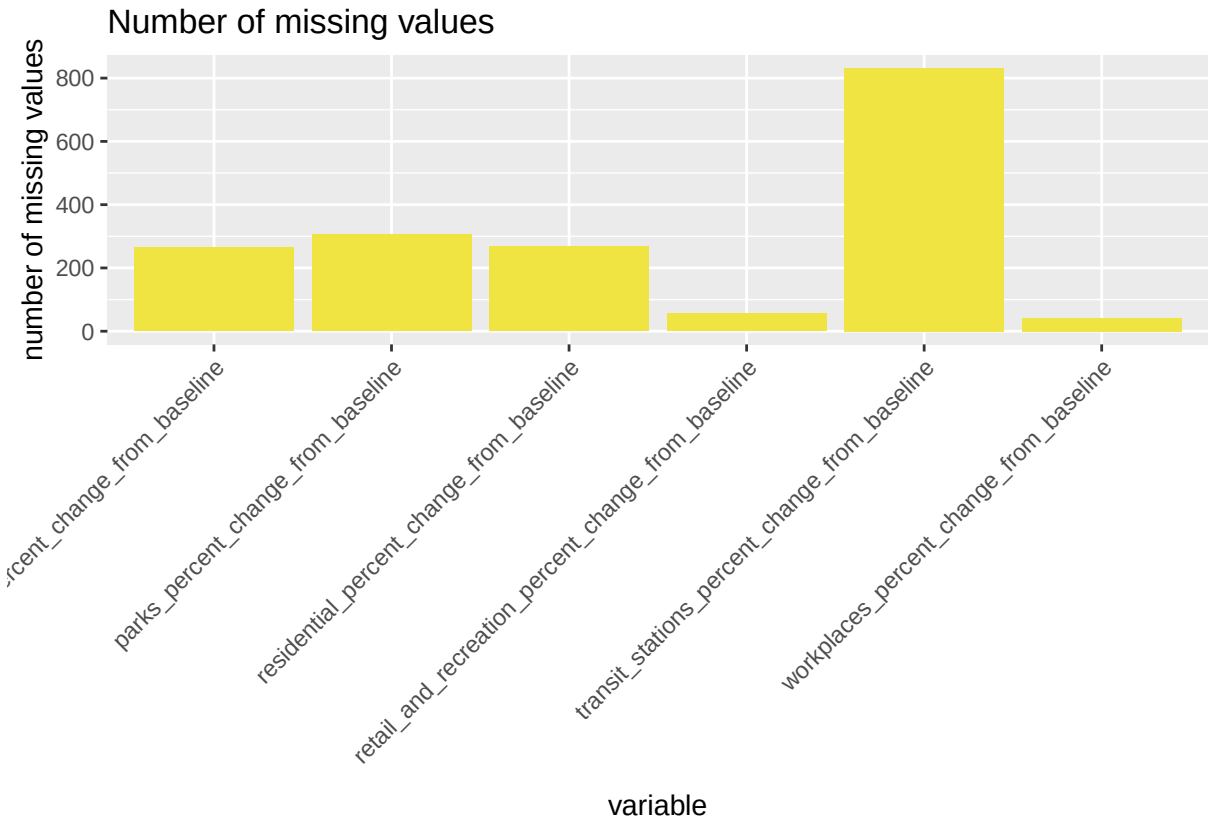
We check missing values.

```
aggr(Google, col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(Google), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```



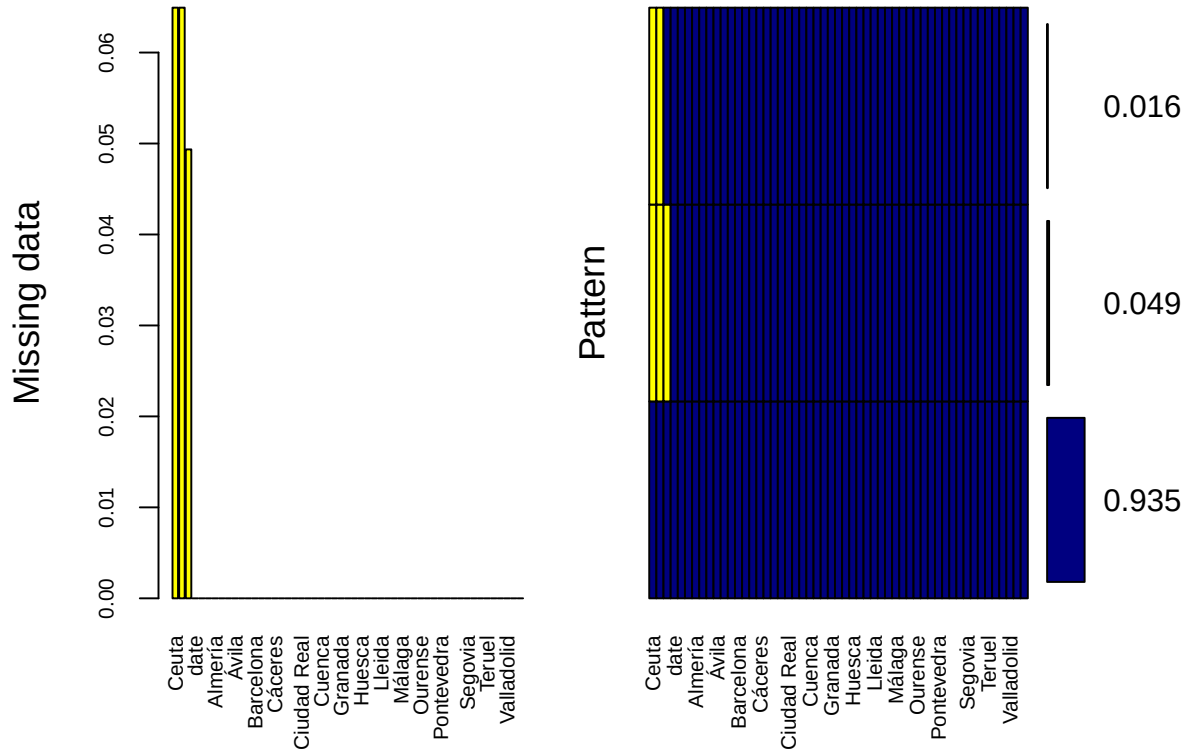

```
##
##  Variables sorted by number of missings:
##                                              Variable       Count
##       transit_stations_percent_change_from_baseline 0.041585445
##                    parks_percent_change_from_baseline 0.015244664
##             residential_percent_change_from_baseline 0.013345329
##    grocery_and_pharmacy_percent_change_from_baseline 0.013195382
##  retail_and_recreation_percent_change_from_baseline 0.002799020
##              workplaces_percent_change_from_baseline 0.002099265
##                                          sub_region_1 0.000000000
##                                          sub_region_2 0.000000000
##                                       iso_3166_2_code 0.000000000
##                                                  date 0.000000000
```

```
Google %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



We generate 6 new dataframes from the 6 features stated in order to imput missing values by province using the approach stated at "imputeTS" library (and also used at EM3).

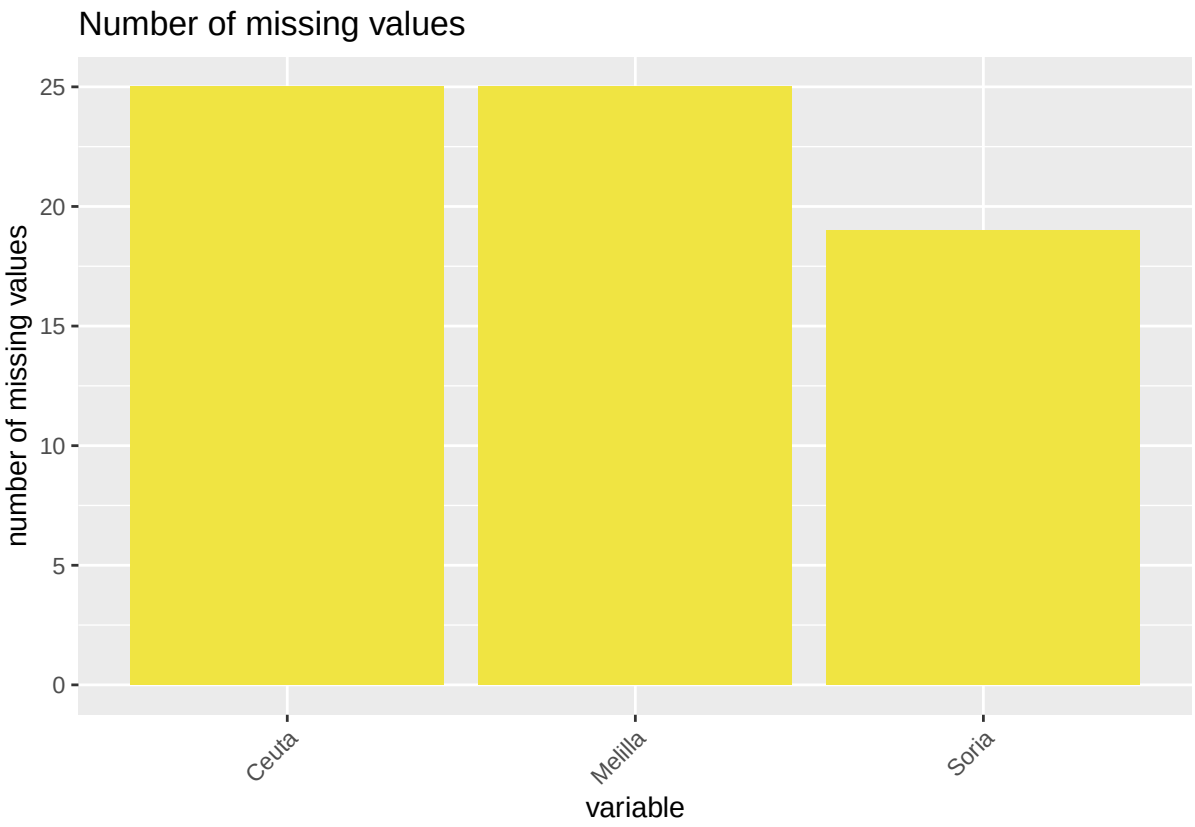
```
# Transpose dataframe
Google_retail<-Google[c(2,4,5)]
Google_t_retail<-dcast(Google_retail, date~sub_region_2, fill=NA)

# Visualize missing values
aggr(Google_t_retail, col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(Google_t_retail), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```



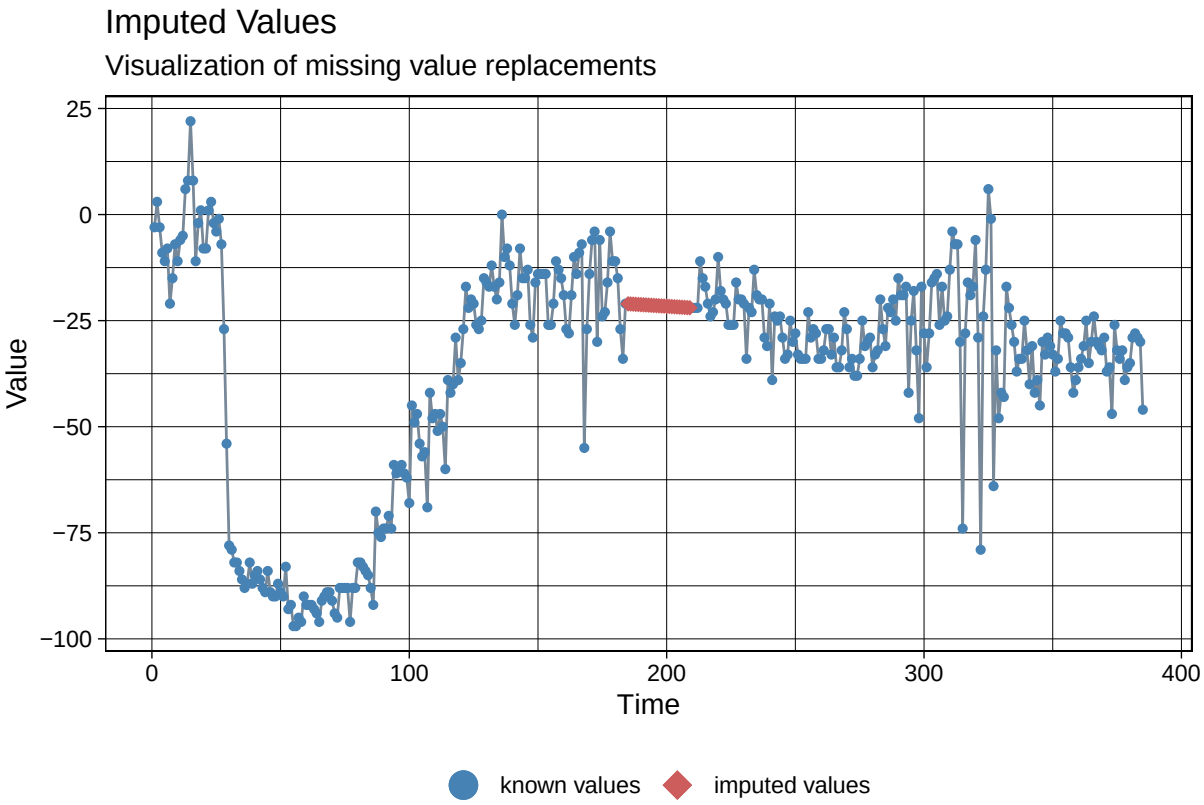

```
##
##  Variables sorted by number of missings:
##                 Variable      Count
##                    Ceuta 0.06493506
##                  Melilla 0.06493506
##                    Soria 0.04935065
##                     date 0.00000000
##                 Albacete 0.00000000
##         Alicante/Alacant 0.00000000
##                  Almería 0.00000000
##              Araba/Álava 0.00000000
##                 Asturias 0.00000000
##                    Ávila 0.00000000
##                  Badajoz 0.00000000
##           Balears, Illes 0.00000000
##                Barcelona 0.00000000
##                  Bizkaia 0.00000000
##                   Burgos 0.00000000
##                  Cáceres 0.00000000
##                    Cádiz 0.00000000
##                Cantabria 0.00000000
##       Castellón/Castelló 0.00000000
##              Ciudad Real 0.00000000
##                  Córdoba 0.00000000
##                Coruña, A 0.00000000
##                   Cuenca 0.00000000
```

```
##                Gipuzkoa 0.00000000
##                  Girona 0.00000000
##                 Granada 0.00000000
##             Guadalajara 0.00000000
##                  Huelva 0.00000000
##                  Huesca 0.00000000
##                   Jaén 0.00000000
##                   León 0.00000000
##                  Lleida 0.00000000
##                   Lugo 0.00000000
##                  Madrid 0.00000000
##                  Málaga 0.00000000
##                  Murcia 0.00000000
##                 Navarra 0.00000000
##                 Ourense 0.00000000
##                Palencia 0.00000000
##             Palmas, Las 0.00000000
##              Pontevedra 0.00000000
##               Rioja, La 0.00000000
##                Salamanca 0.00000000
##  Santa Cruz de Tenerife 0.00000000
##                  Segovia 0.00000000
##                  Sevilla 0.00000000
##                Tarragona 0.00000000
##                   Teruel 0.00000000
##                  Toledo 0.00000000
##       Valencia/València 0.00000000
##              Valladolid 0.00000000
##                  Zamora 0.00000000
##                Zaragoza 0.00000000
```

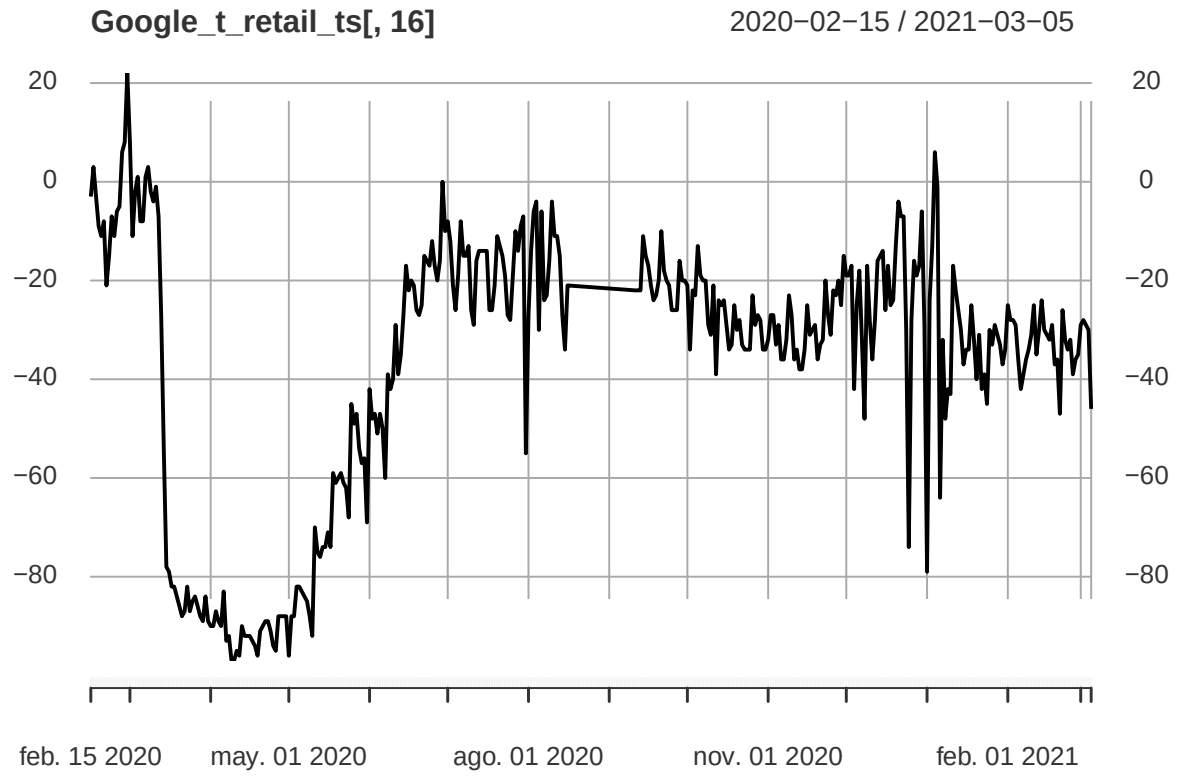
```
Google_t_retail %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



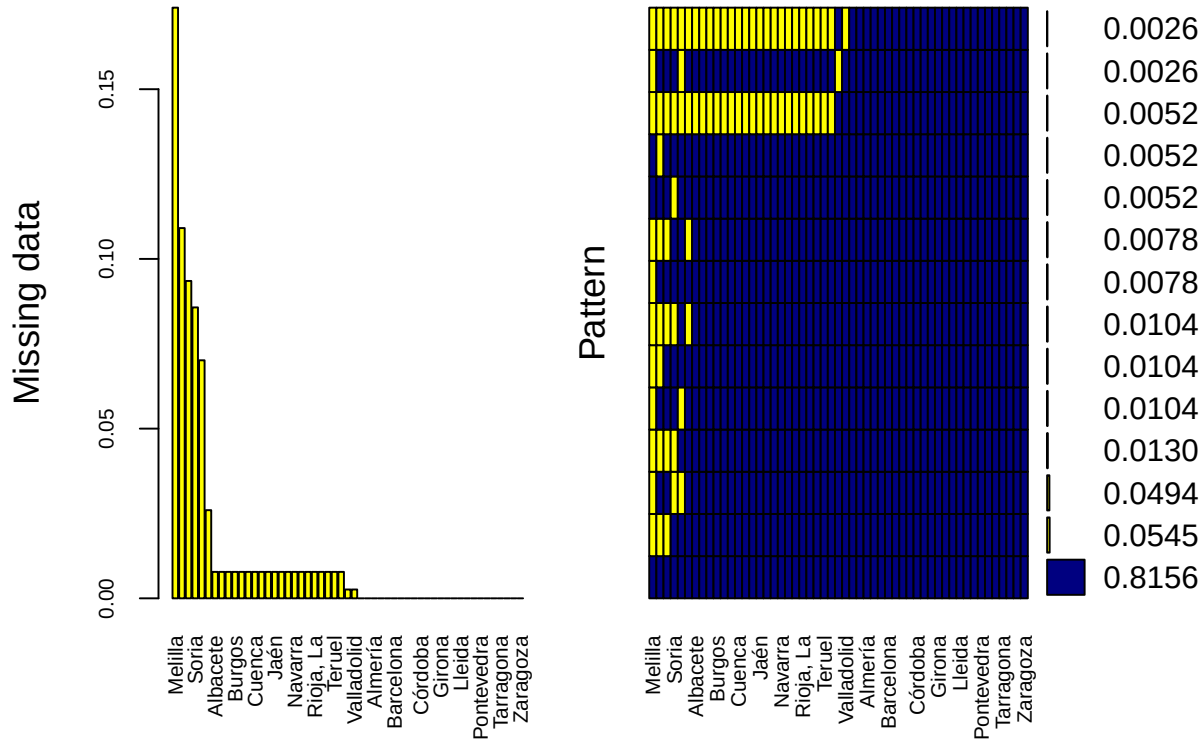

```
# Convert dataframe to ts object
Google_t_retail_ts<-xts(Google_t_retail[-1],Google_t_retail$date)

# Impute the missing values with na_seadec (i.e Ceuta)
imp5 <- na_seadec(Google_t_retail_ts[,16])
ggplot_na_imputations(Google_t_retail_ts[,16], imp5)
```




```
# We select na_seadec for the dataset
Google_t_retail_ts <- na_seadec(Google_t_retail_ts)
plot(Google_t_retail_ts[,16])
```



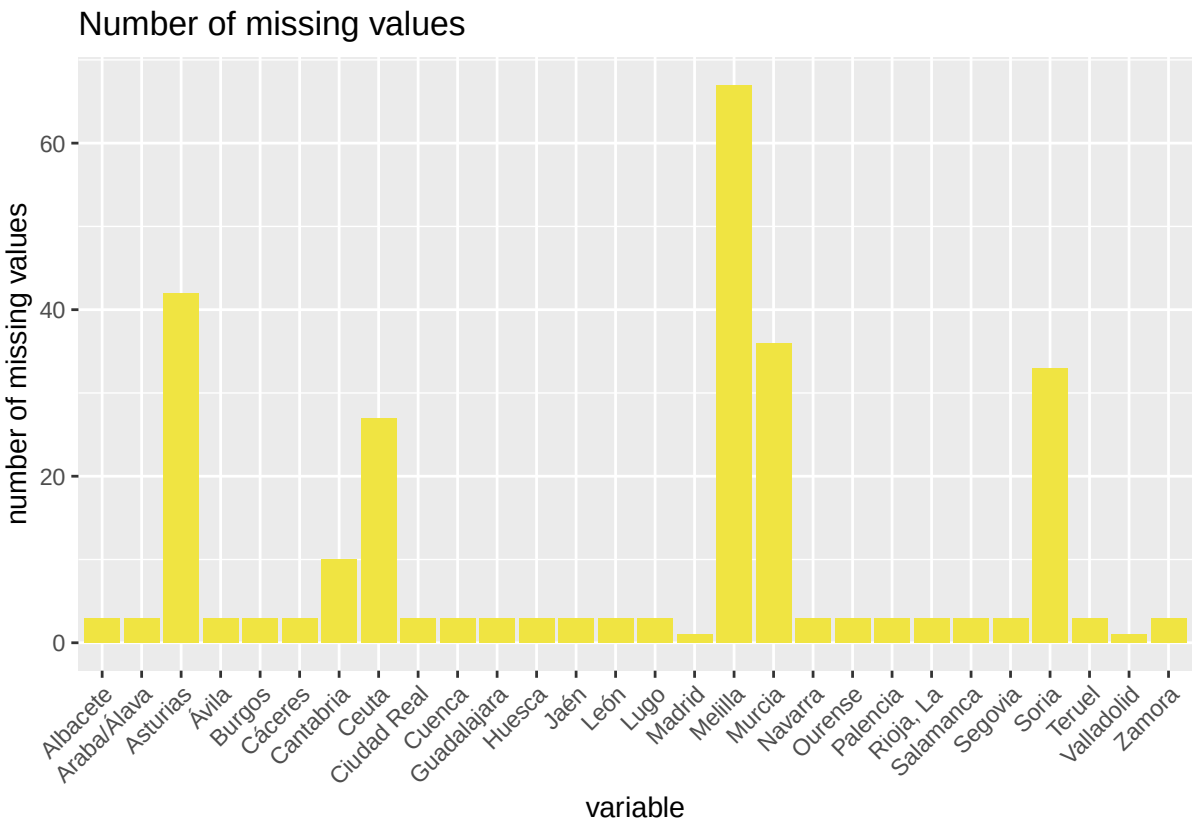

```
# We convert the time series object to a dataframe
Google_retail <- ts_df(Google_t_retail_ts)

names(Google_retail)[names(Google_retail) == "id"] <- "sub_region_2"
names(Google_retail)[names(Google_retail) == "time"] <- "Date"
names(Google_retail)[names(Google_retail) == "value"] <-
  "retail_and_recreation_percent_change_from_baseline"

#####################################################################
# Transpose dataframe
Google_grocery<-Google[c(2,4,6)]
Google_t_grocery<-dcast(Google_grocery, date~sub_region_2, fill=NA)

# Visualize missing values
aggr(Google_t_grocery, col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(Google_t_grocery), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```



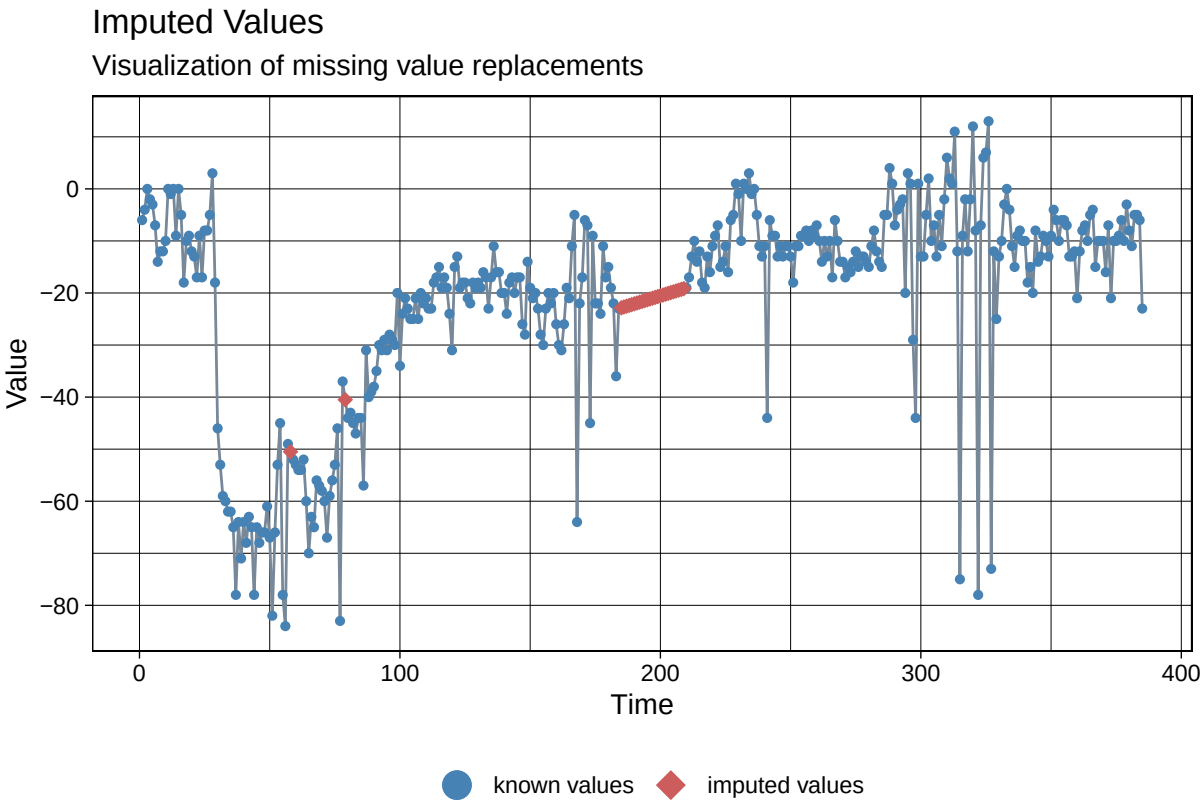

```
##
##  Variables sorted by number of missings:
##                 Variable       Count
##                  Melilla 0.174025974
##                 Asturias 0.109090909
##                   Murcia 0.093506494
##                    Soria 0.085714286
##                    Ceuta 0.070129870
##                Cantabria 0.025974026
##                 Albacete 0.007792208
##              Araba/Álava 0.007792208
##                    Ávila 0.007792208
##                   Burgos 0.007792208
##                  Cáceres 0.007792208
##              Ciudad Real 0.007792208
##                   Cuenca 0.007792208
##              Guadalajara 0.007792208
##                   Huesca 0.007792208
##                     Jaén 0.007792208
##                     León 0.007792208
##                     Lugo 0.007792208
##                  Navarra 0.007792208
##                  Ourense 0.007792208
##                 Palencia 0.007792208
##                Rioja, La 0.007792208
##                Salamanca 0.007792208
```

```
##                 Segovia 0.007792208
##                  Teruel 0.007792208
##                  Zamora 0.007792208
##                  Madrid 0.002597403
##              Valladolid 0.002597403
##                    date 0.000000000
##       Alicante/Alacant 0.000000000
##                 Almería 0.000000000
##                 Badajoz 0.000000000
##          Balears, Illes 0.000000000
##               Barcelona 0.000000000
##                 Bizkaia 0.000000000
##                   Cádiz 0.000000000
##      Castellón/Castelló 0.000000000
##                 Córdoba 0.000000000
##               Coruña, A 0.000000000
##                Gipuzkoa 0.000000000
##                  Girona 0.000000000
##                 Granada 0.000000000
##                  Huelva 0.000000000
##                  Lleida 0.000000000
##                  Málaga 0.000000000
##             Palmas, Las 0.000000000
##              Pontevedra 0.000000000
##  Santa Cruz de Tenerife 0.000000000
##                 Sevilla 0.000000000
##               Tarragona 0.000000000
##                  Toledo 0.000000000
##       Valencia/València 0.000000000
##                 Zaragoza 0.000000000
```

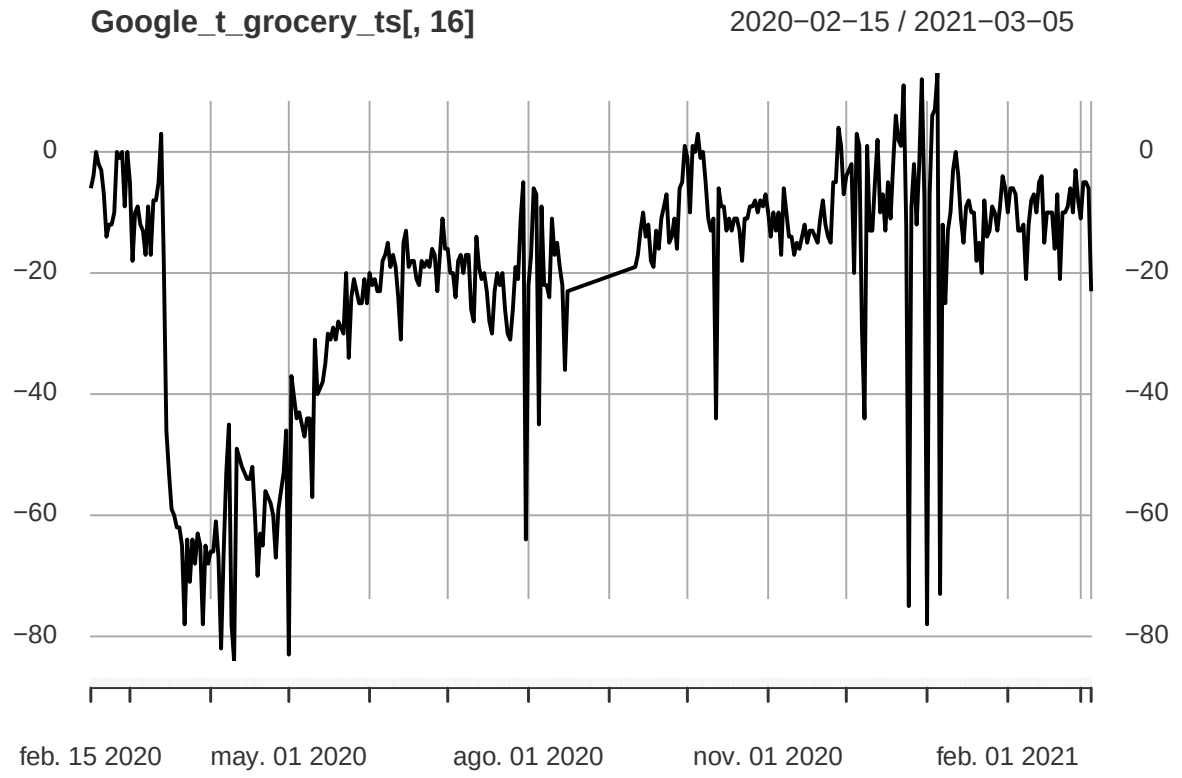
```
Google_t_grocery %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



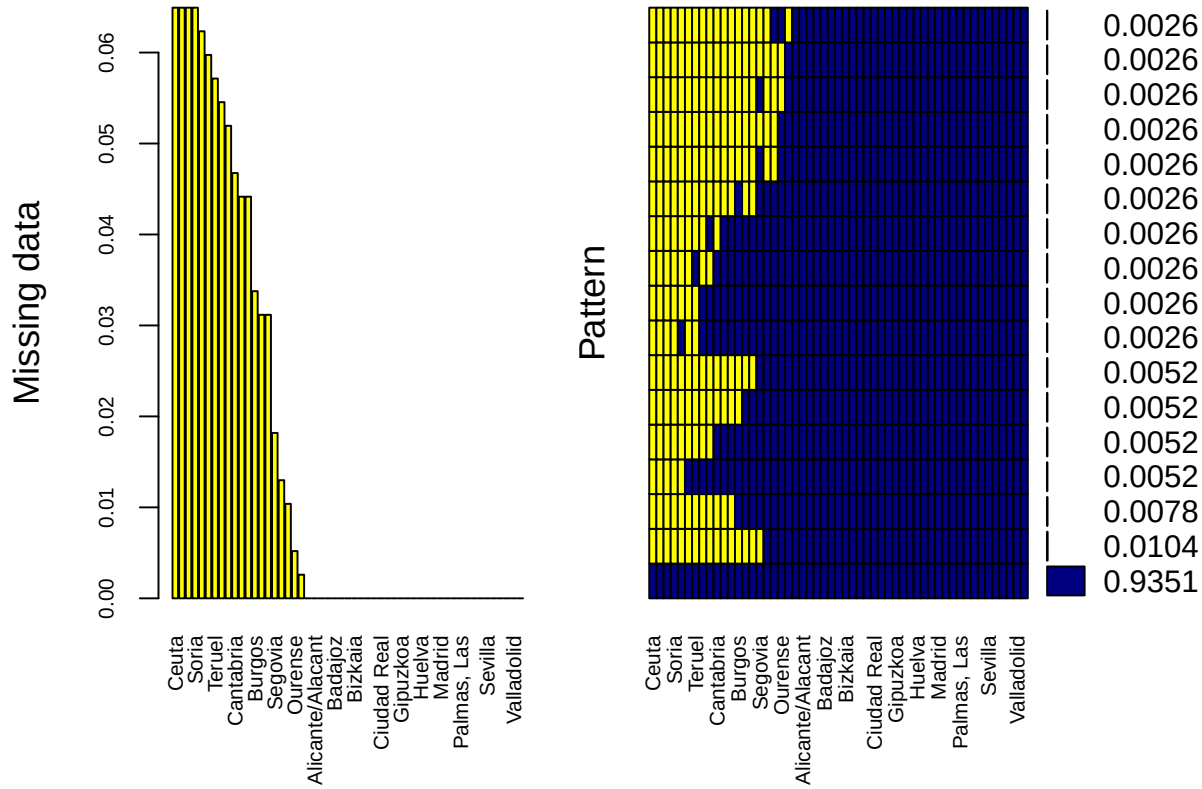

```
# Convert dataframe to ts object
Google_t_grocery_ts<-xts(Google_t_grocery[-1],Google_t_grocery$date)

# Impute the missing values with na_seadec (i.e Ceuta)
imp6 <- na_seadec(Google_t_grocery_ts[,16])
ggplot_na_imputations(Google_t_grocery_ts[,16], imp6)
```




```
# We select na_seadec for the dataset
Google_t_grocery_ts <- na_seadec(Google_t_grocery_ts)
plot(Google_t_grocery_ts[,16])
```



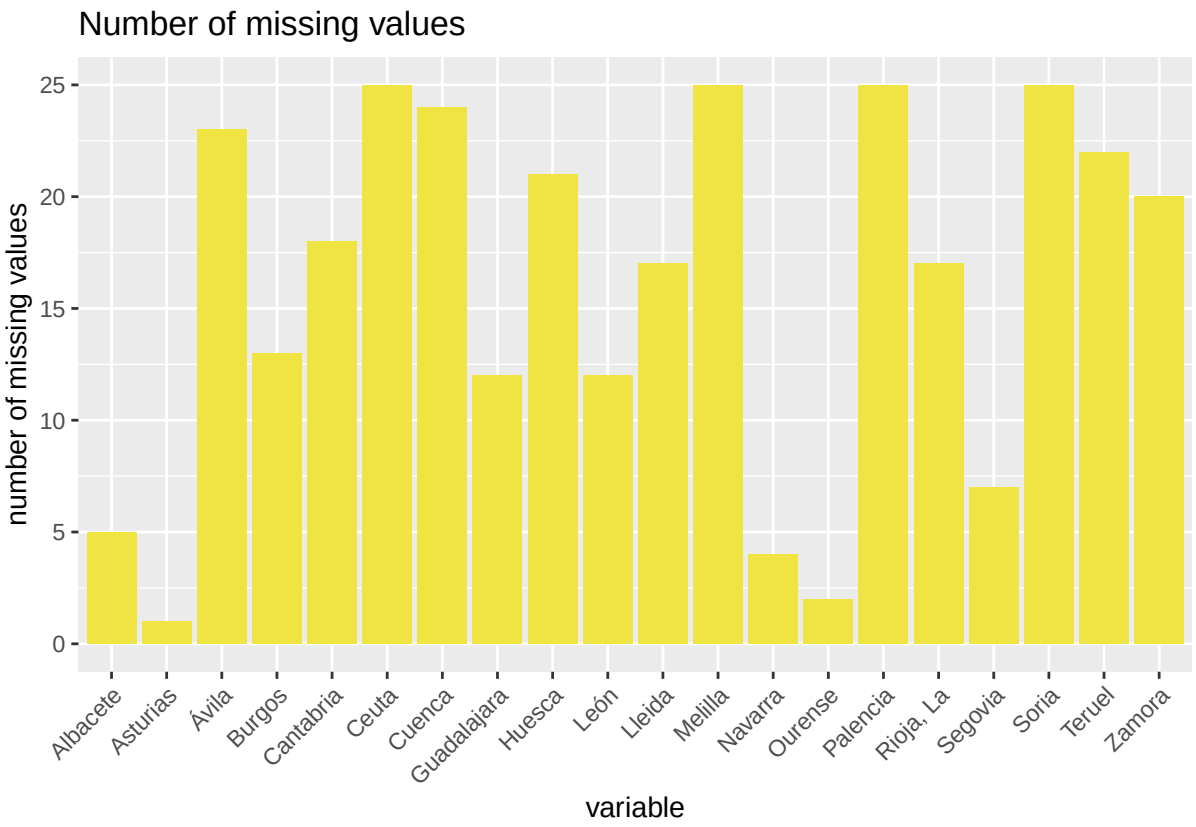

```
# We convert the time series object to a dataframe
Google_grocery <- ts_df(Google_t_grocery_ts)

names(Google_grocery)[names(Google_grocery) == "id"] <- "sub_region_2"
names(Google_grocery)[names(Google_grocery) == "time"] <- "Date"
names(Google_grocery)[names(Google_grocery) == "value"] <-
  "grocery_and_pharmacy_percent_change_from_baseline"

#####################################################################
# Transpose dataframe
Google_parks<-Google[c(2,4,7)]
Google_t_parks<-dcast(Google_parks, date~sub_region_2, fill=NA)

# Visualize missing values
aggr(Google_t_parks, col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(Google_t_parks), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```



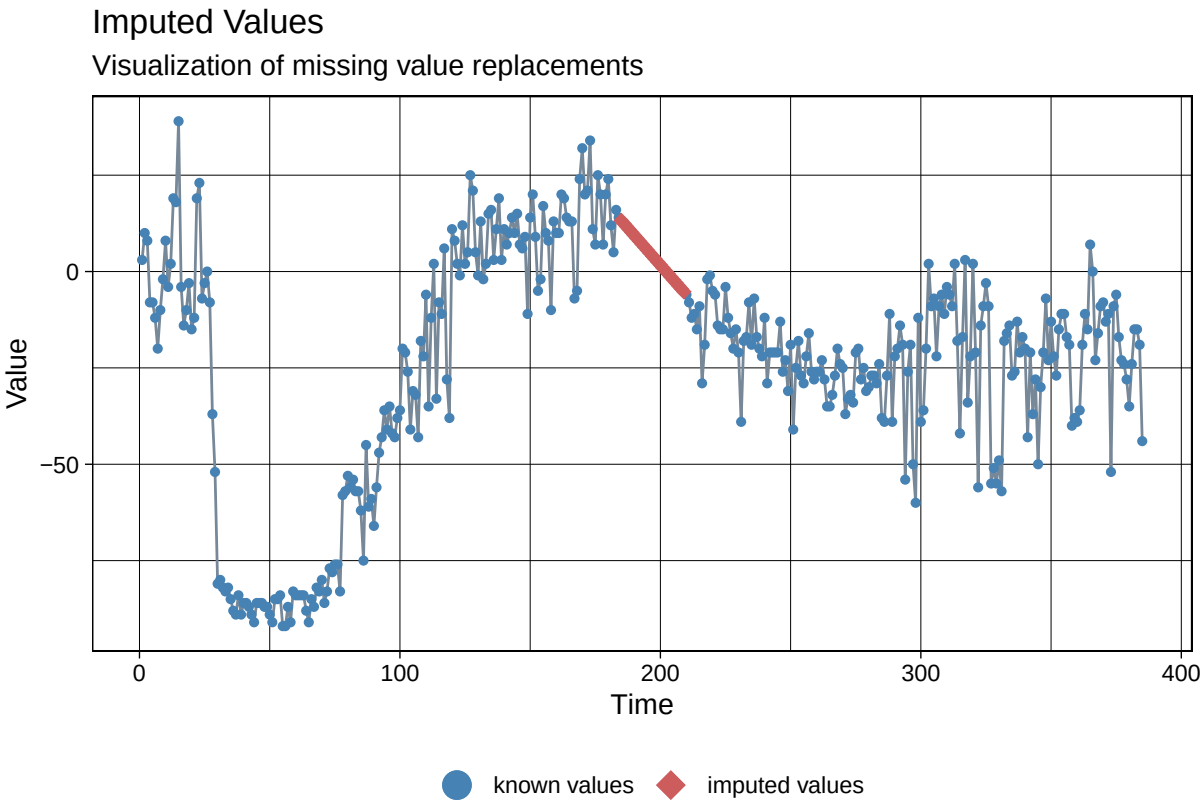

```
##
##  Variables sorted by number of missings:
##                 Variable       Count
##                    Ceuta 0.064935065
##                  Melilla 0.064935065
##                 Palencia 0.064935065
##                    Soria 0.064935065
##                   Cuenca 0.062337662
##                    Ávila 0.059740260
##                   Teruel 0.057142857
##                   Huesca 0.054545455
##                   Zamora 0.051948052
##                Cantabria 0.046753247
##                   Lleida 0.044155844
##                Rioja, La 0.044155844
##                   Burgos 0.033766234
##               Guadalajara 0.031168831
##                    León 0.031168831
##                  Segovia 0.018181818
##                 Albacete 0.012987013
##                  Navarra 0.010389610
##                  Ourense 0.005194805
##                 Asturias 0.002597403
##                     date 0.000000000
##         Alicante/Alacant 0.000000000
##                  Almería 0.000000000
```

```
##              Araba/Álava 0.000000000
##                  Badajoz 0.000000000
##           Balears, Illes 0.000000000
##                Barcelona 0.000000000
##                  Bizkaia 0.000000000
##                  Cáceres 0.000000000
##                    Cádiz 0.000000000
##       Castellón/Castelló 0.000000000
##              Ciudad Real 0.000000000
##                  Córdoba 0.000000000
##                Coruña, A 0.000000000
##                 Gipuzkoa 0.000000000
##                   Girona 0.000000000
##                  Granada 0.000000000
##                   Huelva 0.000000000
##                     Jaén 0.000000000
##                     Lugo 0.000000000
##                   Madrid 0.000000000
##                   Málaga 0.000000000
##                   Murcia 0.000000000
##              Palmas, Las 0.000000000
##               Pontevedra 0.000000000
##                Salamanca 0.000000000
##  Santa Cruz de Tenerife 0.000000000
##                  Sevilla 0.000000000
##                Tarragona 0.000000000
##                   Toledo 0.000000000
##        Valencia/València 0.000000000
##              Valladolid 0.000000000
##                 Zaragoza 0.000000000
```

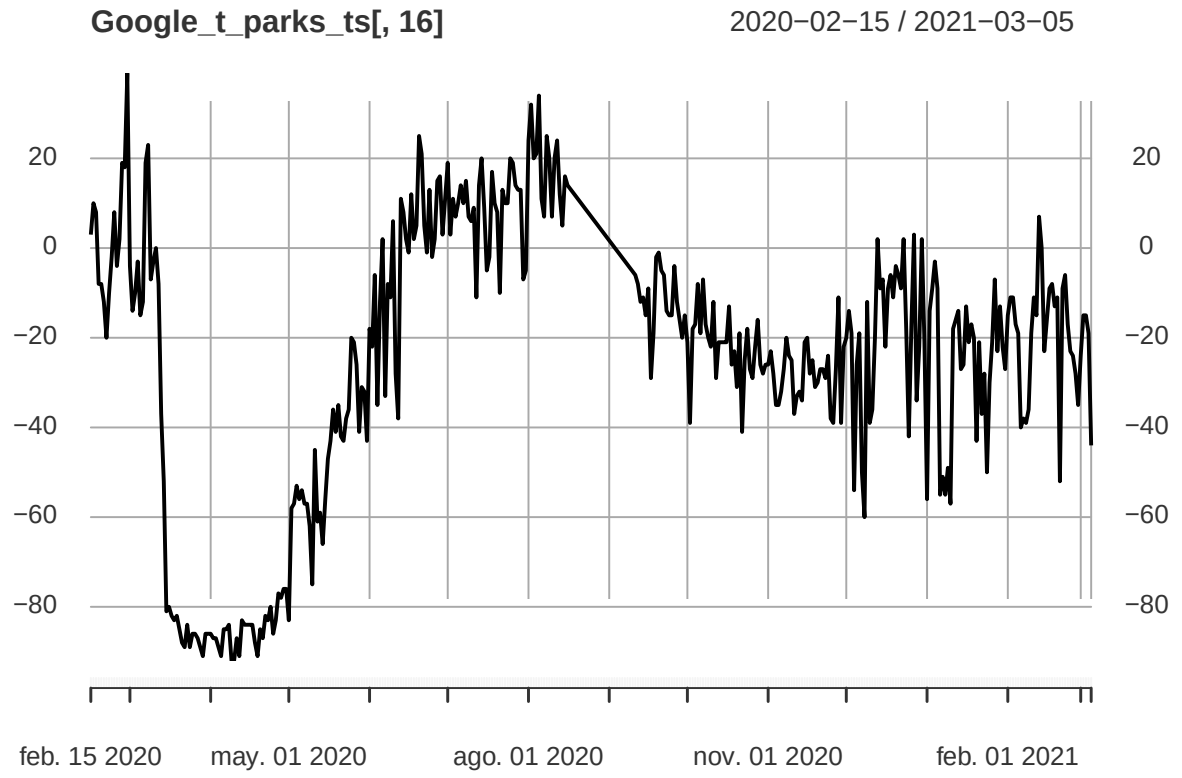
```
Google_t_parks %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



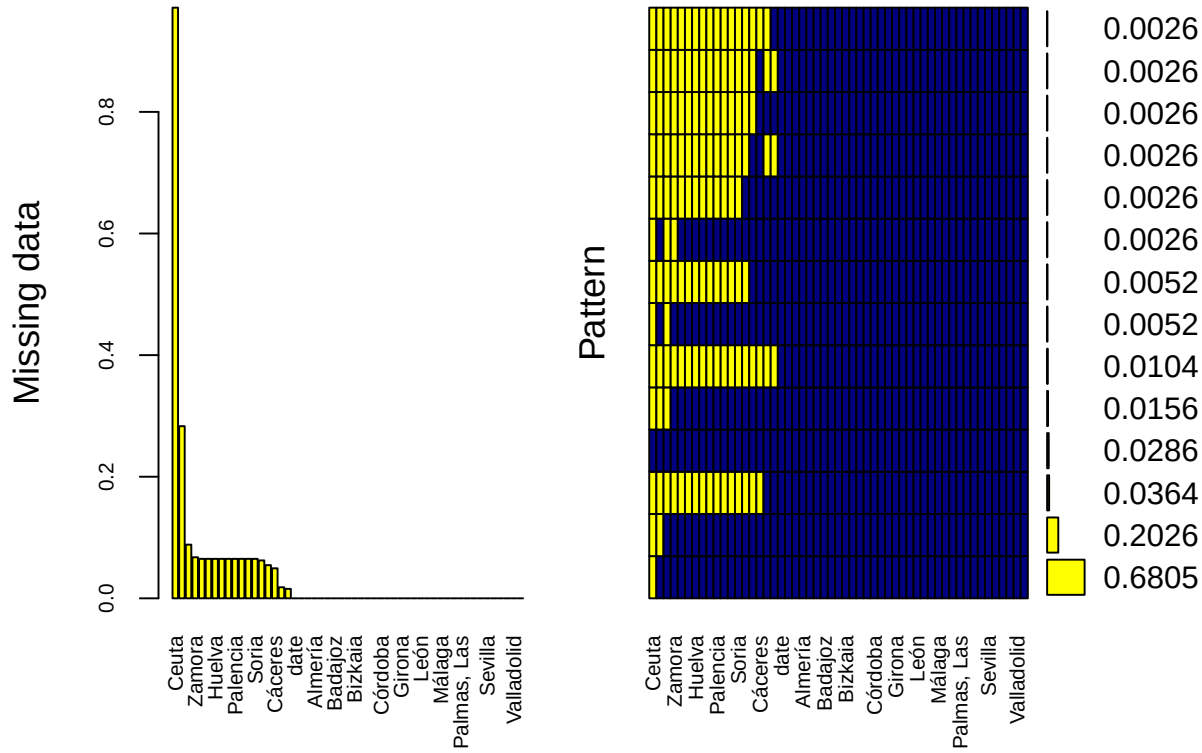

```
# Convert dataframe to ts object
Google_t_parks_ts<-xts(Google_t_parks[-1],Google_t_parks$date)

# Impute the missing values with na_seadec (i.e Ceuta)
imp7 <- na_seadec(Google_t_parks_ts[,16])
ggplot_na_imputations(Google_t_parks_ts[,16], imp7)
```




```
# We select na_seadec for the dataset
Google_t_parks_ts <- na_seadec(Google_t_parks_ts)
plot(Google_t_parks_ts[,16])
```



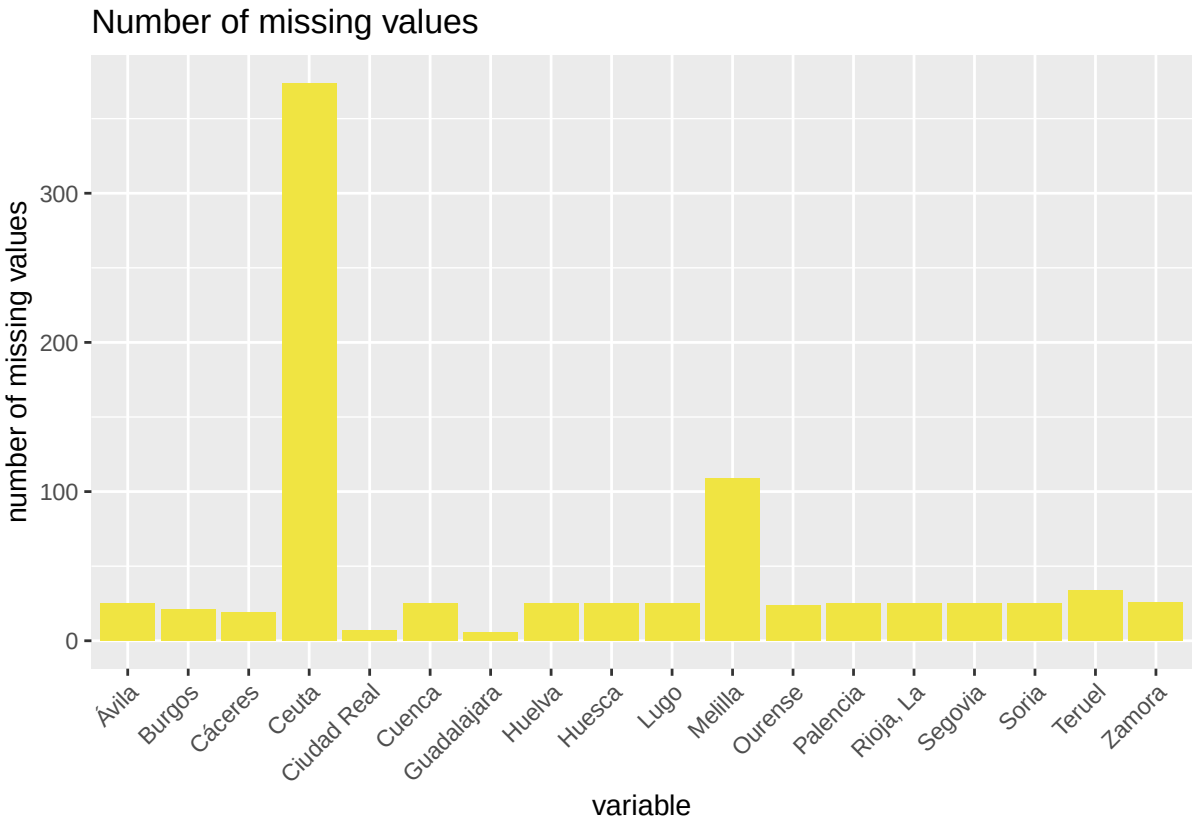

```
# We convert the time series object to a dataframe
Google_parks <- ts_df(Google_t_parks_ts)

names(Google_parks)[names(Google_parks) == "id"] <- "sub_region_2"
names(Google_parks)[names(Google_parks) == "time"] <- "Date"
names(Google_parks)[names(Google_parks) == "value"] <-
  "parks_percent_change_from_baseline"

#####################################################################
# Transpose dataframe
Google_transit<-Google[c(2,4,8)]
Google_t_transit<-dcast(Google_transit, date~sub_region_2, fill=NA)

# Visualize missing values
aggr(Google_t_transit, col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(Google_t_transit), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```



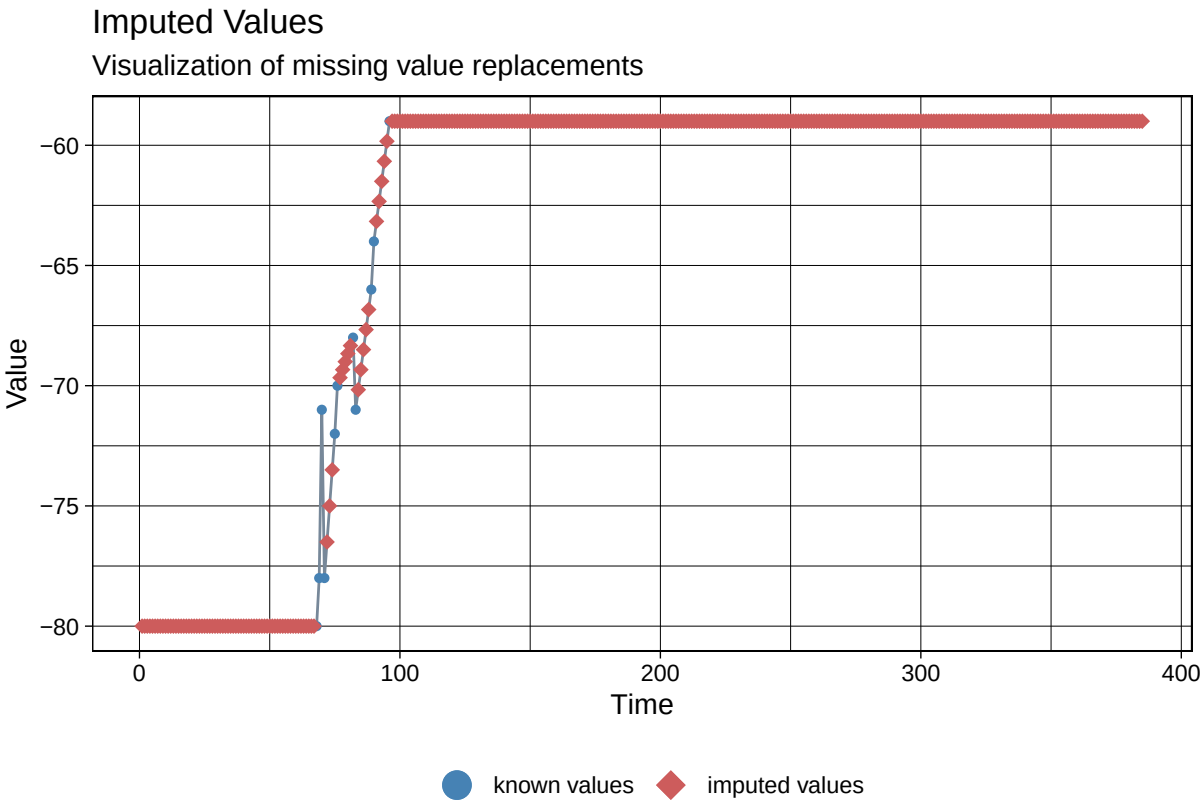

```
##
##  Variables sorted by number of missings:
##                 Variable      Count
##                    Ceuta 0.97142857
##                  Melilla 0.28311688
##                   Teruel 0.08831169
##                   Zamora 0.06753247
##                    Ávila 0.06493506
##                   Cuenca 0.06493506
##                   Huelva 0.06493506
##                   Huesca 0.06493506
##                     Lugo 0.06493506
##                 Palencia 0.06493506
##                Rioja, La 0.06493506
##                  Segovia 0.06493506
##                    Soria 0.06493506
##                  Ourense 0.06233766
##                   Burgos 0.05454545
##                 Cáceres 0.04935065
##              Ciudad Real 0.01818182
##              Guadalajara 0.01558442
##                     date 0.00000000
##                 Albacete 0.00000000
##         Alicante/Alacant 0.00000000
##                  Almería 0.00000000
##              Araba/Álava 0.00000000
```

```
##                Asturias 0.00000000
##                 Badajoz 0.00000000
##          Balears, Illes 0.00000000
##               Barcelona 0.00000000
##                 Bizkaia 0.00000000
##                   Cádiz 0.00000000
##               Cantabria 0.00000000
##      Castellón/Castelló 0.00000000
##                 Córdoba 0.00000000
##               Coruña, A 0.00000000
##                Gipuzkoa 0.00000000
##                  Girona 0.00000000
##                 Granada 0.00000000
##                    Jaén 0.00000000
##                   León 0.00000000
##                  Lleida 0.00000000
##                  Madrid 0.00000000
##                  Málaga 0.00000000
##                  Murcia 0.00000000
##                 Navarra 0.00000000
##             Palmas, Las 0.00000000
##              Pontevedra 0.00000000
##               Salamanca 0.00000000
##  Santa Cruz de Tenerife 0.00000000
##                 Sevilla 0.00000000
##               Tarragona 0.00000000
##                  Toledo 0.00000000
##       Valencia/València 0.00000000
##              Valladolid 0.00000000
##                Zaragoza 0.00000000
```

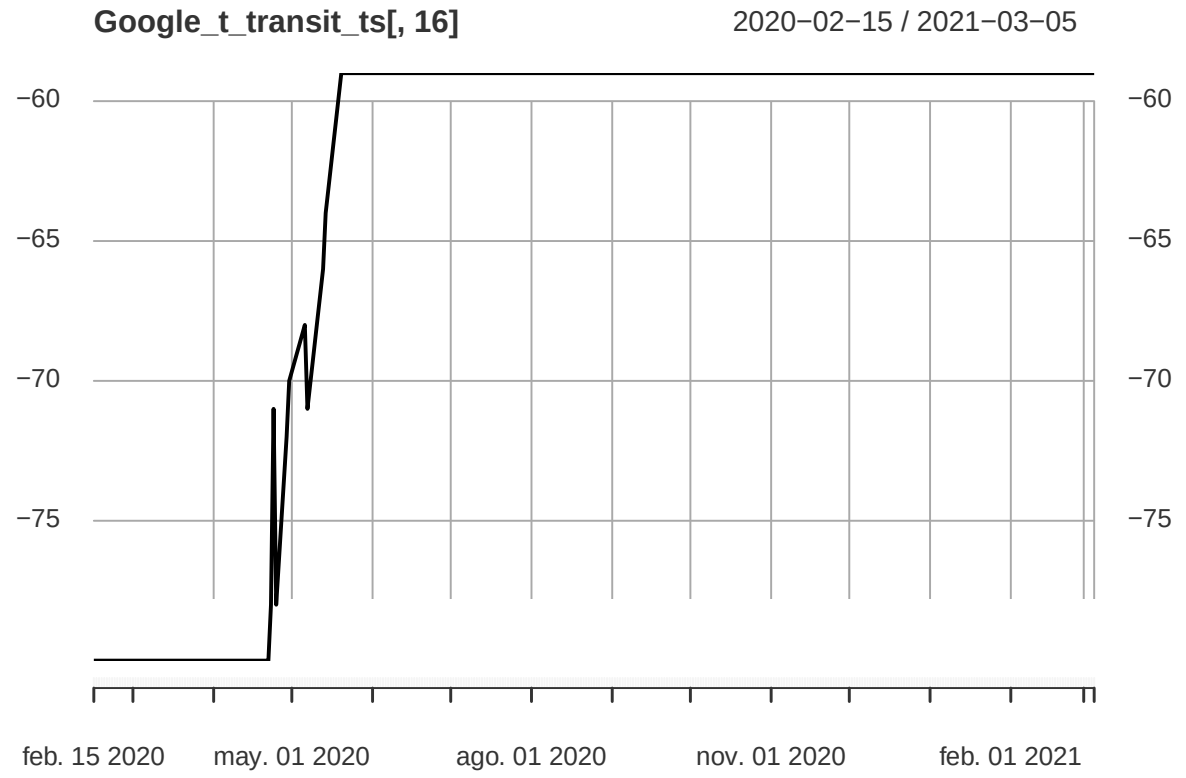
```
Google_t_transit %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



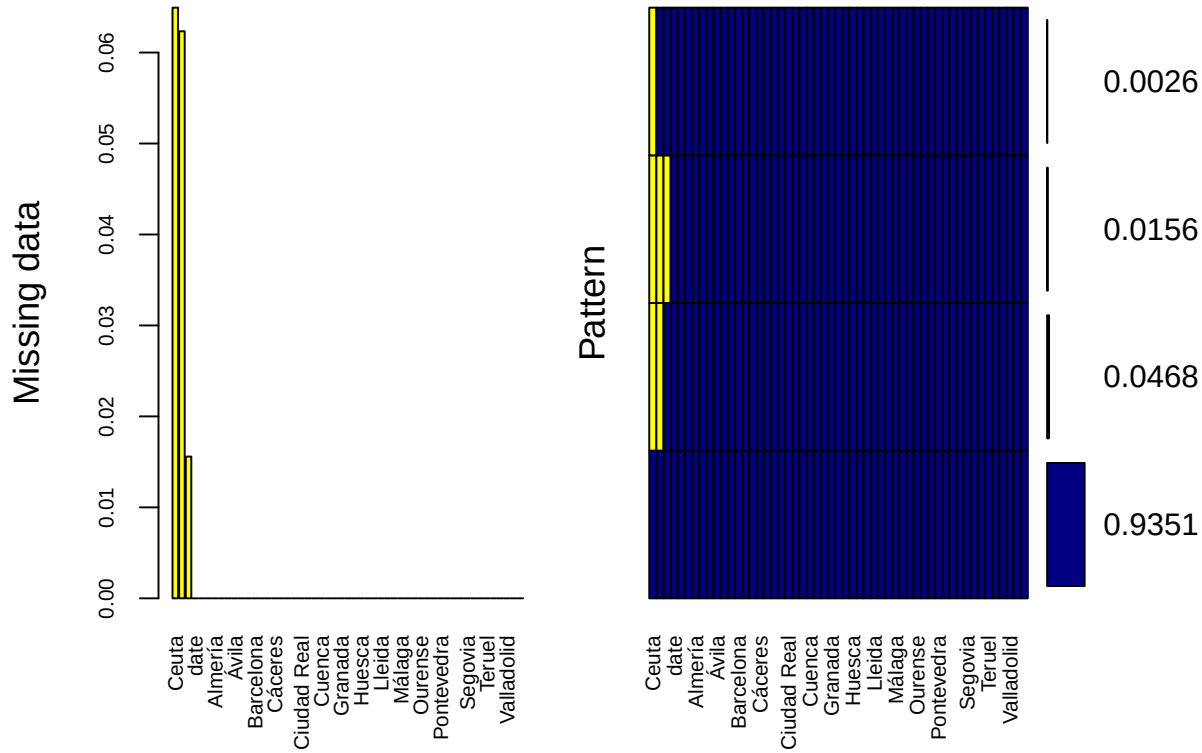

```
# Convert dataframe to ts object
Google_t_transit_ts<-xts(Google_t_transit[-1],Google_t_transit$date)

# Impute the missing values with na_seadec (i.e Ceuta)
imp8 <- na_seadec(Google_t_transit_ts[,16])
ggplot_na_imputations(Google_t_transit_ts[,16], imp8)
```




```
# We select na_seadec for the dataset
Google_t_transit_ts <- na_seadec(Google_t_transit_ts)
plot(Google_t_transit_ts[,16])
```



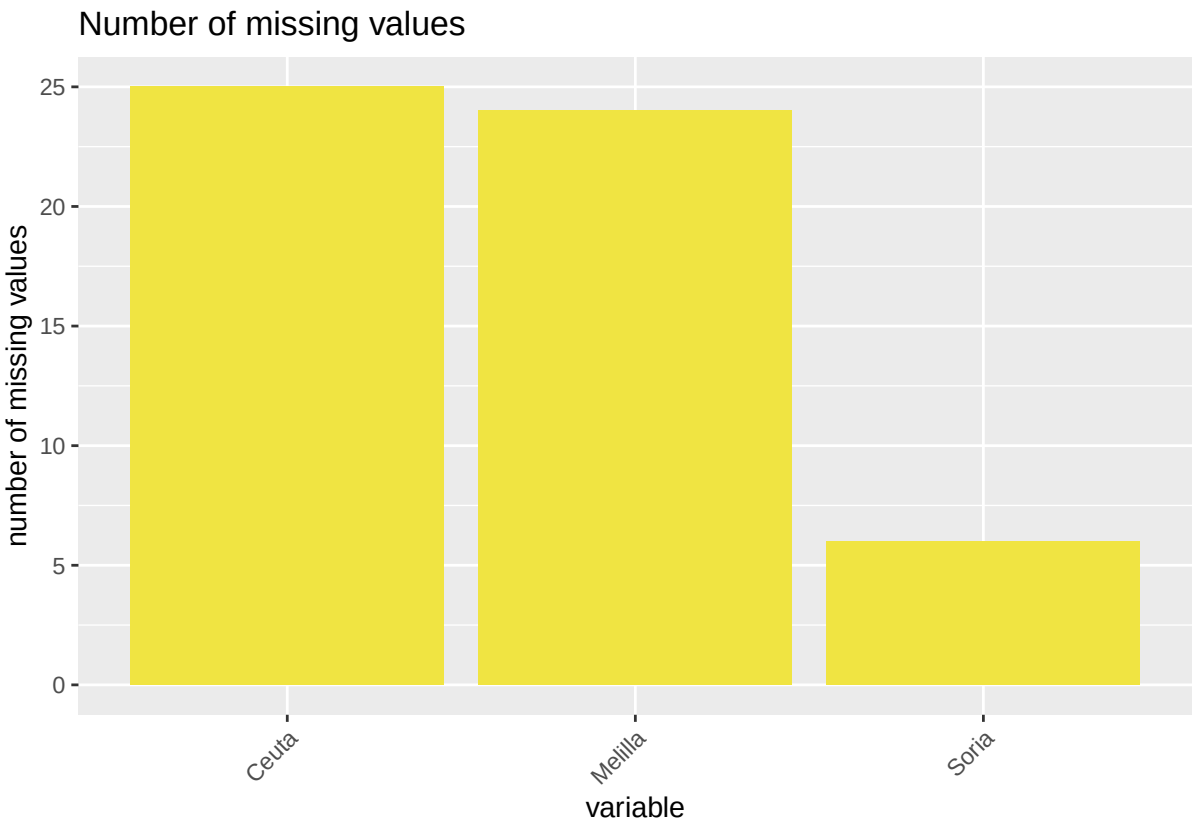

```
# We convert the time series object to a dataframe
Google_transit <- ts_df(Google_t_transit_ts)

names(Google_transit)[names(Google_transit) == "id"] <- "sub_region_2"
names(Google_transit)[names(Google_transit) == "time"] <- "Date"
names(Google_transit)[names(Google_transit) == "value"] <-
  "transit_stations_percent_change_from_baseline"

######################################################################
# Transpose dataframe
Google_workplaces<-Google[c(2,4,9)]
Google_t_workplaces<-dcast(Google_workplaces, date~sub_region_2, fill=NA)

# Visualize missing values
aggr(Google_t_workplaces, col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(Google_t_workplaces), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```



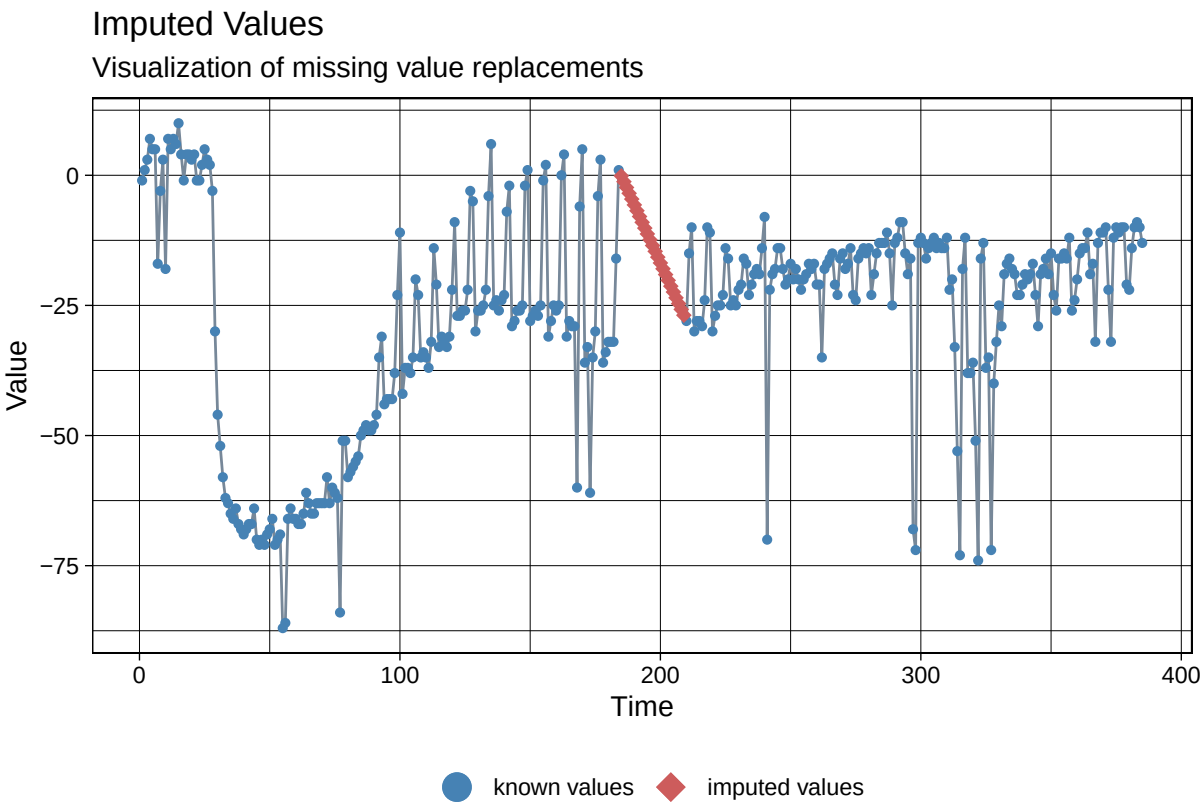

```
##
##  Variables sorted by number of missings:
##                 Variable      Count
##                    Ceuta 0.06493506
##                  Melilla 0.06233766
##                    Soria 0.01558442
##                     date 0.00000000
##                 Albacete 0.00000000
##         Alicante/Alacant 0.00000000
##                  Almería 0.00000000
##              Araba/Álava 0.00000000
##                 Asturias 0.00000000
##                    Ávila 0.00000000
##                  Badajoz 0.00000000
##           Balears, Illes 0.00000000
##                Barcelona 0.00000000
##                  Bizkaia 0.00000000
##                   Burgos 0.00000000
##                  Cáceres 0.00000000
##                    Cádiz 0.00000000
##                Cantabria 0.00000000
##       Castellón/Castelló 0.00000000
##              Ciudad Real 0.00000000
##                  Córdoba 0.00000000
##                Coruña, A 0.00000000
##                   Cuenca 0.00000000
```

```
##                Gipuzkoa 0.00000000
##                  Girona 0.00000000
##                 Granada 0.00000000
##             Guadalajara 0.00000000
##                  Huelva 0.00000000
##                  Huesca 0.00000000
##                   Jaén 0.00000000
##                   León 0.00000000
##                  Lleida 0.00000000
##                   Lugo 0.00000000
##                  Madrid 0.00000000
##                  Málaga 0.00000000
##                  Murcia 0.00000000
##                 Navarra 0.00000000
##                 Ourense 0.00000000
##                Palencia 0.00000000
##             Palmas, Las 0.00000000
##              Pontevedra 0.00000000
##               Rioja, La 0.00000000
##                Salamanca 0.00000000
##  Santa Cruz de Tenerife 0.00000000
##                  Segovia 0.00000000
##                  Sevilla 0.00000000
##                Tarragona 0.00000000
##                   Teruel 0.00000000
##                  Toledo 0.00000000
##       Valencia/València 0.00000000
##              Valladolid 0.00000000
##                  Zamora 0.00000000
##                Zaragoza 0.00000000
```

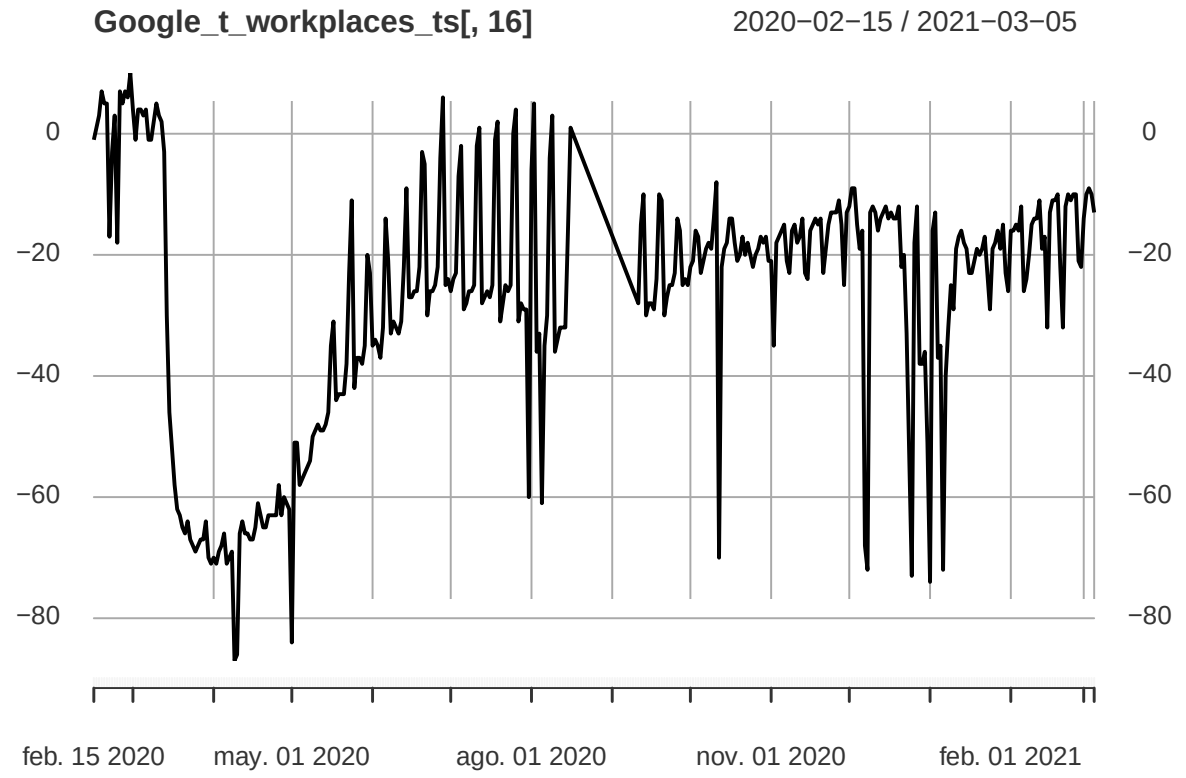
```
Google_t_workplaces %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



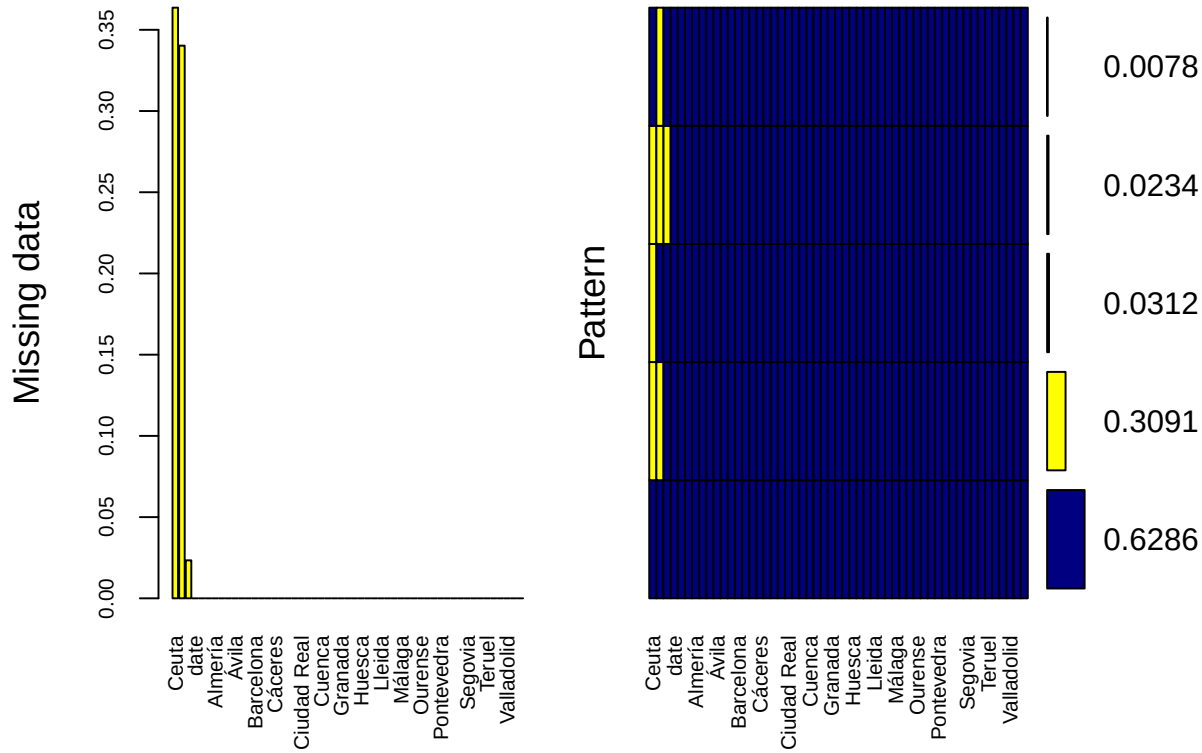

```
# Convert dataframe to ts object
Google_t_workplaces_ts<-xts(Google_t_workplaces[-1],Google_t_workplaces$date)

# Impute the missing values with na_seadec (i.e Ceuta)
imp9 <- na_seadec(Google_t_workplaces_ts[,16])
ggplot_na_imputations(Google_t_workplaces_ts[,16], imp9)
```




```
# We select na_seadec for the dataset
Google_t_workplaces_ts <- na_seadec(Google_t_workplaces_ts)
plot(Google_t_workplaces_ts[,16])
```



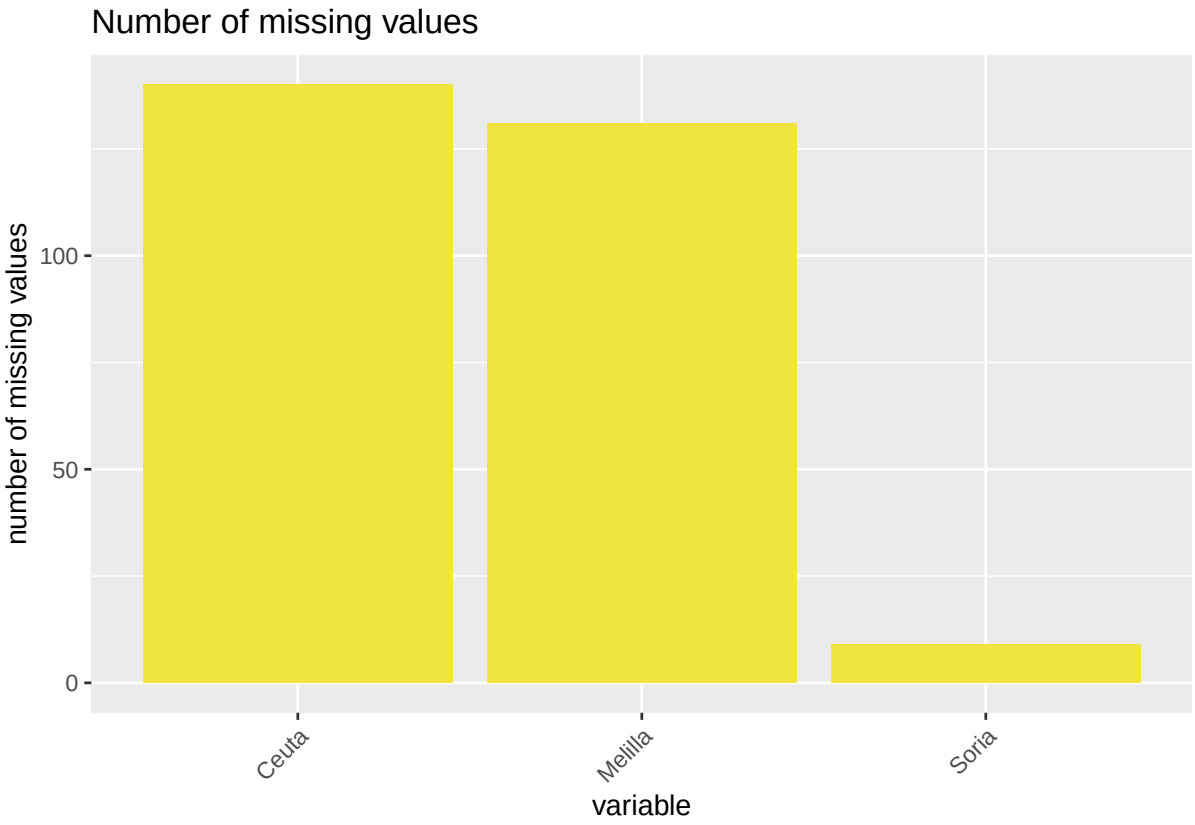

```
# We convert the time series object to a dataframe
Google_workplaces <- ts_df(Google_t_workplaces_ts)

names(Google_workplaces)[names(Google_workplaces) == "id"] <- "sub_region_2"
names(Google_workplaces)[names(Google_workplaces) == "time"] <- "Date"
names(Google_workplaces)[names(Google_workplaces) == "value"] <-
  "workplaces_percent_change_from_baseline"

#####################################################################
# Transpose dataframe
Google_residential<-Google[c(2,4,10)]
Google_t_residential<-dcast(Google_residential, date~sub_region_2, fill=NA)

# Visualize missing values
aggr(Google_t_residential, col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(Google_t_residential), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```



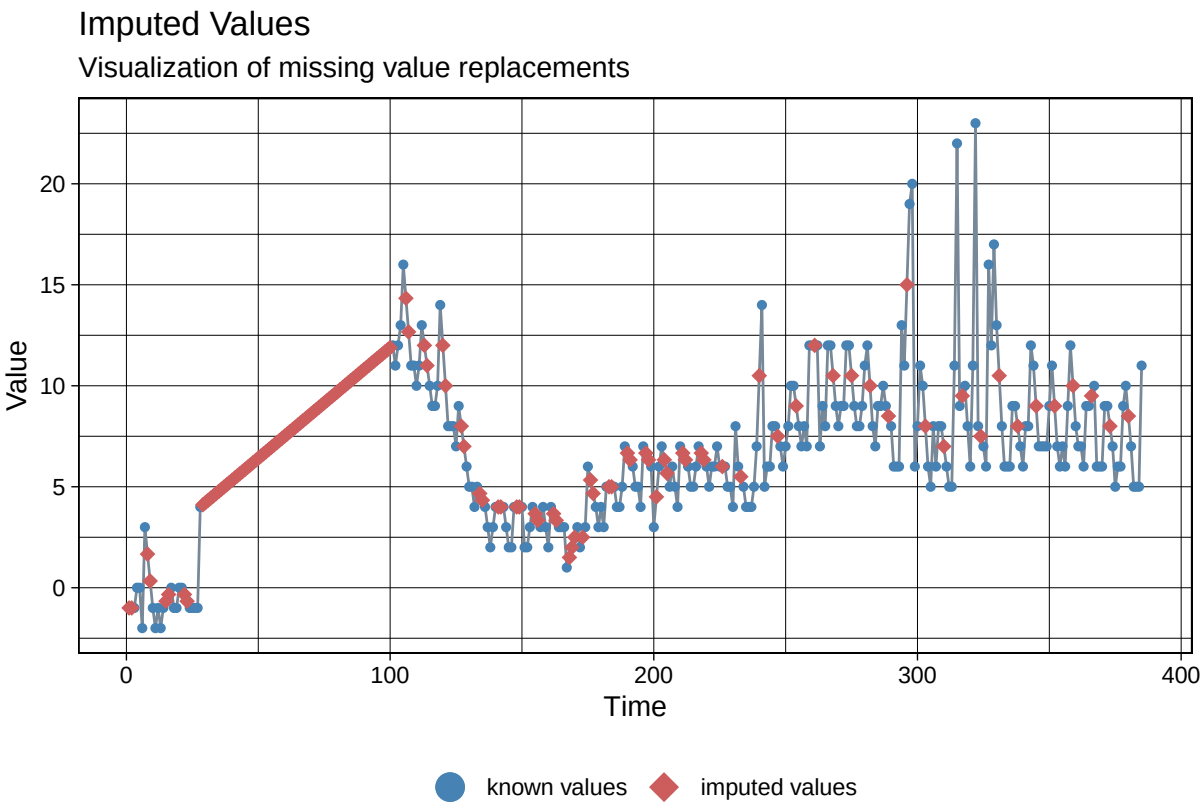

```
##
##  Variables sorted by number of missings:
##                 Variable      Count
##                    Ceuta 0.36363636
##                  Melilla 0.34025974
##                    Soria 0.02337662
##                     date 0.00000000
##                 Albacete 0.00000000
##         Alicante/Alacant 0.00000000
##                  Almería 0.00000000
##              Araba/Álava 0.00000000
##                 Asturias 0.00000000
##                    Ávila 0.00000000
##                  Badajoz 0.00000000
##           Balears, Illes 0.00000000
##                Barcelona 0.00000000
##                  Bizkaia 0.00000000
##                   Burgos 0.00000000
##                  Cáceres 0.00000000
##                    Cádiz 0.00000000
##                Cantabria 0.00000000
##       Castellón/Castelló 0.00000000
##              Ciudad Real 0.00000000
##                  Córdoba 0.00000000
##                Coruña, A 0.00000000
##                   Cuenca 0.00000000
```

```
##                Gipuzkoa 0.00000000
##                  Girona 0.00000000
##                 Granada 0.00000000
##             Guadalajara 0.00000000
##                  Huelva 0.00000000
##                  Huesca 0.00000000
##                   Jaén 0.00000000
##                   León 0.00000000
##                  Lleida 0.00000000
##                   Lugo 0.00000000
##                  Madrid 0.00000000
##                  Málaga 0.00000000
##                  Murcia 0.00000000
##                 Navarra 0.00000000
##                 Ourense 0.00000000
##                Palencia 0.00000000
##             Palmas, Las 0.00000000
##              Pontevedra 0.00000000
##               Rioja, La 0.00000000
##                Salamanca 0.00000000
##  Santa Cruz de Tenerife 0.00000000
##                 Segovia 0.00000000
##                 Sevilla 0.00000000
##               Tarragona 0.00000000
##                  Teruel 0.00000000
##                  Toledo 0.00000000
##       Valencia/València 0.00000000
##              Valladolid 0.00000000
##                  Zamora 0.00000000
##                Zaragoza 0.00000000
```

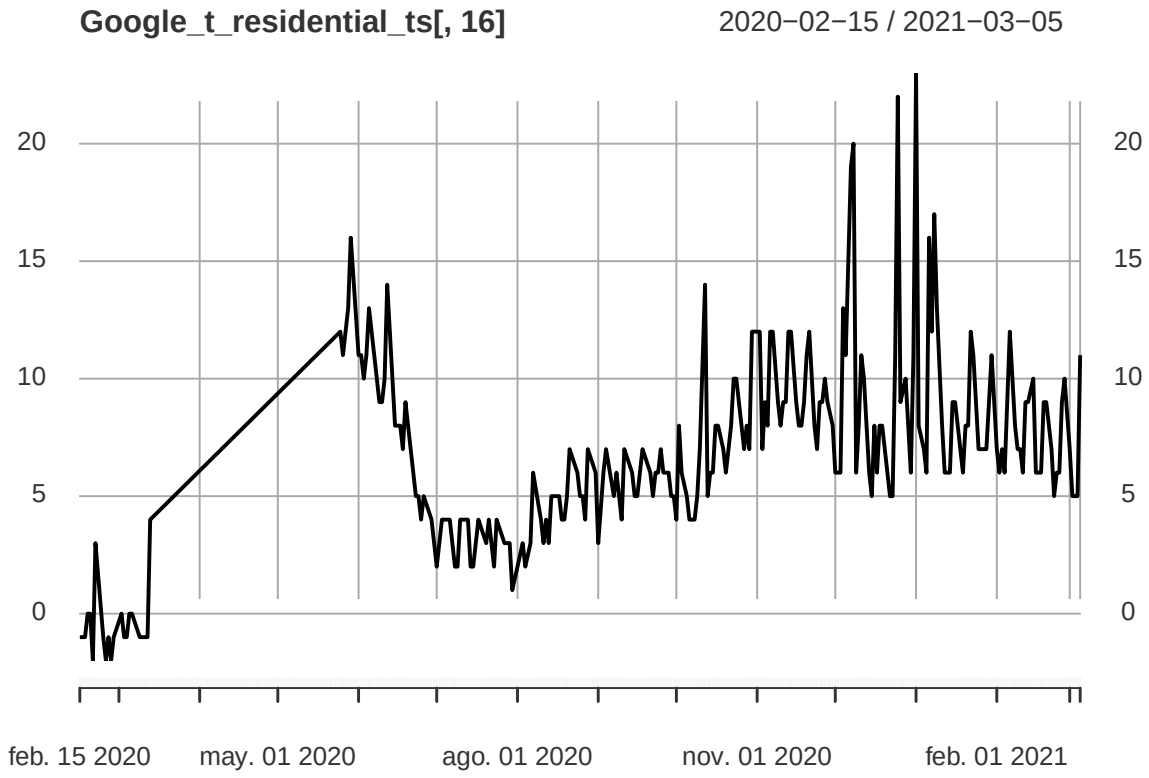
```
Google_t_residential %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
# Convert dataframe to ts object
Google_t_residential_ts<-xts(Google_t_residential[-1],Google_t_residential$date)

# Impute the missing values with na_seadec (i.e Ceuta)
imp10 <- na_seadec(Google_t_residential_ts[,16])
ggplot_na_imputations(Google_t_residential_ts[,16], imp10)
```




```
# We select na_seadec for the dataset
Google_t_residential_ts <- na_seadec(Google_t_residential_ts)
plot(Google_t_residential_ts[,16])
```




```
# We convert the time series object to a dataframe
Google_residential <- ts_df(Google_t_residential_ts)

names(Google_residential)[names(Google_residential) == "id"] <- "sub_region_2"
names(Google_residential)[names(Google_residential) == "time"] <- "Date"
names(Google_residential)[names(Google_residential) == "value"] <-
  "residential_percent_change_from_baseline"
```

Now we merge the previous dataframes into new one with the imputed vaules and we add the ISO code for the province.

```
# New dataframe Google_b
# This approach assumes that the column names are the same and that there's the same number of rows (ou
# Any duplicated columns are automatically eliminated used in the merging process.
Google_b <- merge(Google_retail, Google_grocery) %>%
              merge(Google_parks) %>%
              merge(Google_transit) %>%
              merge(Google_workplaces) %>%
              merge(Google_residential)

# We add the iso code for the province
Google_b$iso_code <- NA
Google_b$iso_code<-Google[match(Google_b$sub_region_2, Google$sub_region_2),3]
rm("Google")
Google<-Google_b
rm("Google_b")
```

```r
# Check table
head(Google,5)
```

```
##   sub_region_2       Date retail_and_recreation_percent_change_from_baseline
## 1     Albacete 2020-02-15                                                  3
## 2     Albacete 2020-02-16                                                  5
## 3     Albacete 2020-02-17                                                 -2
## 4     Albacete 2020-02-18                                                 -3
## 5     Albacete 2020-02-19                                                  0
##   grocery_and_pharmacy_percent_change_from_baseline
## 1                                                -5
## 2                                                 1
## 3                                                 3
## 4                                                -1
## 5                                                 1
##   parks_percent_change_from_baseline
## 1                                 35
## 2                                 40
## 3                                  7
## 4                                 -4
## 5                                  7
##   transit_stations_percent_change_from_baseline
## 1                                            13
## 2                                            18
## 3                                            20
## 4                                             6
## 5                                             9
##   workplaces_percent_change_from_baseline
## 1                                       1
## 2                                       0
## 3                                       5
## 4                                       4
## 5                                       4
##   residential_percent_change_from_baseline iso_code
## 1                                       -3       AB
## 2                                       -4       AB
## 3                                       -1       AB
## 4                                       -1       AB
## 5                                       -1       AB
```

```r
table(Google$sub_region_2)
```

```
##
##               Albacete       Alicante/Alacant                Almería
##                    385                    385                    385
##            Araba/Álava               Asturias                  Ávila
##                    385                    385                    385
##                Badajoz        Balears, Illes              Barcelona
##                    385                    385                    385
##                Bizkaia                 Burgos                Cáceres
##                    385                    385                    385
##                  Cádiz              Cantabria     Castellón/Castelló
##                    385                    385                    385
##                  Ceuta           Ciudad Real                Córdoba
```

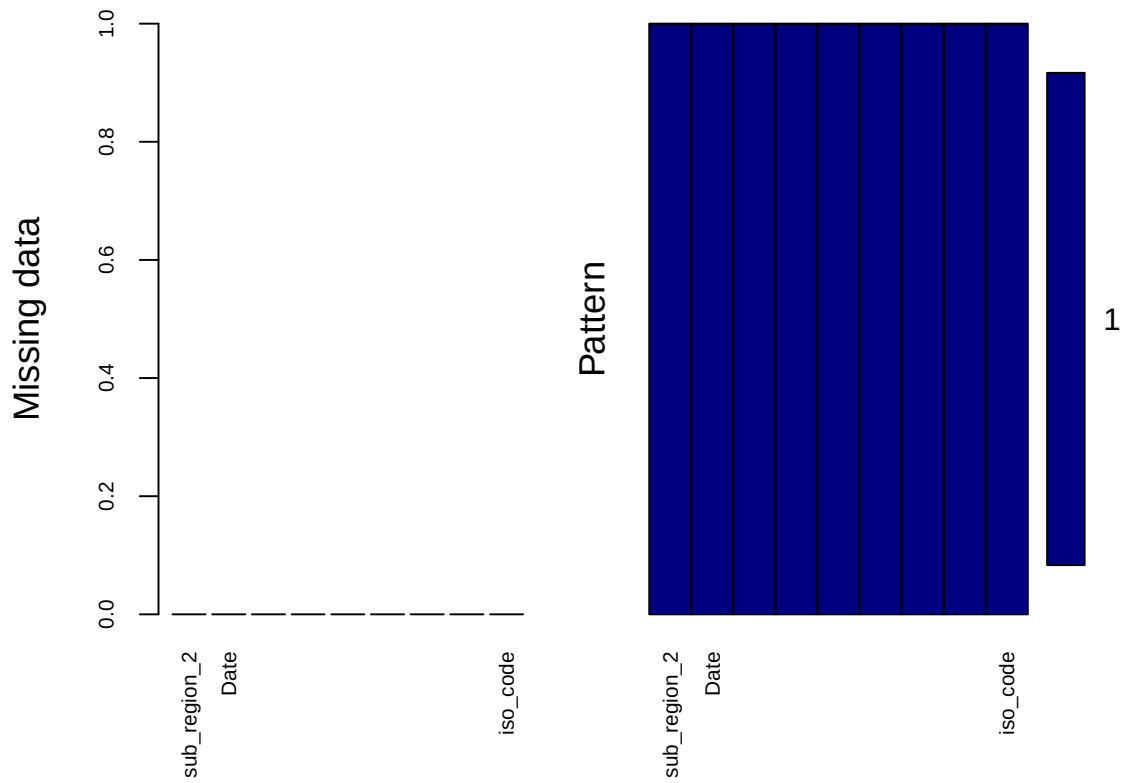
```
##                   385                   385                   385
##             Coruña, A                 Cuenca              Gipuzkoa
##                   385                   385                   385
##                Girona                Granada           Guadalajara
##                   385                   385                   385
##                Huelva                Huesca                 Jaén
##                   385                   385                   385
##                 León                 Lleida                 Lugo
##                   385                   385                   385
##                Madrid                Málaga               Melilla
##                   385                   385                   385
##                Murcia                Navarra               Ourense
##                   385                   385                   385
##              Palencia          Palmas, Las            Pontevedra
##                   385                   385                   385
##             Rioja, La            Salamanca Santa Cruz de Tenerife
##                   385                   385                   385
##               Segovia               Sevilla                 Soria
##                   385                   385                   385
##             Tarragona                Teruel                Toledo
##                   385                   385                   385
##     Valencia/València            Valladolid                Zamora
##                   385                   385                   385
##               Zaragoza
##                   385
```

```
table(Google$iso_code)
```

```
##
##   A  AB  AL  AV   B  BA  BI  BU   C  CA  CC  CE  CO  CR  CS  CU  GC  GI  GR  GU
## 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385
##   H  HU   J   L  LE  LO  LU   M  MA  ML  MU  NA   O  OR   P  PM  PO   S  SA  SE
## 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385
##  SG  SO  SS   T  TE  TF  TO   V  VA  VI   Z  ZA
## 385 385 385 385 385 385 385 385 385 385 385 385
```

We check missing values. We should obtain zero missing values.

```
aggr(Google, col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(Google), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```




```
##
##  Variables sorted by number of missings:
##                                                Variable Count
##                                              sub_region_2     0
##                                                      Date     0
##  retail_and_recreation_percent_change_from_baseline     0
##   grocery_and_pharmacy_percent_change_from_baseline     0
##                  parks_percent_change_from_baseline     0
##       transit_stations_percent_change_from_baseline     0
##             workplaces_percent_change_from_baseline     0
##           residential_percent_change_from_baseline     0
##                                              iso_code     0
```

```
Google %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

2.1.9 CNE review

The CSV files are provided per "imputed date" (fecha)":

- **cases_technic_province.csv** - Number of cases by diagnostic technique and province (of residence)
- **cases_hosp_uci_def_sexo_edad_provres.csv** - Number of hospitalizations, number of ICU admissions and number of deaths by sex, age and province of residence.

```
summary(CNE_tecnica)
```

```
##  provincia_iso         fecha              num_casos      num_casos_prueba_pcr
##  Length:23426       Length:23426       Min.   :   0.0   Min.   :   0.0
##  Class :character   Class :character   1st Qu.:   2.0   1st Qu.:   2.0
##  Mode  :character   Mode  :character   Median :  32.0   Median :  26.0
##                                        Mean   : 136.9   Mean   : 109.6
##                                        3rd Qu.: 120.0   3rd Qu.: 100.0
##                                        Max.   :6972.0   Max.   :6546.0
##  num_casos_prueba_test_ac num_casos_prueba_ag num_casos_prueba_elisa
##  Min.   : 0.0000          Min.   :   0.00     Min.   : 0.0000
##  1st Qu.: 0.0000          1st Qu.:   0.00     1st Qu.: 0.0000
##  Median : 0.0000          Median :   0.00     Median : 0.0000
##  Mean   : 0.2037          Mean   :  26.21     Mean   : 0.1602
##  3rd Qu.: 0.0000          3rd Qu.:   9.00     3rd Qu.: 0.0000
##  Max.   :32.0000          Max.   :3267.00     Max.   :71.0000
##  num_casos_prueba_desconocida
##  Min.   :  0.0000
##  1st Qu.:  0.0000
```

```
##  Median :  0.0000
##  Mean   :  0.7122
##  3rd Qu.:  0.0000
##  Max.   :505.0000
```

```
head(str(CNE_tecnica, vec.len=3))
```

```
## 'data.frame':    23426 obs. of  8 variables:
##  $ provincia_iso               : chr  "A" "AB" "AL" ...
##  $ fecha                       : chr  "2020-01-01" "2020-01-01" "2020-01-01" ...
##  $ num_casos                   : int  0 0 0 0 0 0 0 0 ...
##  $ num_casos_prueba_pcr        : int  0 0 0 0 0 0 0 0 ...
##  $ num_casos_prueba_test_ac    : int  0 0 0 0 0 0 0 0 ...
##  $ num_casos_prueba_ag         : int  0 0 0 0 0 0 0 0 ...
##  $ num_casos_prueba_elisa      : int  0 0 0 0 0 0 0 0 ...
##  $ num_casos_prueba_desconocida: int  0 0 0 0 0 0 0 0 ...
```

```
## NULL
```

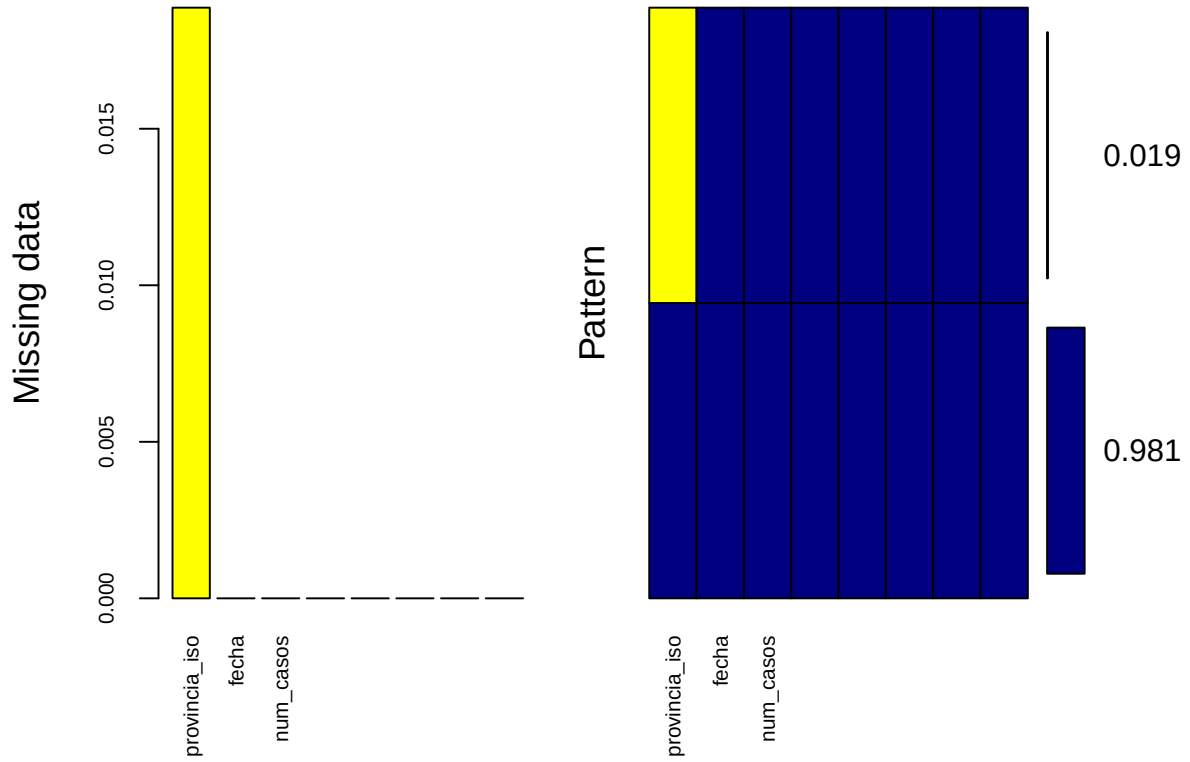
```
table(CNE_tecnica$provincia_iso)
```

```
##
##   A  AB  AL  AV   B  BA  BI  BU   C  CA  CC  CE  CO  CR  CS  CU  GC  GI  GR  GU
## 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442
##   H  HU   J   L  LE  LO  LU   M  MA  ML  MU  NC   O  OR   P  PM  PO   S  SA  SE
## 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442
##  SG  SO  SS   T  TE  TF  TO   V  VA  VI   Z  ZA
## 442 442 442 442 442 442 442 442 442 442 442 442
```

2.1.10 CNE review missing values & impute

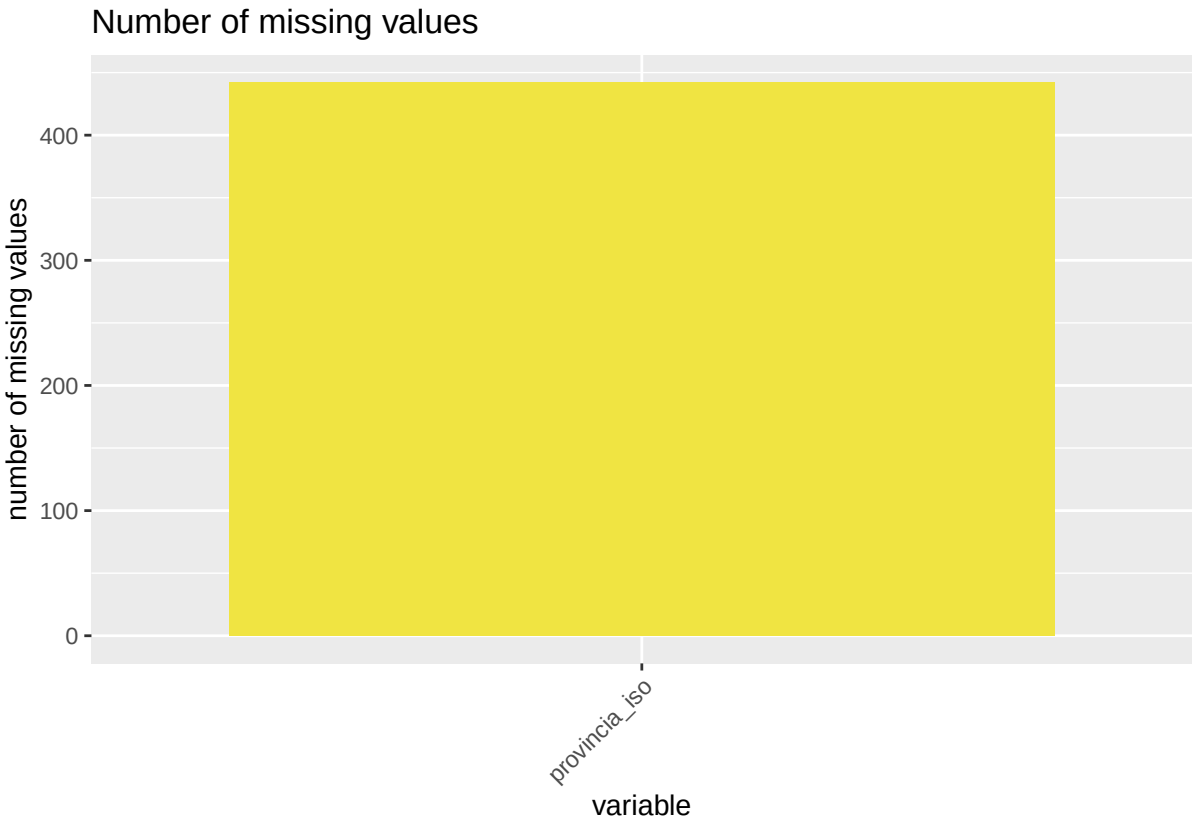
We check missing values for CNE_tecnica. In this case we omit the NA values.

```
aggr(CNE_tecnica, col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(CNE_tecnica), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```




```
##
##  Variables sorted by number of missings:
##                       Variable      Count
##                 provincia_iso 0.01886792
##                         fecha 0.00000000
##                     num_casos 0.00000000
##          num_casos_prueba_pcr 0.00000000
##      num_casos_prueba_test_ac 0.00000000
##           num_casos_prueba_ag 0.00000000
##         num_casos_prueba_elisa 0.00000000
##  num_casos_prueba_desconocida 0.00000000
```

```
CNE_tecnica %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

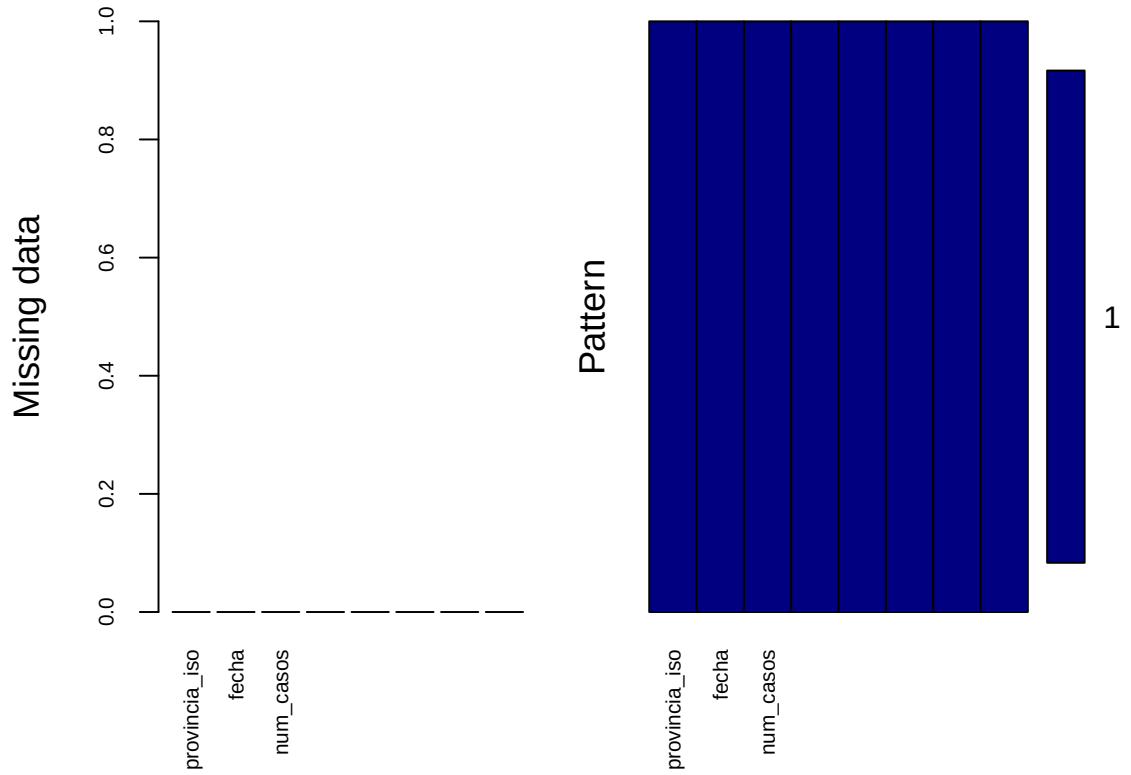



```
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## List of 1
##  $ axis.text.x:List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 1
##   ..$ vjust        : NULL
##   ..$ angle        : num 45
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

```
####################################
CNE_tecnica <- na.omit(CNE_tecnica)
####################################

aggr(CNE_tecnica, col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(CNE_tecnica), cex.axis=.7,
```

```
     gap=3, ylab=c("Missing data","Pattern"))
```




```
##
##  Variables sorted by number of missings:
##                       Variable Count
##                  provincia_iso     0
##                          fecha     0
##                      num_casos     0
##           num_casos_prueba_pcr     0
##       num_casos_prueba_test_ac     0
##            num_casos_prueba_ag     0
##          num_casos_prueba_elisa     0
##  num_casos_prueba_desconocida     0
```

```
CNE_tecnica %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
```

```
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


```
summary(CNE_casos)
```

```
##  provincia_iso          sexo             grupo_edad          fecha
##  Length:702780      Length:702780      Length:702780      Length:702780
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    num_casos          num_hosp           num_uci            num_def
##  Min.   :  0.000   Min.   :  0.0000   Min.   : 0.00000   Min.   :  0.0000
##  1st Qu.:  0.000   1st Qu.:  0.0000   1st Qu.: 0.00000   1st Qu.:  0.0000
##  Median :  0.000   Median :  0.0000   Median : 0.00000   Median :  0.0000
##  Mean   :  4.562   Mean   :  0.4611   Mean   : 0.04117   Mean   :  0.1036
##  3rd Qu.:  2.000   3rd Qu.:  0.0000   3rd Qu.: 0.00000   3rd Qu.:  0.0000
##  Max.   :771.000   Max.   :269.0000   Max.   :35.00000   Max.   :100.0000
```

```
head(str(CNE_casos,vec.len=3))
```

```
## 'data.frame':    702780 obs. of  8 variables:
##  $ provincia_iso: chr  "A" "A" "A" ...
##  $ sexo         : chr  "H" "H" "H" ...
##  $ grupo_edad   : chr  "0-9" "10-19" "20-29" ...
##  $ fecha        : chr  "2020-01-01" "2020-01-01" "2020-01-01" ...
##  $ num_casos    : int  0 0 0 0 0 0 0 0 ...
```

```
##  $ num_hosp     : int  0 0 0 0 0 0 0 0 ...
##  $ num_uci      : int  0 0 0 0 0 0 0 0 ...
##  $ num_def      : int  0 0 0 0 0 0 0 0 ...
```

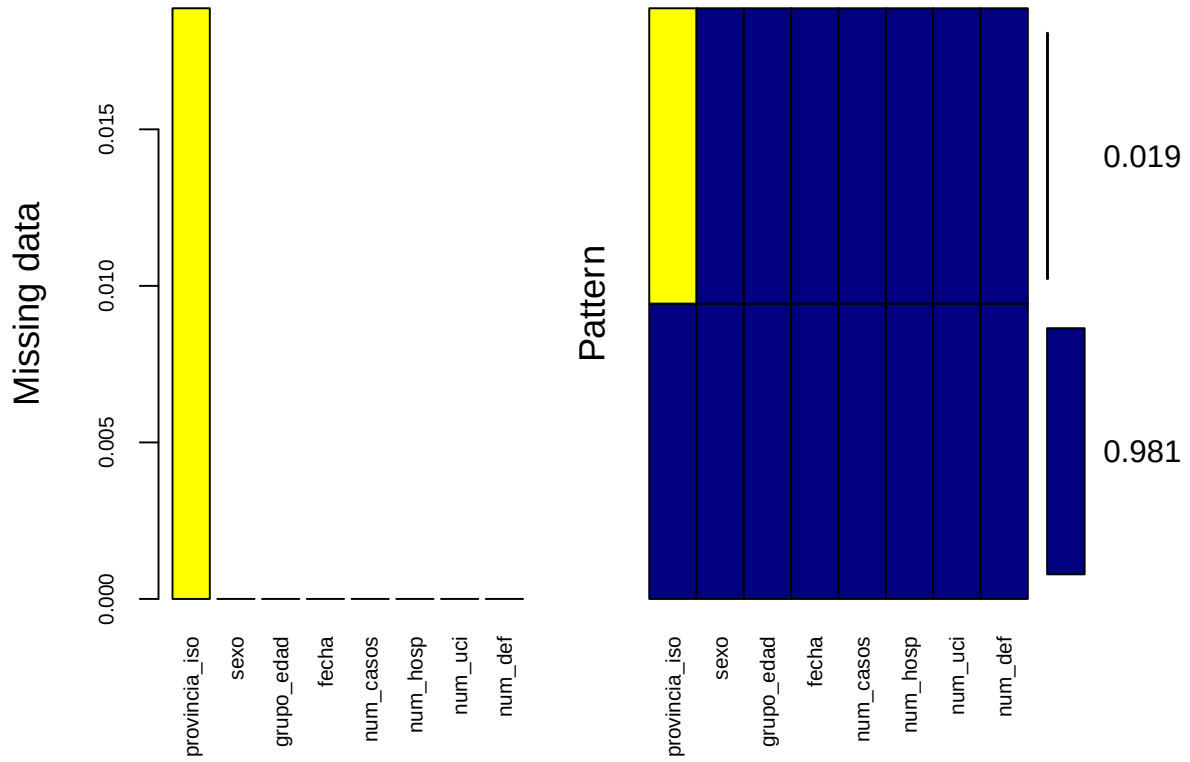
```
## NULL
```

```
table(CNE_casos$provincia_iso)
```

```
##
##     A    AB    AL    AV     B    BA    BI    BU     C    CA    CC    CE    CO
## 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260
##    CR    CS    CU    GC    GI    GR    GU     H    HU     J     L    LE    LO
## 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260
##    LU     M    MA    ML    MU    NC     O    OR     P    PM    PO     S    SA
## 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260
##    SE    SG    SO    SS     T    TE    TF    TO     V    VA    VI     Z    ZA
## 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260
```

We check missing values for CNE_casos. In this case also we omit the NA values.

```
aggr(CNE_casos, col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(CNE_casos), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```



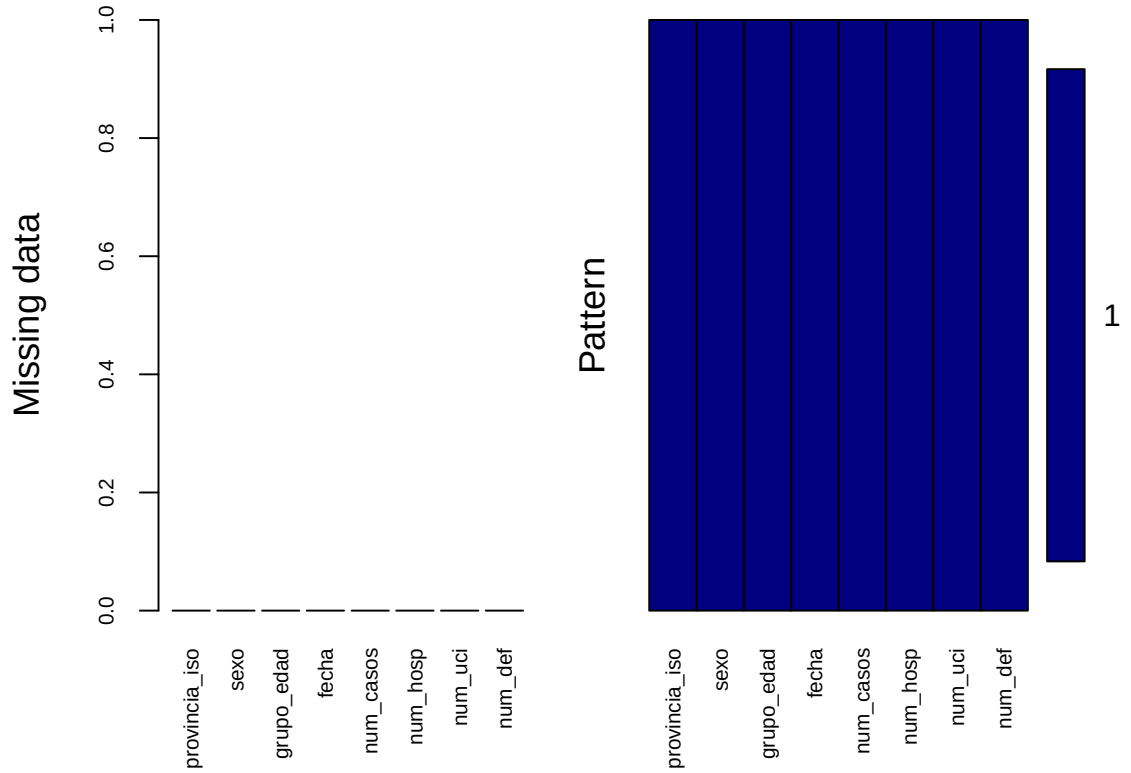

```
##
##  Variables sorted by number of missings:
##       Variable      Count
##  provincia_iso 0.01886792
```

```
##            sexo 0.00000000
##      grupo_edad 0.00000000
##           fecha 0.00000000
##       num_casos 0.00000000
##        num_hosp 0.00000000
##         num_uci 0.00000000
##         num_def 0.00000000
```

```
CNE_casos %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


```
################################
CNE_casos <- na.omit(CNE_casos)
################################

aggr(CNE_casos, col=c('navyblue','yellow'),
```

```
     numbers=TRUE, sortVars=TRUE,
     labels=names(CNE_casos), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```




```
##
##  Variables sorted by number of missings:
##       Variable Count
##  provincia_iso     0
##           sexo     0
##     grupo_edad     0
##          fecha     0
##      num_casos     0
##       num_hosp     0
##        num_uci     0
##        num_def     0
```

```
CNE_casos %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
```

```
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

2.1.11 CNE data transformation

We are going to **transform / eliminate**:

- A - "Fecha" column is transformed (in both datasets) from "character" to "date".
- B - "Grupo_edad" and "Sexo" columns are eliminated from dataset "CNE_casos" due to they are not adding value (mobility does not include this variable).
- C - We change NC iso code to NA (Navarra) in both dataframes.

```
# Transform / eliminate A
CNE_tecnica$fecha <- as.Date(CNE_tecnica$fecha ,format="%Y-%m-%d")
CNE_casos$fecha <- as.Date(CNE_casos$fecha ,format="%Y-%m-%d")

# Transform / eliminate B
CNE_casos<-within(CNE_casos, rm(grupo_edad, sexo))

# Iso code update for Navarra C
CNE_tecnica$provincia_iso[CNE_tecnica$provincia_iso=="NC"] <- "NA"
CNE_casos$provincia_iso[CNE_casos$provincia_iso=="NC"] <- "NA"

# Check table
head(CNE_tecnica,5)
```

```
##   provincia_iso      fecha num_casos num_casos_prueba_pcr
## 1             A 2020-01-01         0                    0
## 2            AB 2020-01-01         0                    0
## 3            AL 2020-01-01         0                    0
## 4            AV 2020-01-01         0                    0
## 5             B 2020-01-01         0                    0
##   num_casos_prueba_test_ac num_casos_prueba_ag num_casos_prueba_elisa
## 1                        0                   0                      0
## 2                        0                   0                      0
## 3                        0                   0                      0
## 4                        0                   0                      0
## 5                        0                   0                      0
##   num_casos_prueba_desconocida
## 1                            0
## 2                            0
## 3                            0
## 4                            0
## 5                            0
```

```
head(CNE_casos,5)
```

```
##   provincia_iso      fecha num_casos num_hosp num_uci num_def
## 1             A 2020-01-01         0        0       0       0
## 2             A 2020-01-01         0        0       0       0
## 3             A 2020-01-01         0        0       0       0
## 4             A 2020-01-01         0        0       0       0
## 5             A 2020-01-01         0        0       0       0
```

We check both dataframes offers the same total results.

```
CNE_tecnica %>%
  group_by(provincia_iso) %>%
  summarise_at(vars(num_casos), sum)
```

```
## # A tibble: 52 x 2
##    provincia_iso num_casos
##    <chr>             <int>
##  1 A                143555
##  2 AB                26916
##  3 AL                47032
##  4 AV                11084
##  5 B                382992
##  6 BA                45886
##  7 BI                80588
##  8 BU                29808
##  9 C                 51272
## 10 CA                70428
## # ... with 42 more rows
```

```
CNE_casos %>%
  group_by(provincia_iso) %>%
  summarise_at(vars(num_casos), sum)
```

```
## # A tibble: 52 x 2
##    provincia_iso num_casos
##    <chr>             <int>
##  1 A                143555
```

```
##  2 AB                26916
##  3 AL                47032
##  4 AV                11084
##  5 B                382992
##  6 BA                45886
##  7 BI                80588
##  8 BU                29808
##  9 C                 51272
## 10 CA                70428
## # ... with 42 more rows
```

2.2 Datasets combinations

We proceed to **combine** the different data sets into one.

2.2.1 CNE_tec_cas

- CNE_casos_g, a groupped dataframe due to the columns eliminated in previous step (grupo_edad, sexo)
- CNE_tec_cas -> CNE_tecnica + CNE_casos_g

Here we merge by columns "provincia_iso","fecha".

```
# CNE_casos_g
CNE_casos_g = CNE_casos %>%
  group_by(provincia_iso, fecha) %>%
  summarise_at(vars(num_casos, num_hosp, num_uci, num_def), sum)
head(CNE_casos_g,5)
```

```
## # A tibble: 5 x 6
## # Groups:   provincia_iso [1]
##   provincia_iso fecha      num_casos num_hosp num_uci num_def
##   <chr>         <date>         <int>    <int>   <int>   <int>
## 1 A             2020-01-01         0        1       0       0
## 2 A             2020-01-02         0        0       0       0
## 3 A             2020-01-03         0        0       0       0
## 4 A             2020-01-04         0        0       0       0
## 5 A             2020-01-05         0        1       0       0
```

```
# New dataframe CNE_tec_cas
CNE_tec_cas<-merge(CNE_tecnica,
                   CNE_casos_g, by.x=c("provincia_iso","fecha"),
                   by.y=c("provincia_iso","fecha"))

# We check both dataframes offers the same total results
CNE_tecnica %>%
  group_by(provincia_iso) %>%
  summarise_at(vars(num_casos), sum)
```

```
## # A tibble: 52 x 2
##    provincia_iso num_casos
##    <chr>             <int>
##  1 A                143555
##  2 AB                26916
##  3 AL                47032
##  4 AV                11084
##  5 B                382992
```

```
##  6 BA                45886
##  7 BI                80588
##  8 BU                29808
##  9 C                 51272
## 10 CA                70428
## # ... with 42 more rows
```

```
CNE_casos_g %>%
  group_by(provincia_iso) %>%
  summarise_at(vars(num_casos), sum)
```

```
## # A tibble: 52 x 2
##    provincia_iso num_casos
##    <chr>             <int>
##  1 A                143555
##  2 AB                26916
##  3 AL                47032
##  4 AV                11084
##  5 B                382992
##  6 BA                45886
##  7 BI                80588
##  8 BU                29808
##  9 C                 51272
## 10 CA                70428
## # ... with 42 more rows
```

```
head(CNE_tec_cas,5)
```

```
##   provincia_iso      fecha num_casos.x num_casos_prueba_pcr
## 1             A 2020-01-01           0                    0
## 2             A 2020-01-02           0                    0
## 3             A 2020-01-03           0                    0
## 4             A 2020-01-04           0                    0
## 5             A 2020-01-05           0                    0
##   num_casos_prueba_test_ac num_casos_prueba_ag num_casos_prueba_elisa
## 1                        0                   0                     0
## 2                        0                   0                     0
## 3                        0                   0                     0
## 4                        0                   0                     0
## 5                        0                   0                     0
##   num_casos_prueba_desconocida num_casos.y num_hosp num_uci num_def
## 1                            0           0        1       0       0
## 2                            0           0        0       0       0
## 3                            0           0        0       0       0
## 4                            0           0        0       0       0
## 5                            0           0        1       0       0
```

```
table(CNE_tec_cas$provincia_iso)
```

```
##
##   A  AB  AL  AV   B  BA  BI  BU   C  CA  CC  CE  CO  CR  CS  CU  GC  GI  GR  GU
## 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442
##   H  HU   J   L  LE  LO  LU   M  MA  ML  MU  NA   O  OR   P  PM  PO   S  SA  SE
## 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442
##  SG  SO  SS   T  TE  TF  TO   V  VA  VI   Z  ZA
## 442 442 442 442 442 442 442 442 442 442 442 442
```

2.2.2 GOG_CNE

- GOG_CNE -> CNE_tec_cas + Google

Here we merge by columns "provincia_iso" / "fecha" and "iso_3166_2_code" / "date".

```
# New dataframe GOG_CNE
GOG_CNE<-merge(CNE_tec_cas,
               Google,
               by.x=c("provincia_iso","fecha"),
               by.y=c("iso_code","Date"))
head(GOG_CNE,5)
```

```
##   provincia_iso      fecha num_casos.x num_casos_prueba_pcr
## 1             A 2020-02-15           1                    1
## 2             A 2020-02-16           1                    1
## 3             A 2020-02-17           1                    1
## 4             A 2020-02-18           1                    1
## 5             A 2020-02-19           1                    1
##   num_casos_prueba_test_ac num_casos_prueba_ag num_casos_prueba_elisa
## 1                        0                   0                     0
## 2                        0                   0                     0
## 3                        0                   0                     0
## 4                        0                   0                     0
## 5                        0                   0                     0
##   num_casos_prueba_desconocida num_casos.y num_hosp num_uci num_def
## 1                            0           0        1       0       0
## 2                            0           0        0       0       0
## 3                            0           0        1       0       0
## 4                            0           0        1       0       0
## 5                            0           0        2       1       0
##       sub_region_2 retail_and_recreation_percent_change_from_baseline
## 1 Alicante/Alacant                                                 3
## 2 Alicante/Alacant                                                -2
## 3 Alicante/Alacant                                                 0
## 4 Alicante/Alacant                                                -5
## 5 Alicante/Alacant                                                 1
##   grocery_and_pharmacy_percent_change_from_baseline
## 1                                                -1
## 2                                                 1
## 3                                                 2
## 4                                                -2
## 5                                                 1
##   parks_percent_change_from_baseline
## 1                                 34
## 2                                  8
## 3                                  9
## 4                                -14
## 5                                 10
##   transit_stations_percent_change_from_baseline
## 1                                             7
## 2                                             5
## 3                                             7
## 4                                            -2
## 5                                             3
##   workplaces_percent_change_from_baseline
```

```
## 1                                      0
## 2                                     -2
## 3                                      3
## 4                                      2
## 5                                      3
##   residential_percent_change_from_baseline
## 1                                       -1
## 2                                       -1
## 3                                        0
## 4                                        1
## 5                                        0
```

```
table(GOG_CNE$provincia_iso)
```

```
##
##   A  AB  AL  AV   B  BA  BI  BU   C  CA  CC  CE  CO  CR  CS  CU  GC  GI  GR  GU
## 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385
##   H  HU   J   L  LE  LO  LU   M  MA  ML  MU  NA   O  OR   P  PM  PO   S  SA  SE
## 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385
##  SG  SO  SS   T  TE  TF  TO   V  VA  VI   Z  ZA
## 385 385 385 385 385 385 385 385 385 385 385 385
```

2.2.3 Total

- Total -> GOG_CNE + EM3

Here we merge by columns "sub_region_2" / "fecha" and "Zonas.de.movilidad" / "Periodo". With this dataset we have 21 features for study.

```
# New dataframe Total
Total<-merge(GOG_CNE,
             EM3,
             by.x=c("sub_region_2","fecha"),
             by.y=c("Zonas.de.movilidad","Periodo"))

head(Total,5)
```

```
##   sub_region_2      fecha provincia_iso num_casos.x num_casos_prueba_pcr
## 1     Albacete 2020-03-16            AB         137                  132
## 2     Albacete 2020-03-17            AB         128                  123
## 3     Albacete 2020-03-18            AB         114                  107
## 4     Albacete 2020-03-19            AB         149                  133
## 5     Albacete 2020-03-20            AB         131                  121
##   num_casos_prueba_test_ac num_casos_prueba_ag num_casos_prueba_elisa
## 1                        5                   0                      0
## 2                        5                   0                      0
## 3                        7                   0                      0
## 4                       16                   0                      0
## 5                       10                   0                      0
##   num_casos_prueba_desconocida num_casos.y num_hosp num_uci num_def
## 1                            0          65       43       3       7
## 2                            0          29       40       4       2
## 3                            0          26       24       7       7
## 4                            0          22       40       5       7
## 5                            0          85       63       4       6
##   retail_and_recreation_percent_change_from_baseline
```

```
## 1                                                -81
## 2                                                -84
## 3                                                -83
## 4                                                -93
## 5                                                -87
##   grocery_and_pharmacy_percent_change_from_baseline
## 1                                               -32
## 2                                               -41
## 3                                               -32
## 4                                               -92
## 5                                               -34
##   parks_percent_change_from_baseline
## 1                                -73
## 2                                -74
## 3                                -70
## 4                                -80
## 5                                -74
##   transit_stations_percent_change_from_baseline
## 1                                           -66
## 2                                           -72
## 3                                           -70
## 4                                           -86
## 5                                           -76
##   workplaces_percent_change_from_baseline
## 1                                     -51
## 2                                     -56
## 3                                     -58
## 4                                     -85
## 5                                     -68
##   residential_percent_change_from_baseline Total
## 1                                       22 9.900
## 2                                       23 9.705
## 3                                       23 9.510
## 4                                       35 9.130
## 5                                       32 8.750
```

```
head(str(Total,vec.len=1))
```

```
## 'data.frame':    15080 obs. of  20 variables:
##  $ sub_region_2                                          : chr  "Albacete" ...
##  $ fecha                                                 : Date, format: "2020-03-16" ...
##  $ provincia_iso                                         : chr  "AB" ...
##  $ num_casos.x                                           : int  137 128 ...
##  $ num_casos_prueba_pcr                                  : int  132 123 ...
##  $ num_casos_prueba_test_ac                              : int  5 5 ...
##  $ num_casos_prueba_ag                                   : int  0 0 ...
##  $ num_casos_prueba_elisa                                : int  0 0 ...
##  $ num_casos_prueba_desconocida                          : int  0 0 ...
##  $ num_casos.y                                           : int  65 29 ...
##  $ num_hosp                                              : int  43 40 ...
##  $ num_uci                                               : int  3 4 ...
##  $ num_def                                               : int  7 2 ...
##  $ retail_and_recreation_percent_change_from_baseline: num  -81 -84 ...
##  $ grocery_and_pharmacy_percent_change_from_baseline : num  -32 -41 ...
##  $ parks_percent_change_from_baseline                : num  -73 -74 ...
```

```
##  $ transit_stations_percent_change_from_baseline     : num  -66 -72 ...
##  $ workplaces_percent_change_from_baseline           : num  -51 -56 ...
##  $ residential_percent_change_from_baseline          : num  22 23 ...
##  $ Total                                             : num  9.9 ...
```

```
## NULL
```

```
summary(Total)
```

```
##  sub_region_2           fecha            provincia_iso       num_casos.x
##  Length:15080       Min.   :2020-03-16   Length:15080       Min.   :   0
##  Class :character   1st Qu.:2020-05-27   Class :character   1st Qu.:   5
##  Mode  :character   Median :2020-08-07   Mode  :character   Median :  39
##                     Mean   :2020-08-07                      Mean   : 126
##                     3rd Qu.:2020-10-19                      3rd Qu.: 120
##                     Max.   :2020-12-30                      Max.   :6565
##  num_casos_prueba_pcr num_casos_prueba_test_ac num_casos_prueba_ag
##  Min.   :   0.0       Min.   : 0.0000          Min.   :   0.00
##  1st Qu.:   5.0       1st Qu.: 0.0000          1st Qu.:   0.00
##  Median :  35.0       Median : 0.0000          Median :   0.00
##  Mean   : 110.2       Mean   : 0.2832          Mean   :  15.19
##  3rd Qu.: 105.0       3rd Qu.: 0.0000          3rd Qu.:   4.00
##  Max.   :6546.0       Max.   :32.0000          Max.   :1465.00
##  num_casos_prueba_elisa num_casos_prueba_desconocida  num_casos.y
##  Min.   : 0.0000        Min.   : 0.0000              Min.   :   0
##  1st Qu.: 0.0000        1st Qu.: 0.0000              1st Qu.:   6
##  Median : 0.0000        Median : 0.0000              Median :  37
##  Mean   : 0.1989        Mean   : 0.1317              Mean   : 127
##  3rd Qu.: 0.0000        3rd Qu.: 0.0000              3rd Qu.: 117
##  Max.   :71.0000        Max.   :65.0000              Max.   :7724
##     num_hosp          num_uci           num_def
##  Min.   :   0.00   Min.   :  0.000   Min.   :  0.000
##  1st Qu.:   1.00   1st Qu.:  0.000   1st Qu.:  0.000
##  Median :   4.00   Median :  0.000   Median :  1.000
##  Mean   :  14.86   Mean   :  1.281   Mean   :  3.437
##  3rd Qu.:  12.00   3rd Qu.:  1.000   3rd Qu.:  3.000
##  Max.   :1930.00   Max.   :135.000   Max.   :334.000
##  retail_and_recreation_percent_change_from_baseline
##  Min.   :-97.00
##  1st Qu.:-57.00
##  Median :-30.00
##  Mean   :-37.29
##  3rd Qu.:-17.00
##  Max.   : 71.00
##  grocery_and_pharmacy_percent_change_from_baseline
##  Min.   :-96.00
##  1st Qu.:-24.00
##  Median : -6.00
##  Mean   :-11.75
##  3rd Qu.:  4.00
##  Max.   :194.00
##  parks_percent_change_from_baseline
##  Min.   :-94.000
##  1st Qu.:-30.000
##  Median : -2.000
```

```
##  Mean   :  5.809
##  3rd Qu.: 30.000
##  Max.   :543.000
##  transit_stations_percent_change_from_baseline
##  Min.   :-100.00
##  1st Qu.: -53.00
##  Median : -31.00
##  Mean   : -35.19
##  3rd Qu.: -17.00
##  Max.   :  74.00
##  workplaces_percent_change_from_baseline
##  Min.   :-92.00
##  1st Qu.:-43.00
##  Median :-26.00
##  Mean   :-29.08
##  3rd Qu.:-13.00
##  Max.   : 55.00
##  residential_percent_change_from_baseline     Total
##  Min.   :-10.00                           Min.   : 1.95
##  1st Qu.:  4.00                           1st Qu.:11.36
##  Median :  7.00                           Median :14.39
##  Mean   : 10.14                           Mean   :14.20
##  3rd Qu.: 14.00                           3rd Qu.:17.11
##  Max.   : 48.00                           Max.   :29.00
```

```
Total$num_casos.x <- as.numeric(Total$num_casos.x)
Total$num_casos_prueba_pcr <- as.numeric(Total$num_casos_prueba_pcr)
Total$num_casos_prueba_test_ac <- as.numeric(Total$num_casos_prueba_test_ac)
Total$num_casos_prueba_ag <- as.numeric(Total$num_casos_prueba_ag)
Total$num_casos_prueba_elisa <- as.numeric(Total$num_casos_prueba_elisa)
Total$num_casos_prueba_desconocida   <- as.numeric(Total$num_casos_prueba_desconocida)
Total$num_casos.y <- as.numeric(Total$num_casos.y)
Total$num_hosp <- as.numeric(Total$num_hosp)
Total$num_uci <- as.numeric(Total$num_uci)
Total$num_def <- as.numeric(Total$num_def)

table(Total$sub_region_2)
```

```
##
##               Albacete       Alicante/Alacant                Almería
##                    290                    290                    290
##           Araba/Álava               Asturias                  Ávila
##                    290                    290                    290
##                Badajoz        Balears, Illes              Barcelona
##                    290                    290                    290
##                Bizkaia                 Burgos                Cáceres
##                    290                    290                    290
##                  Cádiz              Cantabria     Castellón/Castelló
##                    290                    290                    290
##                  Ceuta           Ciudad Real                Córdoba
##                    290                    290                    290
##              Coruña, A                 Cuenca              Gipuzkoa
##                    290                    290                    290
##                 Girona                Granada           Guadalajara
##                    290                    290                    290
```

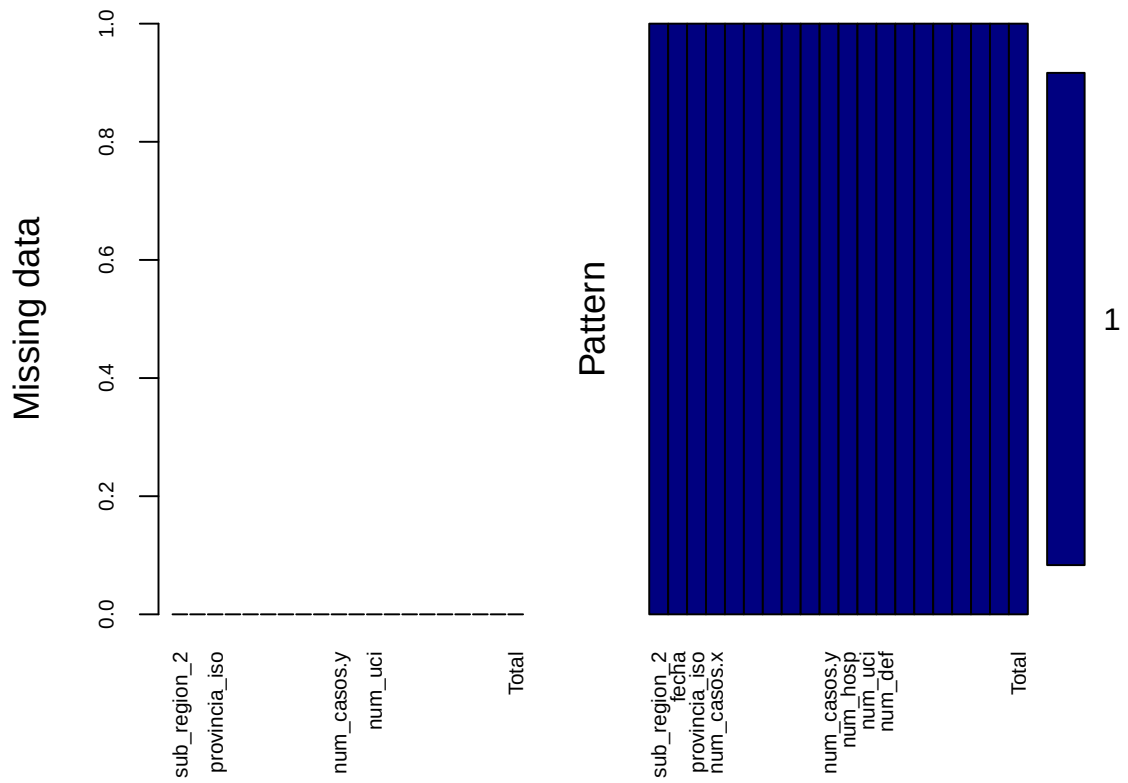
```
##                 Huelva                 Huesca                  Jaén
##                    290                    290                    290
##                  León                 Lleida                  Lugo
##                    290                    290                    290
##                 Madrid                 Málaga                Melilla
##                    290                    290                    290
##                 Murcia                Navarra                Ourense
##                    290                    290                    290
##               Palencia           Palmas, Las             Pontevedra
##                    290                    290                    290
##             Rioja, La              Salamanca Santa Cruz de Tenerife
##                    290                    290                    290
##                Segovia                Sevilla                  Soria
##                    290                    290                    290
##              Tarragona                Teruel                 Toledo
##                    290                    290                    290
##      Valencia/València             Valladolid                 Zamora
##                    290                    290                    290
##               Zaragoza
##                    290
```

```
table(Total$provincia_iso)
```

```
##
##   A  AB  AL  AV   B  BA  BI  BU   C  CA  CC  CE  CO  CR  CS  CU  GC  GI  GR  GU
## 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290
##   H  HU   J   L  LE  LO  LU   M  MA  ML  MU  NA   O  OR   P  PM  PO   S  SA  SE
## 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290
##  SG  SO  SS   T  TE  TF  TO   V  VA  VI   Z  ZA
## 290 290 290 290 290 290 290 290 290 290 290 290
```

We check the missing values. We should have zero missing values

```
aggr(Total, col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(Total), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```




```
##
##  Variables sorted by number of missings:
##                                                 Variable Count
##                                               sub_region_2     0
##                                                      fecha     0
##                                                provincia_iso     0
##                                                  num_casos.x     0
##                                          num_casos_prueba_pcr     0
##                                      num_casos_prueba_test_ac     0
##                                           num_casos_prueba_ag     0
##                                         num_casos_prueba_elisa     0
##                                 num_casos_prueba_desconocida     0
##                                                  num_casos.y     0
##                                                     num_hosp     0
##                                                      num_uci     0
##                                                      num_def     0
##  retail_and_recreation_percent_change_from_baseline     0
##   grocery_and_pharmacy_percent_change_from_baseline     0
##                  parks_percent_change_from_baseline     0
##       transit_stations_percent_change_from_baseline     0
##            workplaces_percent_change_from_baseline     0
##           residential_percent_change_from_baseline     0
##                                                 Total     0
```

```
Total %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
```

```
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


```
# Review results
# Discrepancies due to different time-frames when merge CNE dataframes (see previous checks)
Total %>%
  group_by(provincia_iso) %>%
  summarise_at(vars(num_casos.x,num_casos.y), sum)
```

```
## # A tibble: 52 x 3
##    provincia_iso num_casos.x num_casos.y
##    <chr>               <dbl>       <dbl>
##  1 A                   56493       55068
##  2 AB                  16459       16626
##  3 AL                  21488       21372
##  4 AV                   6525        6681
##  5 B                  257034      261208
##  6 BA                  23165       22612
```

```
##  7 BI                  56867       57728
##  8 BU                  22742       22978
##  9 C                   25641       25604
## 10 CA                  31593       31225
## # ... with 42 more rows
```

```
# CSV file generation
head(Total,5)
```

```
##   sub_region_2      fecha provincia_iso num_casos.x num_casos_prueba_pcr
## 1     Albacete 2020-03-16            AB         137                  132
## 2     Albacete 2020-03-17            AB         128                  123
## 3     Albacete 2020-03-18            AB         114                  107
## 4     Albacete 2020-03-19            AB         149                  133
## 5     Albacete 2020-03-20            AB         131                  121
##   num_casos_prueba_test_ac num_casos_prueba_ag num_casos_prueba_elisa
## 1                        5                   0                      0
## 2                        5                   0                      0
## 3                        7                   0                      0
## 4                       16                   0                      0
## 5                       10                   0                      0
##   num_casos_prueba_desconocida num_casos.y num_hosp num_uci num_def
## 1                            0          65       43       3       7
## 2                            0          29       40       4       2
## 3                            0          26       24       7       7
## 4                            0          22       40       5       7
## 5                            0          85       63       4       6
##   retail_and_recreation_percent_change_from_baseline
## 1                                                -81
## 2                                                -84
## 3                                                -83
## 4                                                -93
## 5                                                -87
##   grocery_and_pharmacy_percent_change_from_baseline
## 1                                               -32
## 2                                               -41
## 3                                               -32
## 4                                               -92
## 5                                               -34
##   parks_percent_change_from_baseline
## 1                                -73
## 2                                -74
## 3                                -70
## 4                                -80
## 5                                -74
##   transit_stations_percent_change_from_baseline
## 1                                           -66
## 2                                           -72
## 3                                           -70
## 4                                           -86
## 5                                           -76
##   workplaces_percent_change_from_baseline
## 1                                     -51
## 2                                     -56
## 3                                     -58
```

```
## 4                                    -85
## 5                                    -68
##   residential_percent_change_from_baseline Total
## 1                                       22 9.900
## 2                                       23 9.705
## 3                                       23 9.510
## 4                                       35 9.130
## 5                                       32 8.750
```

```
head(str(Total,vec.len=1))
```

```
## 'data.frame':    15080 obs. of  20 variables:
##  $ sub_region_2                                      : chr  "Albacete" ...
##  $ fecha                                             : Date, format: "2020-03-16" ...
##  $ provincia_iso                                     : chr  "AB" ...
##  $ num_casos.x                                       : num  137 128 ...
##  $ num_casos_prueba_pcr                              : num  132 123 ...
##  $ num_casos_prueba_test_ac                          : num  5 5 ...
##  $ num_casos_prueba_ag                               : num  0 0 ...
##  $ num_casos_prueba_elisa                            : num  0 0 ...
##  $ num_casos_prueba_desconocida                      : num  0 0 ...
##  $ num_casos.y                                       : num  65 29 ...
##  $ num_hosp                                          : num  43 40 ...
##  $ num_uci                                           : num  3 4 ...
##  $ num_def                                           : num  7 2 ...
##  $ retail_and_recreation_percent_change_from_baseline: num  -81 -84 ...
##  $ grocery_and_pharmacy_percent_change_from_baseline : num  -32 -41 ...
##  $ parks_percent_change_from_baseline                : num  -73 -74 ...
##  $ transit_stations_percent_change_from_baseline     : num  -66 -72 ...
##  $ workplaces_percent_change_from_baseline           : num  -51 -56 ...
##  $ residential_percent_change_from_baseline          : num  22 23 ...
##  $ Total                                             : num  9.9 ...
```

```
## NULL
```

```
summary(Total)
```

```
##  sub_region_2           fecha            provincia_iso       num_casos.x
##  Length:15080       Min.   :2020-03-16   Length:15080       Min.   :   0
##  Class :character   1st Qu.:2020-05-27   Class :character   1st Qu.:   5
##  Mode  :character   Median :2020-08-07   Mode  :character   Median :  39
##                     Mean   :2020-08-07                      Mean   : 126
##                     3rd Qu.:2020-10-19                      3rd Qu.: 120
##                     Max.   :2020-12-30                      Max.   :6565
##  num_casos_prueba_pcr num_casos_prueba_test_ac num_casos_prueba_ag
##  Min.   :   0.0       Min.   : 0.0000          Min.   :   0.00
##  1st Qu.:   5.0       1st Qu.: 0.0000          1st Qu.:   0.00
##  Median :  35.0       Median : 0.0000          Median :   0.00
##  Mean   : 110.2       Mean   : 0.2832          Mean   :  15.19
##  3rd Qu.: 105.0       3rd Qu.: 0.0000          3rd Qu.:   4.00
##  Max.   :6546.0       Max.   :32.0000          Max.   :1465.00
##  num_casos_prueba_elisa num_casos_prueba_desconocida  num_casos.y
##  Min.   : 0.0000        Min.   : 0.0000              Min.   :   0
##  1st Qu.: 0.0000        1st Qu.: 0.0000              1st Qu.:   6
##  Median : 0.0000        Median : 0.0000              Median :  37
##  Mean   : 0.1989        Mean   : 0.1317              Mean   : 127
```

```
##  3rd Qu.: 0.0000        3rd Qu.: 0.0000              3rd Qu.: 117
##  Max.   :71.0000        Max.   :65.0000              Max.   :7724
##     num_hosp          num_uci           num_def
##  Min.   :   0.00   Min.   :  0.000   Min.   :  0.000
##  1st Qu.:   1.00   1st Qu.:  0.000   1st Qu.:  0.000
##  Median :   4.00   Median :  0.000   Median :  1.000
##  Mean   :  14.86   Mean   :  1.281   Mean   :  3.437
##  3rd Qu.:  12.00   3rd Qu.:  1.000   3rd Qu.:  3.000
##  Max.   :1930.00   Max.   :135.000   Max.   :334.000
##  retail_and_recreation_percent_change_from_baseline
##  Min.   :-97.00
##  1st Qu.:-57.00
##  Median :-30.00
##  Mean   :-37.29
##  3rd Qu.:-17.00
##  Max.   : 71.00
##  grocery_and_pharmacy_percent_change_from_baseline
##  Min.   :-96.00
##  1st Qu.:-24.00
##  Median : -6.00
##  Mean   :-11.75
##  3rd Qu.:  4.00
##  Max.   :194.00
##  parks_percent_change_from_baseline
##  Min.   :-94.000
##  1st Qu.:-30.000
##  Median : -2.000
##  Mean   :  5.809
##  3rd Qu.: 30.000
##  Max.   :543.000
##  transit_stations_percent_change_from_baseline
##  Min.   :-100.00
##  1st Qu.: -53.00
##  Median : -31.00
##  Mean   : -35.19
##  3rd Qu.: -17.00
##  Max.   :  74.00
##  workplaces_percent_change_from_baseline
##  Min.   :-92.00
##  1st Qu.:-43.00
##  Median :-26.00
##  Mean   :-29.08
##  3rd Qu.:-13.00
##  Max.   : 55.00
##  residential_percent_change_from_baseline     Total
##  Min.   :-10.00                           Min.   : 1.95
##  1st Qu.:  4.00                           1st Qu.:11.36
##  Median :  7.00                           Median :14.39
##  Mean   : 10.14                           Mean   :14.20
##  3rd Qu.: 14.00                           3rd Qu.:17.11
##  Max.   : 48.00                           Max.   :29.00
```

```
table(Total$provincia_iso)
```

```
##
```

```
##   A  AB  AL  AV   B  BA  BI  BU   C  CA  CC  CE  CO  CR  CS  CU  GC  GI  GR  GU
## 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290
##   H  HU   J   L  LE  LO  LU   M  MA  ML  MU  NA   O  OR   P  PM  PO   S  SA  SE
## 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290
##  SG  SO  SS   T  TE  TF  TO   V  VA  VI   Z  ZA
## 290 290 290 290 290 290 290 290 290 290 290 290
```

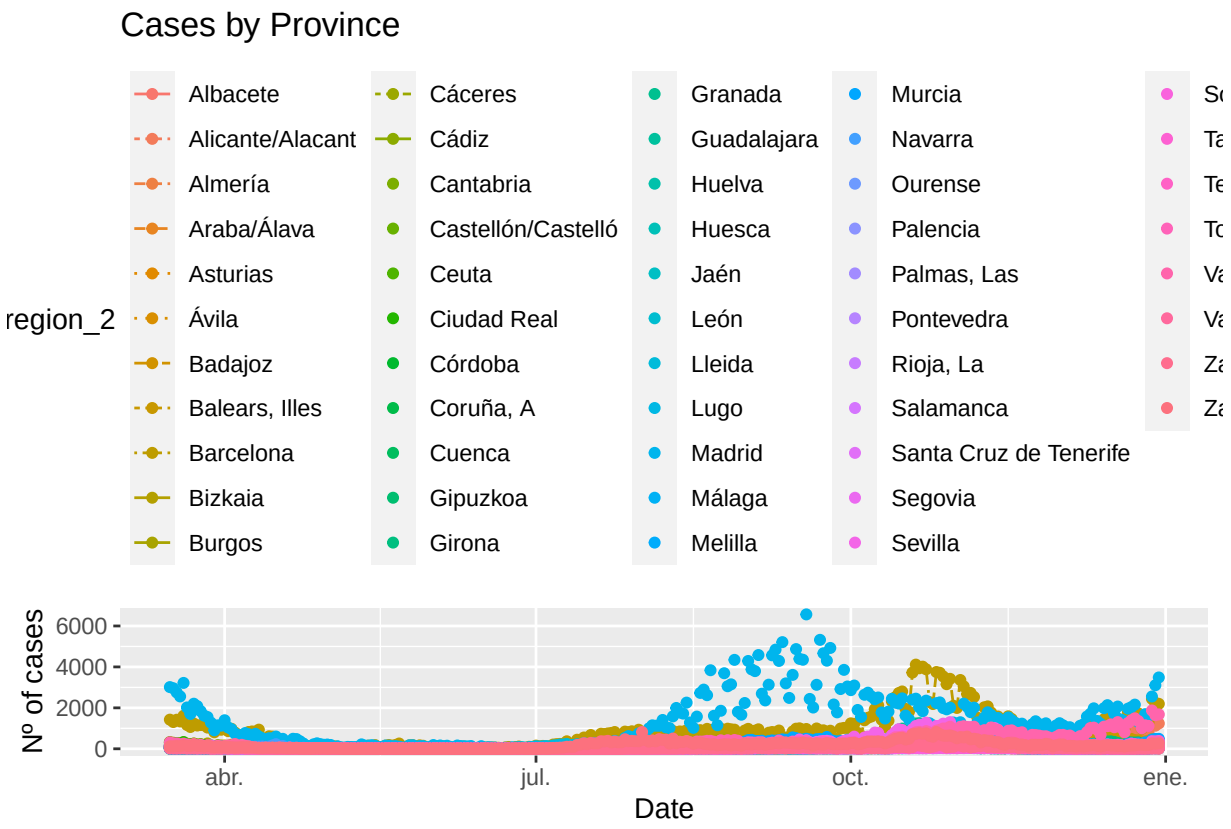
```
write.csv2(Total,"D:\\UOC Master Data Science\\_ M2.882 - TFM - Área 5\\UOC - Guia - PECS\\Pec3\\Total.
           row.names = FALSE)
```

2.3 Visual analysis

2.3.1 Dataframe plots (Málaga, Sevilla and Cádiz)

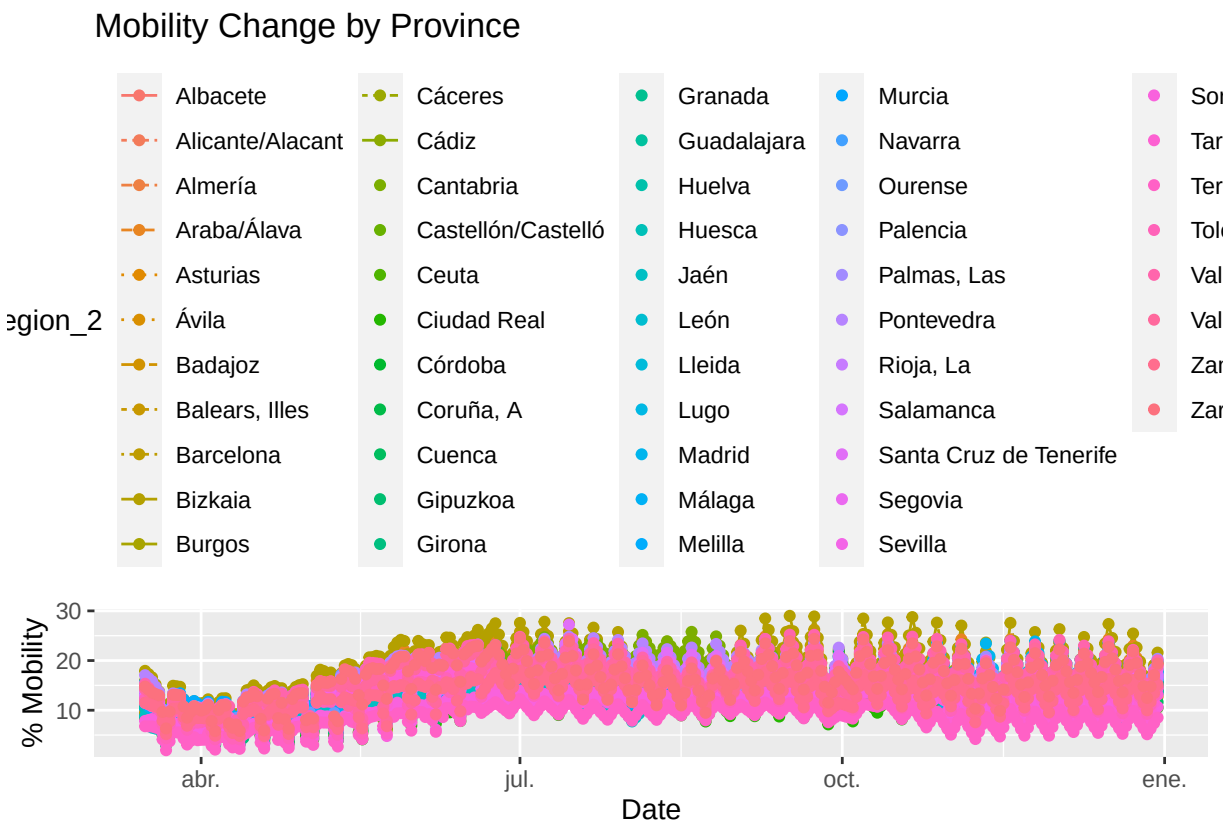
We have generated some plots from the **dataframe** object generated.

```
# Line plots
# All num_casos.x
ggplot(Total, aes(x=fecha, y=num_casos.x, group=sub_region_2)) +
  geom_line(aes(linetype=sub_region_2, color=sub_region_2))+
  geom_point(aes(color=sub_region_2))+
  theme(legend.position="top") +
  labs(title="Cases by Province",
       x ="Date", y = "Nº of cases")
```

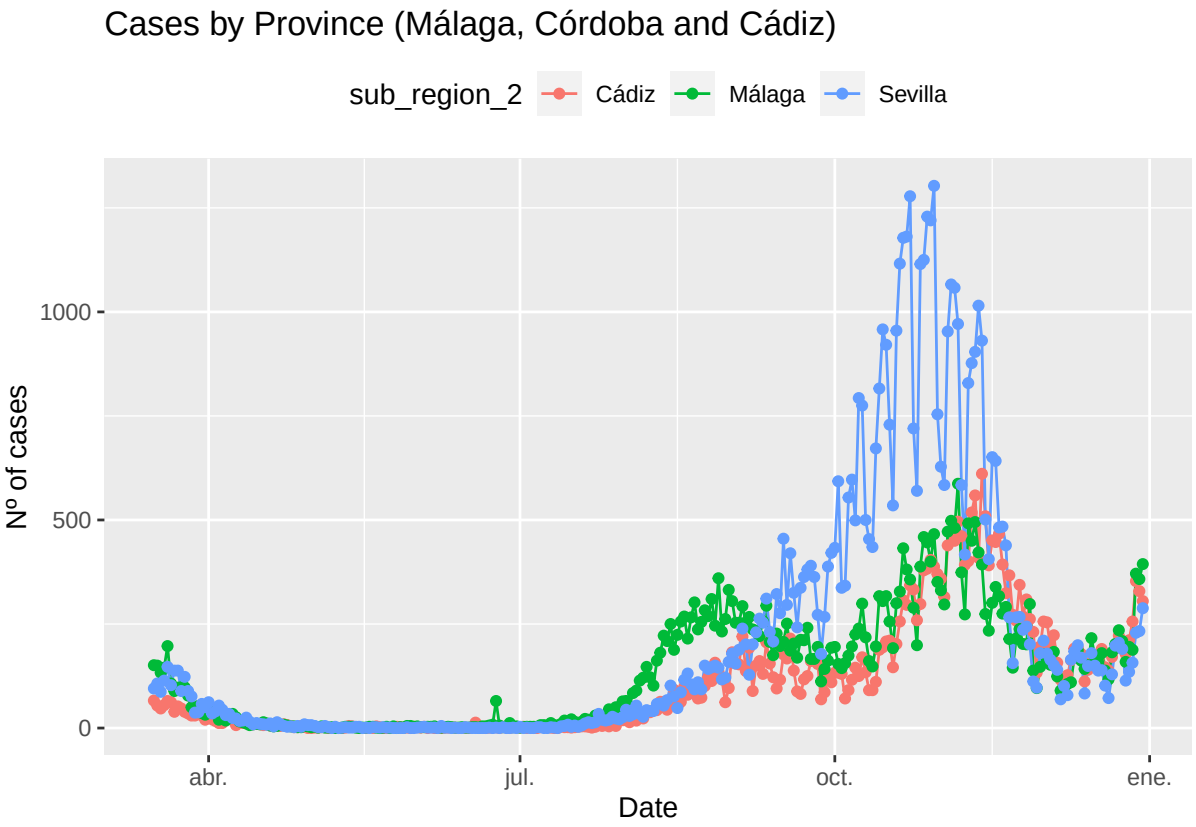



```
# All Total (mobility)
ggplot(Total, aes(x=fecha, y=Total, group=sub_region_2)) +
  geom_line(aes(linetype=sub_region_2, color=sub_region_2))+
```

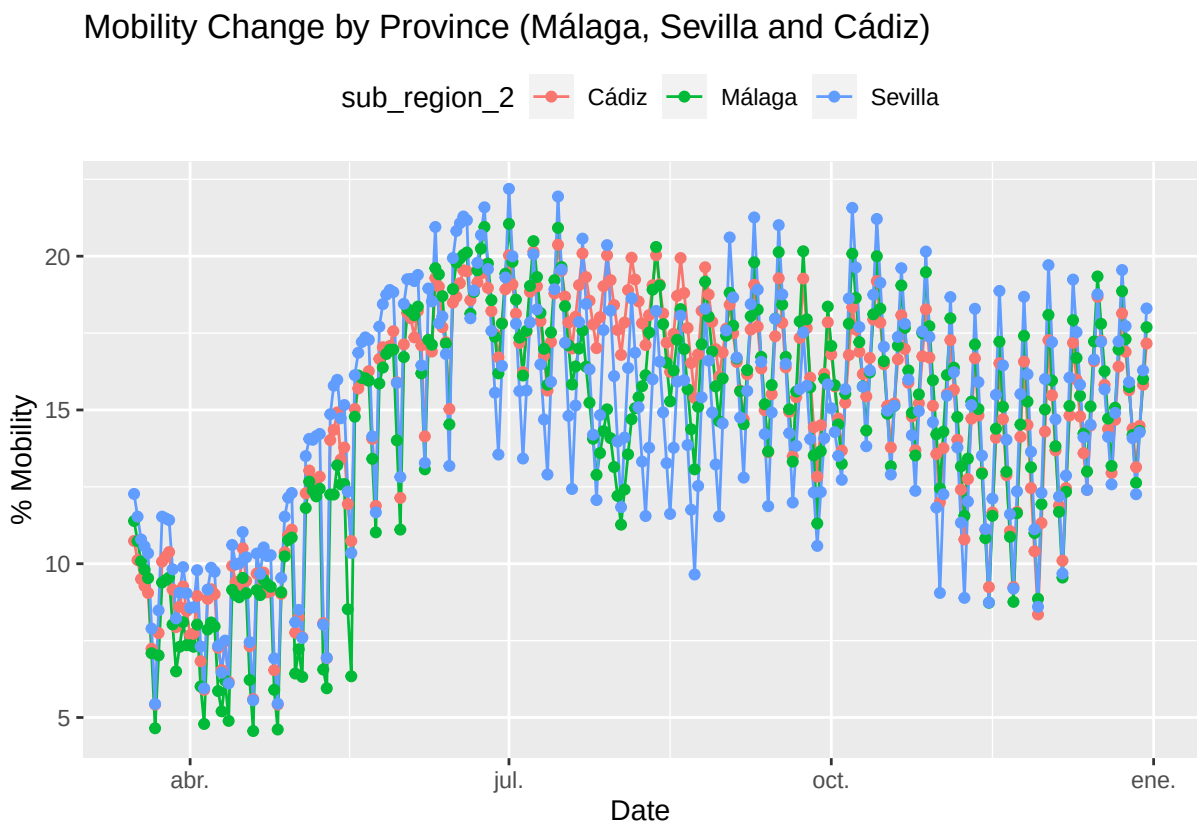
```
  geom_point(aes(color=sub_region_2))+
  theme(legend.position="top") +
  labs(title="Mobility Change by Province",
       x ="Date", y = "% Mobility")
```




```
# Mal, Sev and Cad - num_casos.x
Total %>%
  filter(sub_region_2 == "Málaga" | sub_region_2 == "Cádiz" |
         sub_region_2 == "Sevilla") %>%
  ggplot(aes(x=fecha, y=num_casos.x))+
    geom_line(aes(color=sub_region_2))+
    geom_point(aes(color=sub_region_2))+
    theme(legend.position="top") +
    labs(title="Cases by Province (Málaga, Córdoba and Cádiz)",
        x ="Date", y = "Nº of cases")
```




```
# Mal, Sev and Cad - Total (mobility)
Total %>%
  filter(sub_region_2 == "Málaga" | sub_region_2 == "Cádiz" |
         sub_region_2 == "Sevilla") %>%
  ggplot(aes(x=fecha, y=Total))+
    geom_line(aes(color=sub_region_2))+
    geom_point(aes(color=sub_region_2))+
    theme(legend.position="top") +
  labs(title="Mobility Change by Province (Málaga, Sevilla and Cádiz)",
       x ="Date", y = "% Mobility")
```



2.3.2 Time-series plots (Barcelona, Madrid, Málaga, Sevilla and Cádiz)

We have generated some plots from the **time-series** object generated. We use tsibble().

```
# Convert dataframe to ts object
#library(fpp3)
Total_ts <- Total[-3] %>%
  mutate(Dia_c = as_date(fecha)) %>%
  select(-fecha) %>%
  as_tsibble(key = c(sub_region_2),
             index = Dia_c)

# Filter for Bar, Mad, Mal, Cor and, Cad
Total_ts %>% filter(sub_region_2 == "Barcelona" | sub_region_2 == "Madrid" |
                    sub_region_2 == "Málaga" | sub_region_2 == "Sevilla" |
                    sub_region_2 == "Cádiz") -> Total_ts_b

##############################################
Total_ts
```

```
## # A tsibble: 15,080 x 19 [1D]
## # Key:       sub_region_2 [52]
##    sub_region_2 num_casos.x num_casos_prueb~ num_casos_prueb~ num_casos_prueb~
##    <chr>              <dbl>            <dbl>            <dbl>            <dbl>
##  1 Albacete             137              132                5                0
##  2 Albacete             128              123                5                0
##  3 Albacete             114              107                7                0
```

```
##  4 Albacete             149              133               16                0
##  5 Albacete             131              121               10                0
##  6 Albacete             129              120                9                0
##  7 Albacete             125              112               13                0
##  8 Albacete             112              103                9                0
##  9 Albacete             107               91               16                0
## 10 Albacete              78               64               14                0
## # ... with 15,070 more rows, and 14 more variables:
## #   num_casos_prueba_elisa <dbl>, num_casos_prueba_desconocida <dbl>,
## #   num_casos.y <dbl>, num_hosp <dbl>, num_uci <dbl>, num_def <dbl>,
## #   retail_and_recreation_percent_change_from_baseline <dbl>,
## #   grocery_and_pharmacy_percent_change_from_baseline <dbl>,
## #   parks_percent_change_from_baseline <dbl>,
## #   transit_stations_percent_change_from_baseline <dbl>,
## #   workplaces_percent_change_from_baseline <dbl>,
## #   residential_percent_change_from_baseline <dbl>, Total <dbl>, Dia_c <date>
```

```
Total_ts_b
```

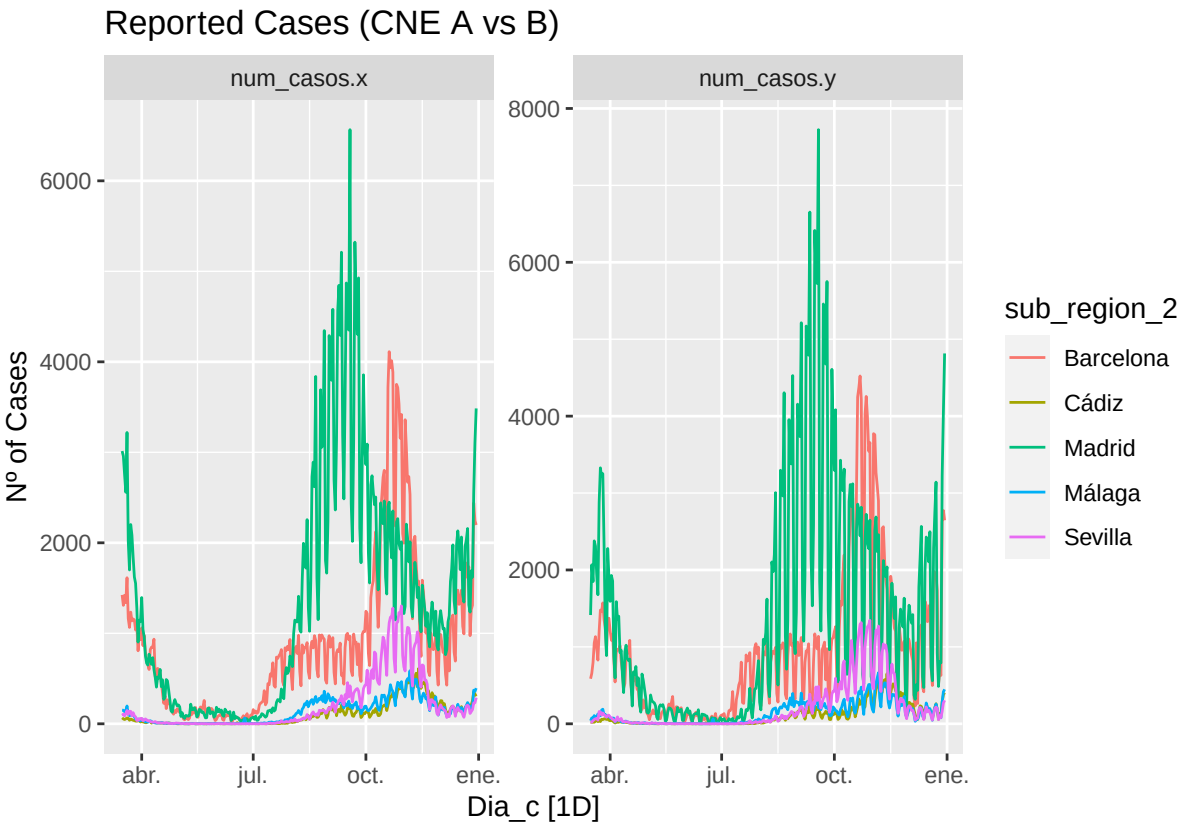
```
## # A tsibble: 1,450 x 19 [1D]
## # Key:       sub_region_2 [5]
##    sub_region_2 num_casos.x num_casos_prueb~ num_casos_prueb~ num_casos_prueb~
##    <chr>              <dbl>            <dbl>            <dbl>            <dbl>
##  1 Barcelona           1424             1351                0                0
##  2 Barcelona           1309             1273                0                0
##  3 Barcelona           1420             1379                0                0
##  4 Barcelona           1338             1289                0                0
##  5 Barcelona           1614             1557                0                0
##  6 Barcelona           1164             1136                0                0
##  7 Barcelona           1065             1032                0                0
##  8 Barcelona           1234             1187                0                0
##  9 Barcelona           1137             1095                0                0
## 10 Barcelona           1159             1131                0                0
## # ... with 1,440 more rows, and 14 more variables:
## #   num_casos_prueba_elisa <dbl>, num_casos_prueba_desconocida <dbl>,
## #   num_casos.y <dbl>, num_hosp <dbl>, num_uci <dbl>, num_def <dbl>,
## #   retail_and_recreation_percent_change_from_baseline <dbl>,
## #   grocery_and_pharmacy_percent_change_from_baseline <dbl>,
## #   parks_percent_change_from_baseline <dbl>,
## #   transit_stations_percent_change_from_baseline <dbl>,
## #   workplaces_percent_change_from_baseline <dbl>,
## #   residential_percent_change_from_baseline <dbl>, Total <dbl>, Dia_c <date>
```

```
Total_ts_b %>% distinct(sub_region_2)
```

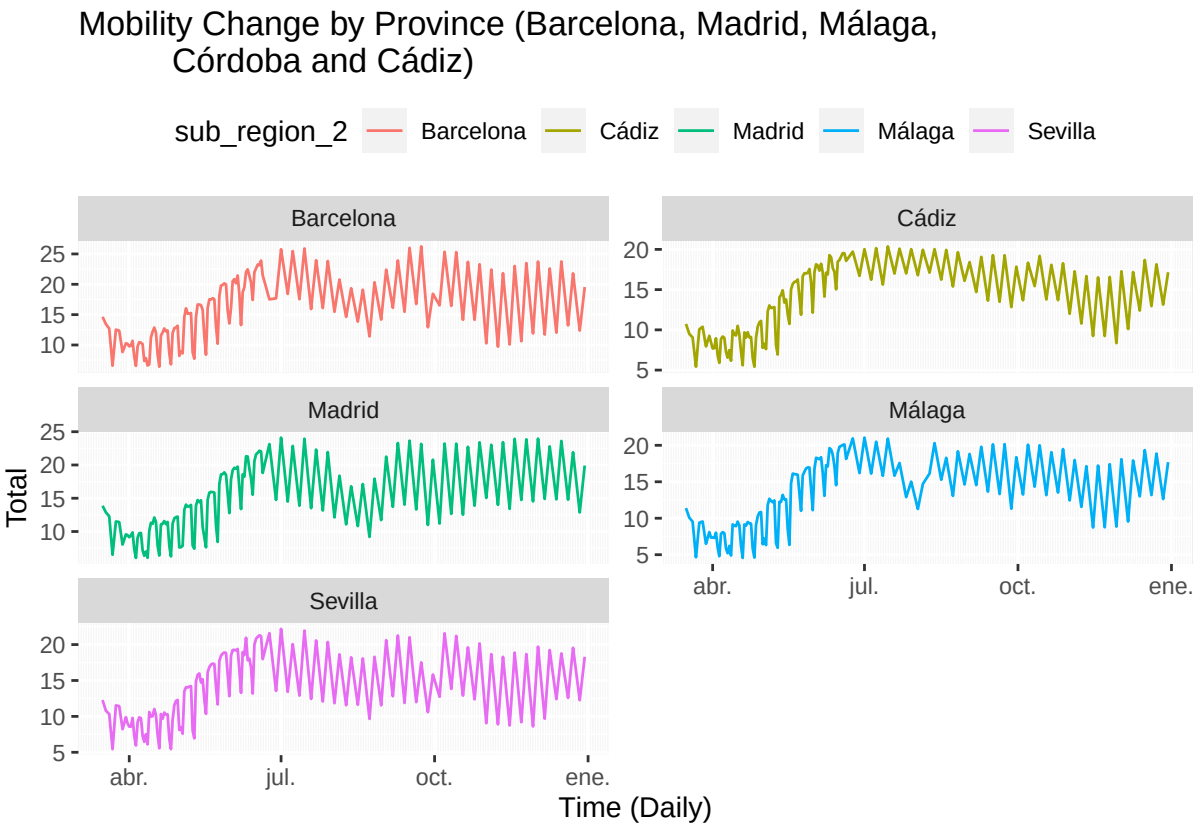
```
## # A tibble: 5 x 1
##   sub_region_2
##   <chr>
## 1 Barcelona
## 2 Cádiz
## 3 Madrid
## 4 Málaga
## 5 Sevilla
```

```
##########################################
```

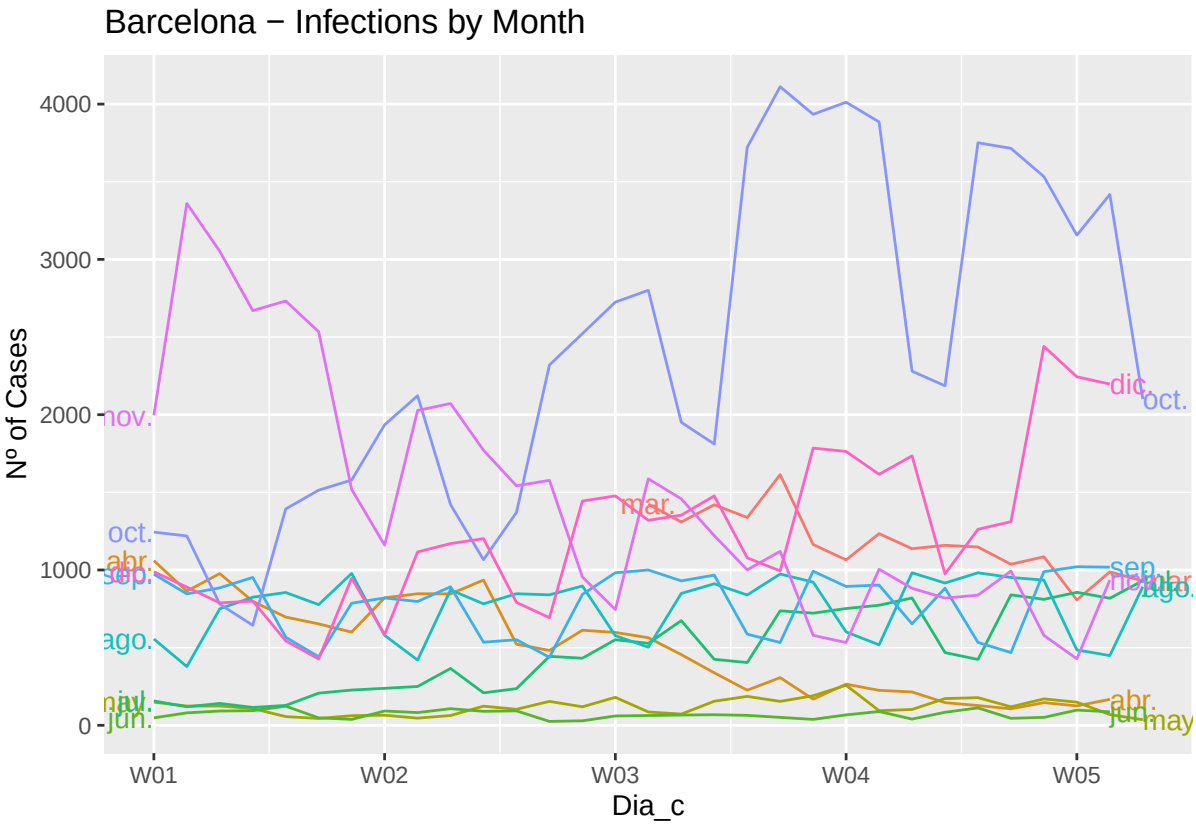
```
# Plots
# A num_casos.x,num_casos.y
autoplot(Total_ts_b, vars(num_casos.x,num_casos.y)) +
  labs(y = "Nº of Cases",
       title = "Reported Cases (CNE A vs B)")
```




```
# B Total (mobility)
autoplot(Total_ts_b, Total) +
  facet_wrap(~sub_region_2, scales = "free_y", ncol=2) +
  theme(legend.position = "top") +
  scale_x_date(date_minor_breaks = "1 day", name = "Time (Daily)") +
  ggtitle(label = "Mobility Change by Province (Barcelona, Madrid, Málaga,
          Córdoba and Cádiz)")
```

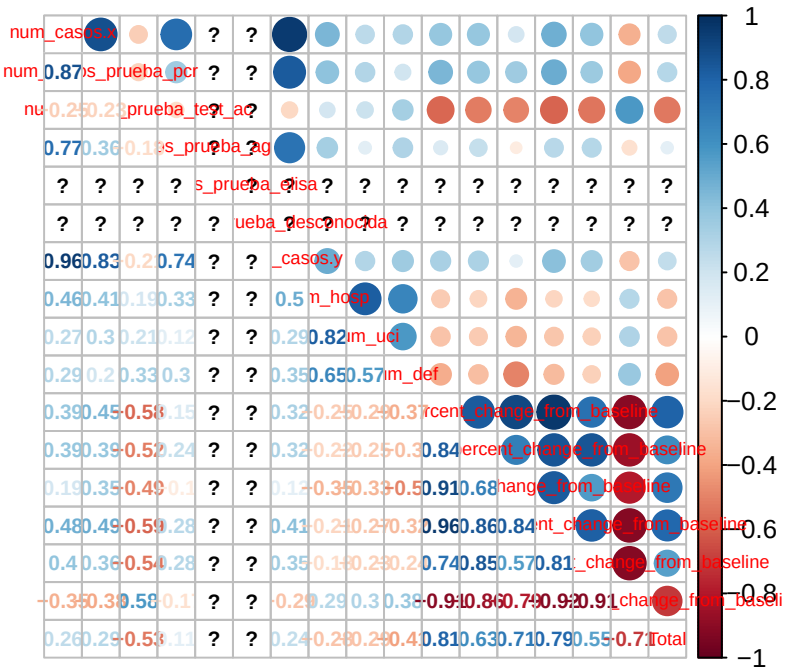



```
# C sub_region_2 == "Barcelona" by month
Total_ts %>% filter(sub_region_2 == "Barcelona") %>%
  gg_season(num_casos.x, period = "month", labels = "both") +
  theme(legend.position = "top") +
  labs(y="Nº of Cases", title="Barcelona - Infections by Month")
```

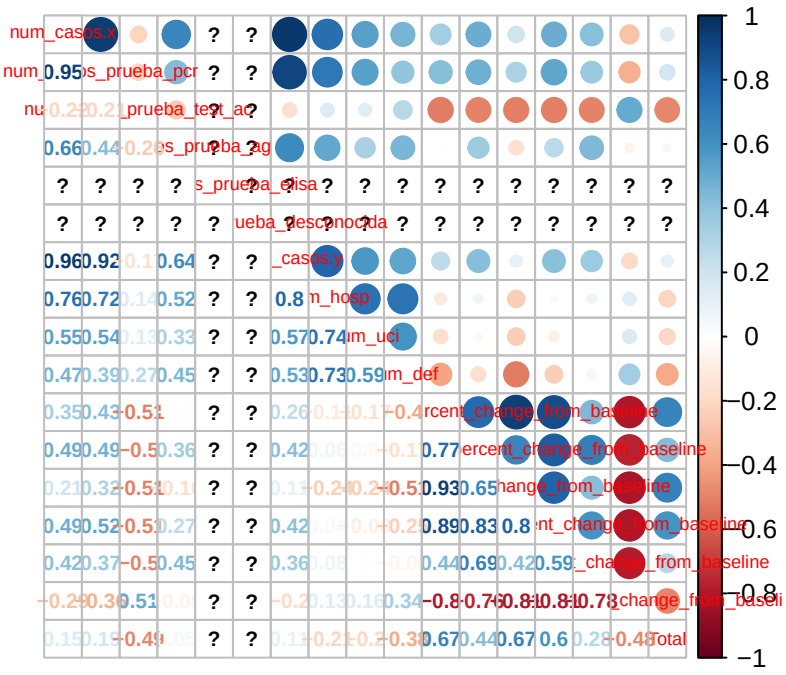


2.3.3 Correlation plots (from dataframe)

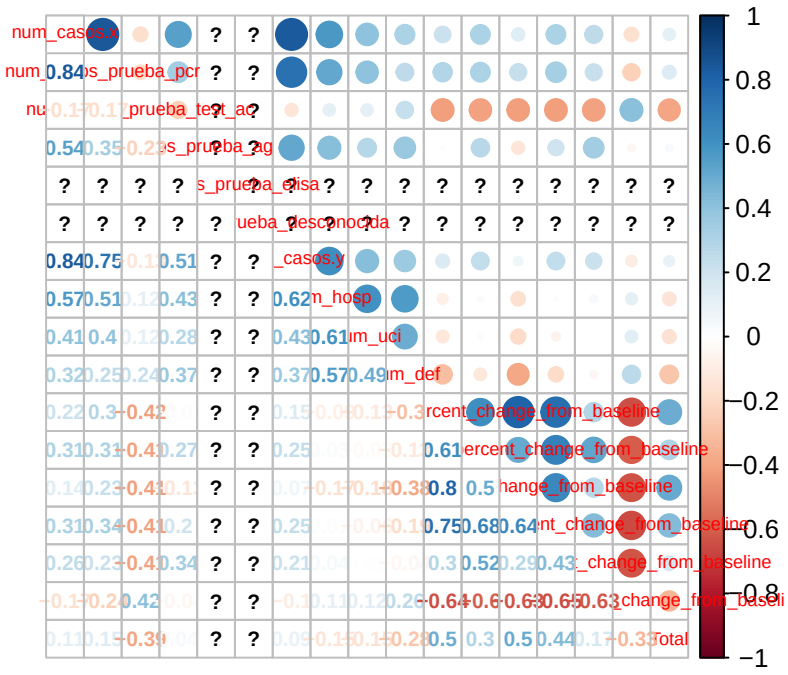
```
# Filter to "sub_region_2" == "Barcelona" or "Málaga"
# Character / date columns are eliminated
#library(corrplot)
#### Málaga
# pearson
Total.res<-Total %>%
  filter(sub_region_2 == "Málaga")
Total.res<-cor(Total.res[,c(-1,-2,-3)],method="pearson")
corrplot.mixed(Total.res,upper="circle",number.cex=.65,tl.cex=.6, title="Málaga -
               pearson ")
```



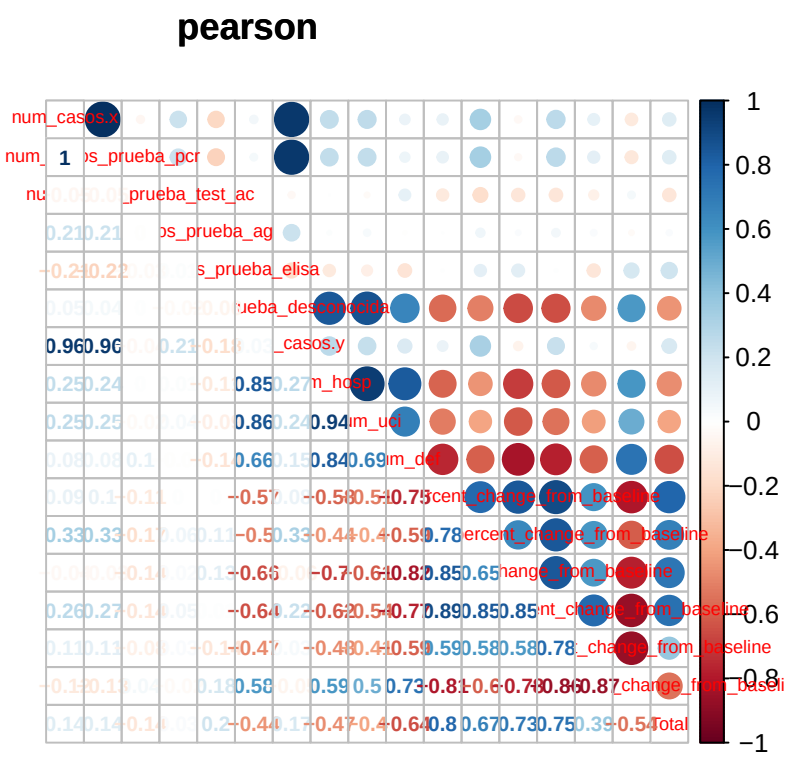

```
# spearman
Total.res<-Total %>%
  filter(sub_region_2 == "Málaga")
Total.res<-cor(Total.res[,c(-1,-2,-3)],method="spearman")
corrplot.mixed(Total.res,upper="circle",number.cex=.65,tl.cex=.6, title="Málaga -
               spearman ")
```



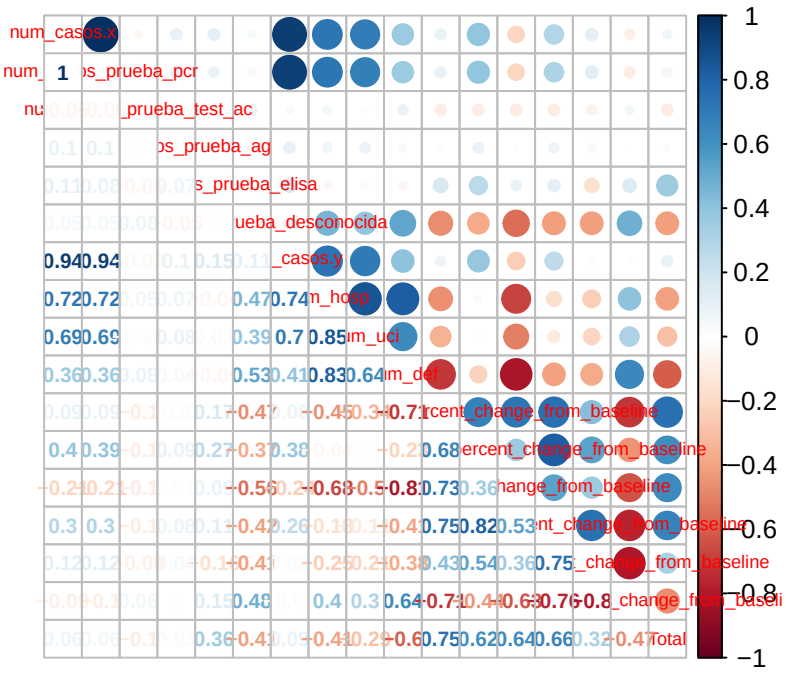

```
# kendall
Total.res<-Total %>%
  filter(sub_region_2 == "Málaga")
Total.res<-cor(Total.res[,c(-1,-2,-3)],method="kendall")
corrplot.mixed(Total.res,upper="circle",number.cex=.65,tl.cex=.6, title="Málaga -
               kendall ")
```



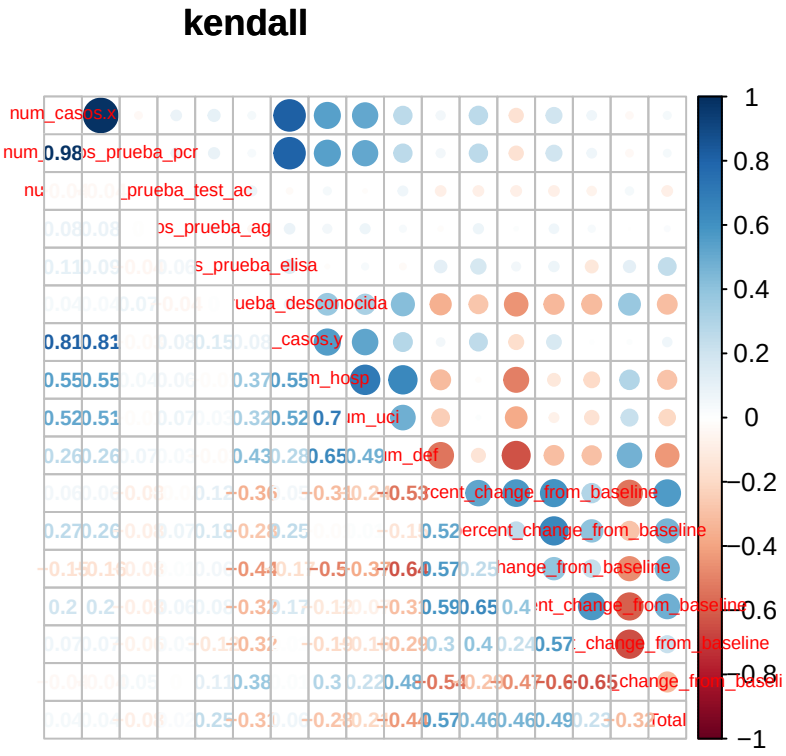

```
#### Barcelona
# pearson
Total.res<-Total %>%
  filter(sub_region_2 == "Barcelona")
Total.res<-cor(Total.res[,c(-1,-2,-3)],method="pearson")
corrplot.mixed(Total.res,upper="circle",number.cex=.65,tl.cex=.6, title="Barcelona -
               pearson ")
```




```
# spearman
Total.res<-Total %>%
  filter(sub_region_2 == "Barcelona")
Total.res<-cor(Total.res[,c(-1,-2,-3)],method="spearman")
corrplot.mixed(Total.res,upper="circle",number.cex=.65,tl.cex=.6, title="Barcelona -
               spearman ")
```




```
# kendall
Total.res<-Total %>%
  filter(sub_region_2 == "Barcelona")
Total.res<-cor(Total.res[,c(-1,-2,-3)],method="kendall")
corrplot.mixed(Total.res,upper="circle",number.cex=.65,tl.cex=.6, title="Barcelona -
               kendall ")
```



2.3.4 PCA (Barcelona)

```
pca <- prcomp(Total.res, scale = T)
summary(pca)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     3.0590 1.8914 1.12142 1.02409 0.83504 0.57825 0.46266
## Proportion of Variance 0.5504 0.2104 0.07398 0.06169 0.04102 0.01967 0.01259
## Cumulative Proportion  0.5504 0.7609 0.83485 0.89655 0.93756 0.95723 0.96982
##                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.41006 0.33620 0.28445 0.21986 0.17734 0.16911 0.15372
## Proportion of Variance 0.00989 0.00665 0.00476 0.00284 0.00185 0.00168 0.00139
## Cumulative Proportion  0.97971 0.98636 0.99112 0.99397 0.99582 0.99750 0.99889
##                           PC15    PC16      PC17
## Standard deviation     0.1369 0.01274 1.826e-17
## Proportion of Variance 0.0011 0.00001 0.000e+00
## Cumulative Proportion  1.0000 1.00000 1.000e+00
```

```
pca$rotation
```

```
##                                                  PC1          PC2
## num_casos.x                                -0.1354826901 -0.469400471
## num_casos_prueba_pcr                       -0.1356873829 -0.469237915
## num_casos_prueba_test_ac                   -0.0782958432  0.237502756
## num_casos_prueba_ag                         0.0005323686  0.008790618
```

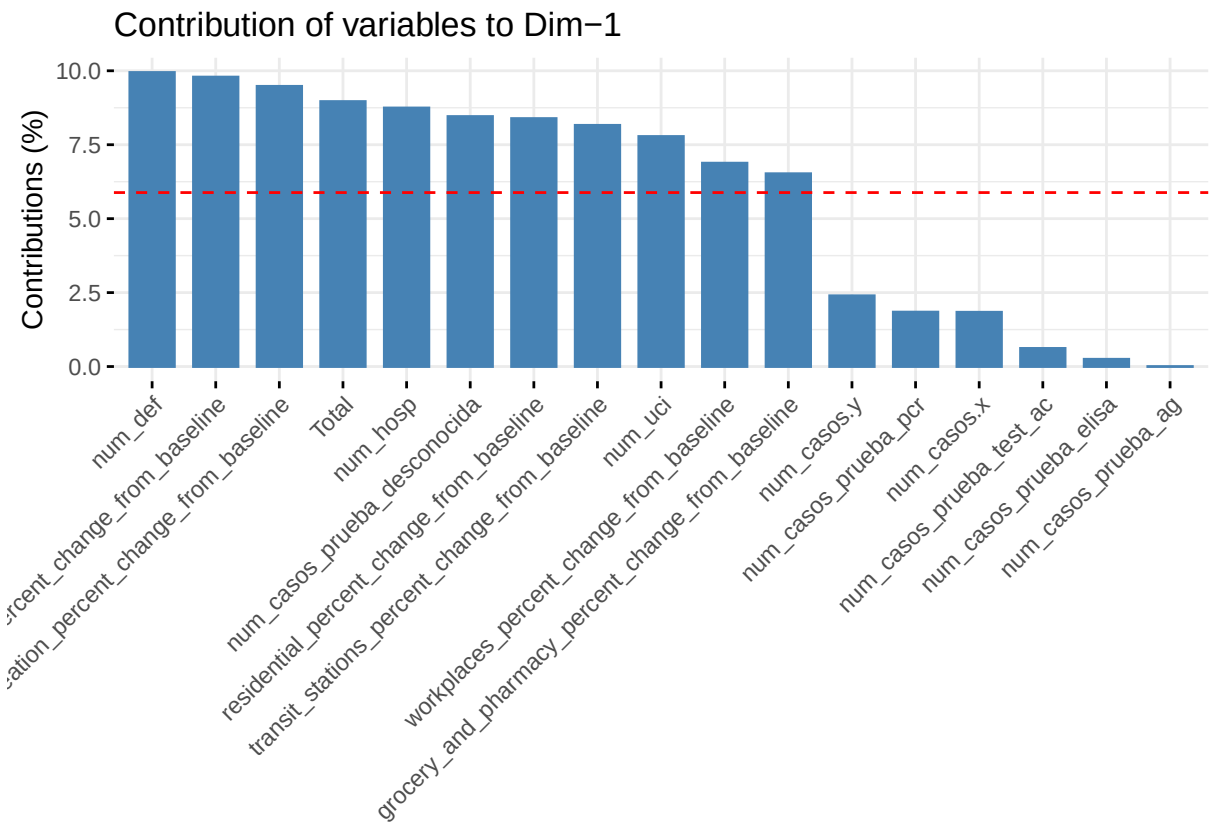
```
## num_casos_prueba_elisa                                0.0495912937  0.004385553
## num_casos_prueba_desconocida                         -0.2907430443  0.113686282
## num_casos.y                                          -0.1546876637 -0.450080493
## num_hosp                                             -0.2956764064 -0.199498016
## num_uci                                              -0.2788863801 -0.230508957
## num_def                                              -0.3153112111 -0.030904089
## retail_and_recreation_percent_change_from_baseline    0.3077977291 -0.116093786
## grocery_and_pharmacy_percent_change_from_baseline     0.2553316609 -0.261732823
## parks_percent_change_from_baseline                    0.3128404870  0.012776948
## transit_stations_percent_change_from_baseline         0.2855716515 -0.218931952
## workplaces_percent_change_from_baseline               0.2622358364 -0.170838668
## residential_percent_change_from_baseline             -0.2895648649  0.166332144
## Total                                                 0.2993175108 -0.081611685
##                                                              PC3          PC4
## num_casos.x                                          -0.002875288 -0.028445983
## num_casos_prueba_pcr                                 -0.012292151 -0.026795983
## num_casos_prueba_test_ac                             -0.392712472 -0.268237834
## num_casos_prueba_ag                                  -0.034460138  0.950253303
## num_casos_prueba_elisa                                0.811470962 -0.060011278
## num_casos_prueba_desconocida                          0.019674417 -0.078105637
## num_casos.y                                           0.041668217 -0.036946432
## num_hosp                                             -0.051419434 -0.010413059
## num_uci                                              -0.018126448  0.006323566
## num_def                                              -0.031133681  0.009747098
## retail_and_recreation_percent_change_from_baseline    0.033200491 -0.075398568
## grocery_and_pharmacy_percent_change_from_baseline     0.027494995 -0.030995413
## parks_percent_change_from_baseline                    0.029299658 -0.011018427
## transit_stations_percent_change_from_baseline        -0.099387689 -0.018021579
## workplaces_percent_change_from_baseline              -0.303553292  0.028981871
## residential_percent_change_from_baseline              0.211975965  0.003369575
## Total                                                 0.173800838 -0.065471155
##                                                              PC5          PC6
## num_casos.x                                          -0.15653475  0.16549669
## num_casos_prueba_pcr                                 -0.15256567  0.17028866
## num_casos_prueba_test_ac                             -0.81534671 -0.10365823
## num_casos_prueba_ag                                  -0.25969423 -0.01360833
## num_casos_prueba_elisa                               -0.34805371 -0.37185910
## num_casos_prueba_desconocida                          0.20580692 -0.11429059
## num_casos.y                                          -0.15827091  0.15297764
## num_hosp                                              0.03855580 -0.11688342
## num_uci                                               0.06172358 -0.06173530
## num_def                                               0.08791553 -0.26888200
## retail_and_recreation_percent_change_from_baseline -0.02731712  0.18700422
## grocery_and_pharmacy_percent_change_from_baseline   0.01445198 -0.34534708
## parks_percent_change_from_baseline                 -0.01705528  0.39173313
## transit_stations_percent_change_from_baseline       0.03520761 -0.26003250
## workplaces_percent_change_from_baseline             0.10943614 -0.48832683
## residential_percent_change_from_baseline            0.03104587  0.13040337
## Total                                              -0.05468582  0.19337305
##                                                              PC7          PC8
## num_casos.x                                           0.11538168  0.07515692
## num_casos_prueba_pcr                                  0.11048811  0.07713501
## num_casos_prueba_test_ac                             -0.08510797 -0.12117026
## num_casos_prueba_ag                                  -0.03926649 -0.14832473
```

```
## num_casos_prueba_elisa                                0.24859208  0.01374512
## num_casos_prueba_desconocida                          0.19845835 -0.75630827
## num_casos.y                                           0.04117476 -0.02860659
## num_hosp                                             -0.11367668 -0.07611704
## num_uci                                              -0.02723992 -0.20583241
## num_def                                              -0.20890547  0.15436921
## retail_and_recreation_percent_change_from_baseline -0.10130686 -0.31700087
## grocery_and_pharmacy_percent_change_from_baseline  -0.63305922 -0.04093499
## parks_percent_change_from_baseline                    0.15076652  0.04293438
## transit_stations_percent_change_from_baseline      -0.06852096 -0.19335324
## workplaces_percent_change_from_baseline            0.37588850  0.19906444
## residential_percent_change_from_baseline           -0.39494991  0.23079899
## Total                                                -0.26415699 -0.26912855
##                                                             PC9         PC10
## num_casos.x                                          -0.204798939 -0.059620970
## num_casos_prueba_pcr                                 -0.206074966 -0.057649643
## num_casos_prueba_test_ac                              0.031408747  0.010328003
## num_casos_prueba_ag                                  -0.027397924 -0.004813823
## num_casos_prueba_elisa                                0.085159377 -0.047189805
## num_casos_prueba_desconocida                         -0.349092565 -0.125363275
## num_casos.y                                          -0.130984317  0.031574368
## num_hosp                                              0.283065654  0.067094187
## num_uci                                               0.717067401  0.169083317
## num_def                                              -0.009119052  0.013475201
## retail_and_recreation_percent_change_from_baseline  0.099531350 -0.317145228
## grocery_and_pharmacy_percent_change_from_baseline  -0.110976209 -0.296856620
## parks_percent_change_from_baseline                    0.283874181 -0.240873233
## transit_stations_percent_change_from_baseline         0.094114251 -0.074265913
## workplaces_percent_change_from_baseline              -0.125028631  0.240735868
## residential_percent_change_from_baseline             -0.178700468  0.013386663
## Total                                                -0.118124812  0.793981499
##                                                             PC11          PC12
## num_casos.x                                          -0.049617465 -0.0988357734
## num_casos_prueba_pcr                                 -0.044726847 -0.1058640000
## num_casos_prueba_test_ac                             -0.042562355 -0.0389190655
## num_casos_prueba_ag                                   0.003758095 -0.0008550931
## num_casos_prueba_elisa                                0.023430627 -0.0013867741
## num_casos_prueba_desconocida                         -0.154872738 -0.1615965585
## num_casos.y                                           0.033429358  0.1359821524
## num_hosp                                              0.302911164  0.0671255962
## num_uci                                              -0.400459562  0.0477114671
## num_def                                               0.553043315 -0.1950704924
## retail_and_recreation_percent_change_from_baseline  0.286708424  0.6275693247
## grocery_and_pharmacy_percent_change_from_baseline  -0.348092535 -0.0594956464
## parks_percent_change_from_baseline                   -0.082227368 -0.4958947989
## transit_stations_percent_change_from_baseline         0.264287632 -0.4433548371
## workplaces_percent_change_from_baseline              -0.216784094  0.2008296543
## residential_percent_change_from_baseline             -0.278596900  0.0313367951
## Total                                                 0.074087201 -0.0830357636
##                                                             PC13        PC14
## num_casos.x                                           0.3503737964 -0.03447444
## num_casos_prueba_pcr                                  0.3691954013 -0.03383423
## num_casos_prueba_test_ac                             -0.0023721487 -0.01431289
## num_casos_prueba_ag                                   0.0009495141 -0.01895981
```

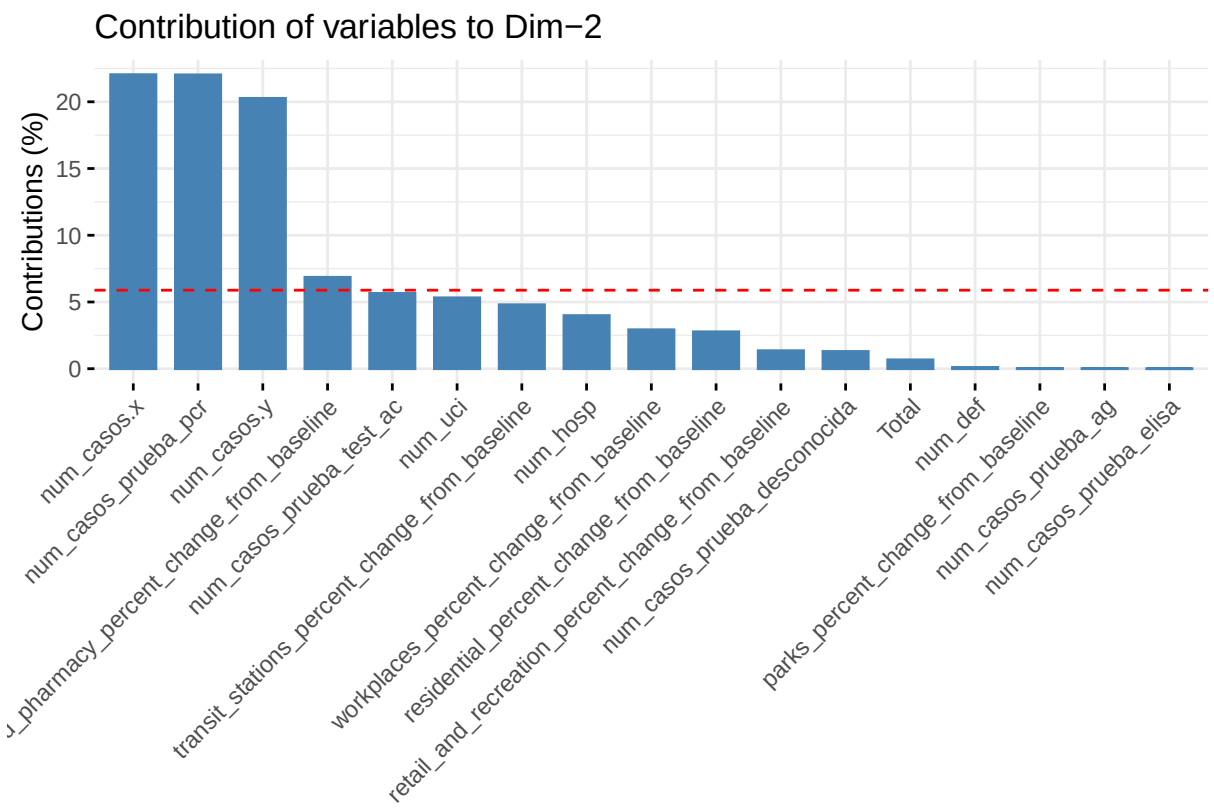
```
## num_casos_prueba_elisa                                0.0189278939 -0.03470038
## num_casos_prueba_desconocida                         -0.0341231415 -0.11380732
## num_casos.y                                          -0.8087585769  0.15136854
## num_hosp                                             -0.0093409355 -0.32920173
## num_uci                                               0.0905105752  0.12476243
## num_def                                              -0.0179131471 -0.20425404
## retail_and_recreation_percent_change_from_baseline    0.1591640063  0.05248551
## grocery_and_pharmacy_percent_change_from_baseline    -0.0689717137 -0.30356466
## parks_percent_change_from_baseline                   -0.1967768342 -0.33947866
## transit_stations_percent_change_from_baseline         0.0373896719  0.66046288
## workplaces_percent_change_from_baseline              -0.0493969622 -0.09490414
## residential_percent_change_from_baseline              0.0240600499  0.34670086
## Total                                                 0.0590314137 -0.13712447
##                                                             PC15          PC16
## num_casos.x                                           0.021886709 -0.710827710
## num_casos_prueba_pcr                                  0.020989875  0.696614910
## num_casos_prueba_test_ac                              0.002128322 -0.009045443
## num_casos_prueba_ag                                  -0.001031653 -0.006169593
## num_casos_prueba_elisa                               -0.019875868  0.004055133
## num_casos_prueba_desconocida                         -0.000227830 -0.014131490
## num_casos.y                                           0.065935929  0.012850260
## num_hosp                                             -0.738022542 -0.005532761
## num_uci                                               0.291185664 -0.005062923
## num_def                                               0.517898894 -0.032682612
## retail_and_recreation_percent_change_from_baseline    0.103782290 -0.031882076
## grocery_and_pharmacy_percent_change_from_baseline     0.033876519  0.014999252
## parks_percent_change_from_baseline                   -0.032945065 -0.039812604
## transit_stations_percent_change_from_baseline        -0.169052448 -0.008281975
## workplaces_percent_change_from_baseline              -0.059237981 -0.040905347
## residential_percent_change_from_baseline             -0.223342981 -0.056238440
## Total                                                 0.040900018 -0.009746369
##                                                             PC17
## num_casos.x                                          -0.04694046
## num_casos_prueba_pcr                                  0.08912178
## num_casos_prueba_test_ac                              0.09544863
## num_casos_prueba_ag                                   0.05828430
## num_casos_prueba_elisa                                0.04498527
## num_casos_prueba_desconocida                          0.16508413
## num_casos.y                                          -0.01242631
## num_hosp                                              0.04584112
## num_uci                                               0.06322805
## num_def                                               0.30912666
## retail_and_recreation_percent_change_from_baseline    0.33484091
## grocery_and_pharmacy_percent_change_from_baseline    -0.13871900
## parks_percent_change_from_baseline                    0.41803506
## transit_stations_percent_change_from_baseline         0.04834211
## workplaces_percent_change_from_baseline               0.44739326
## residential_percent_change_from_baseline              0.57600490
## Total                                                 0.07657033
```

```
if(!require(FactoMineR)){
    install.packages('FactoMineR', repos='http://cran.us.r-project.org')
    library(FactoMineR)}
if(!require(factoextra)){
```

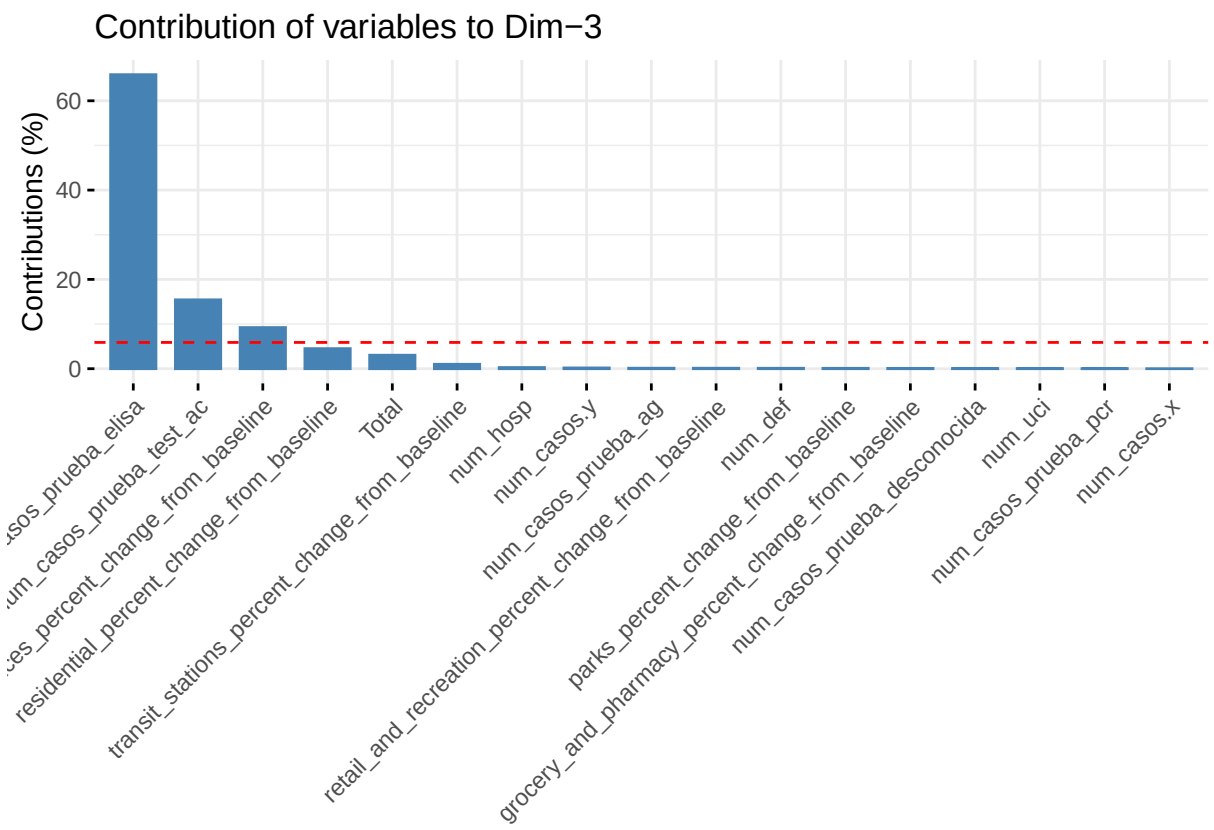
```
    install.packages('factoextra', repos='http://cran.us.r-project.org')
    library(factoextra)}

# Var contribution for PC1-PC5
fviz_contrib(pca, choice = "var", axes = 1)
```



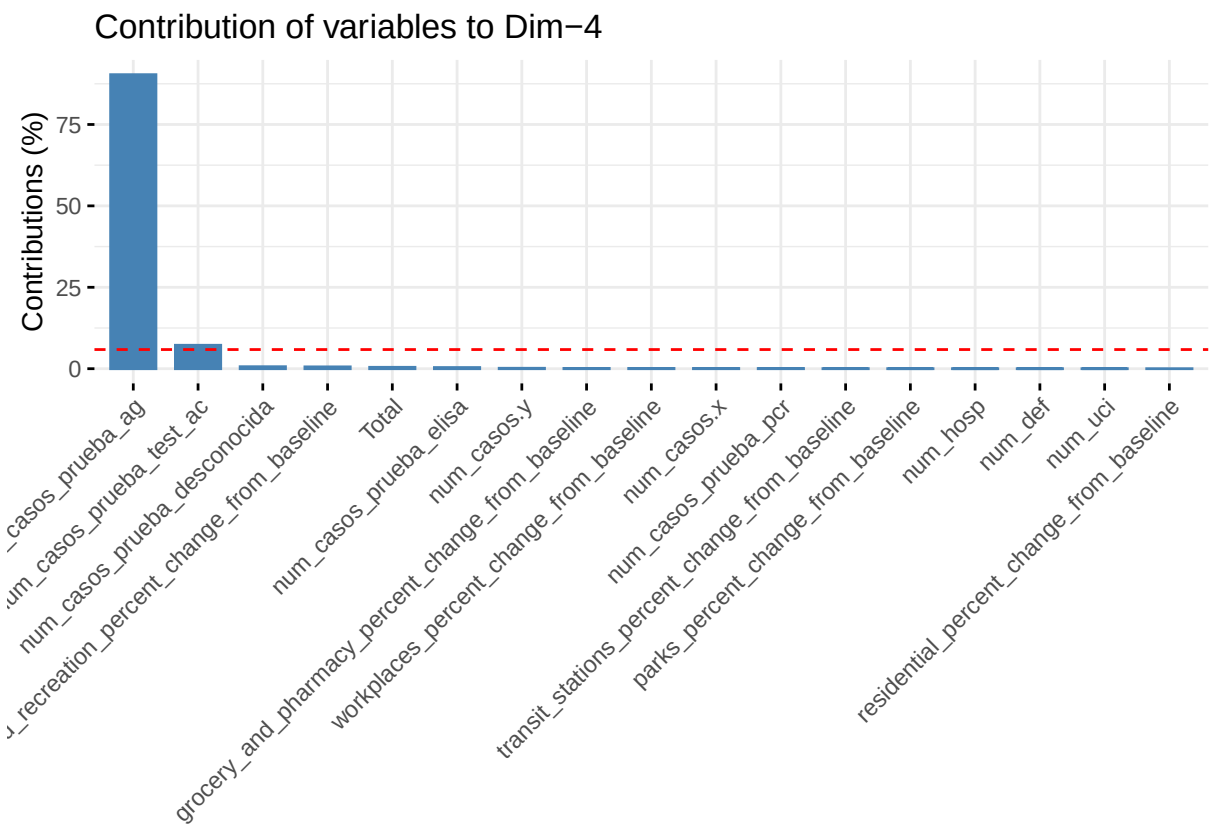

```
fviz_contrib(pca, choice = "var", axes = 2)
```



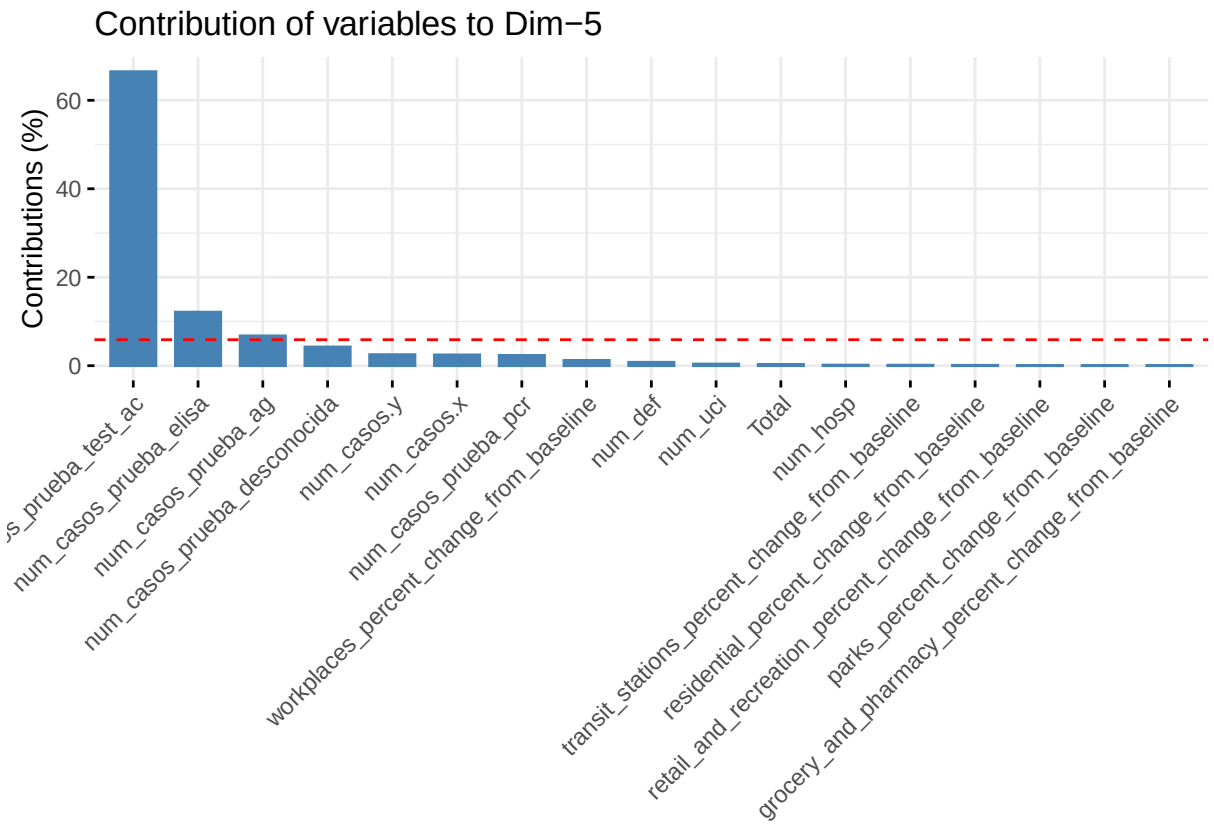

```
fviz_contrib(pca, choice = "var", axes = 3)
```




```
fviz_contrib(pca, choice = "var", axes = 4)
```

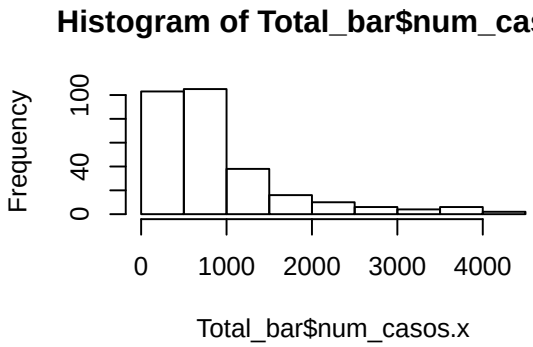


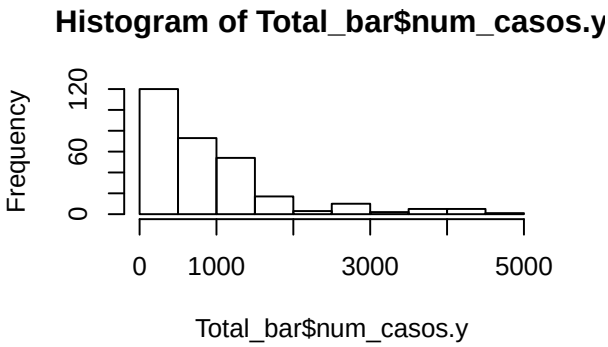

```
fviz_contrib(pca, choice = "var", axes = 5)
```

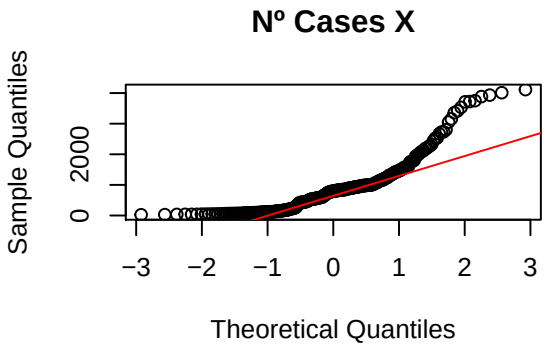


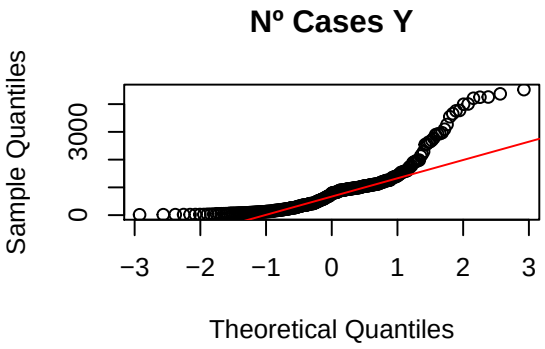
2.3.5 Review normality (Barcelona)

```
# Check for Barcelona
# Raw
Total %>%
  filter(sub_region_2 == "Barcelona") -> Total_bar

par(mfrow=c(2,2))

hist(Total_bar$num_casos.x)
hist(Total_bar$num_casos.y)

qqnorm(Total_bar$num_casos.x, main="Nº Cases X")
qqline(Total_bar$num_casos.x,col=2)

qqnorm(Total_bar$num_casos.y, main="Nº Cases Y")
qqline(Total_bar$num_casos.y,col=2)
```

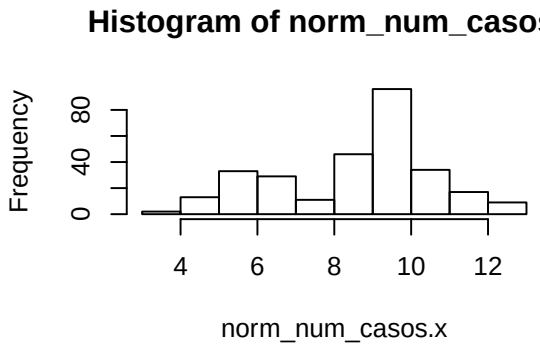


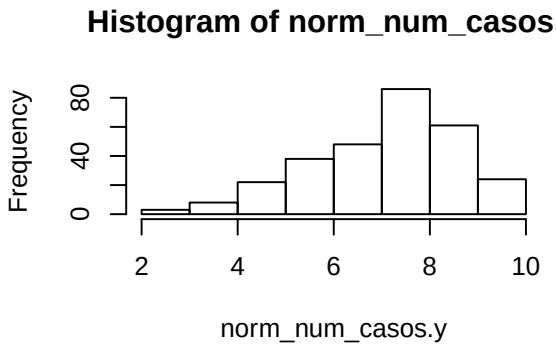


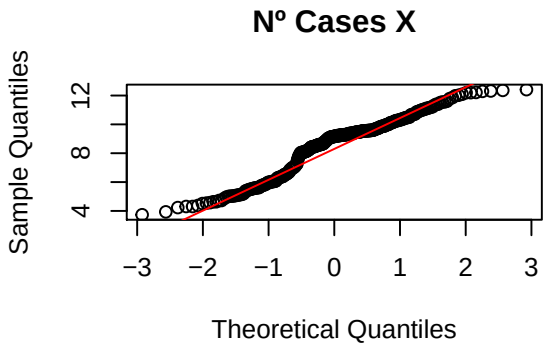


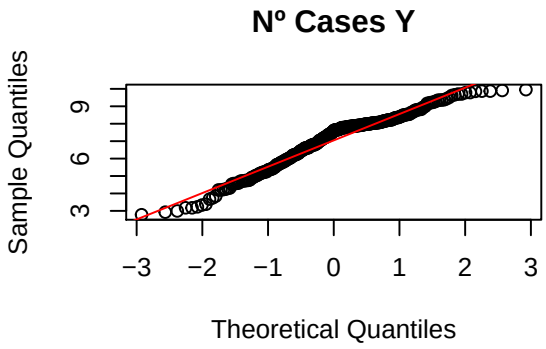



```r
# Normalize
library(DescTools)
norm_num_casos.x <- BoxCox(Total_bar$num_casos.x, lambda =
                           BoxCoxLambda(Total_bar$num_casos.x))
norm_num_casos.y <- BoxCox(Total_bar$num_casos.y, lambda =
                           BoxCoxLambda(Total_bar$num_casos.y))

hist(norm_num_casos.x)
hist(norm_num_casos.y)

qqnorm(norm_num_casos.x, main="Nº Cases X")
qqline(norm_num_casos.x,col=2)

qqnorm(norm_num_casos.y, main="Nº Cases Y")
qqline(norm_num_casos.y,col=2)
```










```
# Columns removal according PCA results and SME knowledge
Total <- Total[c(-3,-6,-7,-8,-10)]
Total_bar <- Total_bar[c(-3,-6,-7,-8,-10)]
Total_ts <- Total_ts[c(-4,-5,-6,-8)]
Total_ts_b <- Total_ts_b[c(-4,-5,-6,-8)]
table(Total_ts$sub_region_2)
```

```
##
##              Albacete      Alicante/Alacant               Almería
##                   290                   290                   290
##           Araba/Álava              Asturias                 Ávila
##                   290                   290                   290
##               Badajoz        Balears, Illes              Barcelona
##                   290                   290                   290
##               Bizkaia                Burgos               Cáceres
##                   290                   290                   290
##                 Cádiz             Cantabria    Castellón/Castelló
##                   290                   290                   290
##                 Ceuta           Ciudad Real               Córdoba
##                   290                   290                   290
##            Coruña, A               Cuenca              Gipuzkoa
##                   290                   290                   290
##               Girona               Granada           Guadalajara
##                   290                   290                   290
##               Huelva               Huesca                 Jaén
##                   290                   290                   290
```

```
##                 León               Lleida                 Lugo
##                  290                  290                  290
##               Madrid               Málaga              Melilla
##                  290                  290                  290
##               Murcia              Navarra              Ourense
##                  290                  290                  290
##             Palencia         Palmas, Las           Pontevedra
##                  290                  290                  290
##           Rioja, La           Salamanca Santa Cruz de Tenerife
##                  290                  290                  290
##              Segovia              Sevilla               Soria
##                  290                  290                  290
##            Tarragona               Teruel              Toledo
##                  290                  290                  290
##     Valencia/València           Valladolid               Zamora
##                  290                  290                  290
##             Zaragoza
##                  290
```

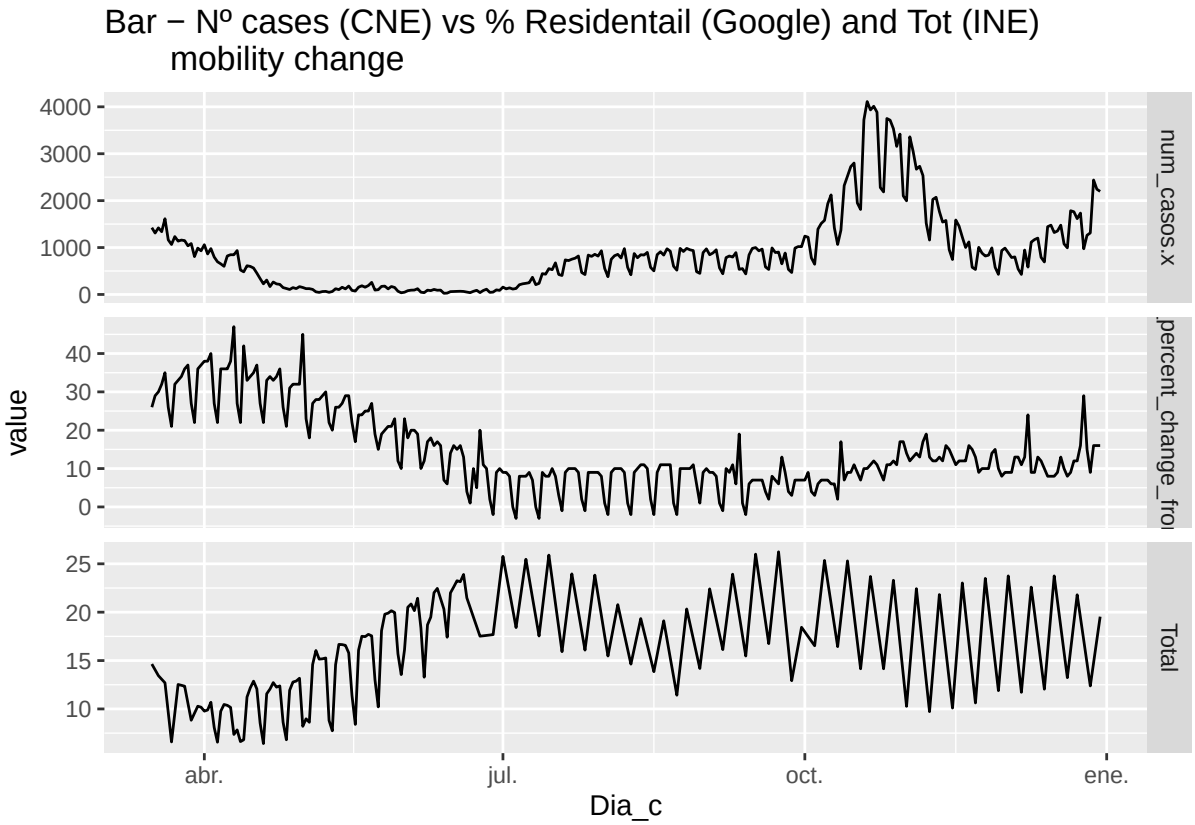
```
#str(Total_ts)
summary(Total_ts)
```

```
##  sub_region_2        num_casos.x    num_casos_prueba_pcr
##  Length:15080       Min.   :   0   Min.   :   0.0
##  Class :character   1st Qu.:   5   1st Qu.:   5.0
##  Mode  :character   Median :  39   Median :  35.0
##                     Mean   : 126   Mean   : 110.2
##                     3rd Qu.: 120   3rd Qu.: 105.0
##                     Max.   :6565   Max.   :6546.0
##  num_casos_prueba_desconocida    num_hosp          num_uci
##  Min.   : 0.0000              Min.   :   0.00   Min.   :  0.000
##  1st Qu.: 0.0000              1st Qu.:   1.00   1st Qu.:  0.000
##  Median : 0.0000              Median :   4.00   Median :  0.000
##  Mean   : 0.1317              Mean   :  14.86   Mean   :  1.281
##  3rd Qu.: 0.0000              3rd Qu.:  12.00   3rd Qu.:  1.000
##  Max.   :65.0000              Max.   :1930.00   Max.   :135.000
##     num_def        retail_and_recreation_percent_change_from_baseline
##  Min.   :  0.000   Min.   :-97.00
##  1st Qu.:  0.000   1st Qu.:-57.00
##  Median :  1.000   Median :-30.00
##  Mean   :  3.437   Mean   :-37.29
##  3rd Qu.:  3.000   3rd Qu.:-17.00
##  Max.   :334.000   Max.   : 71.00
##  grocery_and_pharmacy_percent_change_from_baseline
##  Min.   :-96.00
##  1st Qu.:-24.00
##  Median : -6.00
##  Mean   :-11.75
##  3rd Qu.:  4.00
##  Max.   :194.00
##  parks_percent_change_from_baseline
##  Min.   :-94.000
##  1st Qu.:-30.000
##  Median : -2.000
##  Mean   :  5.809
```

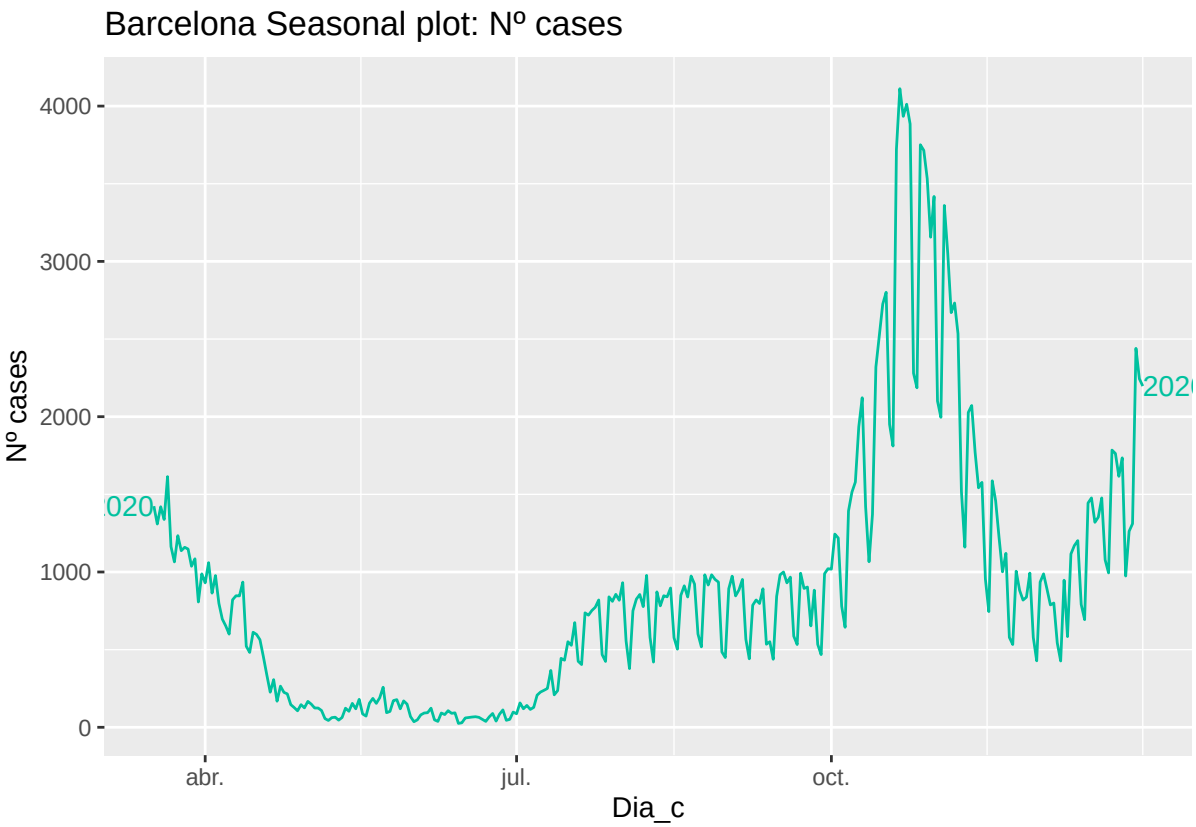
```
##  3rd Qu.: 30.000
##  Max.   :543.000
##  transit_stations_percent_change_from_baseline
##  Min.   :-100.00
##  1st Qu.: -53.00
##  Median : -31.00
##  Mean   : -35.19
##  3rd Qu.: -17.00
##  Max.   :  74.00
##  workplaces_percent_change_from_baseline
##  Min.   :-92.00
##  1st Qu.:-43.00
##  Median :-26.00
##  Mean   :-29.08
##  3rd Qu.:-13.00
##  Max.   : 55.00
##  residential_percent_change_from_baseline     Total           Dia_c
##  Min.   :-10.00                           Min.   : 1.95   Min.   :2020-03-16
##  1st Qu.:  4.00                           1st Qu.:11.36   1st Qu.:2020-05-27
##  Median :  7.00                           Median :14.39   Median :2020-08-07
##  Mean   : 10.14                           Mean   :14.20   Mean   :2020-08-07
##  3rd Qu.: 14.00                           3rd Qu.:17.11   3rd Qu.:2020-10-19
##  Max.   : 48.00                           Max.   :29.00   Max.   :2020-12-30
```

2.3.6 Final plots (Barcelona and others)

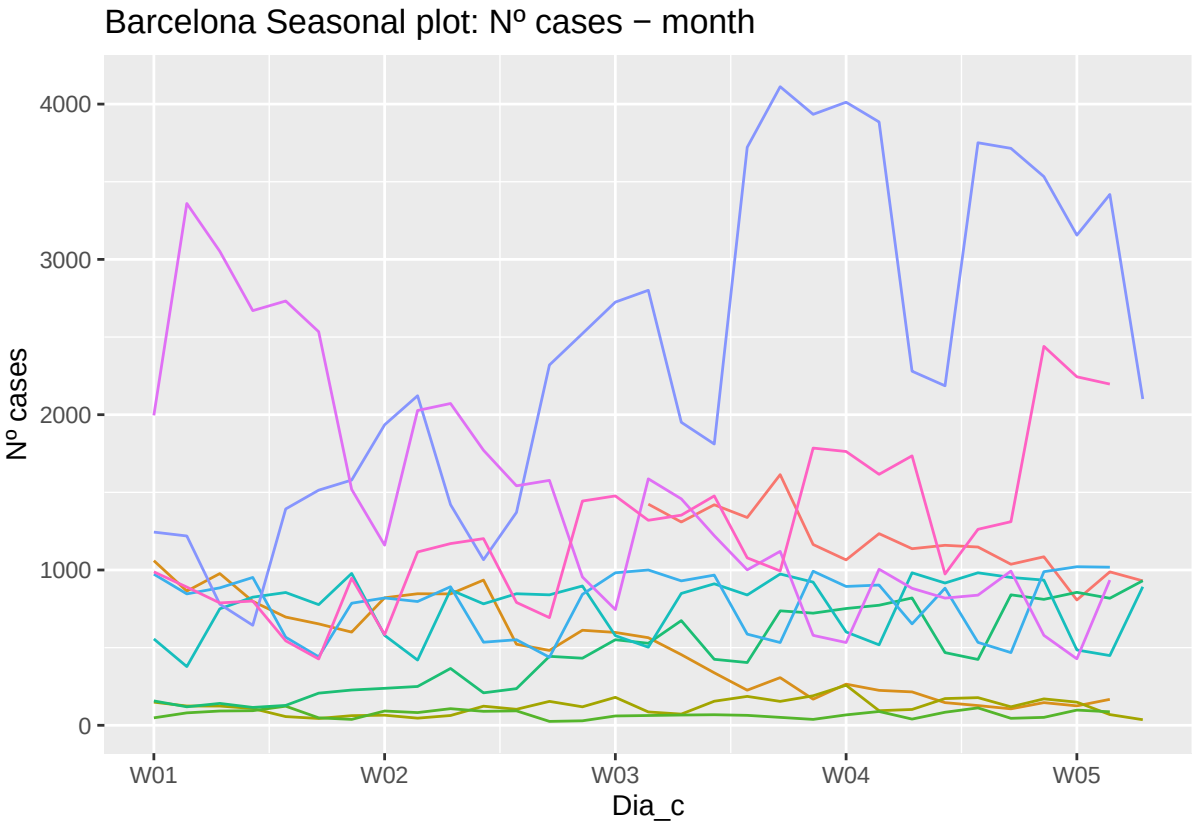
```
# Total_ts plots for the selected variables
# % of mobility reported by INE and Google (EM3 study)
Total_ts_b %>%
  filter(sub_region_2 == "Barcelona") %>%
  pivot_longer(c(2,13,14)) %>%
  ggplot(aes(x = Dia_c, y = value)) +
  geom_line() +
  facet_grid(vars(name), scales = "free_y")+
  labs(title = "Bar - Nº cases (CNE) vs % Residentail (Google) and Tot (INE)
       mobility change")
```



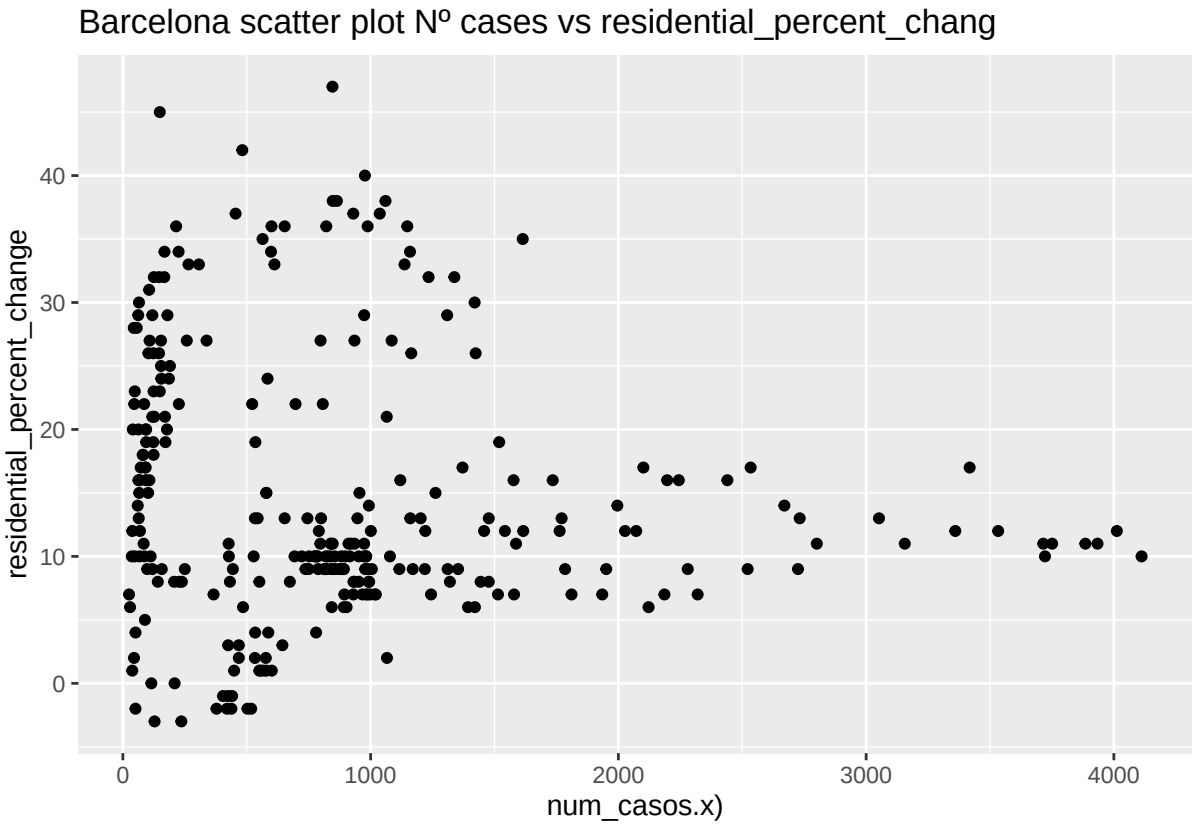

```
# Barcelona Seasonal plot: Nº cases
Total_ts_b %>%
  filter(sub_region_2 == "Barcelona") %>%
  gg_season(num_casos.x, labels = "both") +
  labs(y = "Nº cases",
       title = "Barcelona Seasonal plot: Nº cases")
```



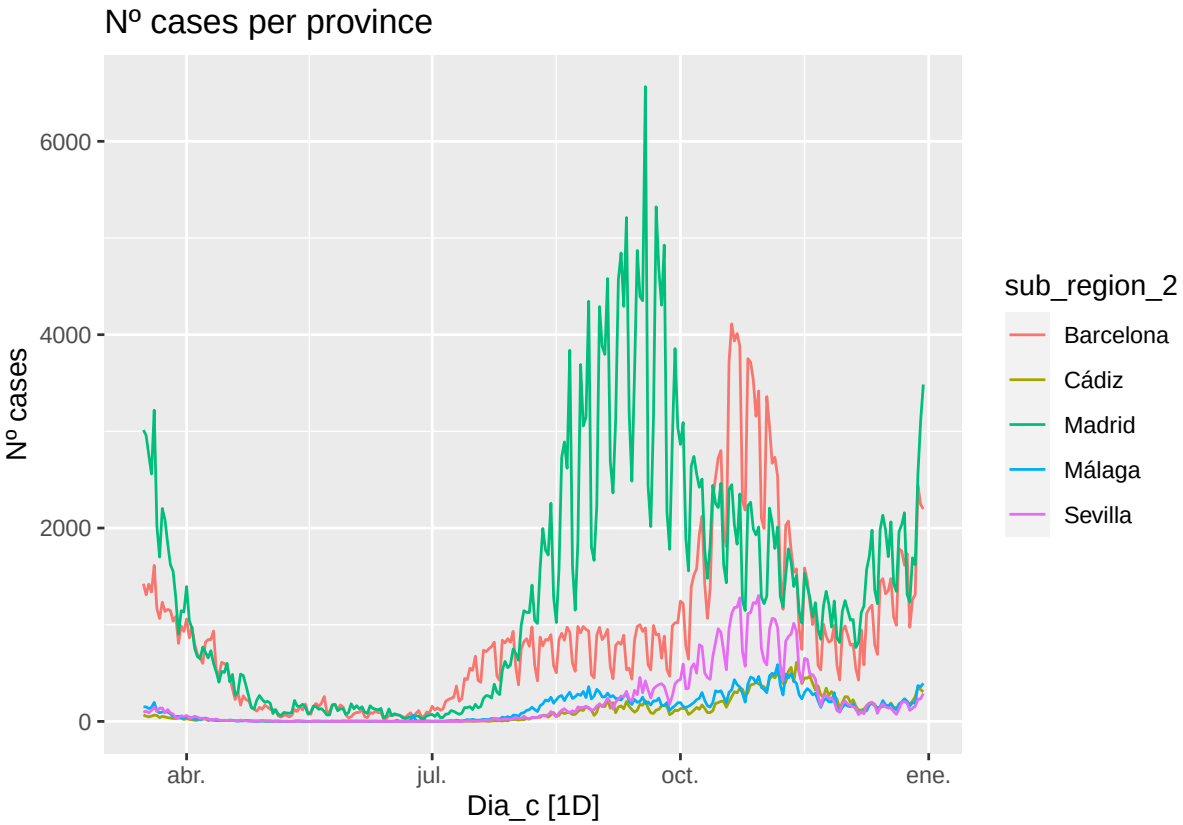

```
# Barcelona Seasonal plot: Nº cases by month
Total_ts_b %>%
  filter(sub_region_2 == "Barcelona") %>%
  gg_season(num_casos.x, period = "month") +
  theme(legend.position = "none") +
  labs(y="Nº cases", title="Barcelona Seasonal plot: Nº cases - month")
```



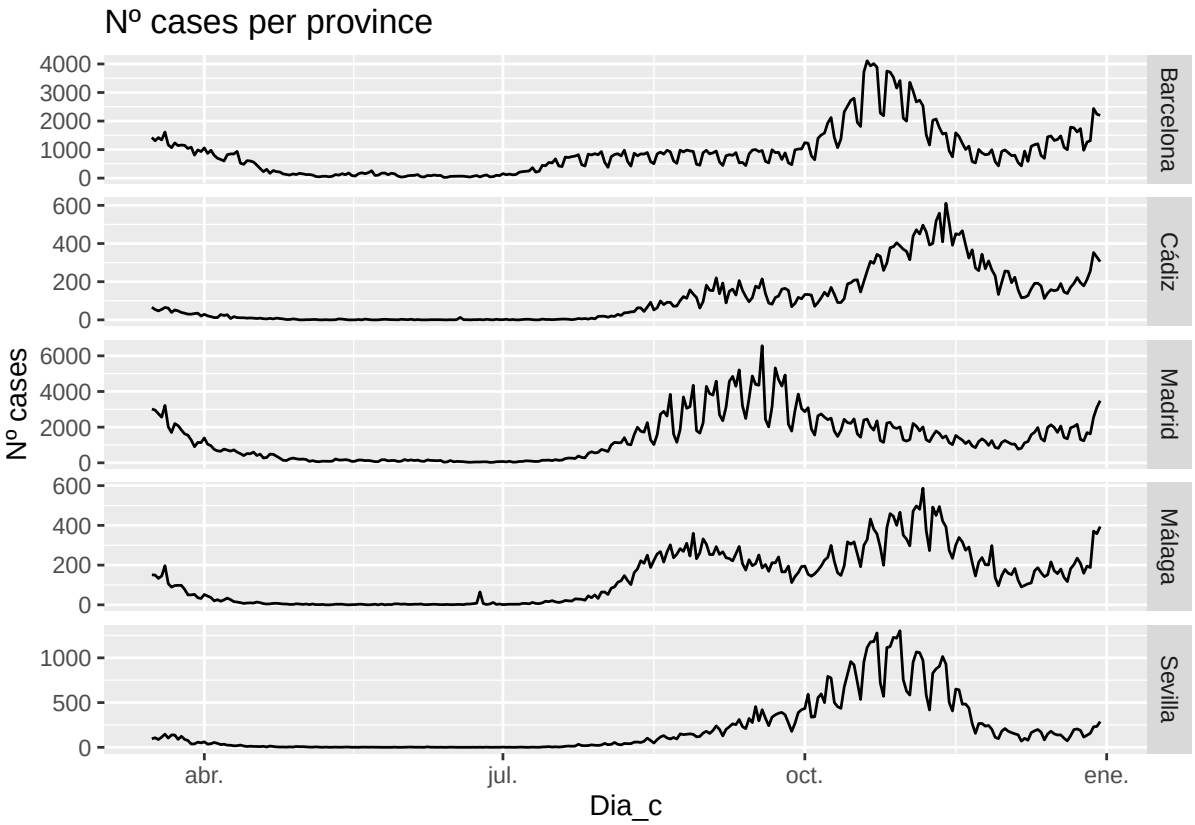

```
# Barcelona scatter plot Nº cases vs residential_percent_change
Total_ts_b %>%
  filter(sub_region_2 == "Barcelona") %>%
  ggplot(aes(x = num_casos.x, y = residential_percent_change_from_baseline )) +
  geom_point() +
  labs(x = "num_casos.x)",
       y = "residential_percent_change",
       title="Barcelona scatter plot Nº cases vs residential_percent_chang")
```



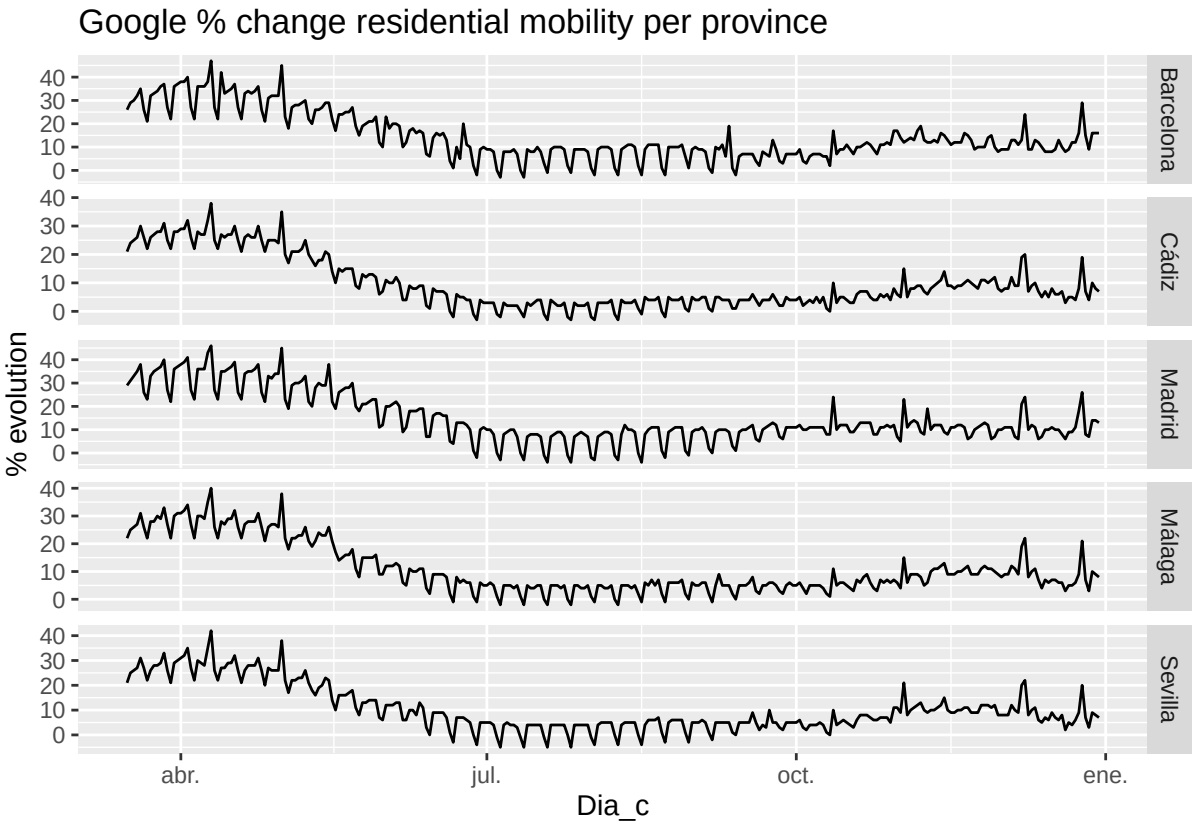

```
########################################
# A - Nº cases per province (Barcelona, Madrid, Málaga, Córdoba and Cádiz)
autoplot(Total_ts_b, num_casos.x) +
  labs(y = "Nº cases",
       title = "Nº cases per province")
```



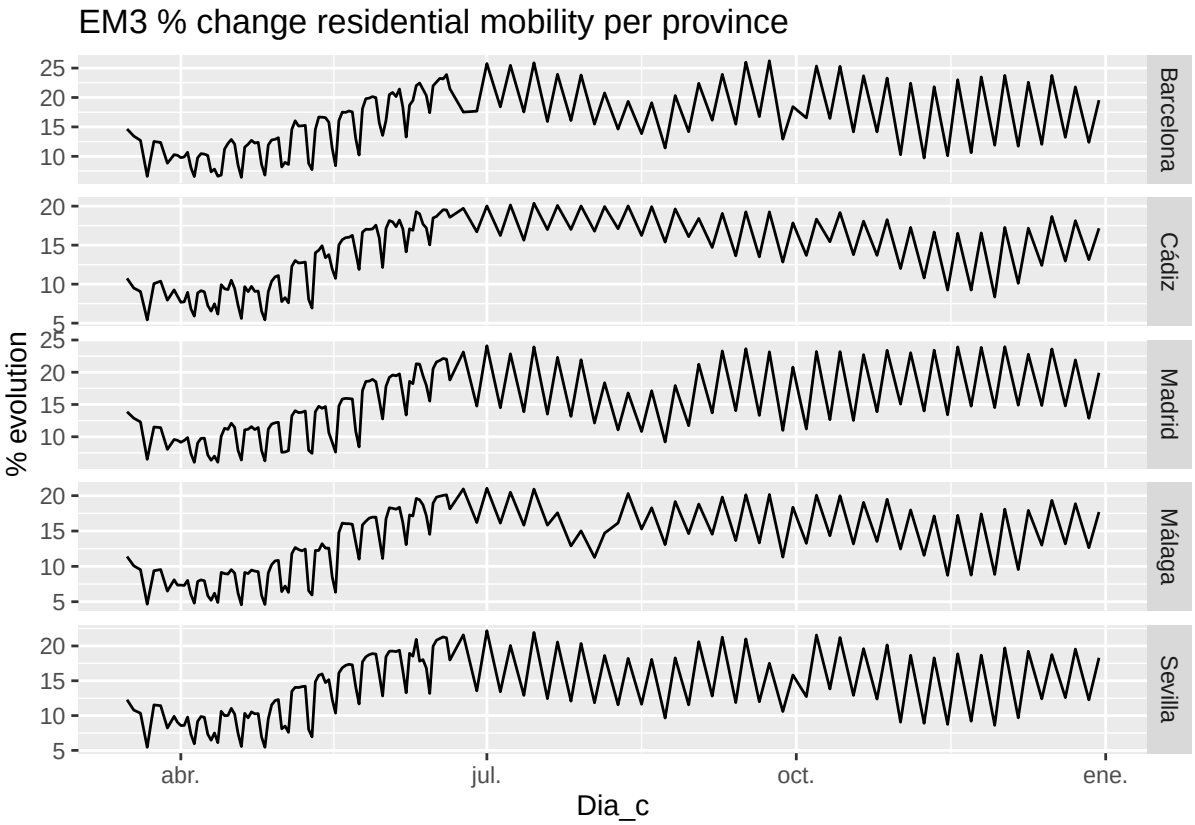

```
# B - Nº cases per province (Barcelona, Madrid, Málaga, Córdoba and Cádiz)
Total_ts_b %>%
  group_by(sub_region_2) %>%
  summarise(CASOS = sum(num_casos.x))%>%
  ggplot(aes(x = Dia_c, y = CASOS)) +
  geom_line() +
  facet_grid(vars(sub_region_2), scales = "free_y") +
  labs(title = "Nº cases per province", y= "Nº cases")
```



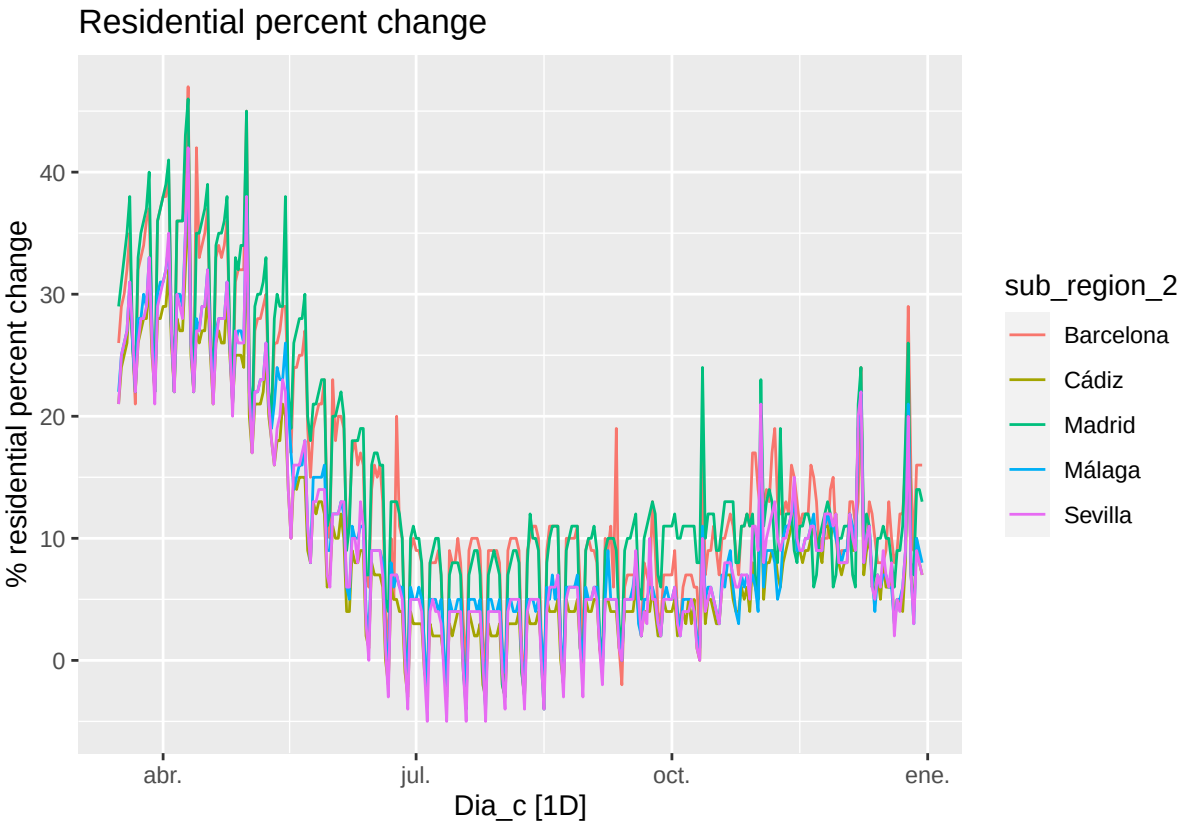

```
# B.b - Google % change residential mobility per province (Barcelona, Madrid, Málaga, Cádiz and Sevilla
Total_ts_b %>%
  group_by(sub_region_2) %>%
  summarise(per_c = (residential_percent_change_from_baseline))%>%
  ggplot(aes(x = Dia_c, y = per_c)) +
  geom_line() +
  facet_grid(vars(sub_region_2), scales = "free_y") +
  labs(title = "Google % change residential mobility per province", y= "% evolution")
```



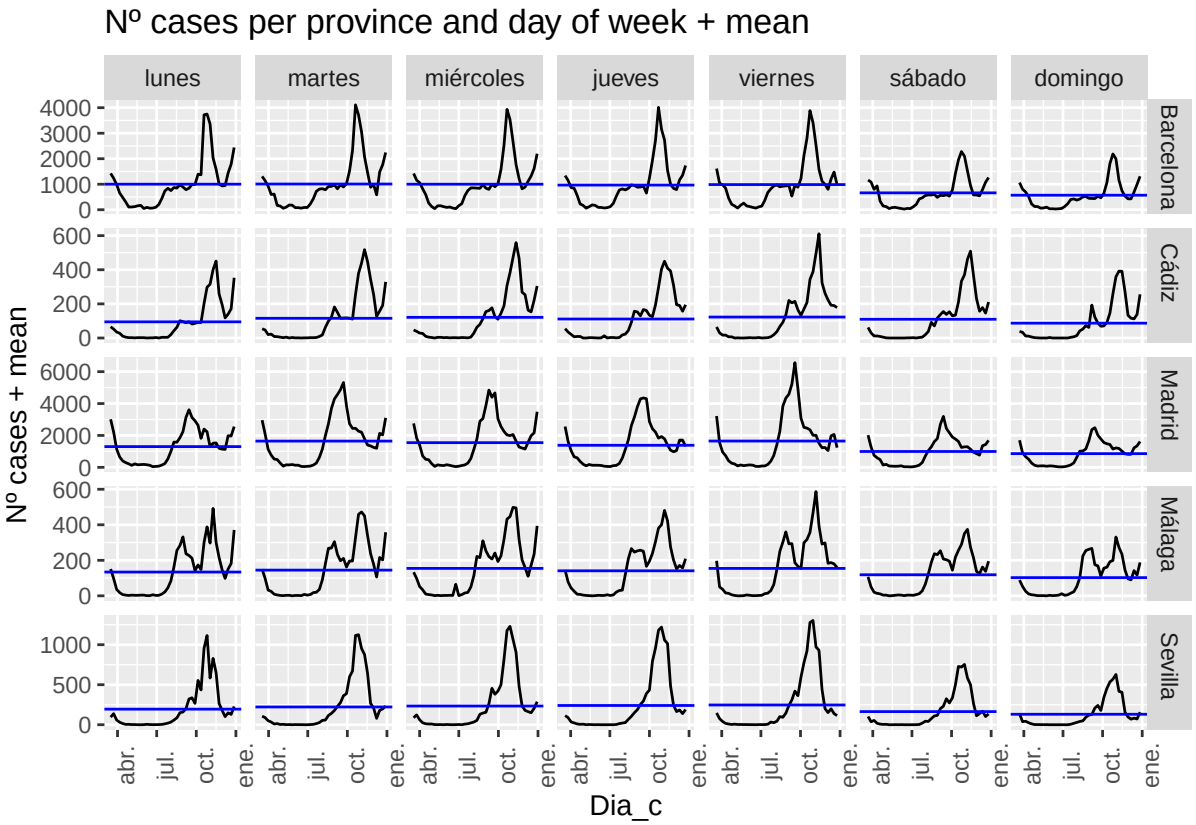

```
# B.c - EM3 % change residential mobility per province (Barcelona, Madrid, Málaga, Cádiz and Sevilla)
Total_ts_b %>%
  group_by(sub_region_2) %>%
  summarise(per_c = (Total))%>%
  ggplot(aes(x = Dia_c, y = per_c)) +
  geom_line() +
  facet_grid(vars(sub_region_2), scales = "free_y") +
  labs(title = "EM3 % change residential mobility per province", y= "% evolution")
```



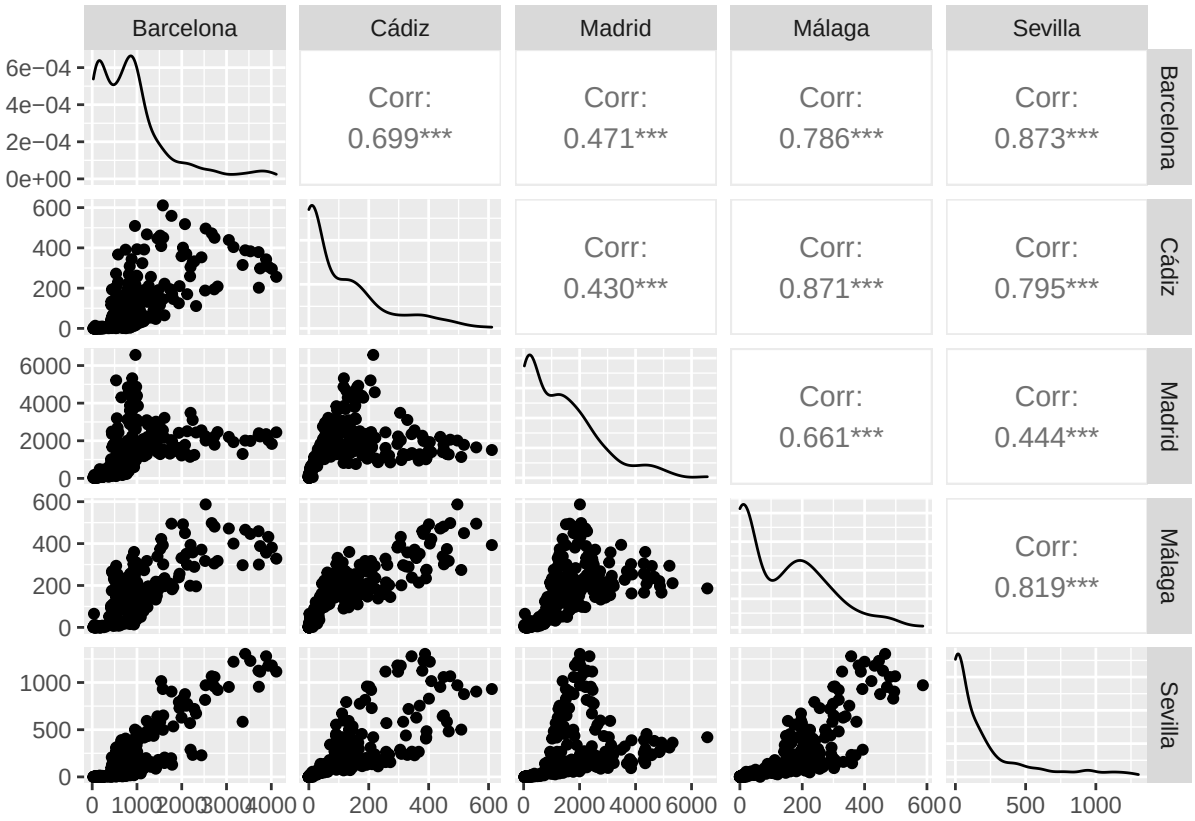

```
# % residential percent change (Barcelona, Madrid, Málaga, Cádiz and Sevilla)
autoplot(Total_ts_b, residential_percent_change_from_baseline ) +
  labs(y = "% residential percent change",
       title = "Residential percent change")
```



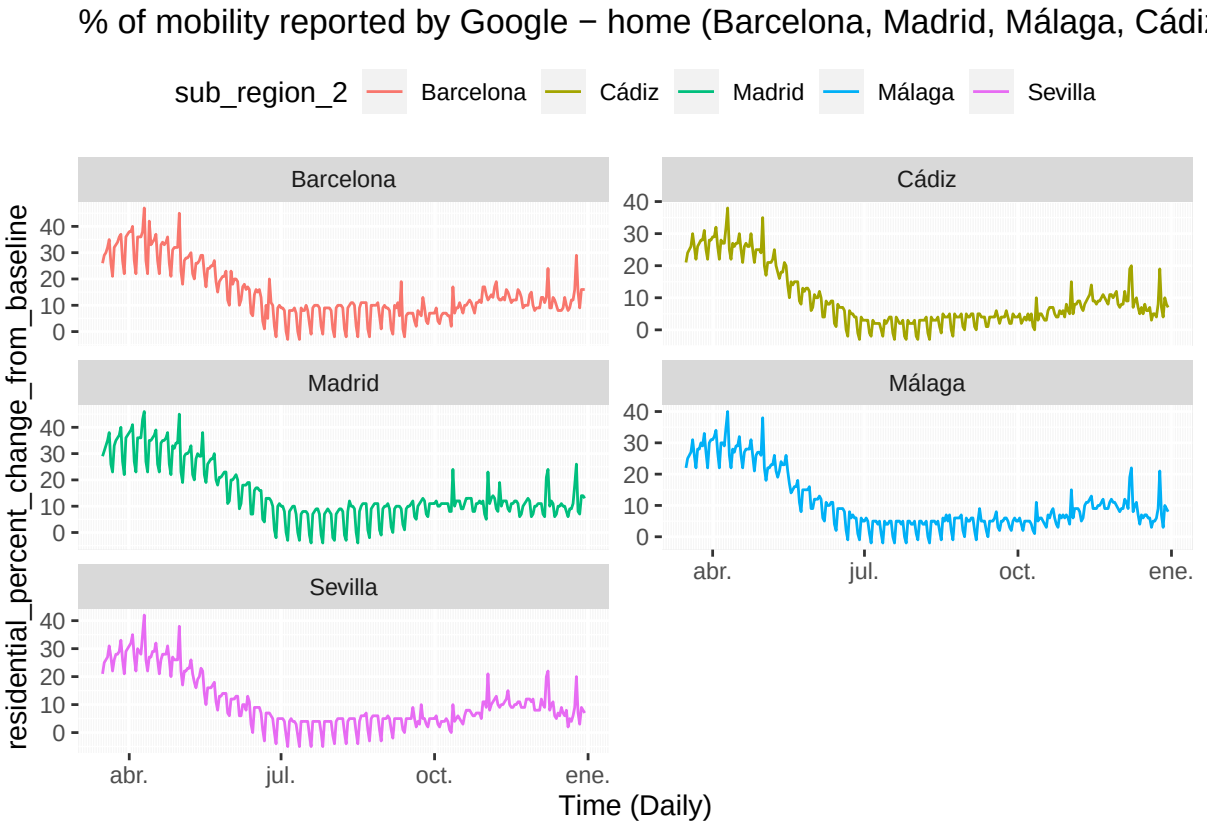

```
# Nº cases per province and day of week + mean
Total_ts_b %>%
  gg_subseries(num_casos.x, period = "week") +
  labs(y = "Nº cases + mean",
       title = "Nº cases per province and day of week + mean")
```




```
# Correlation plot / nº cases by province
Total_ts_b %>%
  group_by(sub_region_2) %>%
  summarise(CASOS = sum(num_casos.x))%>%
  pivot_wider(values_from=CASOS, names_from=sub_region_2) %>%
  GGally::ggpairs(2:6)
```




```
# % of mobility reported by Google - residential_percent_change
autoplot(Total_ts_b, residential_percent_change_from_baseline) +
  facet_wrap(~sub_region_2, scales = "free_y", ncol=2) +
  theme(legend.position = "top") +
  scale_x_date(date_minor_breaks = "1 day", name = "Time (Daily)") +
  ggtitle(label = "% of mobility reported by Google - home (Barcelona, Madrid, Málaga, Cádiz and Sevill
```



3 ARIMA - fpp3 library

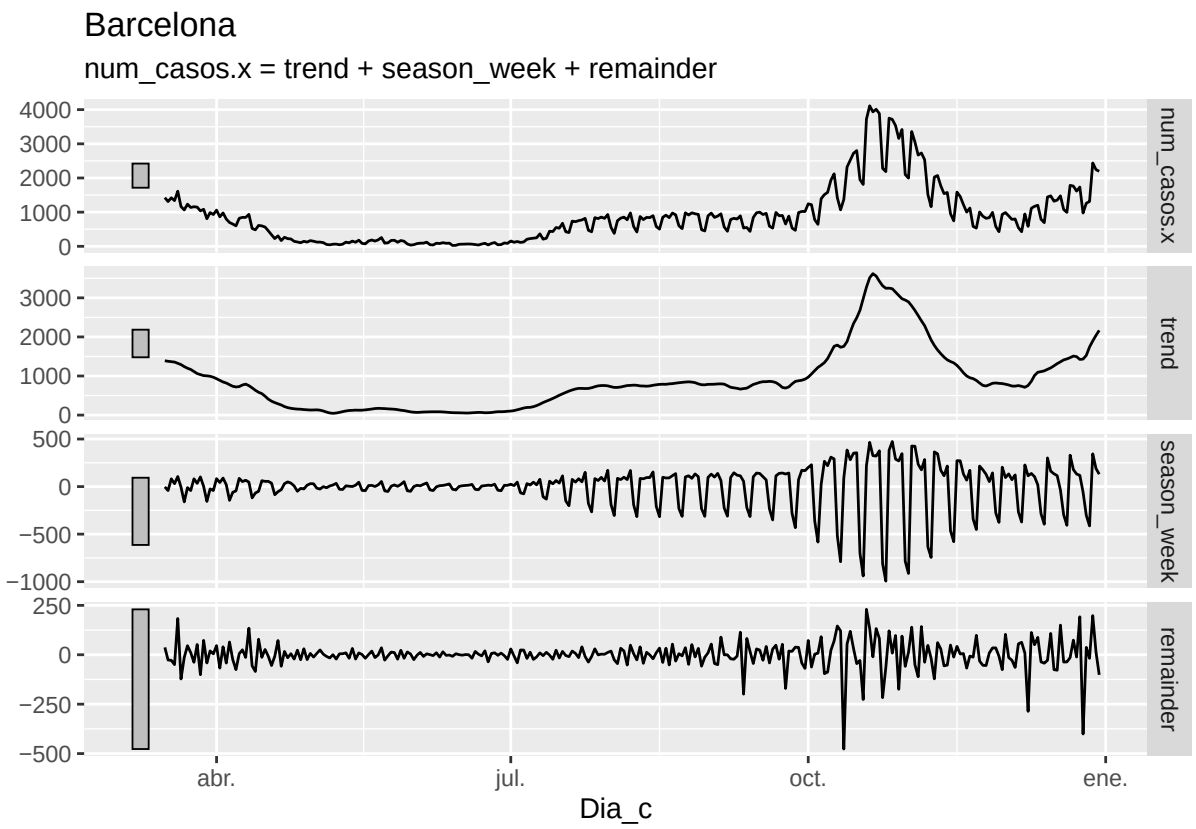
3.1 STL (Seasonal and Trend decomposition using Loess - Barcelona, Madrid, Málaga, Cádiz and Sevilla)

As stated by (Hyndman and Athanasopoulos 2021)... "STL has several advantages over classical decomposition, and the SEATS and X-11 methods:

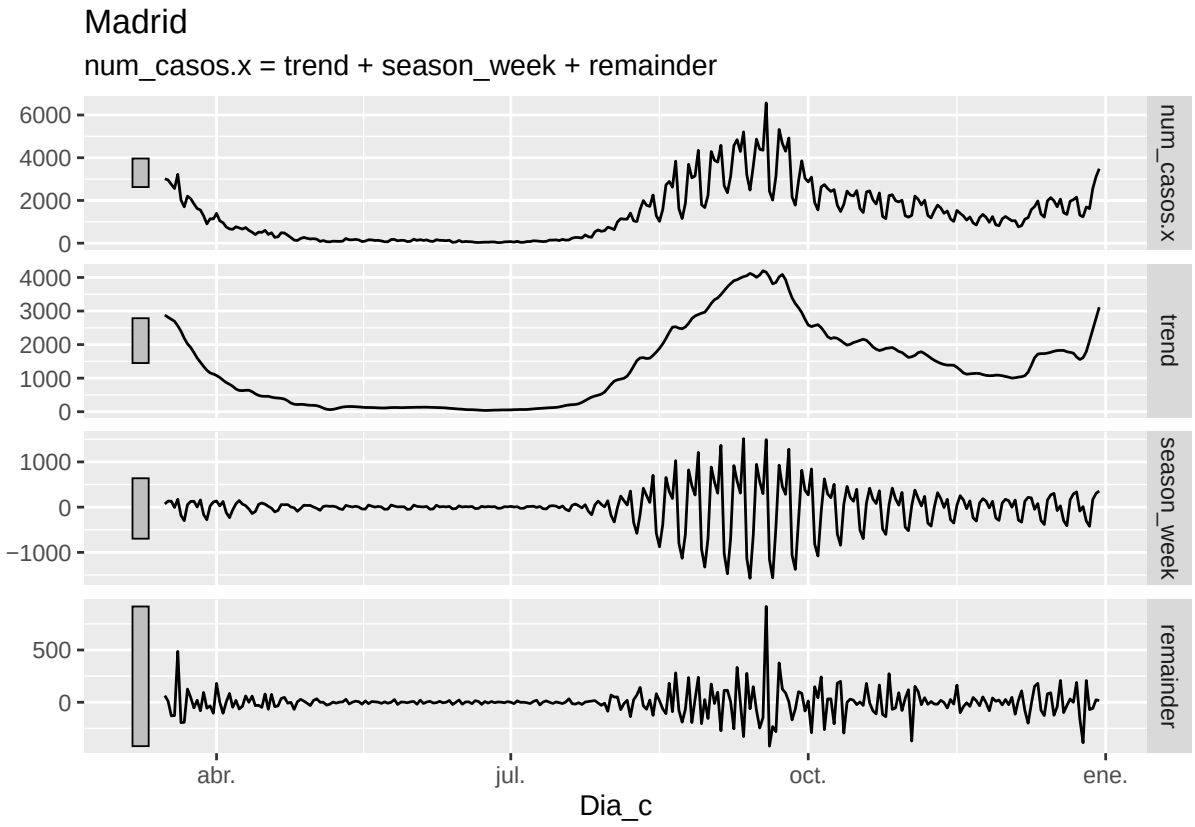
- Unlike SEATS and X-11, STL will handle any type of seasonality, not only monthly and quarterly data.
- The seasonal component is allowed to change over time, and the rate of change can be controlled by the user.
- The smoothness of the trend-cycle can also be controlled by the user.
- It can be robust to outliers (i.e., the user can specify a robust decomposition), so that occasional unusual observations will not affect the estimates of the trend-cycle and seasonal components. They will, however, affect the remainder component"...

```
# Check Seasonal and trend
#Total_ts %>%
#  filter(sub_region_2 == "Barcelona") %>%
#  model(STL(num_casos.x)) %>%
#  components() %>%
#  autoplot()

#Total_ts %>%
#  filter(sub_region_2 == "Barcelona") %>%
#  model(STL(num_casos.x ~ season(window = 7))) %>%
```

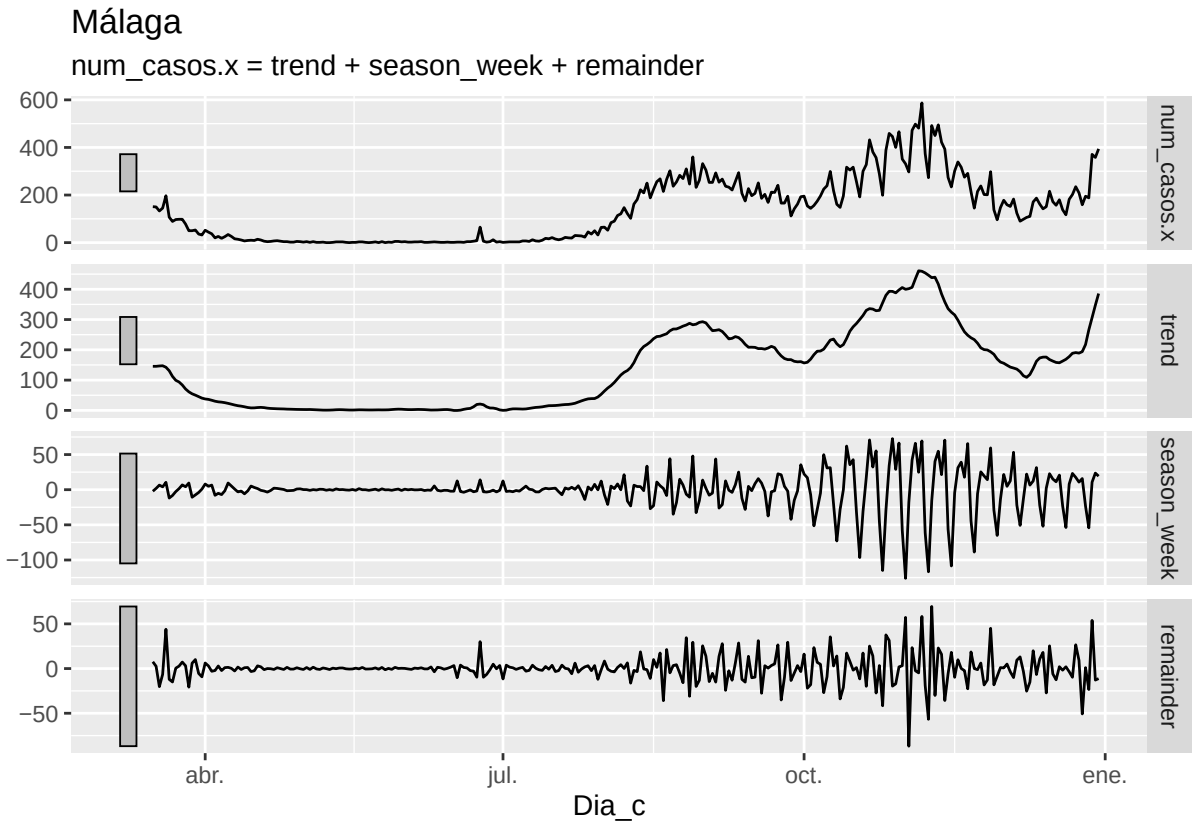
```
#  components() %>%
#  autoplot()

Total_ts %>%
  #filter_index("2020-09-1" ~ "2020-12-31")  %>%
  filter(sub_region_2 == "Barcelona") %>%
  model(STL(num_casos.x ~ season(window = 7) +
              trend(window = 7))) %>%
  components() %>%
  autoplot() + labs(title="Barcelona")
```



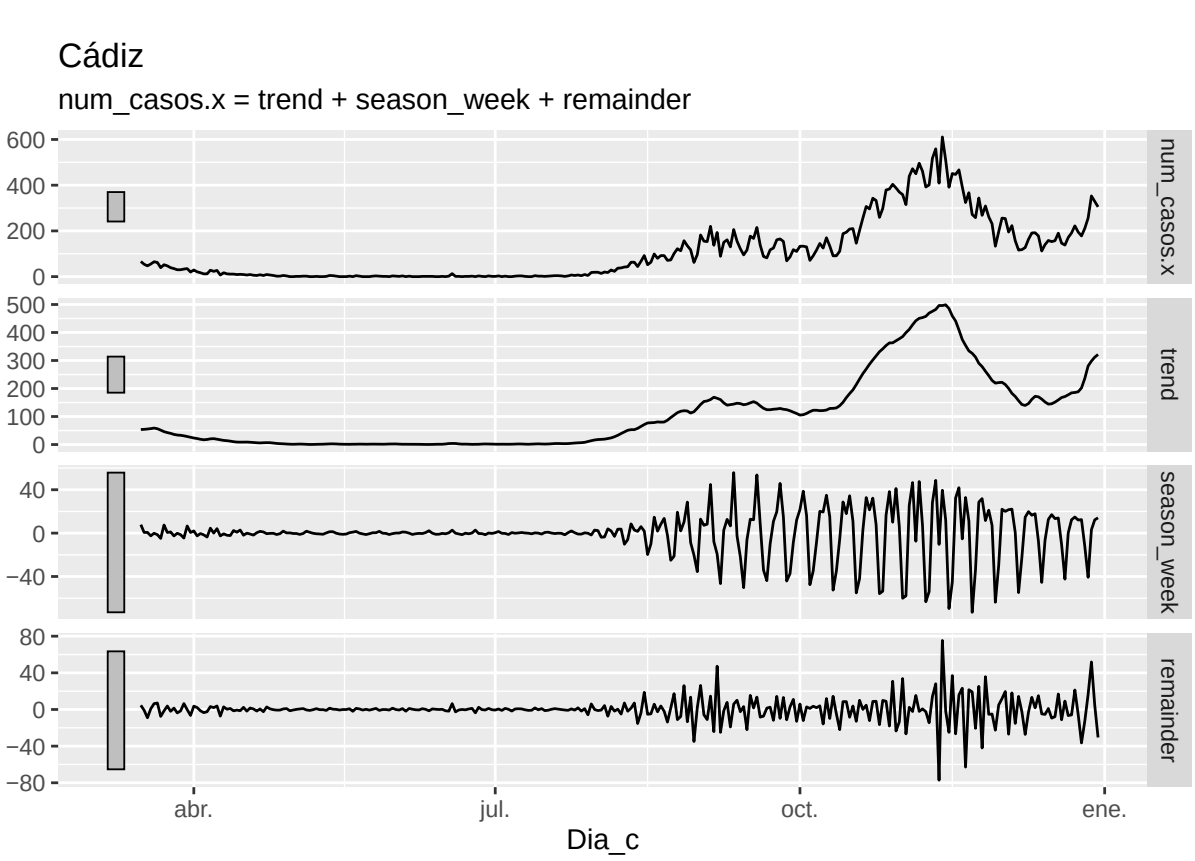

```
Total_ts %>%
  #filter_index("2020-09-1" ~ "2020-12-31")  %>%
  filter(sub_region_2 == "Madrid") %>%
  model(STL(num_casos.x ~ season(window = 7) +
              trend(window = 7))) %>%
  components() %>%
  autoplot() + labs(title="Madrid")
```



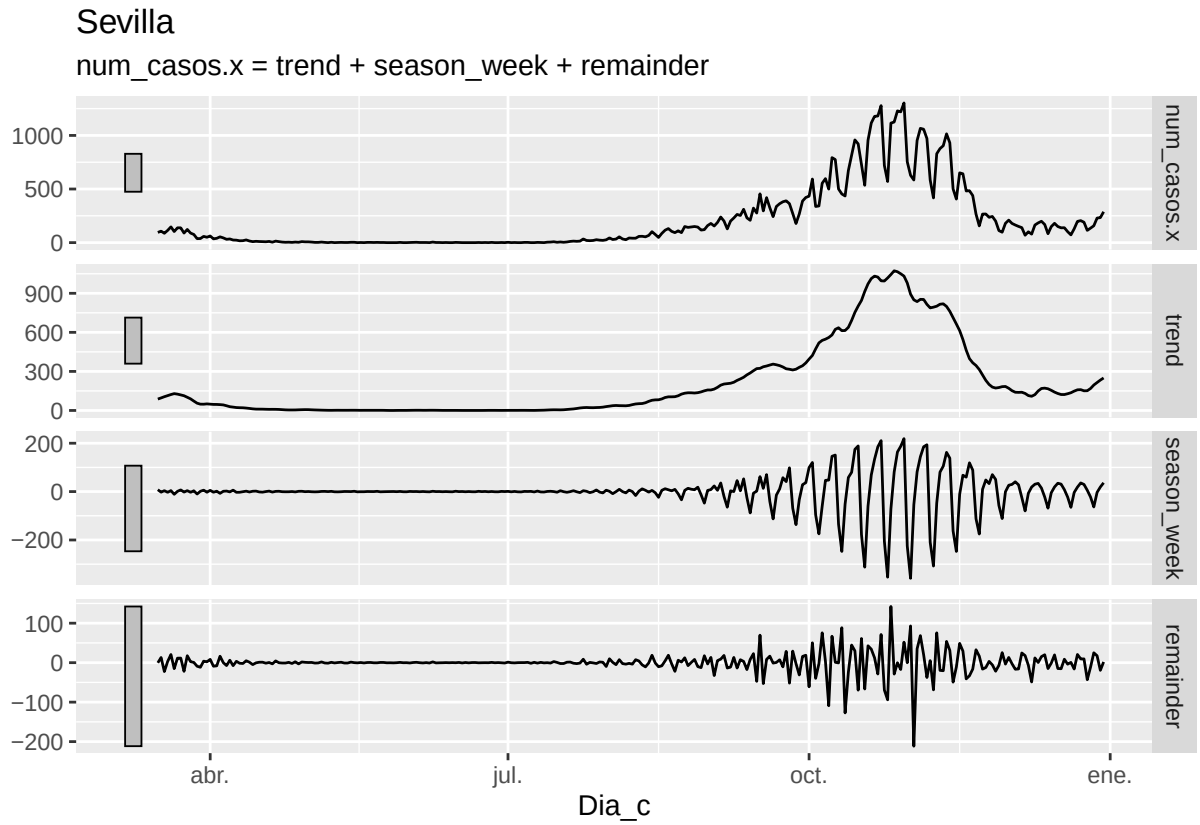

```
Total_ts %>%
  #filter_index("2020-09-1" ~ "2020-12-31")  %>%
  filter(sub_region_2 == "Málaga") %>%
  model(STL(num_casos.x ~ season(window = 7) +
              trend(window = 7))) %>%
  components() %>%
  autoplot() + labs(title="Málaga")
```




```
Total_ts %>%
  #filter_index("2020-09-1" ~ "2020-12-31")  %>%
  filter(sub_region_2 == "Cádiz") %>%
  model(STL(num_casos.x ~ season(window = 7) +
              trend(window = 7))) %>%
  components() %>%
  autoplot() + labs(title="Cádiz")
```




```
Total_ts %>%
  #filter_index("2020-09-1" ~ "2020-12-31")  %>%
  filter(sub_region_2 == "Sevilla") %>%
  model(STL(num_casos.x ~ season(window = 7) +
              trend(window = 7))) %>%
  components() %>%
  autoplot() + labs(title="Sevilla")
```



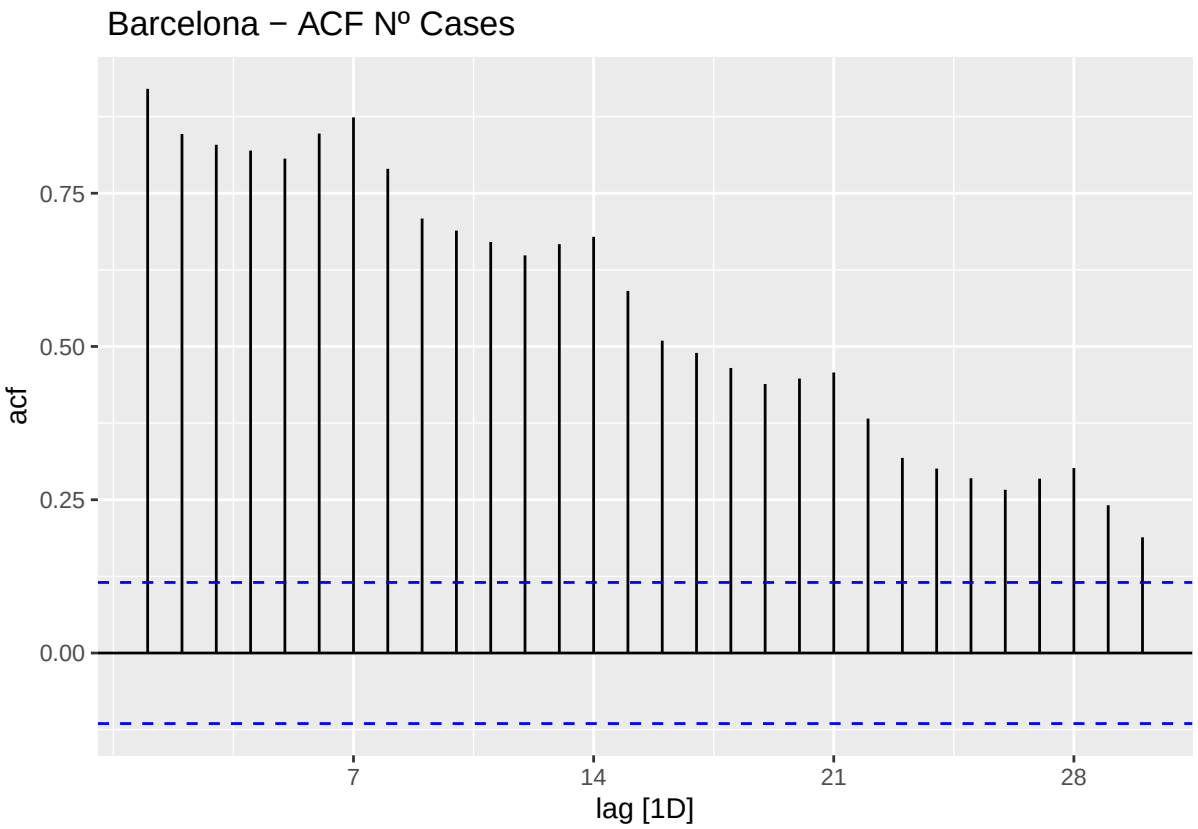
3.2 ACF and PACF (Barcelona, Madrid, Málaga, Córdoba and Cádiz)

As stated by (Hyndman and Athanasopoulos 2021)... "ACF plot is also useful for identifying non-stationary time series. For a stationary time series, the ACF will drop to zero relatively quickly, while the ACF of non-stationary data decreases slowly.

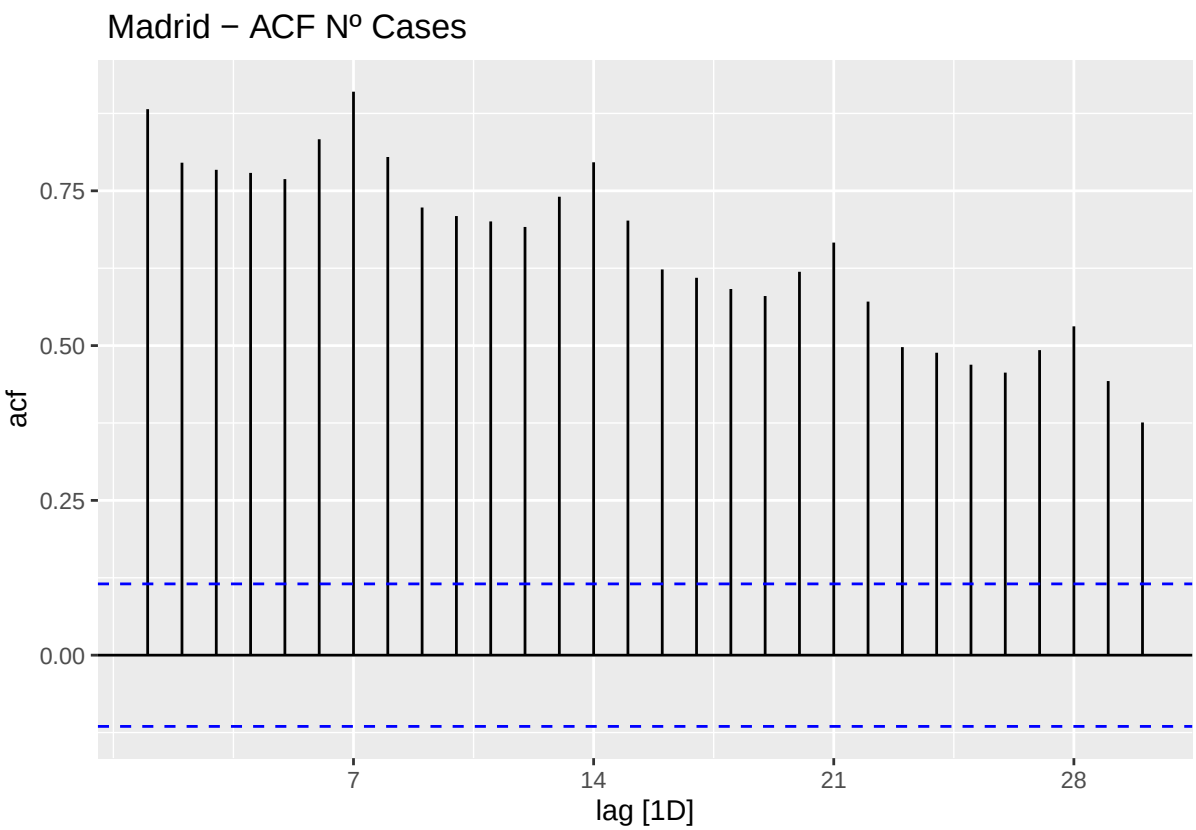
- Our time-series for Bar is non-stationary

```
# New time-series for Bar, Mad, Mal, Cor and, Cad
Total_ts %>%
  filter_index("2020-03-15" ~ "2020-12-31")  %>%
  filter(sub_region_2 == "Barcelona")-> Bar_N_cases #%>%
Total_ts %>%
  filter_index("2020-03-15" ~ "2020-12-31")  %>%
  filter(sub_region_2 == "Madrid")-> Mad_N_cases #%>%
Total_ts %>%
  filter_index("2020-03-15" ~ "2020-12-31")  %>%
  filter(sub_region_2 == "Málaga")-> Mal_N_cases #%>%
Total_ts %>%
  filter_index("2020-03-15" ~ "2020-12-31")  %>%
  filter(sub_region_2 == "Cádiz")-> Cad_N_cases #%>%
Total_ts %>%
  filter_index("2020-03-15" ~ "2020-12-31")  %>%
  filter(sub_region_2 == "Sevilla")-> Sev_N_cases #%>%

# ACF
Bar_N_cases %>%
```

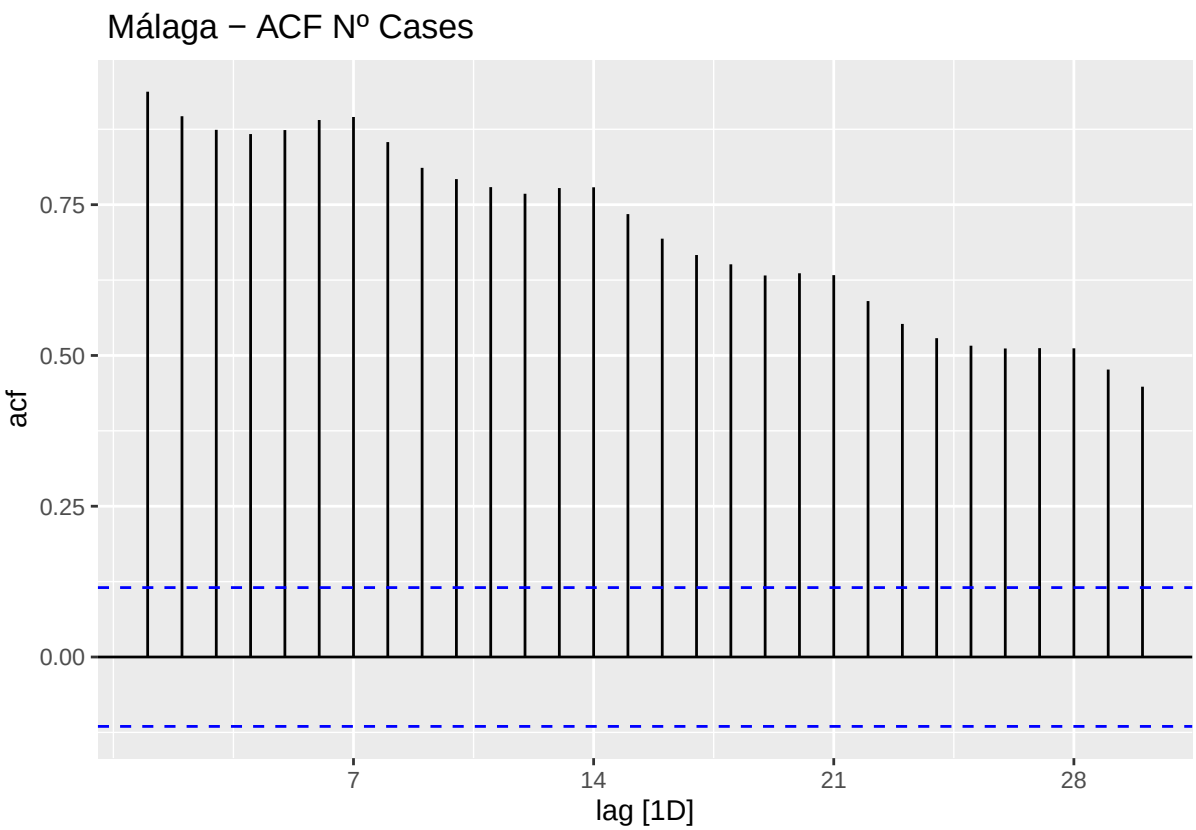
```
  ACF(num_casos.x, lag_max = 30) %>%
  autoplot() +
  labs(title=" Barcelona - ACF Nº Cases")
```



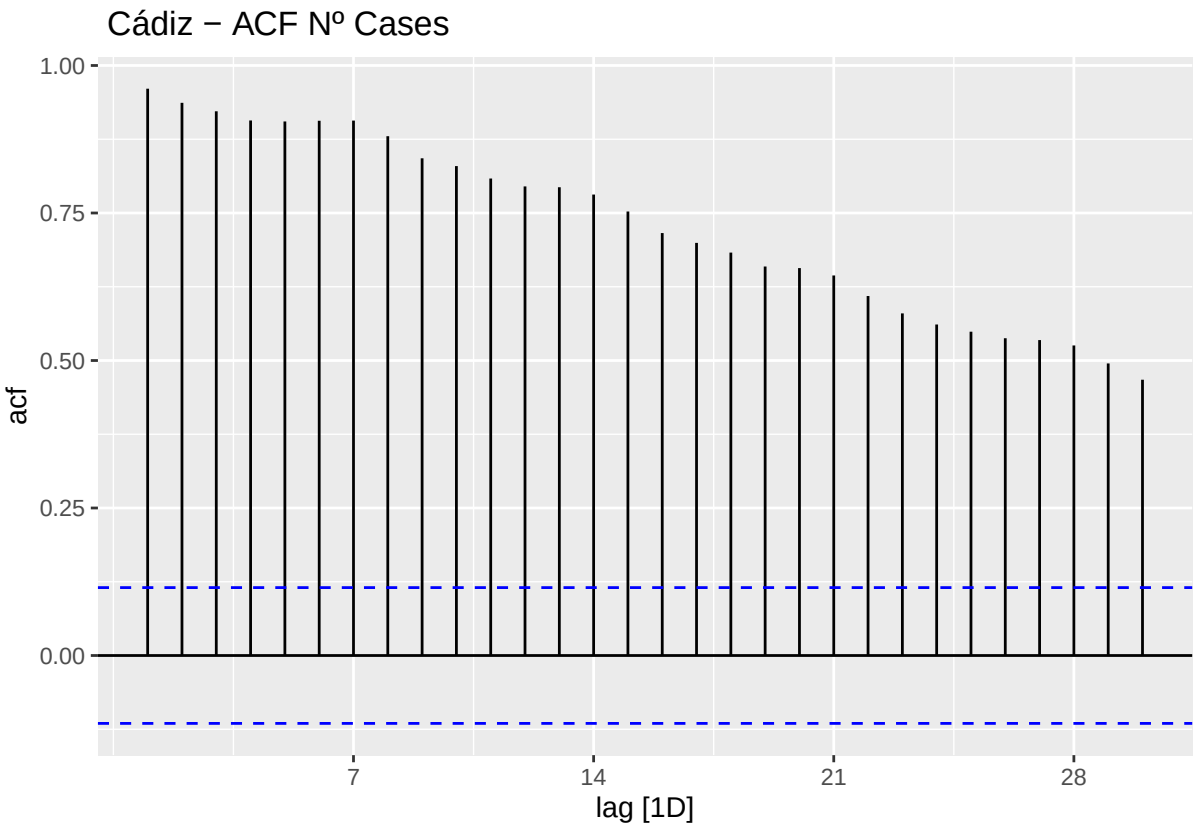

```
Mad_N_cases %>%
  ACF(num_casos.x, lag_max = 30) %>%
  autoplot() +
  labs(title=" Madrid - ACF Nº Cases")
```



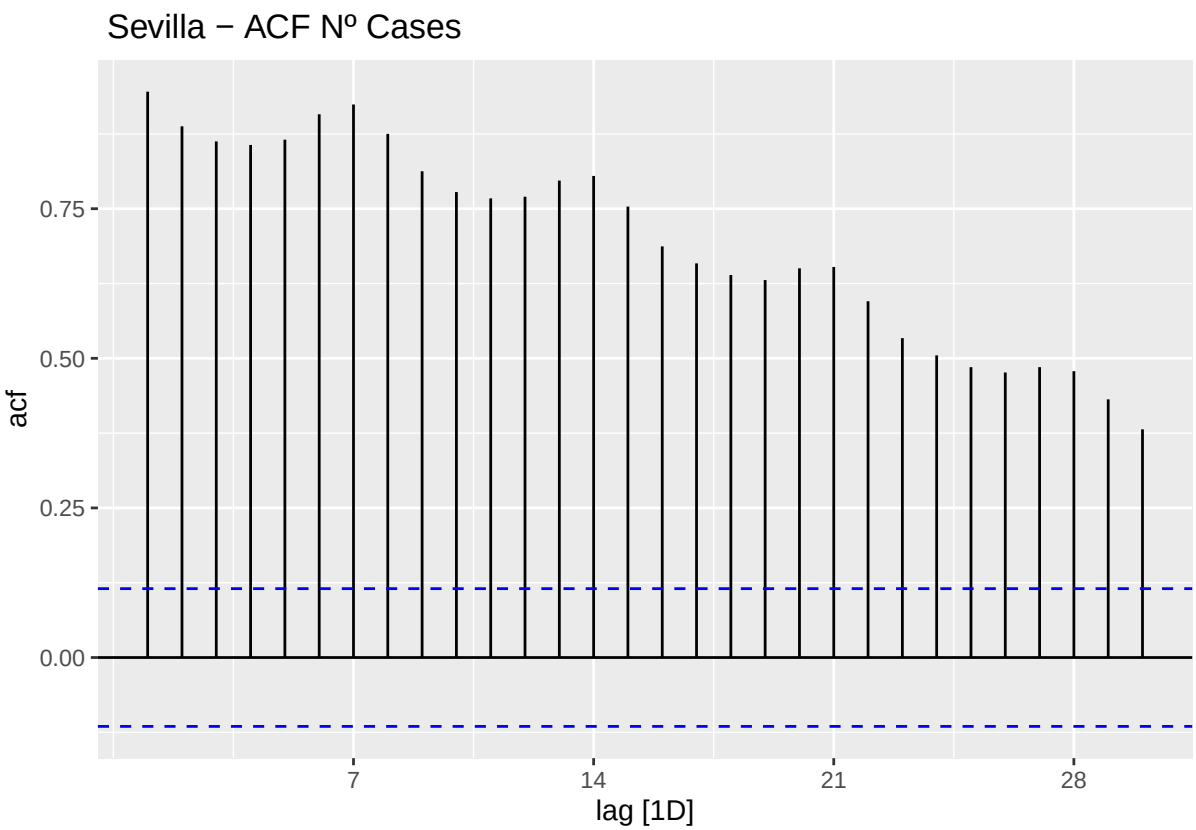

```
Mal_N_cases %>%
  ACF(num_casos.x, lag_max = 30) %>%
  autoplot() +
  labs(title=" Málaga - ACF Nº Cases")
```



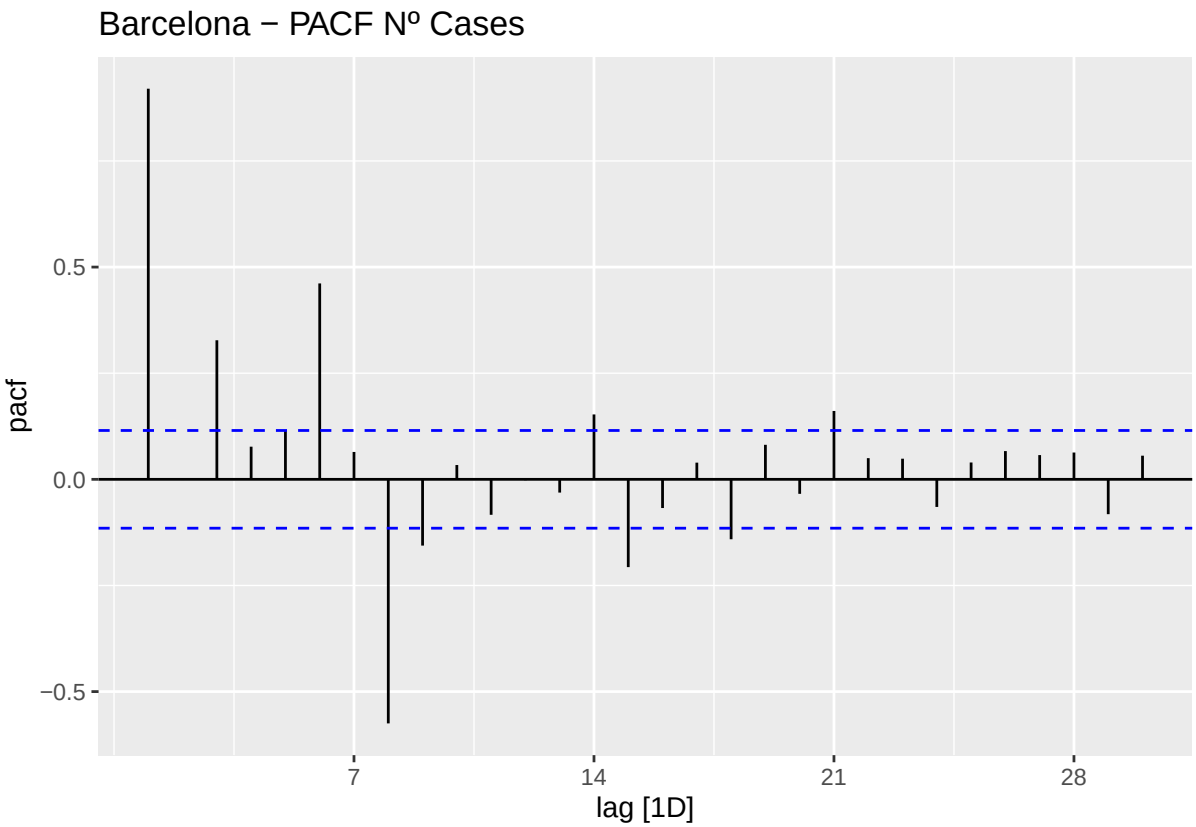

```
Cad_N_cases %>%
  ACF(num_casos.x, lag_max = 30) %>%
  autoplot() +
  labs(title=" Cádiz - ACF Nº Cases")
```



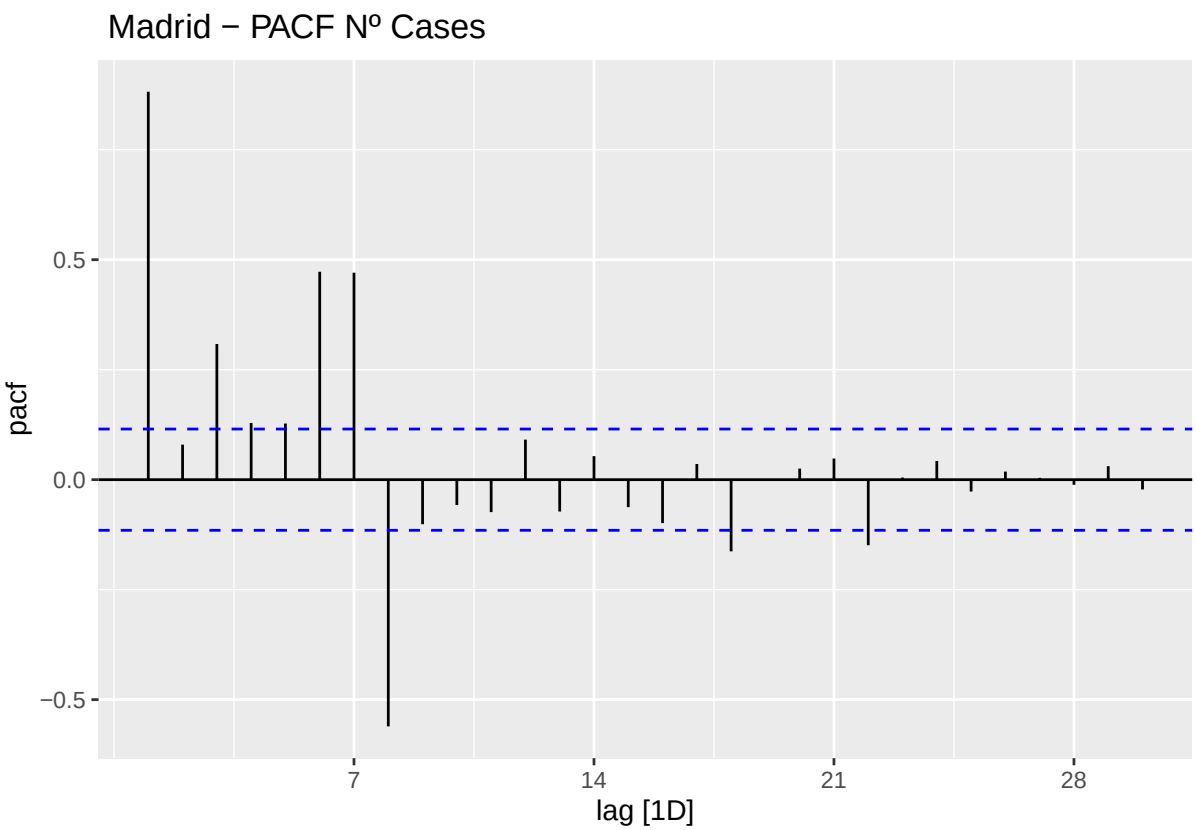

```
Sev_N_cases %>%
  ACF(num_casos.x, lag_max = 30) %>%
  autoplot() +
  labs(title=" Sevilla - ACF Nº Cases")
```



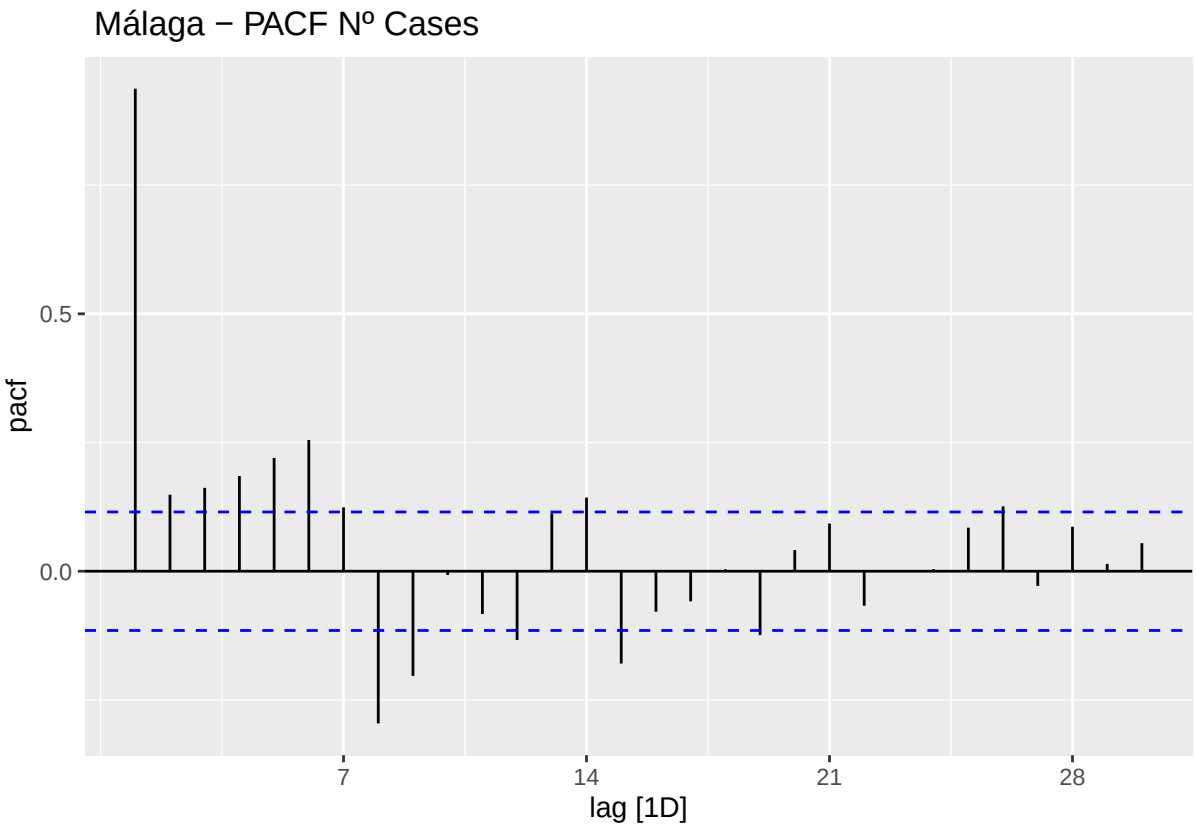

```
# PACF
Bar_N_cases %>%
  PACF(num_casos.x, lag_max = 30) %>%
  autoplot() +
  labs(title="Barcelona - PACF Nº Cases")
```



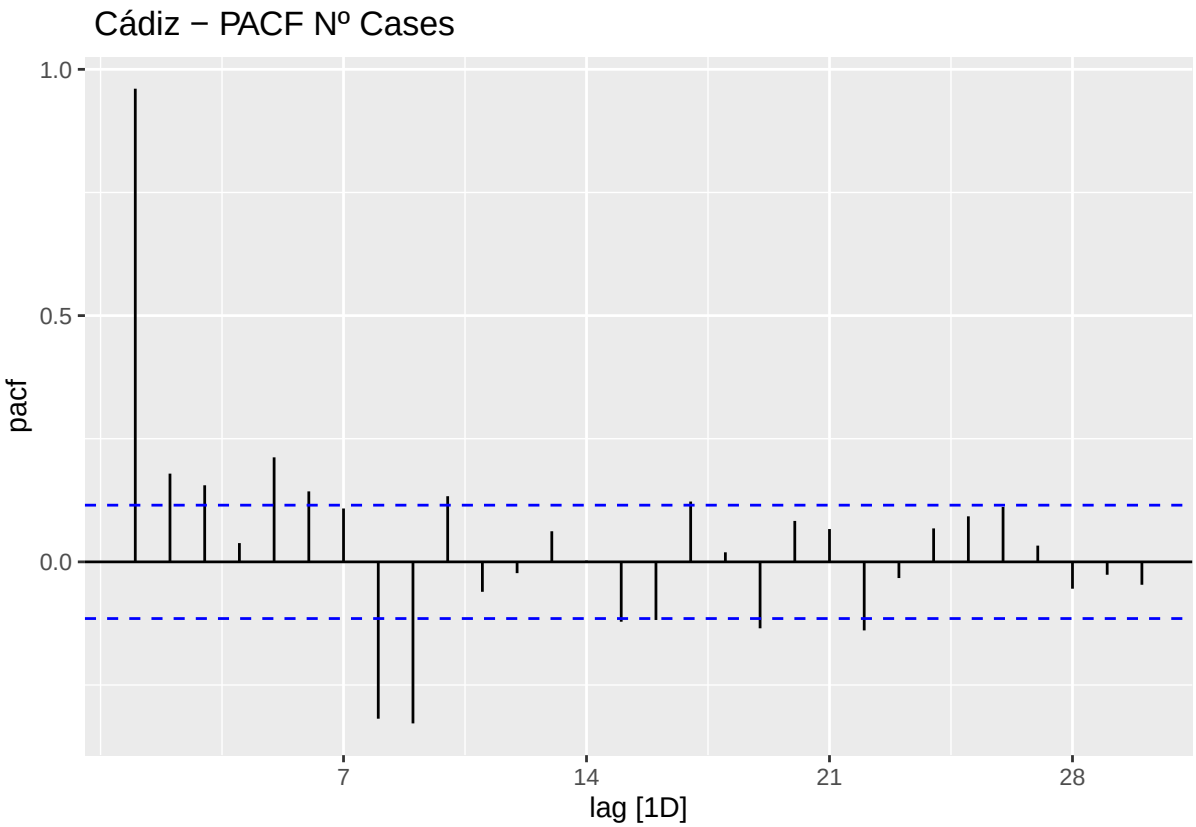

```
Mad_N_cases %>%
  PACF(num_casos.x, lag_max = 30) %>%
  autoplot() +
  labs(title=" Madrid - PACF Nº Cases")
```



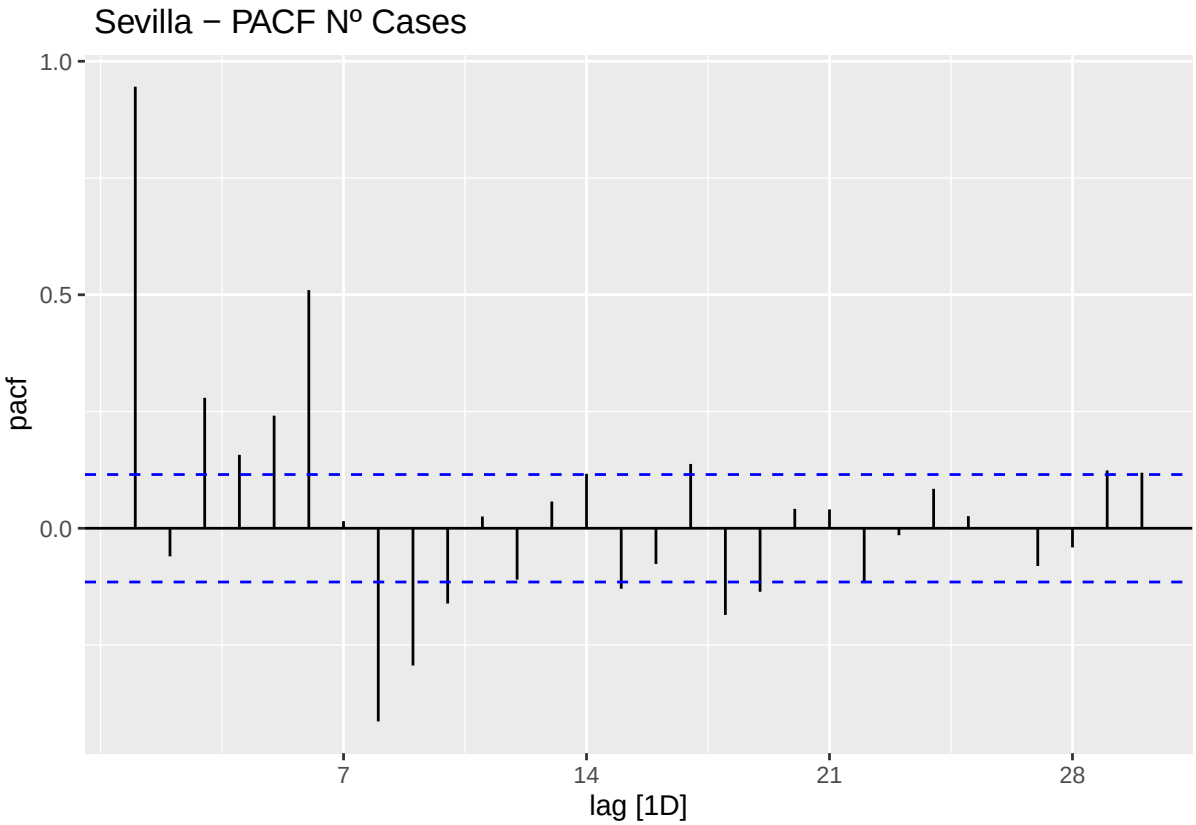

```
Mal_N_cases %>%
  PACF(num_casos.x, lag_max = 30) %>%
  autoplot() +
  labs(title=" Málaga - PACF Nº Cases")
```




```
Cad_N_cases %>%
  PACF(num_casos.x, lag_max = 30) %>%
  autoplot() +
  labs(title=" Cádiz - PACF Nº Cases")
```

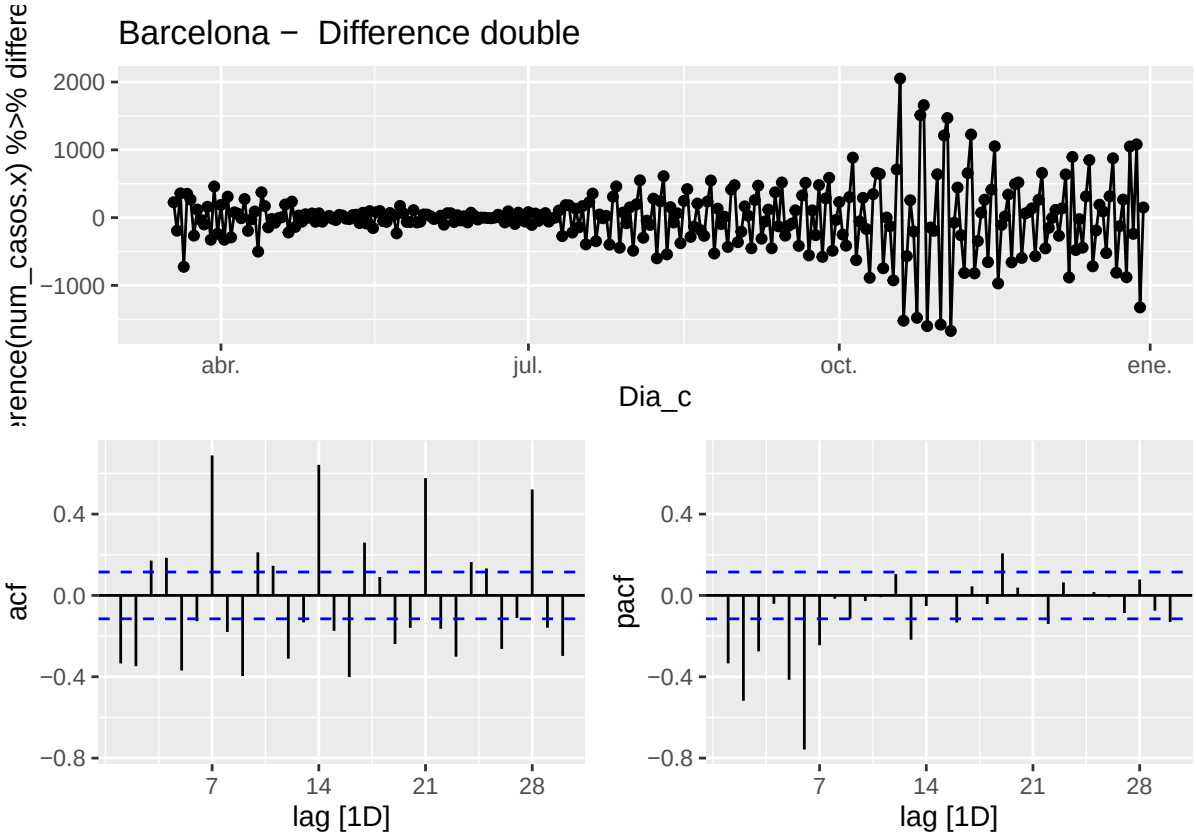



```
Sev_N_cases %>%
  PACF(num_casos.x, lag_max = 30) %>%
  autoplot() +
  labs(title=" Sevilla - PACF Nº Cases")
```

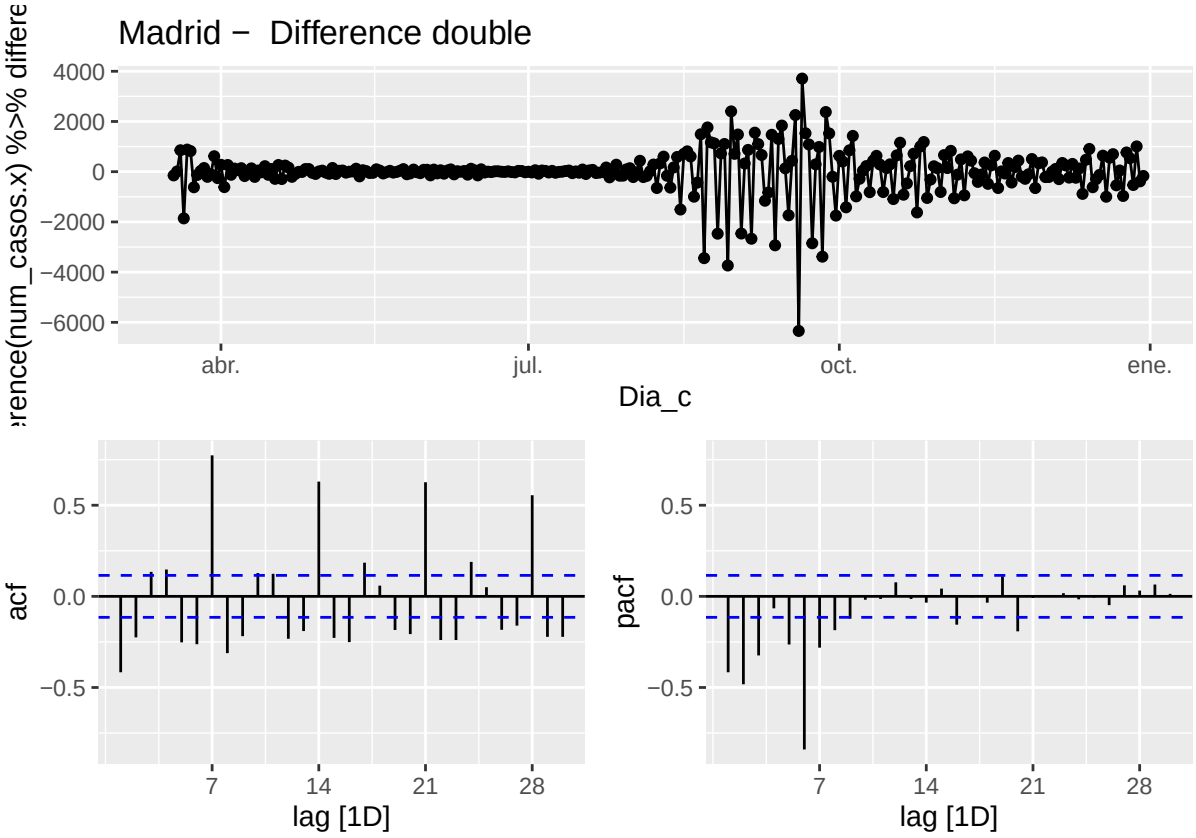


Double difference is plotted.

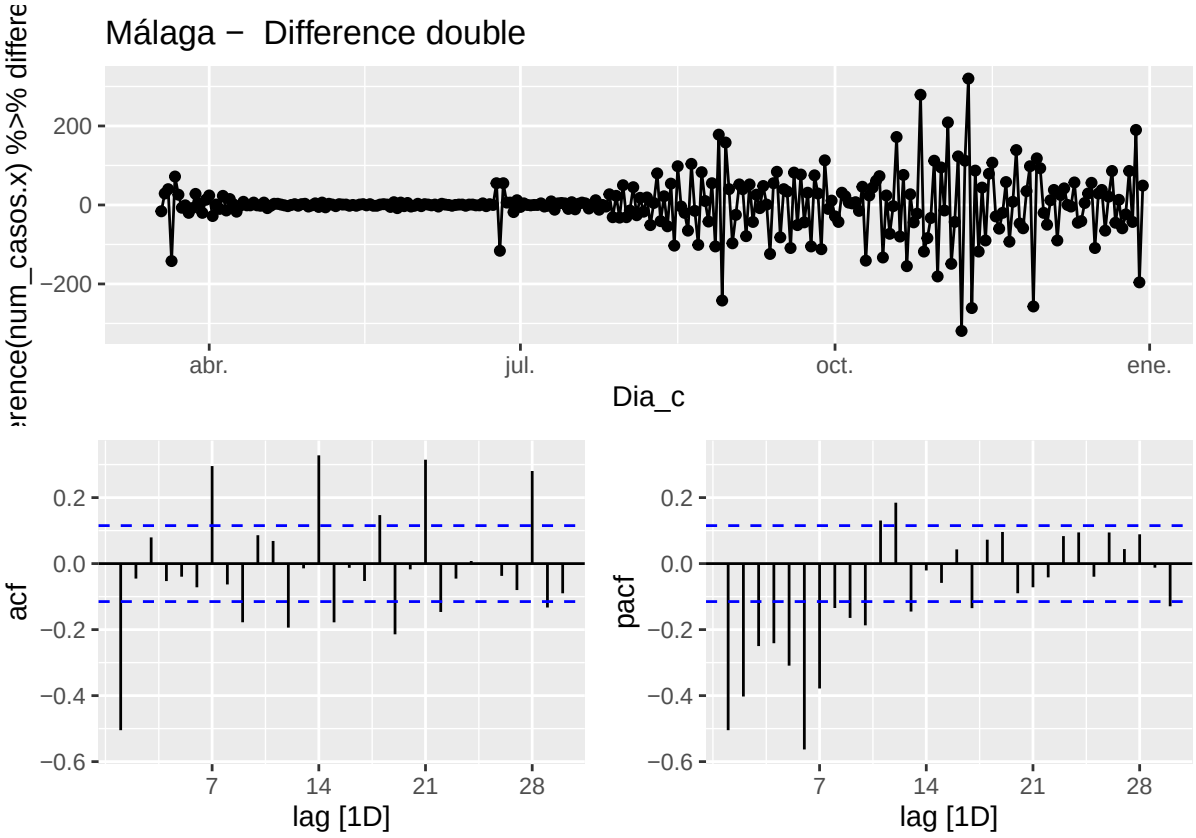
```
# Difference
#Bar_N_cases %>%
#  gg_tsdisplay(difference(num_casos.x),
#               plot_type='partial', lag_max = 30)

#Bar_N_cases %>%
#  gg_tsdisplay(difference(log(num_casos.x)),
#               plot_type='partial', lag_max = 30)

# Difference double
Bar_N_cases %>%
  gg_tsdisplay(difference(num_casos.x) %>%
               difference(),
               plot_type='partial', lag_max = 30)+
  labs(title="Barcelona -  Difference double")
```



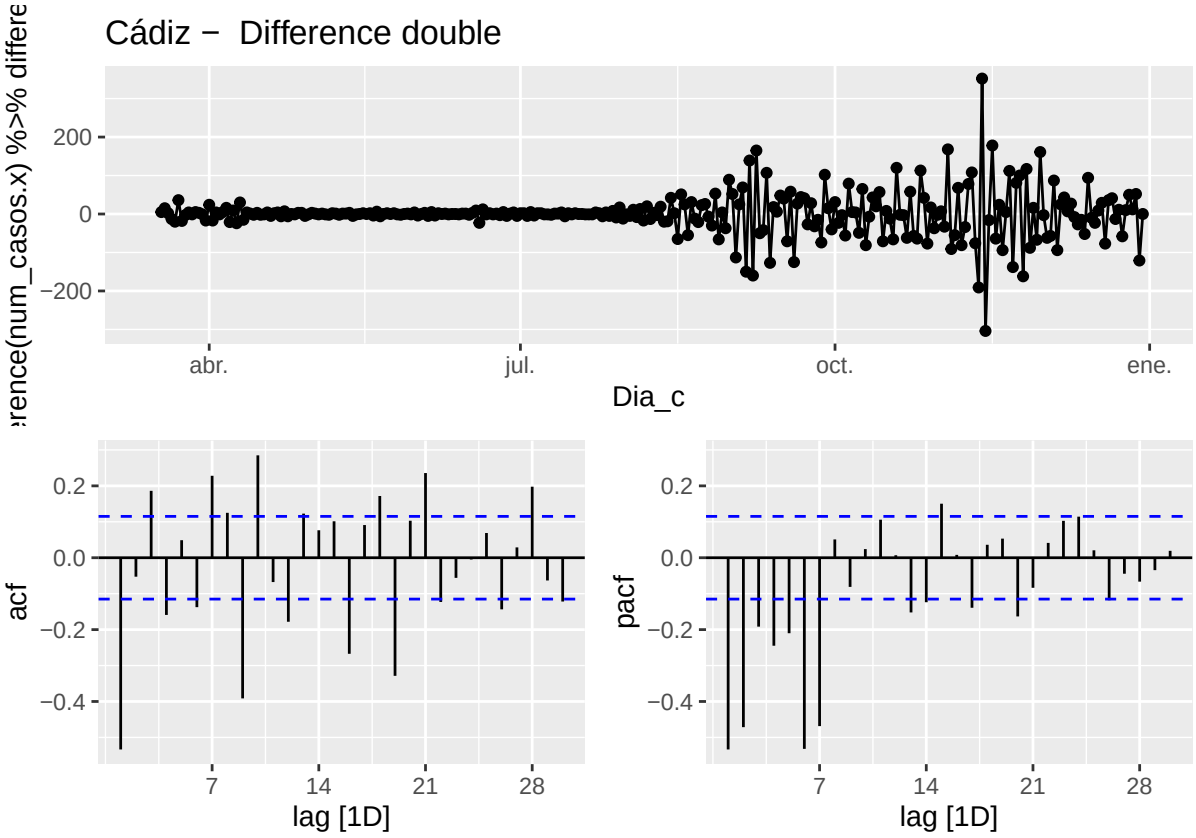

```
Mad_N_cases %>%
  gg_tsdisplay(difference(num_casos.x) %>%
                 difference(),
                 plot_type='partial', lag_max = 30)+
  labs(title="Madrid -  Difference double")
```



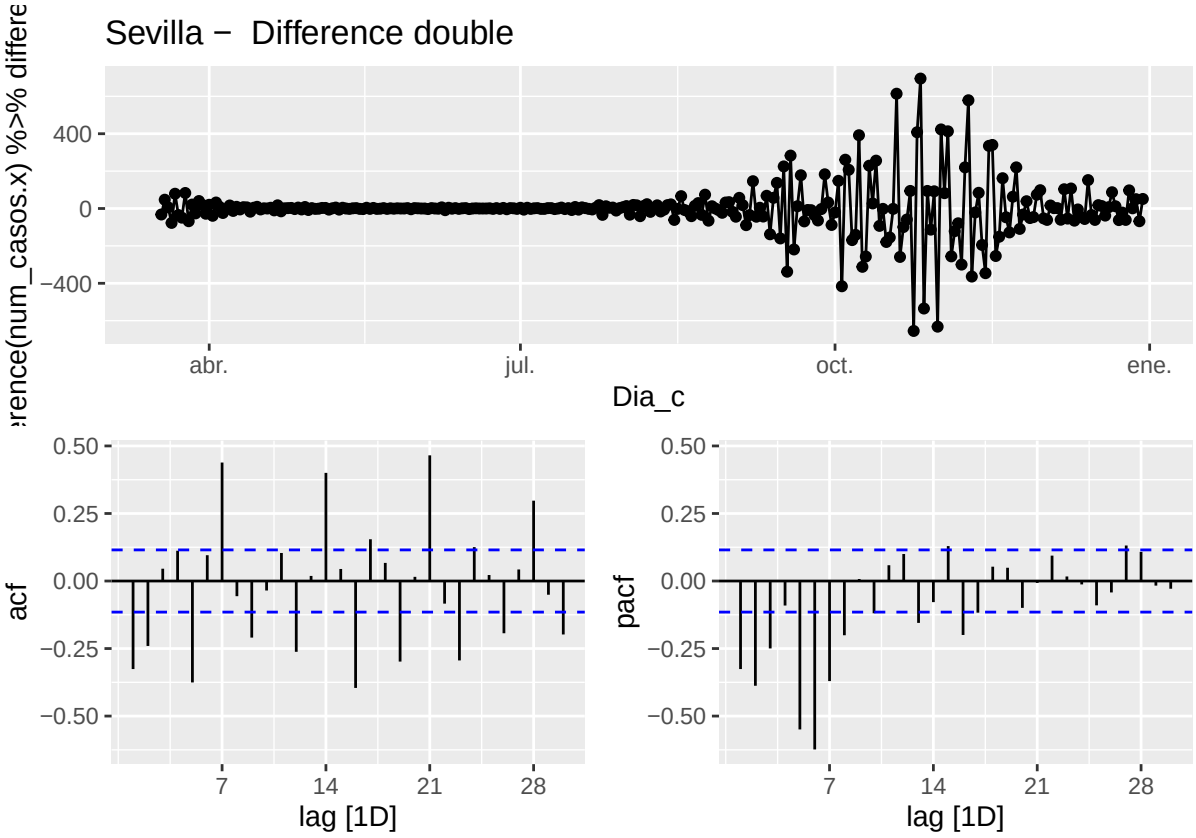

```
Mal_N_cases %>%
  gg_tsdisplay(difference(num_casos.x) %>%
                 difference(),
                 plot_type='partial', lag_max = 30)+
  labs(title="Málaga -  Difference double")
```




```
Cad_N_cases %>%
  gg_tsdisplay(difference(num_casos.x) %>%
                 difference(),
                 plot_type='partial', lag_max = 30)+
  labs(title="Cádiz -  Difference double")
```




```
Sev_N_cases %>%
  gg_tsdisplay(difference(num_casos.x) %>%
                 difference(),
                 plot_type='partial', lag_max = 30)+
  labs(title="Sevilla -  Difference double")
```



3.3 Model and Forecast (Barcelona, Madrid, Málaga, Córdoba and Cádiz)

3.3.1 Univariate (7, 14, 21 days) Barcelona

As stated by (Hyndman and Athanasopoulos 2021)... "The ARIMA() function uses unitroot_nsdiffs() to determine D (the number of seasonal differences to use), and unitroot_ndiffs() to determine d (the number of ordinary differences to use), when these are not specified."

```
# Train and test ts
Bar_N_cases_tr <- Bar_N_cases %>%
  filter_index("2020-03-15" ~ "2020-12-9")
Bar_N_cases_tt <- Bar_N_cases %>%
  filter_index("2020-12-10" ~ "2020-12-31")

# Model train
fit_model <- Bar_N_cases_tr %>%
  model(
    SNaive = SNAIVE(num_casos.x),
    arima_man = ARIMA(num_casos.x ~ pdq(2,1,2) + PDQ(1,1,1)),
    arima_at1 = ARIMA(num_casos.x),
    arima_at2 = ARIMA(num_casos.x, stepwise = FALSE, approx = FALSE))

# Show and report model
fit_model
```

```
## # A mable: 1 x 5
## # Key:     sub_region_2 [1]
```

```
##   sub_region_2   SNaive                arima_man                arima_at1
##   <chr>          <model>                 <model>                  <model>
## 1 Barcelona    <SNAIVE> <ARIMA(2,1,2)(1,1,1)[7]> <ARIMA(1,0,2)(1,1,0)[7]>
## # ... with 1 more variable: arima_at2 <model>
```

```
report(fit_model)
```

```
## # A tibble: 4 x 9
##   sub_region_2 .model     sigma2 log_lik   AIC  AICc   BIC ar_roots  ma_roots
##   <chr>        <chr>       <dbl>   <dbl> <dbl> <dbl> <dbl> <list>    <list>
## 1 Barcelona    SNaive    148096.      NA    NA    NA    NA  <NULL>    <NULL>
## 2 Barcelona    arima_man  35350.  -1735. 3484. 3485. 3509. <cpl [9]> <cpl [9]>
## 3 Barcelona    arima_at1  36260.  -1746. 3502. 3502. 3520. <cpl [8]> <cpl [2]>
## 4 Barcelona    arima_at2  33389.  -1735. 3485. 3485. 3510. <cpl [4]> <cpl [2]>
```

```
# Good model >> Less Sigma / More BIC or AIC
fit_model %>% pivot_longer(!sub_region_2,
                           names_to = "Model name",
                           values_to = "Orders")
```

```
## # A mable: 4 x 3
## # Key:     sub_region_2, Model name [4]
##   sub_region_2 `Model name`                    Orders
##   <chr>        <chr>                          <model>
## 1 Barcelona    SNaive                        <SNAIVE>
## 2 Barcelona    arima_man    <ARIMA(2,1,2)(1,1,1)[7]>
## 3 Barcelona    arima_at1    <ARIMA(1,0,2)(1,1,0)[7]>
## 4 Barcelona    arima_at2    <ARIMA(4,0,2)(0,1,0)[7]>
```

```
glance(fit_model) %>% arrange(AICc) %>% select(.model:BIC)
```

```
## # A tibble: 4 x 6
##   .model     sigma2 log_lik   AIC  AICc   BIC
##   <chr>       <dbl>   <dbl> <dbl> <dbl> <dbl>
## 1 arima_man  35350.  -1735. 3484. 3485. 3509.
## 2 arima_at2  33389.  -1735. 3485. 3485. 3510.
## 3 arima_at1  36260.  -1746. 3502. 3502. 3520.
## 4 SNaive    148096.     NA    NA    NA    NA
```

```
# We use a Ljung-Box test >> large p-value, confirmins residuals are similar to white noise.
augment(fit_model) %>%
  features(.innov, ljung_box, lag=7)
```

```
## # A tibble: 4 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>       <dbl>     <dbl>
## 1 Barcelona    arima_at1    19.6   0.00646
## 2 Barcelona    arima_at2    10.2   0.175
## 3 Barcelona    arima_man    16.2   0.0235
## 4 Barcelona    SNaive      827.    0
```

```
augment(fit_model) %>%
  features(.innov, ljung_box, lag=14)
```

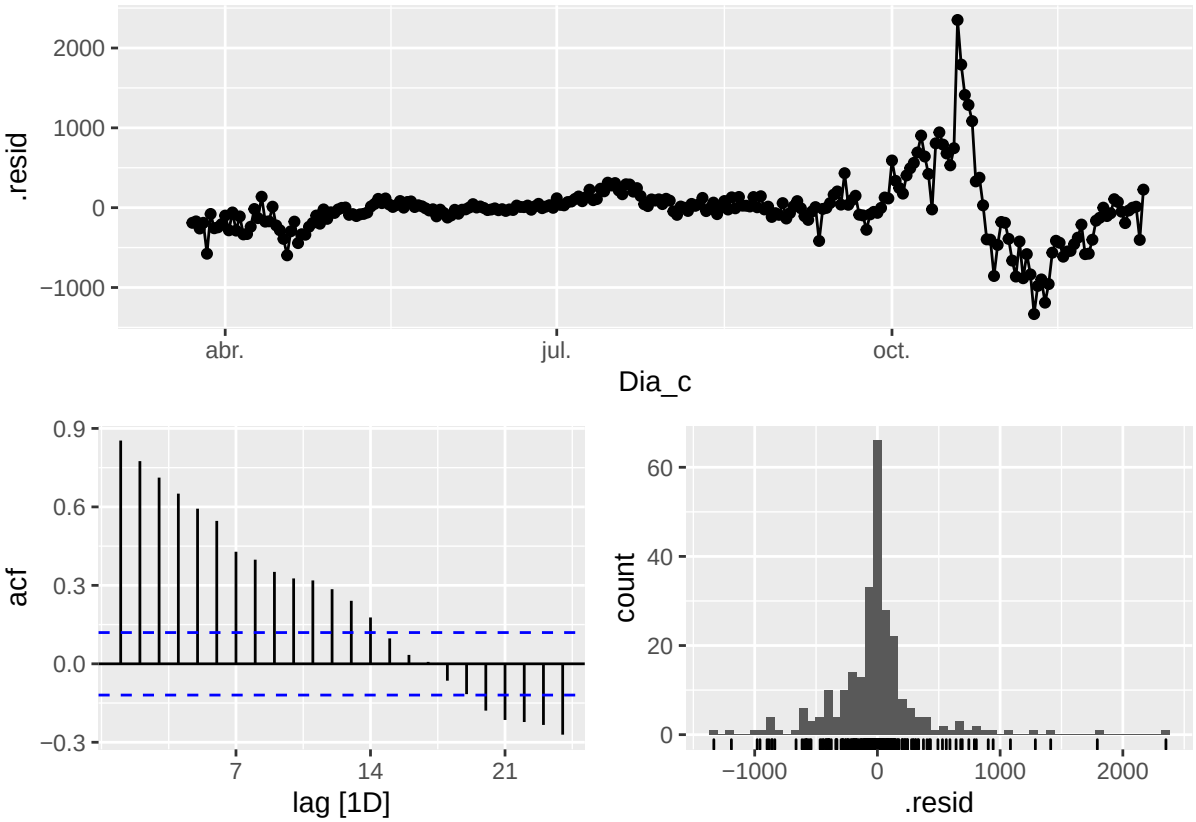
```
## # A tibble: 4 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>       <dbl>     <dbl>
## 1 Barcelona    arima_at1    35.1   0.00140
```

```
## 2 Barcelona    arima_at2    24.2   0.0438
## 3 Barcelona    arima_man    21.4   0.0907
## 4 Barcelona    SNaive     1009.    0
```

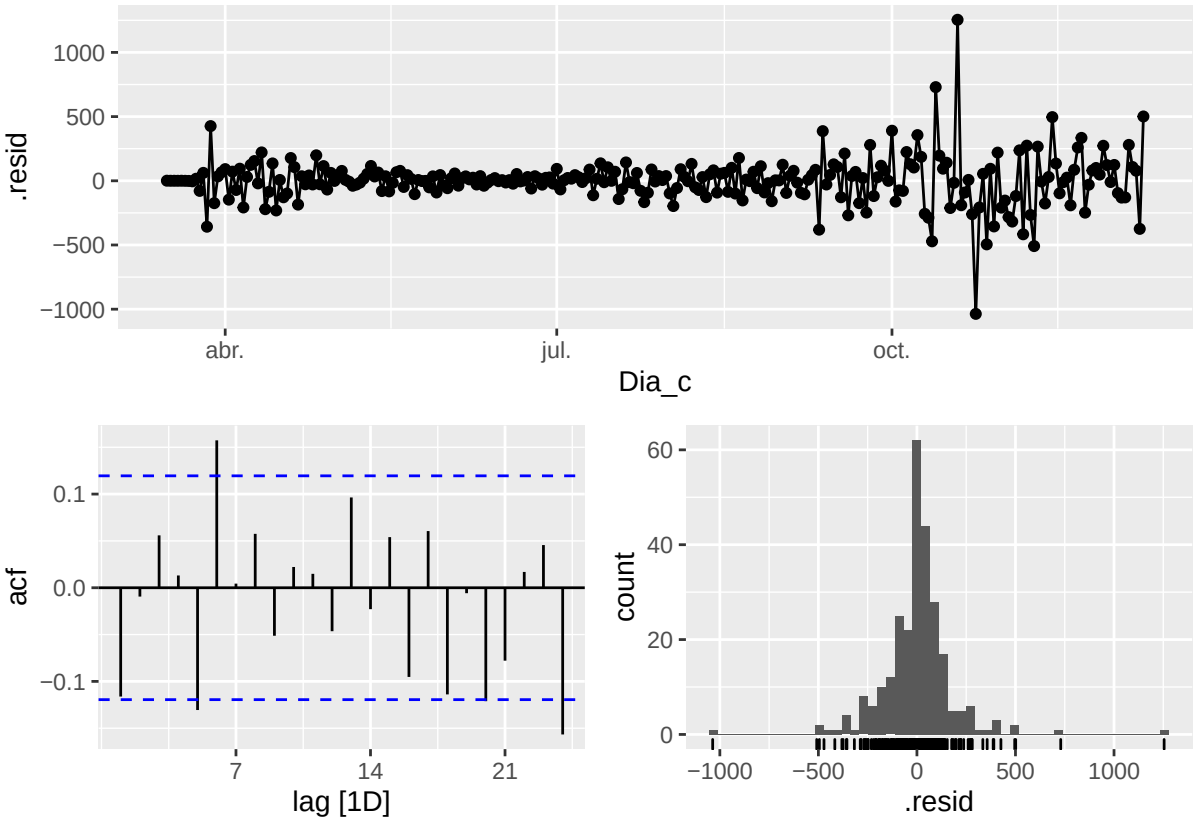
```
augment(fit_model) %>%
  features(.innov, ljung_box, lag=21)
```

```
## # A tibble: 4 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>       <dbl>     <dbl>
## 1 Barcelona    arima_at1    54.0 0.0000974
## 2 Barcelona    arima_at2    38.9 0.0100
## 3 Barcelona    arima_man    35.8 0.0229
## 4 Barcelona    SNaive     1039.  0
```

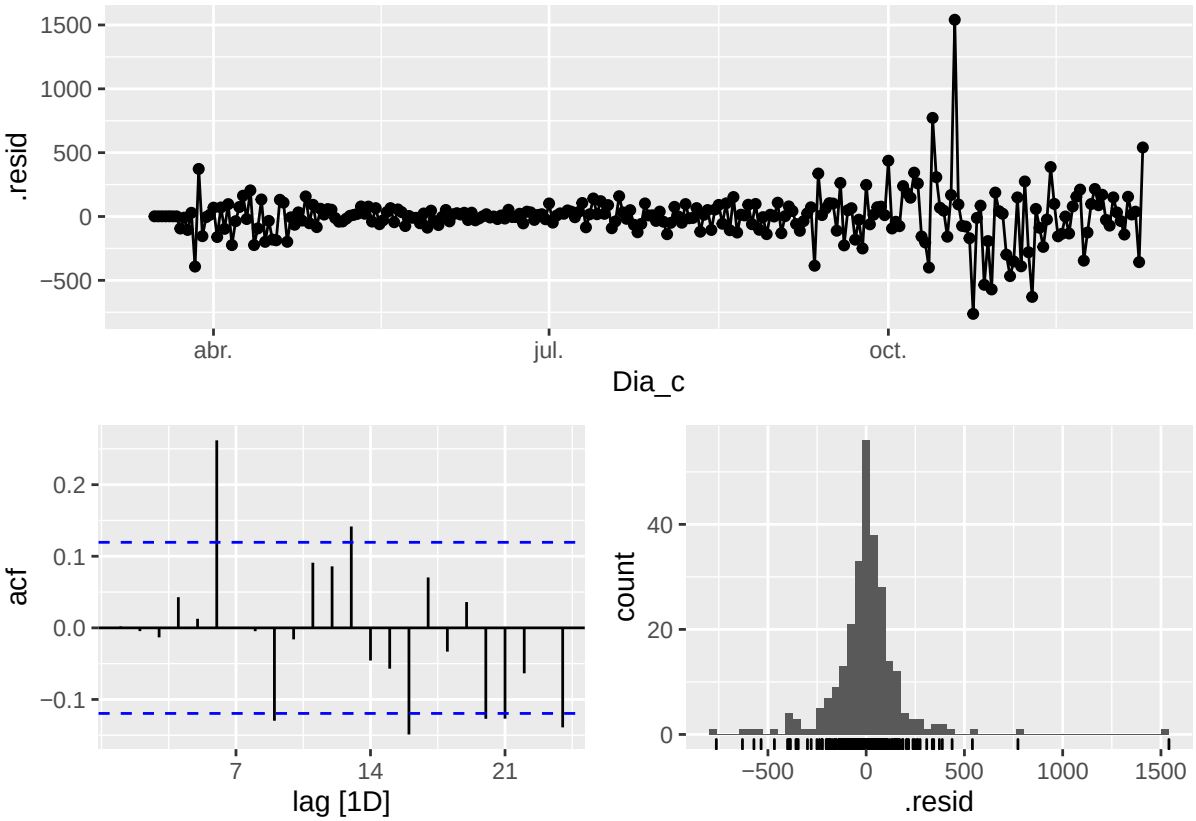
```
fit_model %>% select(SNaive) %>% gg_tsresiduals()
```



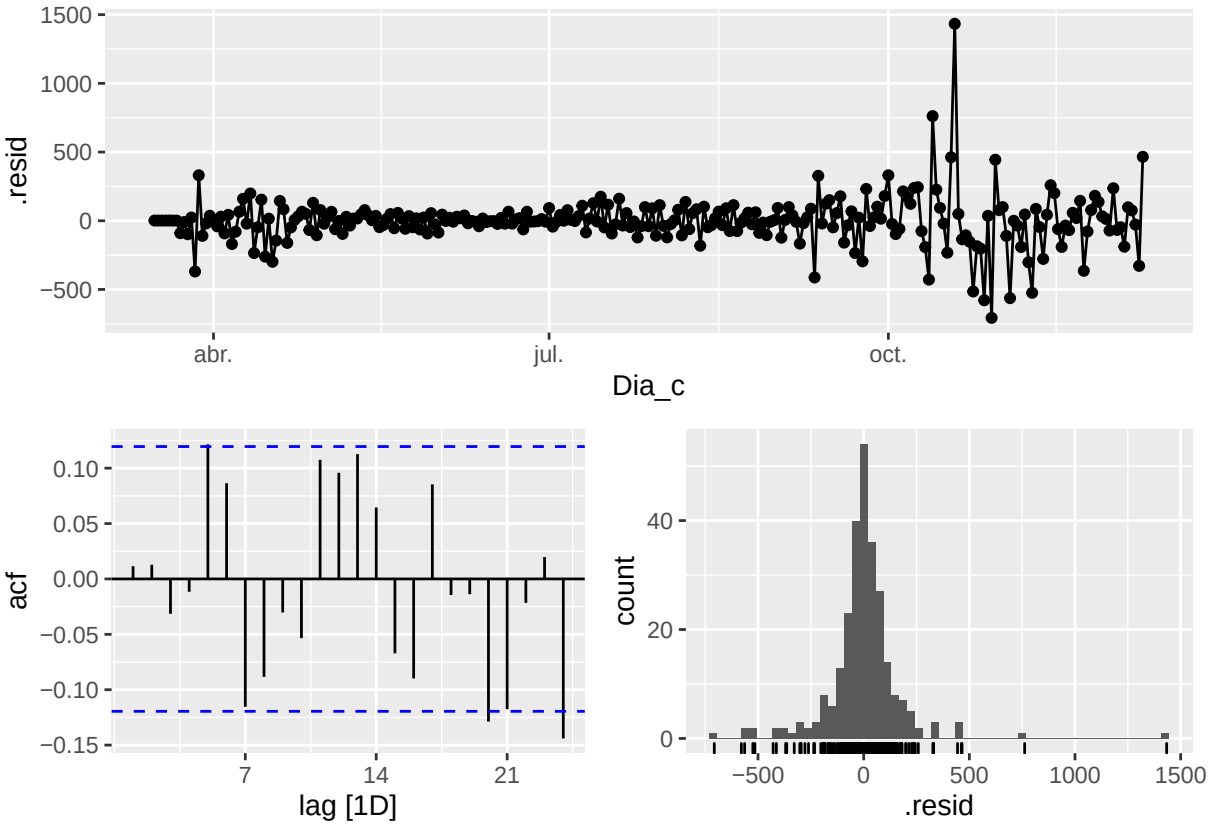

```
fit_model %>% select(arima_man) %>% gg_tsresiduals()
```




```
fit_model %>% select(arima_at1) %>% gg_tsresiduals()
```




```
fit_model %>% select(arima_at2) %>% gg_tsresiduals()
```




```
# Significant spikes (at lag 6, 15, etc.) out of 30 is still consistent with white noise.
# To be sure, use a Ljung-Box test, which has a large p-value, confirming that
# the residuals are similar to white noise.
# Note that the alternative models also pass this test.

# Forecast
fc_h7<-fabletools::forecast(fit_model, h=7)
fc_h14<-fabletools::forecast(fit_model, h=14)
fc_h21<-fabletools::forecast(fit_model, h=21)

# Accuracy
fabletools::accuracy(fc_h7, Bar_N_cases_tt)
```

```
## # A tibble: 4 x 11
##   .model    sub_region_2 .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE    ACF1
##   <chr>     <chr>        <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl>
## 1 arima_at1 Barcelona    Test   285.  339.  285.  23.5  23.5   NaN   NaN  0.107
## 2 arima_at2 Barcelona    Test   333.  390.  333.  26.9  26.9   NaN   NaN  0.0700
## 3 arima_man Barcelona    Test   204.  277.  204.  17.3  17.3   NaN   NaN  0.155
## 4 SNaive    Barcelona    Test   413   467.  413   35.1  35.1   NaN   NaN -0.138
```

```
fabletools::accuracy(fc_h14, Bar_N_cases_tt)
```

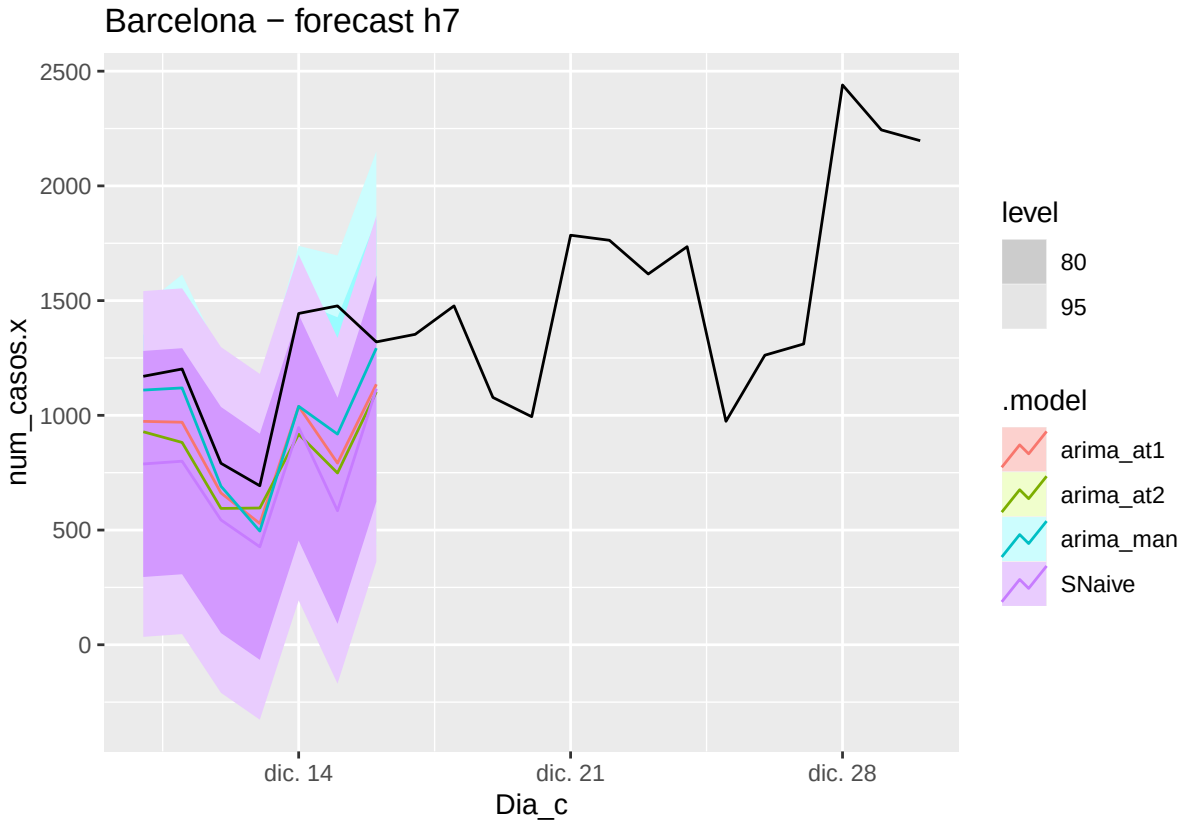
```
## # A tibble: 4 x 11
##   .model    sub_region_2 .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE  ACF1
##   <chr>     <chr>        <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Barcelona    Test   411.  472.  411.  30.2  30.2   NaN   NaN 0.376
```

```
## 2 arima_at2 Barcelona    Test   449.  512.  449.  32.7  32.7   NaN   NaN 0.321
## 3 arima_man Barcelona    Test   293.  367.  293.  22.1  22.1   NaN   NaN 0.355
## 4 SNaive    Barcelona    Test   554.  612.  554.  41.8  41.8   NaN   NaN 0.189
```

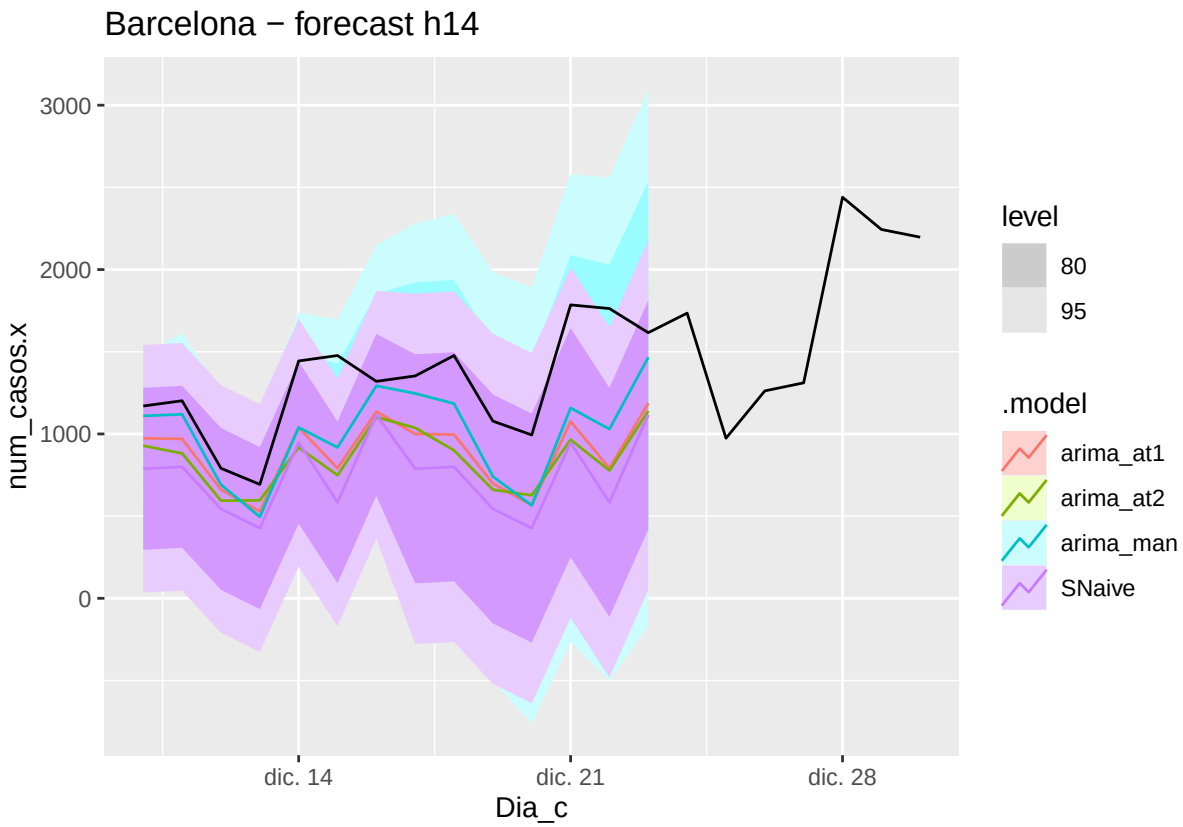
```
fabletools::accuracy(fc_h21, Bar_N_cases_tt)
```

```
## # A tibble: 4 x 11
##   .model    sub_region_2 .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE  ACF1
##   <chr>     <chr>        <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Barcelona    Test   540.  659.  546.  34.0  34.6   NaN   NaN 0.531
## 2 arima_at2 Barcelona    Test   580.  697.  580.  36.6  36.6   NaN   NaN 0.526
## 3 arima_man Barcelona    Test   379.  507.  409.  23.8  26.9   NaN   NaN 0.479
## 4 SNaive    Barcelona    Test   700.  806.  700.  46.0  46.0   NaN   NaN 0.455
```

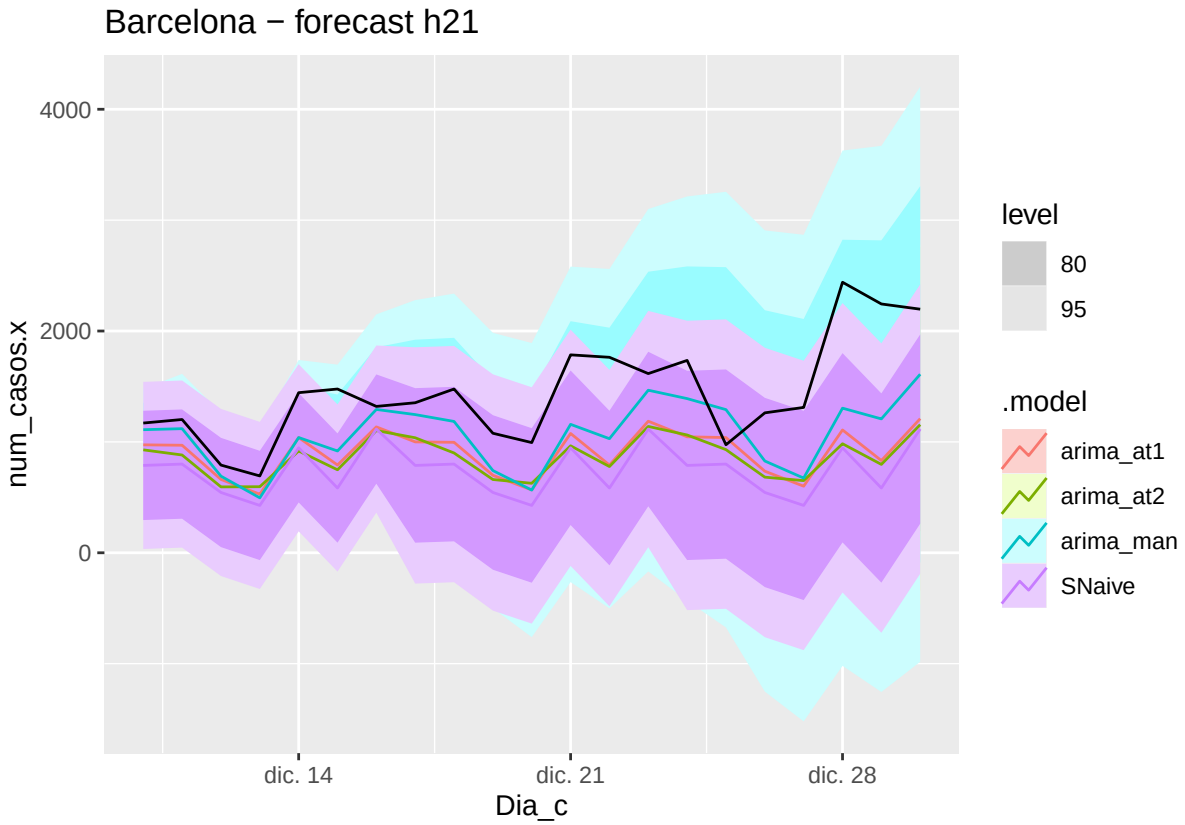
```
# Plots
fc_h7 %>%
  #filter(.model=='arima_at1'|.model=='arima_at2'|.model=='SNaive') %>%
  autoplot(Bar_N_cases_tt) +
  labs(title="Barcelona - forecast h7")
```




```
fc_h14 %>%
  #filter(.model=='arima_at1'|.model=='arima_at2'|.model=='SNaive') %>%
  autoplot(Bar_N_cases_tt) +
  labs(title="Barcelona - forecast h14")
```




```
fc_h21 %>%
  #filter(.model=='arima_at1'|.model=='arima_at2'|.model=='SNaive') %>%
  autoplot(Bar_N_cases_tt) +
  labs(title="Barcelona - forecast h21")
```



3.3.2 Multivariate (7, 14, 21 days) Barcelona

```
# Model train
# We have added mobility variables to models
fit_model <- Bar_N_cases_tr %>%
  model(
    SNaive = SNAIVE(num_casos.x),
    arima_man = ARIMA(num_casos.x ~ retail_and_recreation_percent_change_from_baseline +
                        grocery_and_pharmacy_percent_change_from_baseline +
                        parks_percent_change_from_baseline +
                        transit_stations_percent_change_from_baseline +
                        workplaces_percent_change_from_baseline +
                        residential_percent_change_from_baseline + Total + pdq(2,1,2) +
                        PDQ(1,1,1)),
    arima_at1 = ARIMA(num_casos.x~ retail_and_recreation_percent_change_from_baseline +
                        grocery_and_pharmacy_percent_change_from_baseline +
                        parks_percent_change_from_baseline +
                        transit_stations_percent_change_from_baseline +
                        workplaces_percent_change_from_baseline +
                        residential_percent_change_from_baseline + Total),
    arima_at2 = ARIMA(num_casos.x~ retail_and_recreation_percent_change_from_baseline +
                        grocery_and_pharmacy_percent_change_from_baseline +
                        parks_percent_change_from_baseline +
                        transit_stations_percent_change_from_baseline +
                        workplaces_percent_change_from_baseline +
```

```
                        residential_percent_change_from_baseline + Total,
                      stepwise = FALSE,approx = FALSE))

# Show and report model
fit_model
```

```
## # A mable: 1 x 5
## # Key:     sub_region_2 [1]
##   sub_region_2   SNaive                                arima_man
##   <chr>         <model>                                  <model>
## 1 Barcelona    <SNAIVE> <LM w/ ARIMA(2,1,2)(1,1,1)[7] errors>
## # ... with 2 more variables: arima_at1 <model>, arima_at2 <model>
```

```
report(fit_model)
```

```
## # A tibble: 4 x 9
##   sub_region_2 .model     sigma2 log_lik   AIC  AICc   BIC ar_roots   ma_roots
##   <chr>        <chr>       <dbl>   <dbl> <dbl> <dbl> <dbl> <list>     <list>
## 1 Barcelona    SNaive    148096.      NA    NA    NA    NA  <NULL>     <NULL>
## 2 Barcelona    arima_man  24852.  -1688. 3405. 3407. 3455. <cpl [9]>  <cpl [9]>
## 3 Barcelona    arima_at1  27061.  -1756. 3535. 3536. 3574. <cpl [8]>  <cpl [1]>
## 4 Barcelona    arima_at2  26350.  -1752. 3527. 3528. 3570. <cpl [15]> <cpl [0]>
```

```
# Good model >> Less Sigma - More BIC or AIC
fit_model %>% pivot_longer(!sub_region_2,
                           names_to = "Model name",
                           values_to = "Orders")
```

```
## # A mable: 4 x 3
## # Key:     sub_region_2, Model name [4]
##   sub_region_2 `Model name`                                  Orders
##   <chr>        <chr>                                        <model>
## 1 Barcelona    SNaive                                      <SNAIVE>
## 2 Barcelona    arima_man    <LM w/ ARIMA(2,1,2)(1,1,1)[7] errors>
## 3 Barcelona    arima_at1    <LM w/ ARIMA(1,0,1)(1,0,0)[7] errors>
## 4 Barcelona    arima_at2    <LM w/ ARIMA(1,0,0)(2,0,0)[7] errors>
```

```
glance(fit_model) %>% arrange(AICc) %>% select(.model:BIC)
```

```
## # A tibble: 4 x 6
##   .model     sigma2 log_lik   AIC  AICc   BIC
##   <chr>       <dbl>   <dbl> <dbl> <dbl> <dbl>
## 1 arima_man  24852.  -1688. 3405. 3407. 3455.
## 2 arima_at2  26350.  -1752. 3527. 3528. 3570.
## 3 arima_at1  27061.  -1756. 3535. 3536. 3574.
## 4 SNaive    148096.     NA    NA    NA    NA
```

```
# We use a Ljung-Box test >> large p-value, confirmins residuals are similar to white noise.
augment(fit_model) %>%
  features(.innov, ljung_box, lag=7)
```

```
## # A tibble: 4 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>       <dbl>     <dbl>
## 1 Barcelona    arima_at1   17.3     0.0158
## 2 Barcelona    arima_at2   16.2     0.0232
## 3 Barcelona    arima_man    9.22    0.237
```

```
## 4 Barcelona    SNaive     827.      0
```

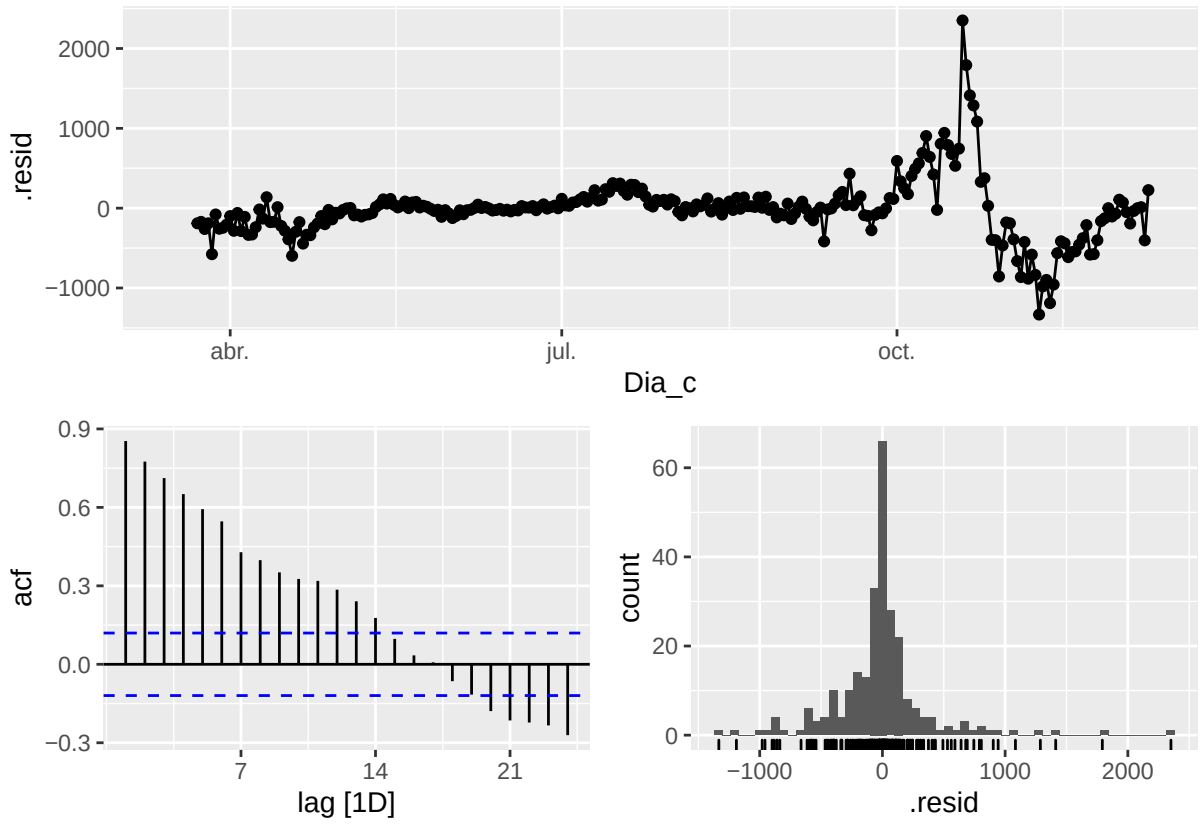
```
augment(fit_model) %>%
  features(.innov, ljung_box, lag=14)
```

```
## # A tibble: 4 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>       <dbl>     <dbl>
## 1 Barcelona    arima_at1    40.4  0.000221
## 2 Barcelona    arima_at2    39.7  0.000284
## 3 Barcelona    arima_man    32.0  0.00394
## 4 Barcelona    SNaive     1009.   0
```

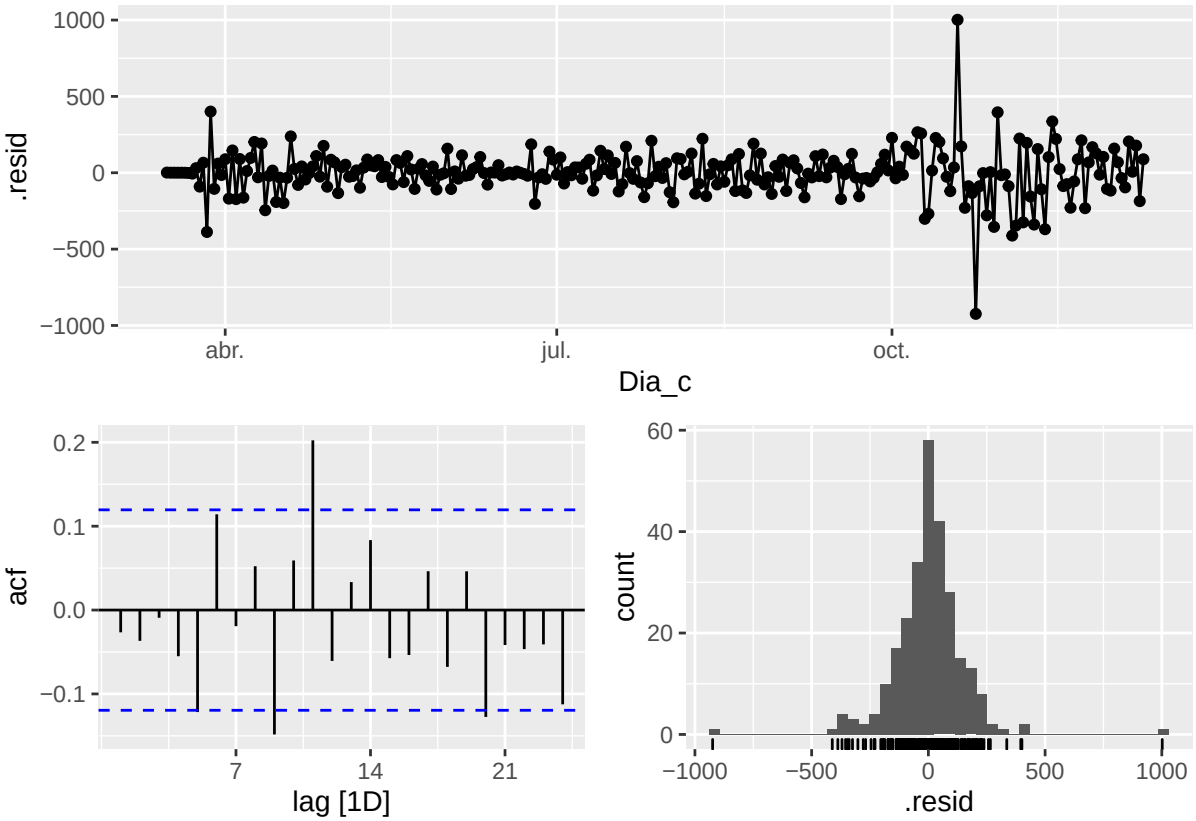
```
augment(fit_model) %>%
  features(.innov, ljung_box, lag=21)
```

```
## # A tibble: 4 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>       <dbl>     <dbl>
## 1 Barcelona    arima_at1    52.0  0.000191
## 2 Barcelona    arima_at2    49.9  0.000372
## 3 Barcelona    arima_man    41.7  0.00464
## 4 Barcelona    SNaive     1039.   0
```

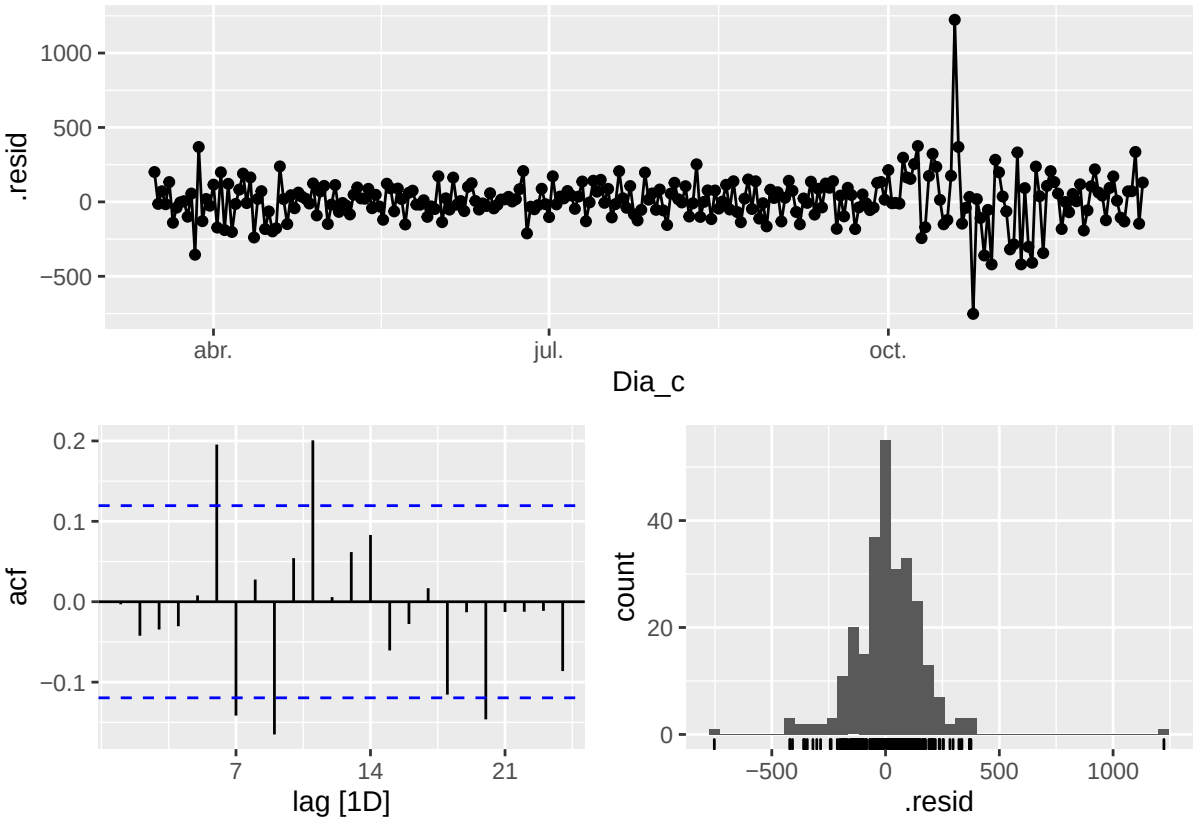
```
fit_model %>% select(SNaive) %>% gg_tsresiduals()
```



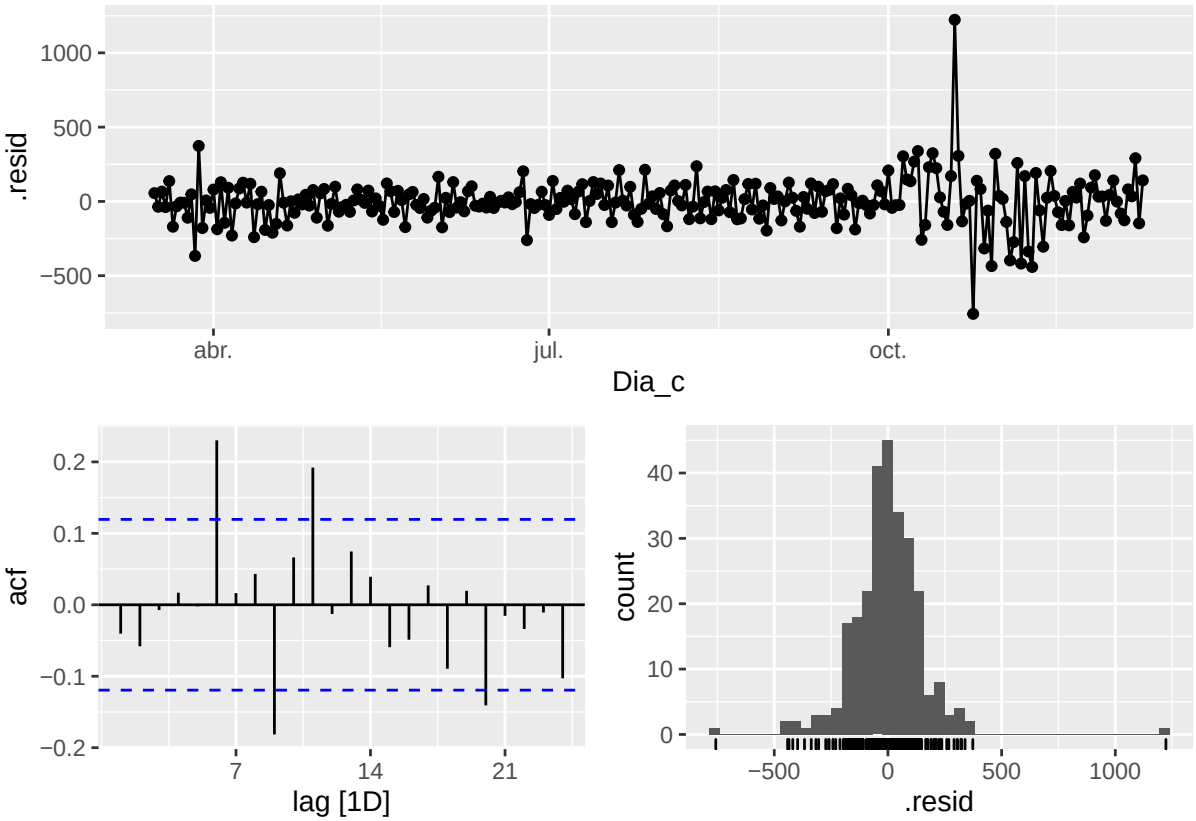

```
fit_model %>% select(arima_man) %>% gg_tsresiduals()
```




```
fit_model %>% select(arima_at1) %>% gg_tsresiduals()
```




```
fit_model %>% select(arima_at2) %>% gg_tsresiduals()
```




```
# Significant spikes out of 30 is still consistent with white noise.
# To be sure, use a Ljung-Box test, which has a large p-value, confirming that the
# residuals are similar to white noise.
# Note that the alternative models also pass this test.

# New data (dynamic regression)
# Here it is needed generate future values for the exogenous variables
# Foir simplicity we select a rand number included into the 2nd and 3rd quantile for
# the variable
# h7
Bar_N_cases_fr7 <- new_data(Bar_N_cases_tr, 7) %>%
  mutate(retail_and_recreation_percent_change_from_baseline =
           runif(7,quantile(Bar_N_cases_tt$retail_and_recreation_percent_change_from_baseline,
                            0.25),
                 quantile(Bar_N_cases_tt$retail_and_recreation_percent_change_from_baseline,
                          0.75)),
         grocery_and_pharmacy_percent_change_from_baseline =
           runif(7,quantile(Bar_N_cases_tt$grocery_and_pharmacy_percent_change_from_baseline,
                            0.25),
                 quantile(Bar_N_cases_tt$grocery_and_pharmacy_percent_change_from_baseline,
                          0.75)),
         parks_percent_change_from_baseline =
           runif(7,quantile(Bar_N_cases_tr$parks_percent_change_from_baseline,
                            0.25),
                 quantile(Bar_N_cases_tt$parks_percent_change_from_baseline,
                          0.75)),
```

```
          transit_stations_percent_change_from_baseline =
            runif(7,quantile(Bar_N_cases_tt$transit_stations_percent_change_from_baseline,
                             0.25),
                    quantile(Bar_N_cases_tt$transit_stations_percent_change_from_baseline,
                             0.75)),
          workplaces_percent_change_from_baseline =
            runif(7,quantile(Bar_N_cases_tt$workplaces_percent_change_from_baseline,
                             0.25),
                    quantile(Bar_N_cases_tt$workplaces_percent_change_from_baseline,
                             0.75)),
          residential_percent_change_from_baseline =
            runif(7,quantile(Bar_N_cases_tt$residential_percent_change_from_baseline,
                             0.25),
                    quantile(Bar_N_cases_tt$residential_percent_change_from_baseline,
                             0.75)),
          Total = runif(7,quantile(Bar_N_cases_tt$Total,0.25),
                        quantile(Bar_N_cases_tt$Total,0.75)))

# h14
Bar_N_cases_fr14 <- new_data(Bar_N_cases_tr, 14) %>%
  mutate(retail_and_recreation_percent_change_from_baseline =
            runif(14,quantile(Bar_N_cases_tt$retail_and_recreation_percent_change_from_baseline,
                              0.25),
                    quantile(Bar_N_cases_tt$retail_and_recreation_percent_change_from_baseline,
                             0.75)),
          grocery_and_pharmacy_percent_change_from_baseline =
            runif(14,quantile(Bar_N_cases_tt$grocery_and_pharmacy_percent_change_from_baseline,
                              0.25),
                    quantile(Bar_N_cases_tt$grocery_and_pharmacy_percent_change_from_baseline,
                             0.75)),
          parks_percent_change_from_baseline =
            runif(14,quantile(Bar_N_cases_tt$parks_percent_change_from_baseline,
                              0.25),
                    quantile(Bar_N_cases_tt$parks_percent_change_from_baseline,
                             0.75)),
          transit_stations_percent_change_from_baseline =
            runif(14,quantile(Bar_N_cases_tt$transit_stations_percent_change_from_baseline,
                              0.25),
                    quantile(Bar_N_cases_tt$transit_stations_percent_change_from_baseline,
                             0.75)),
          workplaces_percent_change_from_baseline =
            runif(14,quantile(Bar_N_cases_tt$workplaces_percent_change_from_baseline,
                              0.25),
                    quantile(Bar_N_cases_tt$workplaces_percent_change_from_baseline,
                             0.75)),
          residential_percent_change_from_baseline =
            runif(14,quantile(Bar_N_cases_tt$residential_percent_change_from_baseline,
                              0.25),
                    quantile(Bar_N_cases_tt$residential_percent_change_from_baseline,
                             0.75)),
          Total = runif(14,quantile(Bar_N_cases_tt$Total,0.25),
                        quantile(Bar_N_cases_tt$Total,0.75)))
```

```
# h21
Bar_N_cases_fr21 <- new_data(Bar_N_cases_tr, 21) %>%
  mutate(retail_and_recreation_percent_change_from_baseline =
           runif(21,quantile(Bar_N_cases_tt$retail_and_recreation_percent_change_from_baseline,
                             0.25),
                 quantile(Bar_N_cases_tt$retail_and_recreation_percent_change_from_baseline,
                          0.75)),
         grocery_and_pharmacy_percent_change_from_baseline =
           runif(21,quantile(Bar_N_cases_tt$grocery_and_pharmacy_percent_change_from_baseline,
                             0.25),
                 quantile(Bar_N_cases_tt$grocery_and_pharmacy_percent_change_from_baseline,
                          0.75)),
         parks_percent_change_from_baseline =
           runif(21,quantile(Bar_N_cases_tt$parks_percent_change_from_baseline,
                             0.25),
                 quantile(Bar_N_cases_tt$parks_percent_change_from_baseline,
                          0.75)),
         transit_stations_percent_change_from_baseline =
           runif(21,quantile(Bar_N_cases_tt$transit_stations_percent_change_from_baseline,
                             0.25),
                 quantile(Bar_N_cases_tt$transit_stations_percent_change_from_baseline,
                          0.75)),
         workplaces_percent_change_from_baseline =
           runif(21,quantile(Bar_N_cases_tt$workplaces_percent_change_from_baseline,
                             0.25),
                 quantile(Bar_N_cases_tt$workplaces_percent_change_from_baseline,
                          0.75)),
         residential_percent_change_from_baseline =
           runif(21,quantile(Bar_N_cases_tt$residential_percent_change_from_baseline,
                             0.25),
                 quantile(Bar_N_cases_tt$residential_percent_change_from_baseline,
                          0.75)),
         Total = runif(21,quantile(Bar_N_cases_tt$Total,0.25),
                       quantile(Bar_N_cases_tt$Total,0.75)))

# Forecast
fc_fh7<-fabletools::forecast(fit_model, new_data = Bar_N_cases_fr7)
fc_fh14<-fabletools::forecast(fit_model, new_data = Bar_N_cases_fr14)
fc_fh21<-fabletools::forecast(fit_model, new_data = Bar_N_cases_fr21)

# Accuracy
fabletools::accuracy(fc_fh7, Bar_N_cases)
```

```
## # A tibble: 4 x 11
##   .model    sub_region_2 .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE    ACF1
##   <chr>     <chr>        <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl>
## 1 arima_at1 Barcelona    Test   242.  272.  242.  20.6  20.6 1.07  0.708  0.0321
## 2 arima_at2 Barcelona    Test   191.  233.  191.  15.5  15.5 0.848 0.605  0.222
## 3 arima_man Barcelona    Test   252.  287.  252.  25.5  25.5 1.12  0.747  0.0519
## 4 SNaive    Barcelona    Test   413   467.  413   35.1  35.1 1.83  1.21  -0.138
```

```
fabletools::accuracy(fc_fh14, Bar_N_cases)
```

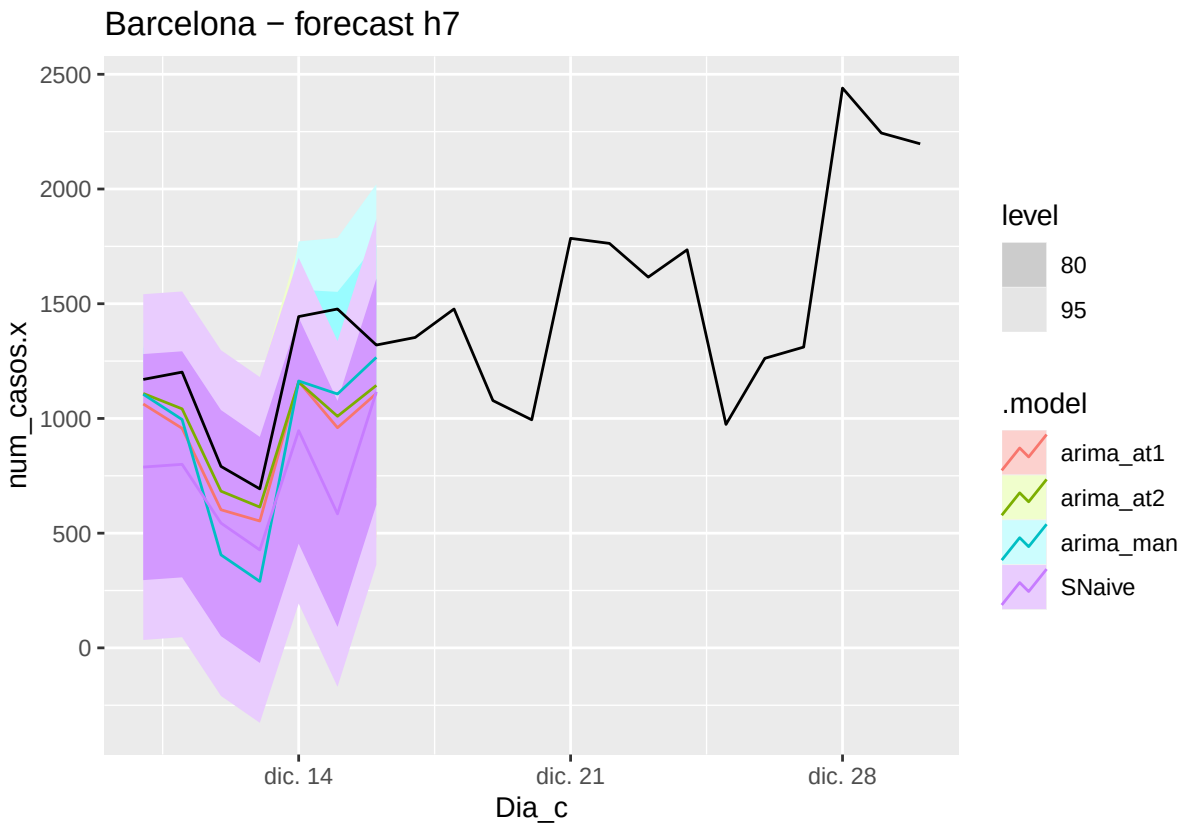
```
## # A tibble: 4 x 11
```

```
##   .model    sub_region_2 .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE  ACF1
##   <chr>     <chr>        <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Barcelona    Test   391.  425.  391.  29.7  29.7  1.74 1.11  0.259
## 2 arima_at2 Barcelona    Test   313.  346.  313.  23.3  23.3  1.39 0.899 0.122
## 3 arima_man Barcelona    Test   412.  453.  412.  35.3  35.3  1.83 1.18  0.550
## 4 SNaive    Barcelona    Test   554.  612.  554.  41.8  41.8  2.46 1.59  0.189
```

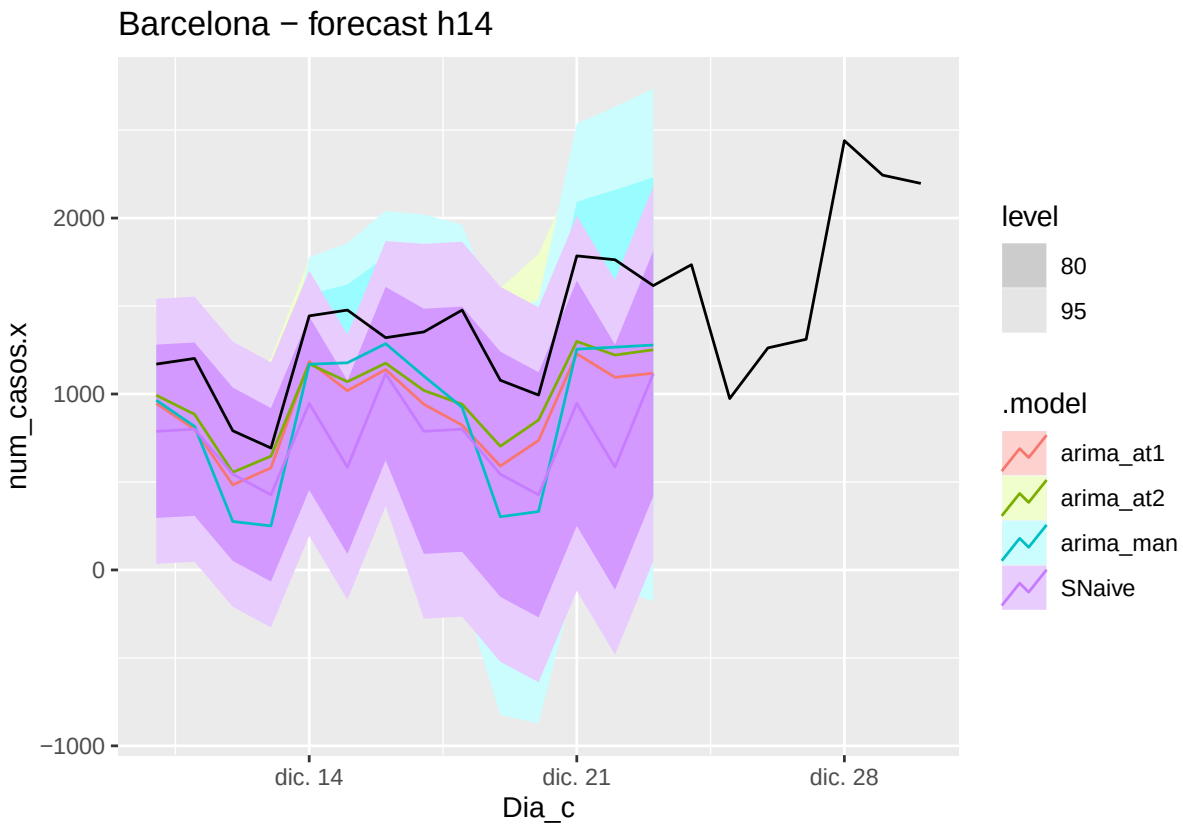
```
fabletools::accuracy(fc_fh21, Bar_N_cases)
```

```
## # A tibble: 4 x 11
##   .model    sub_region_2 .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE  ACF1
##   <chr>     <chr>        <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Barcelona    Test   558.  670.  558.  36.1  36.1  2.48  1.74 0.606
## 2 arima_at2 Barcelona    Test   439.  552.  455.  27.5  29.1  2.02  1.43 0.531
## 3 arima_man Barcelona    Test   557.  663.  575.  39.4  41.2  2.55  1.72 0.524
## 4 SNaive    Barcelona    Test   700.  806.  700.  46.0  46.0  3.11  2.10 0.455
```

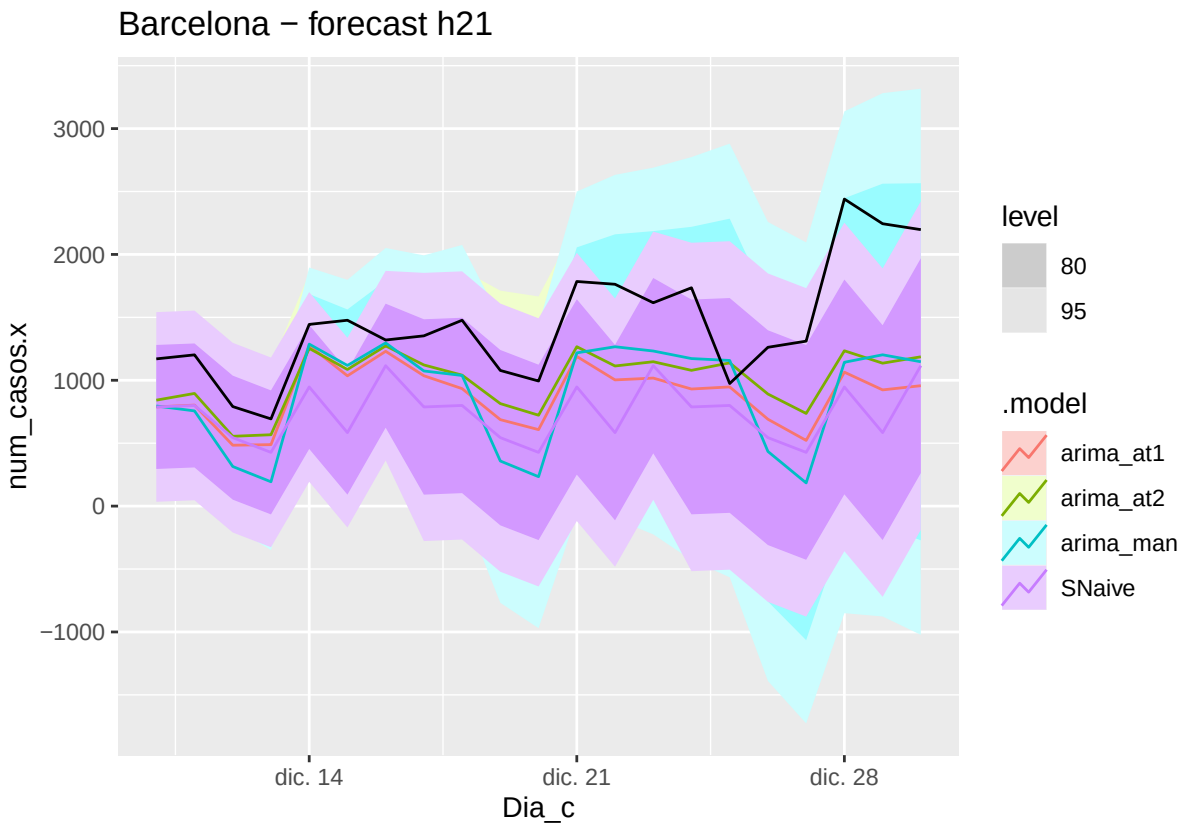
```
# Plots
fc_fh7 %>%
  #filter(.model=='arima_at1'|.model=='arima_at2'|.model=='SNaive') %>%
  autoplot(Bar_N_cases_tt) +
  labs(title="Barcelona - forecast h7")
```




```
fc_fh14 %>%
  #filter(.model=='arima_at1'|.model=='arima_at2'|.model=='SNaive') %>%
  autoplot(Bar_N_cases_tt) +
  labs(title="Barcelona - forecast h14")
```




```
fc_fh21 %>%
  #filter(.model=='arima_at1'|.model=='arima_at2'|.model=='SNaive') %>%
  autoplot(Bar_N_cases_tt) +
  labs(title="Barcelona - forecast h21")
```



3.3.3 Univariate (7, 14, 21 days) Madrid, Málaga, Córdoba and Cádiz

```
# Train and test ts
# Train
Mad_N_cases_tr <- Mad_N_cases %>%
  filter_index("2020-03-15" ~ "2020-12-9")
Mal_N_cases_tr <- Mal_N_cases %>%
  filter_index("2020-03-15" ~ "2020-12-9")
Cad_N_cases_tr <- Cad_N_cases %>%
  filter_index("2020-03-15" ~ "2020-12-9")
Sev_N_cases_tr <- Sev_N_cases %>%
  filter_index("2020-03-15" ~ "2020-12-9")

# Test
Mad_N_cases_tt <- Mad_N_cases %>%
  filter_index("2020-12-10" ~ "2020-12-31")
Mal_N_cases_tt <- Mal_N_cases %>%
  filter_index("2020-12-10" ~ "2020-12-31")
Cad_N_cases_tt <- Cad_N_cases %>%
  filter_index("2020-12-10" ~ "2020-12-31")
Sev_N_cases_tt <- Sev_N_cases %>%
  filter_index("2020-12-10" ~ "2020-12-31")
```

```
####### Madrid #######
# Model train
Mad_fit_model <- Mad_N_cases_tr %>%
```

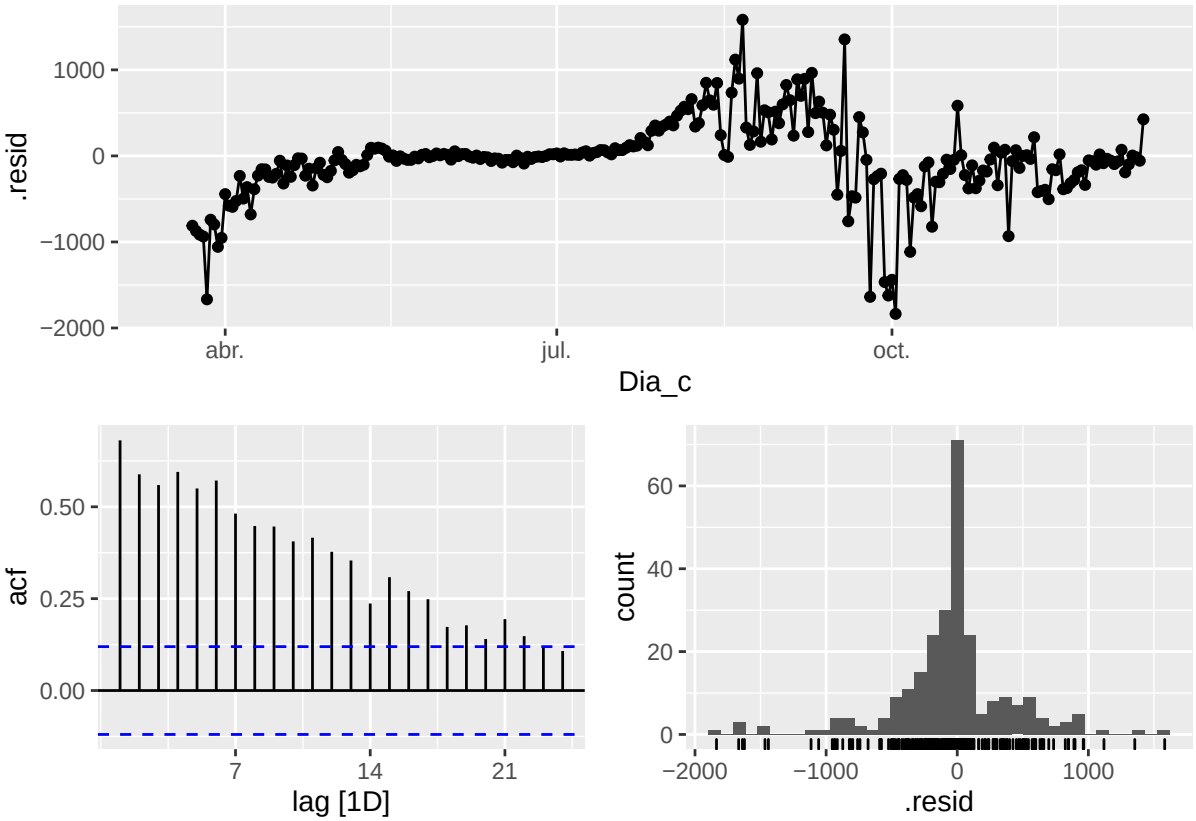
```
  model(
    SNaive = SNAIVE(num_casos.x),
    arima_man = ARIMA(num_casos.x ~ pdq(2,1,2) + PDQ(1,1,1)),
    arima_at1 = ARIMA(num_casos.x),
    arima_at2 = ARIMA(num_casos.x, stepwise = FALSE,approx = FALSE))

Mad_fit_model %>% pivot_longer(!sub_region_2,
                           names_to = "Model name",
                           values_to = "Orders")
```

```
## # A mable: 4 x 3
## # Key:     sub_region_2, Model name [4]
##   sub_region_2 `Model name`                    Orders
##   <chr>        <chr>                          <model>
## 1 Madrid       SNaive                        <SNAIVE>
## 2 Madrid       arima_man    <ARIMA(2,1,2)(1,1,1)[7]>
## 3 Madrid       arima_at1    <ARIMA(2,1,1)(1,1,2)[7]>
## 4 Madrid       arima_at2    <ARIMA(2,1,2)(2,1,0)[7]>
```

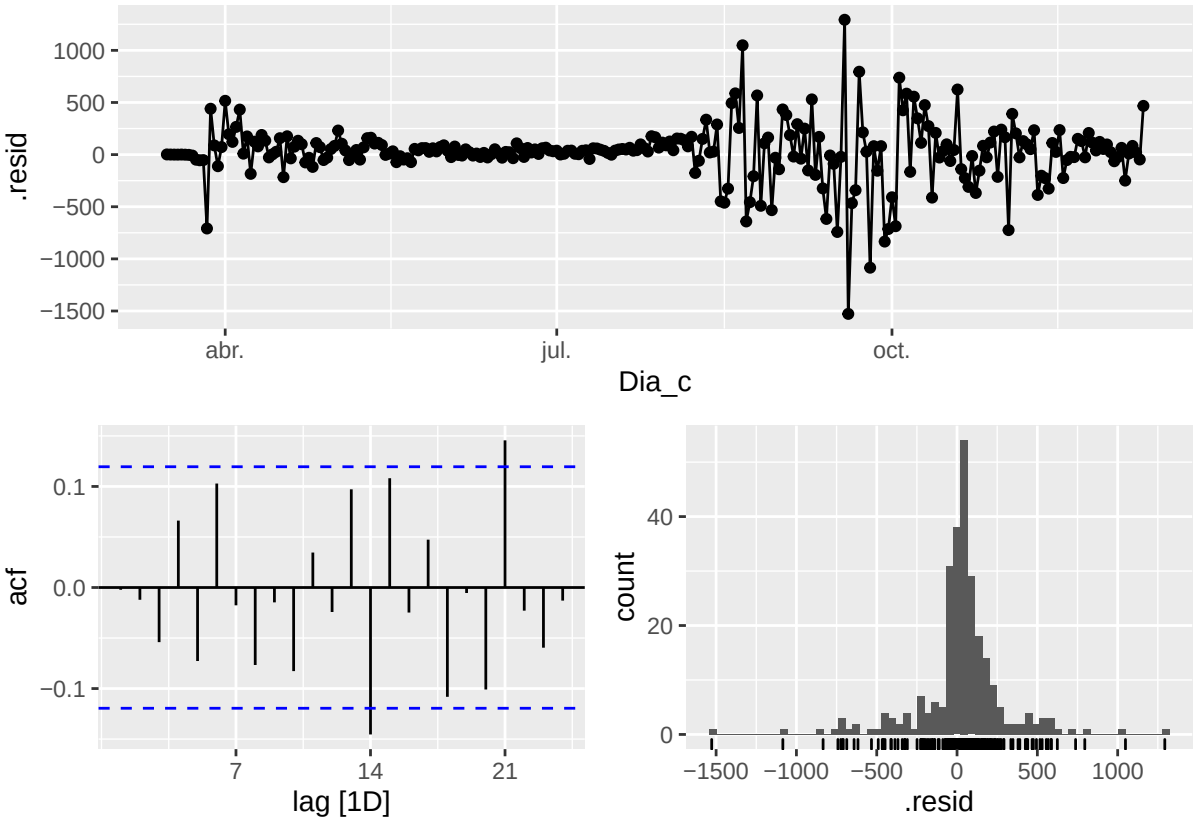
```
# Good model >> Less Sigma / More BIC or AIC
glance(Mad_fit_model) %>% arrange(AICc) %>% select(.model:BIC)
```

```
## # A tibble: 4 x 6
##   .model     sigma2 log_lik   AIC  AICc   BIC
##   <chr>       <dbl>   <dbl> <dbl> <dbl> <dbl>
## 1 arima_at2  84635.  -1849. 3712. 3712. 3737.
## 2 arima_man  82093.  -1850. 3715. 3715. 3740.
## 3 arima_at1  87043.  -1853. 3719. 3719. 3744.
## 4 SNaive    205618.     NA    NA    NA    NA
```

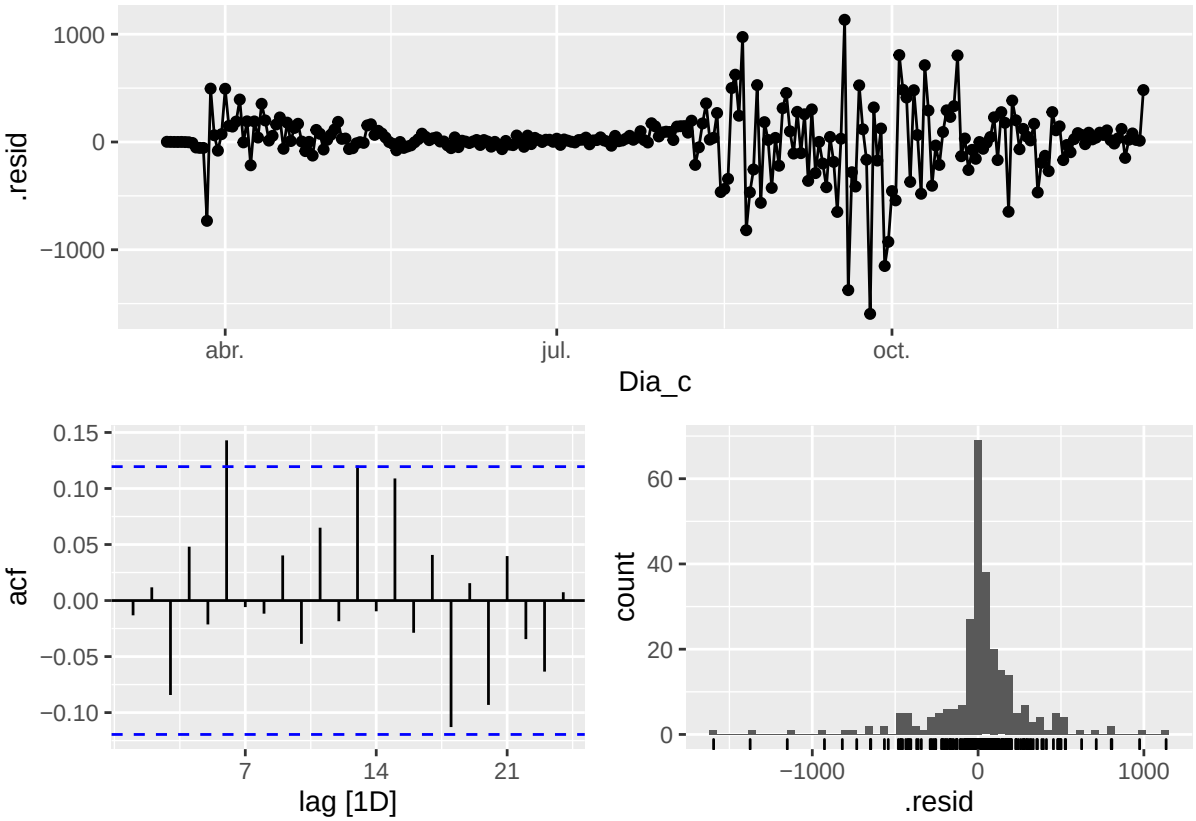
```
Mad_fit_model %>% select(SNaive) %>% gg_tsresiduals()
```



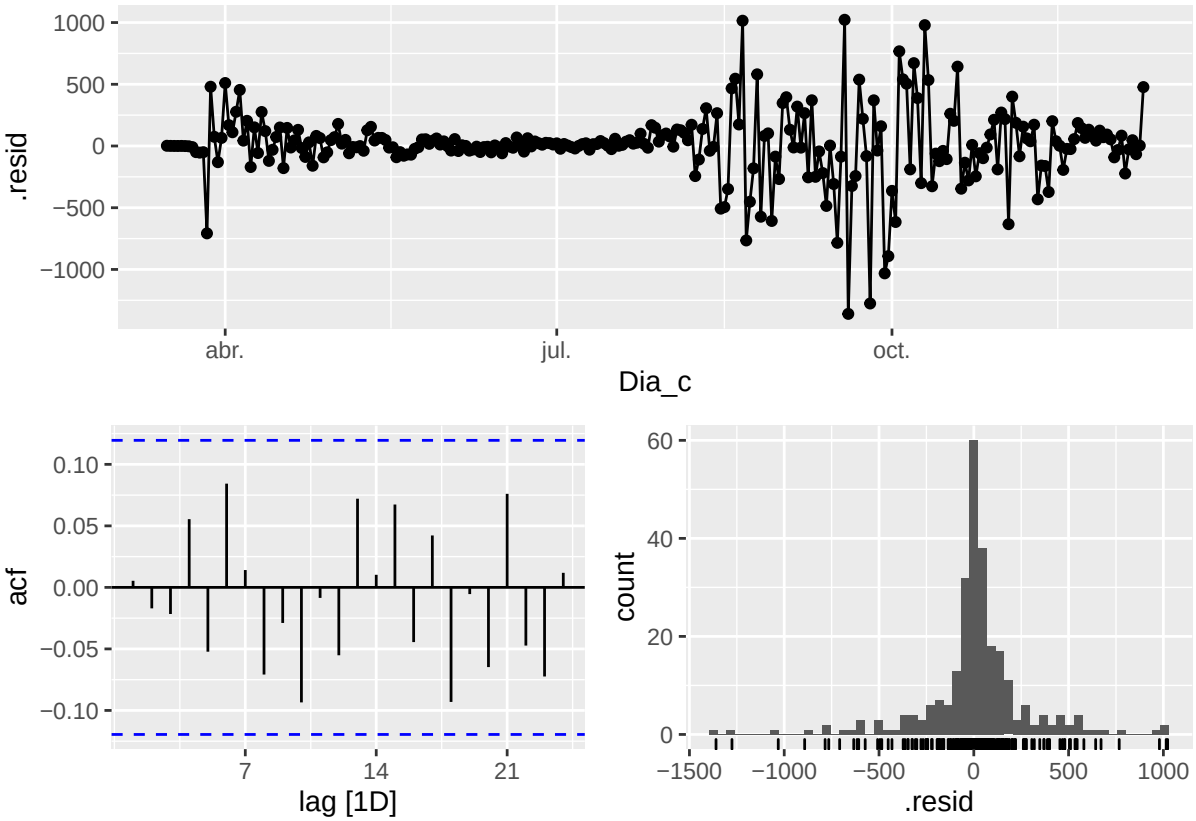

```
Mad_fit_model %>% select(arima_man) %>% gg_tsresiduals()
```




```
Mad_fit_model %>% select(arima_at1) %>% gg_tsresiduals()
```




```
Mad_fit_model %>% select(arima_at2) %>% gg_tsresiduals()
```




```
augment(Mad_fit_model) %>%
  features(.innov, ljung_box, lag=7)
```

```
## # A tibble: 4 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>       <dbl>     <dbl>
## 1 Madrid       arima_at1    8.47     0.293
## 2 Madrid       arima_at2    3.84     0.798
## 3 Madrid       arima_man    6.53     0.480
## 4 Madrid       SNaive     626.       0
```

```
augment(Mad_fit_model) %>%
  features(.innov, ljung_box, lag=14)
```

```
## # A tibble: 4 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>       <dbl>     <dbl>
## 1 Madrid       arima_at1    14.8     0.390
## 2 Madrid       arima_at2    10.3     0.738
## 3 Madrid       arima_man    19.4     0.150
## 4 Madrid       SNaive      918.      0
```

```
augment(Mad_fit_model) %>%
  features(.innov, ljung_box, lag=21)
```

```
## # A tibble: 4 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>       <dbl>     <dbl>
```

```
## 1 Madrid       arima_at1    25.7    0.217
## 2 Madrid       arima_at2    18.2    0.638
## 3 Madrid       arima_man    36.2    0.0208
## 4 Madrid       SNaive     1016.     0
```

```
# Forecast
Mad_fc_h7<-fabletools::forecast(Mad_fit_model, h=7)
Mad_fc_h14<-fabletools::forecast(Mad_fit_model, h=14)
Mad_fc_h21<-fabletools::forecast(Mad_fit_model, h=21)
# Accuracy
fabletools::accuracy(Mad_fc_h7, Mad_N_cases_tt)
```

```
## # A tibble: 4 x 11
##   .model    sub_region_2 .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE   ACF1
##   <chr>     <chr>        <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>
## 1 arima_at1 Madrid       Test   605.  641.  605.  34.0  34.0   NaN   NaN -0.290
## 2 arima_at2 Madrid       Test   621.  653.  621.  35.0  35.0   NaN   NaN -0.299
## 3 arima_man Madrid       Test   620.  647.  620.  35.2  35.2   NaN   NaN -0.287
## 4 SNaive    Madrid       Test   684.  716.  684.  38.6  38.6   NaN   NaN -0.239
```

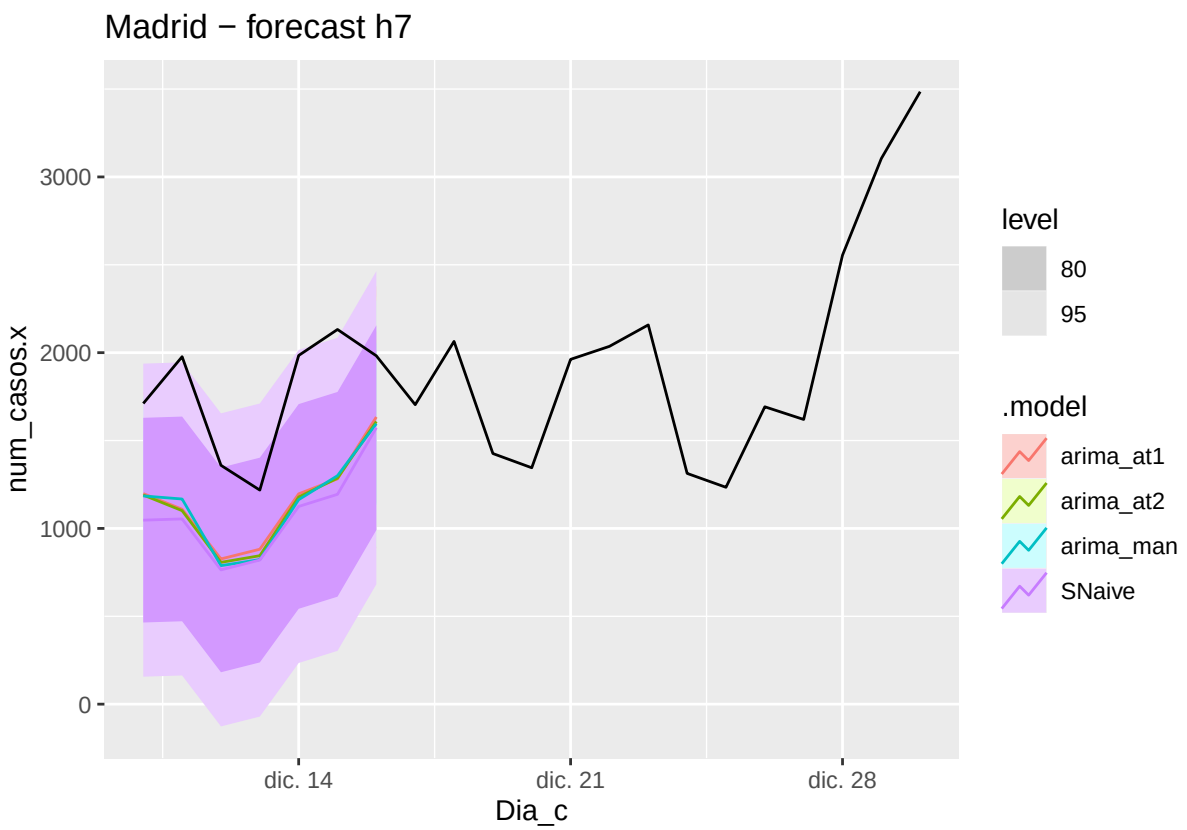
```
fabletools::accuracy(Mad_fc_h14, Mad_N_cases_tt)
```

```
## # A tibble: 4 x 11
##   .model    sub_region_2 .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE   ACF1
##   <chr>     <chr>        <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>
## 1 arima_at1 Madrid       Test   601.  628.  601.  33.4  33.4   NaN   NaN -0.232
## 2 arima_at2 Madrid       Test   595.  619.  595.  33.2  33.2   NaN   NaN -0.208
## 3 arima_man Madrid       Test   604.  624.  604.  34.0  34.0   NaN   NaN -0.193
## 4 SNaive    Madrid       Test   707.  732.  707.  39.6  39.6   NaN   NaN -0.215
```

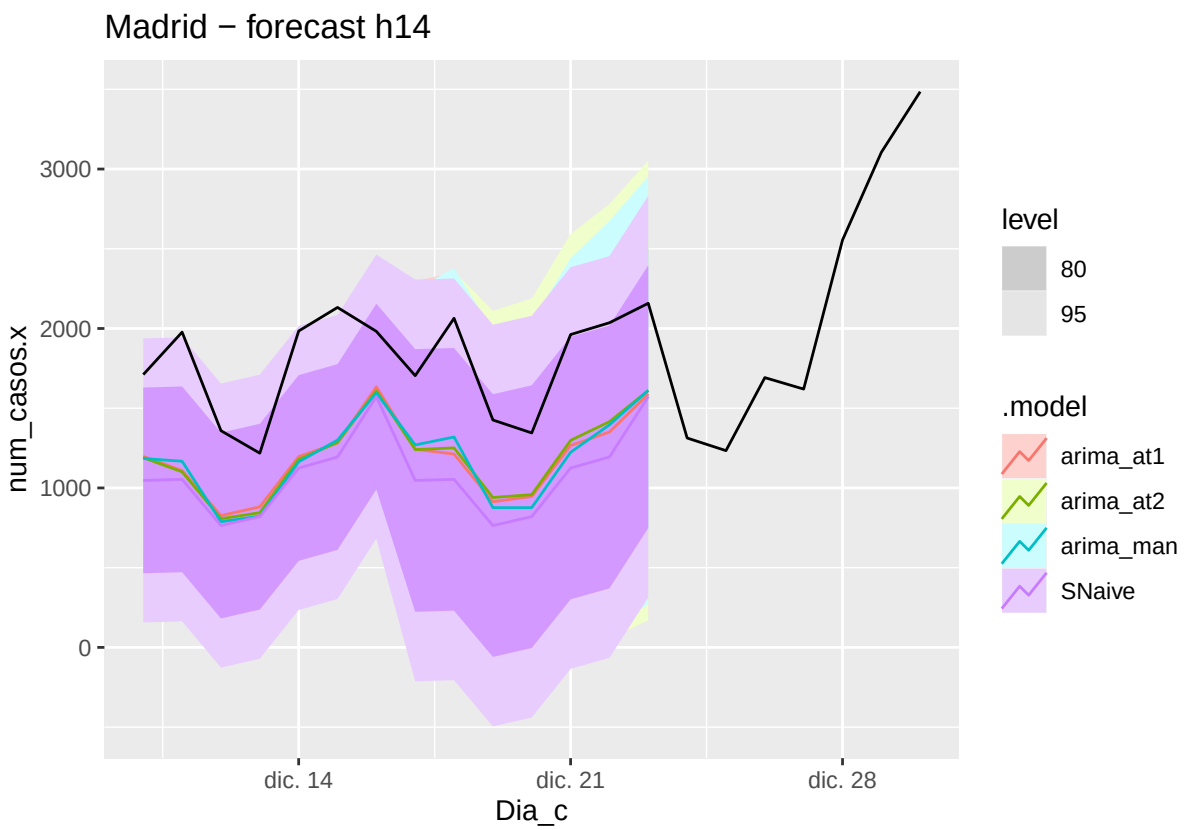
```
fabletools::accuracy(Mad_fc_h21, Mad_N_cases_tt)
```

```
## # A tibble: 4 x 11
##   .model    sub_region_2 .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE  ACF1
##   <chr>     <chr>        <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Madrid       Test   696.  826.  697.  33.8  33.9   NaN   NaN 0.529
## 2 arima_at2 Madrid       Test   662.  788.  674.  32.0  33.0   NaN   NaN 0.519
## 3 arima_man Madrid       Test   695.  834.  714.  33.7  35.2   NaN   NaN 0.551
## 4 SNaive    Madrid       Test   825.  938.  825.  41.2  41.2   NaN   NaN 0.553
```

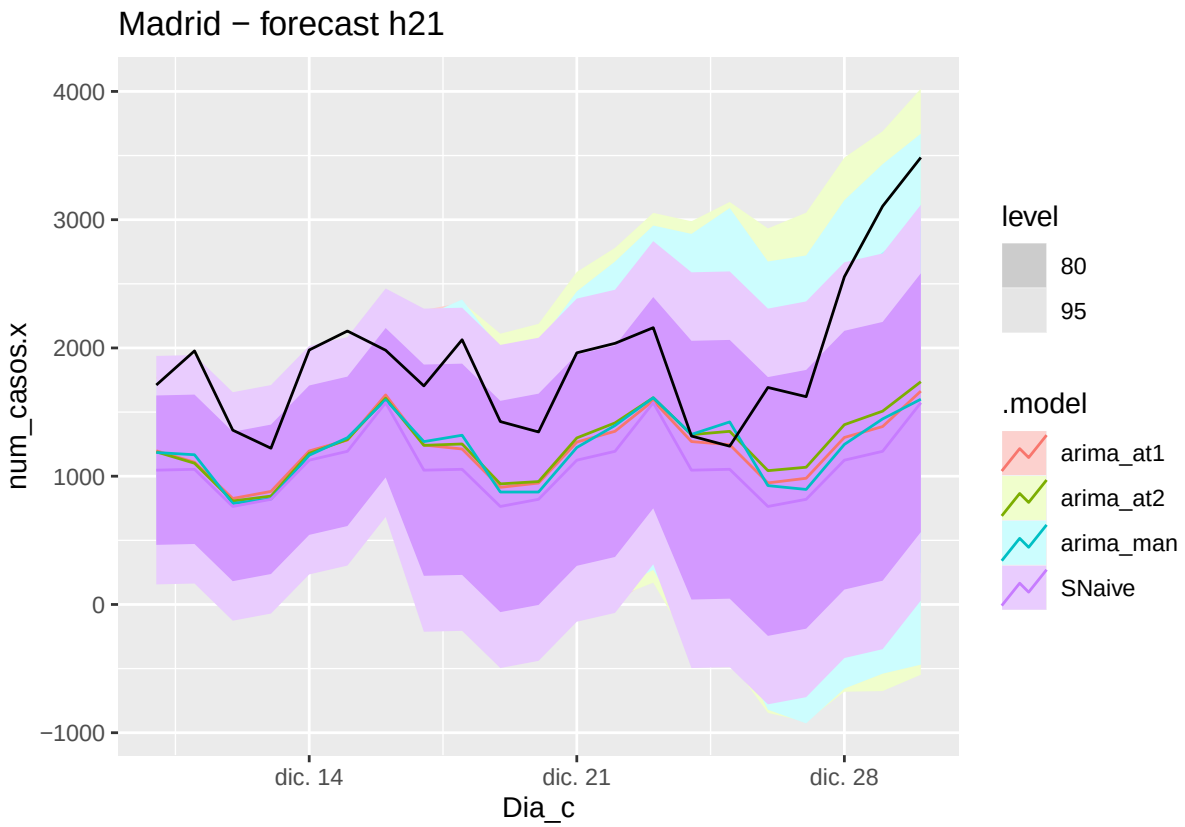
```
# Plots
Mad_fc_h7 %>%
  autoplot(Mad_N_cases_tt) +
  labs(title="Madrid - forecast h7")
```




```
Mad_fc_h14 %>%
  autoplot(Mad_N_cases_tt) +
  labs(title="Madrid - forecast h14")
```




```
Mad_fc_h21 %>%
  autoplot(Mad_N_cases_tt) +
  labs(title="Madrid - forecast h21")
```




```
####### Málaga #######
# Model train
Mal_fit_model <- Mal_N_cases_tr %>%
  model(
    SNaive = SNAIVE(num_casos.x),
    arima_man = ARIMA(num_casos.x ~ pdq(2,1,2) + PDQ(1,1,1)),
    arima_at1 = ARIMA(num_casos.x),
    arima_at2 = ARIMA(num_casos.x, stepwise = FALSE,approx = FALSE))

Mal_fit_model %>% pivot_longer(!sub_region_2,
                           names_to = "Model name",
                           values_to = "Orders")
```

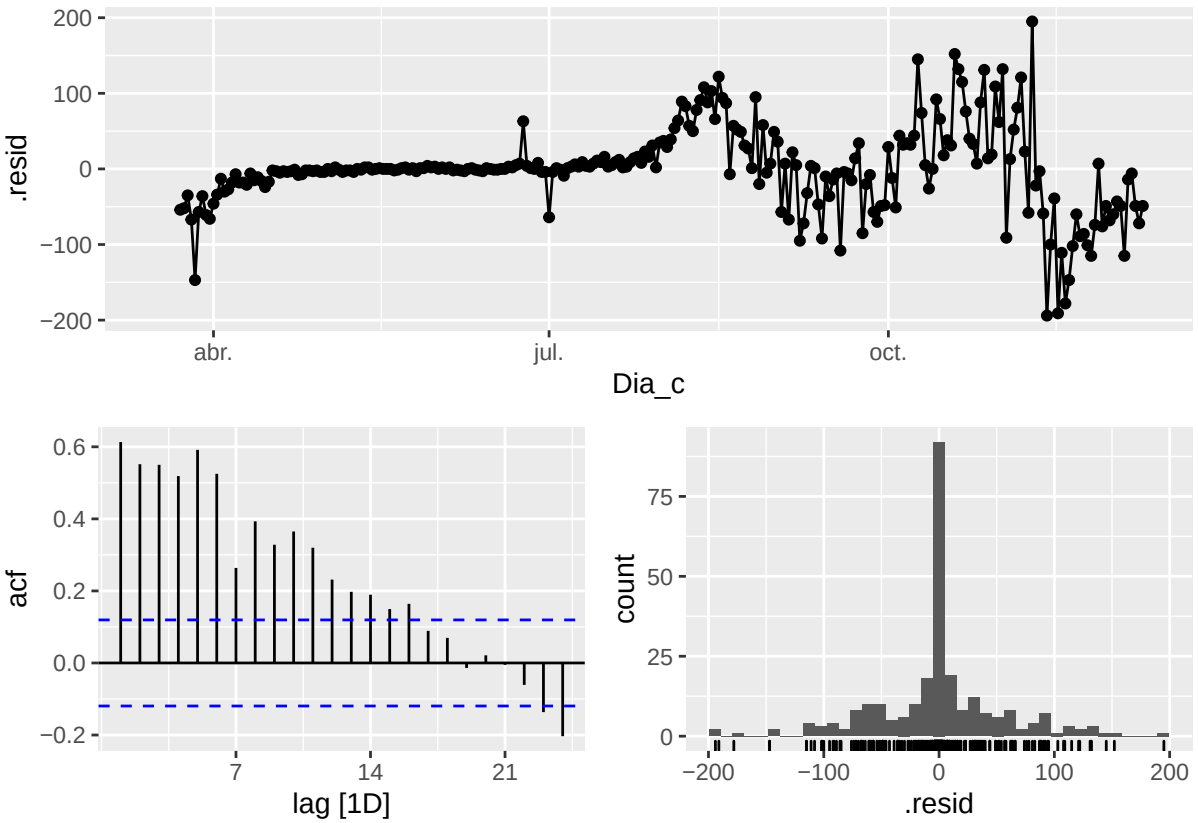
```
## # A mable: 4 x 3
## # Key:     sub_region_2, Model name [4]
##   sub_region_2 `Model name`                    Orders
##   <chr>        <chr>                          <model>
## 1 Málaga       SNaive                        <SNAIVE>
## 2 Málaga       arima_man    <ARIMA(2,1,2)(1,1,1)[7]>
## 3 Málaga       arima_at1    <ARIMA(0,1,1)(2,0,2)[7]>
## 4 Málaga       arima_at2    <ARIMA(1,1,3)(1,0,1)[7]>
```

```
# Good model >> Less Sigma / More BIC or AIC
glance(Mal_fit_model) %>% arrange(AICc) %>% select(.model:BIC)
```

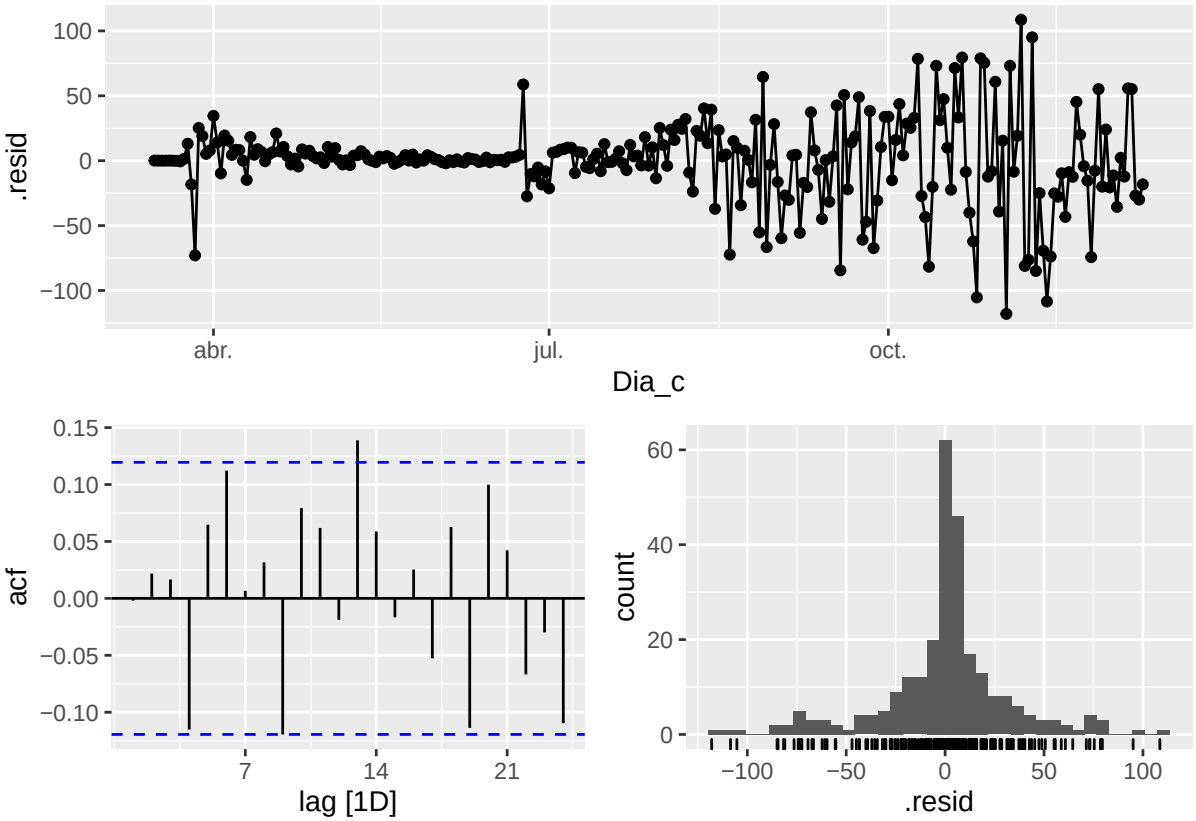
```
## # A tibble: 4 x 6
##   .model    sigma2 log_lik   AIC  AICc   BIC
```

```
##   <chr>      <dbl>   <dbl> <dbl> <dbl> <dbl>
## 1 arima_man  1096.  -1283. 2580. 2581. 2605.
## 2 arima_at2  1078.  -1316. 2645. 2646. 2670.
## 3 arima_at1  1108.  -1320. 2653. 2653. 2674.
## 4 SNaive     2991.     NA    NA    NA    NA
```

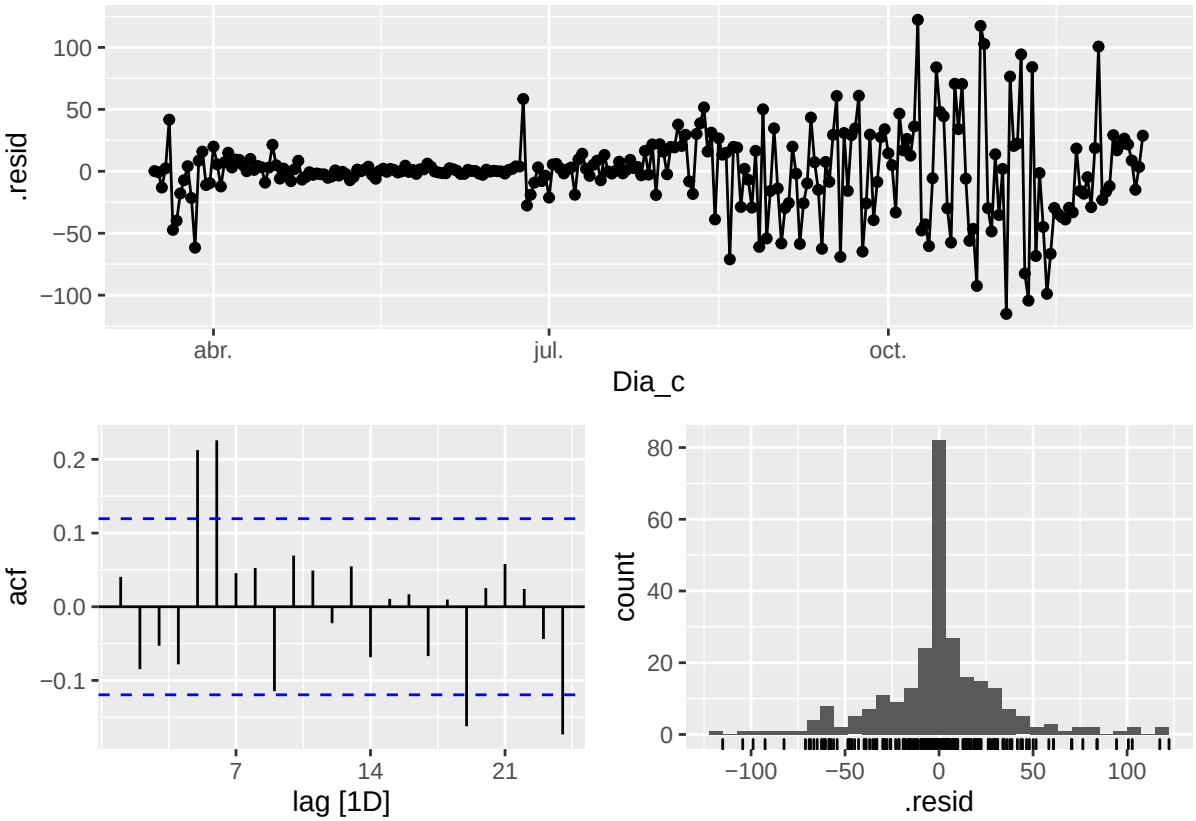
```
Mal_fit_model %>% select(SNaive) %>% gg_tsresiduals()
```



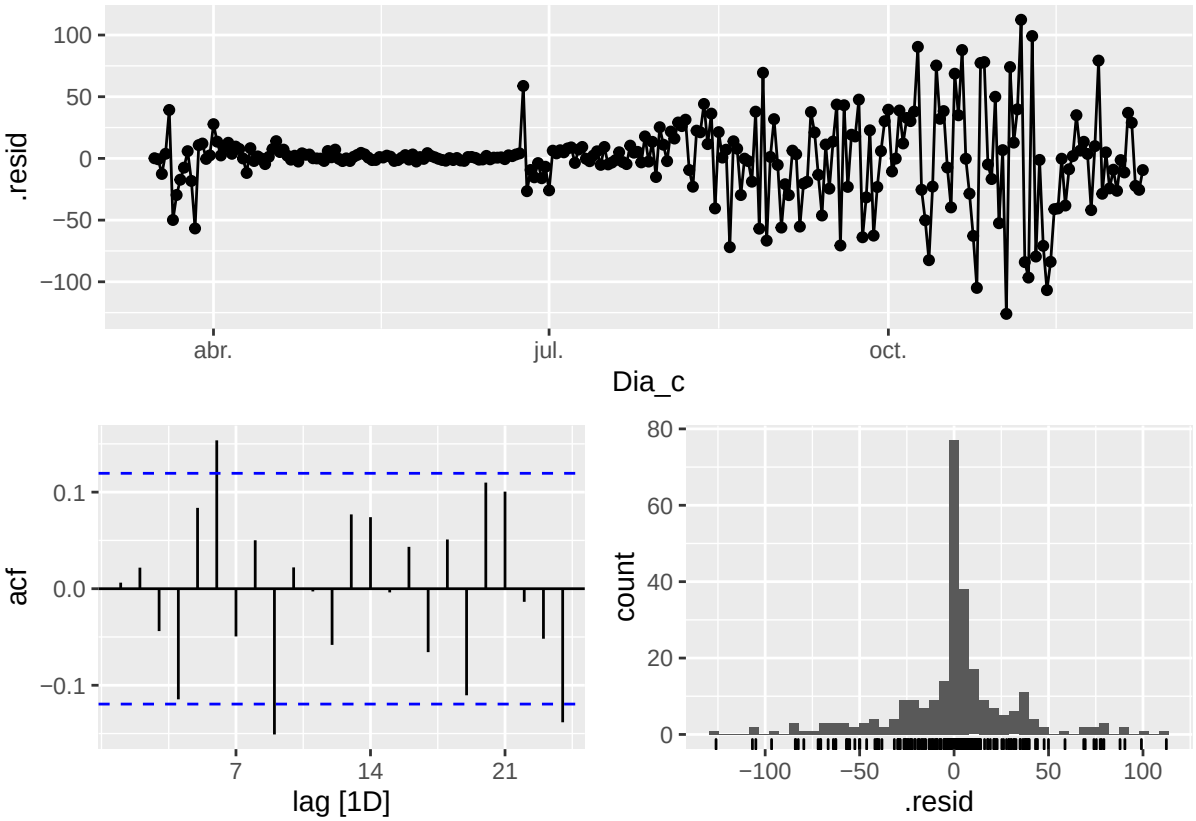

```
Mal_fit_model %>% select(arima_man) %>% gg_tsresiduals()
```




```
Mal_fit_model %>% select(arima_at1) %>% gg_tsresiduals()
```




```
Mal_fit_model %>% select(arima_at2) %>% gg_tsresiduals()
```




```
augment(Mal_fit_model) %>%
  features(.innov, ljung_box, lag=7)
```

```
## # A tibble: 4 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>       <dbl>     <dbl>
## 1 Málaga       arima_at1   32.0  0.0000399
## 2 Málaga       arima_at2   13.4  0.0622
## 3 Málaga       arima_man    8.52 0.289
## 4 Málaga       SNaive     521.   0
```

```
augment(Mal_fit_model) %>%
  features(.innov, ljung_box, lag=14)
```

```
## # A tibble: 4 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>       <dbl>     <dbl>
## 1 Málaga       arima_at1    40.9  0.000187
## 2 Málaga       arima_at2    24.9  0.0358
## 3 Málaga       arima_man    22.2  0.0740
## 4 Málaga       SNaive      693.   0
```

```
augment(Mal_fit_model) %>%
  features(.innov, ljung_box, lag=21)
```

```
## # A tibble: 4 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>       <dbl>     <dbl>
```

```
## 1 Málaga       arima_at1    51.2  0.000252
## 2 Málaga       arima_at2    37.5  0.0148
## 3 Málaga       arima_man    31.7  0.0635
## 4 Málaga       SNaive      711.   0
```

```
# Forecast
Mal_fc_h7<-fabletools::forecast(Mal_fit_model, h=7)
Mal_fc_h14<-fabletools::forecast(Mal_fit_model, h=14)
Mal_fc_h21<-fabletools::forecast(Mal_fit_model, h=21)
# Accuracy
fabletools::accuracy(Mal_fc_h7, Mal_N_cases_tt)
```

```
## # A tibble: 4 x 11
##   .model    sub_region_2 .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE  ACF1
##   <chr>     <chr>        <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Málaga       Test   97.6 102.   97.6  56.8  56.8   NaN   NaN 0.427
## 2 arima_at2 Málaga       Test   98.0 102.   98.0  58.0  58.0   NaN   NaN 0.257
## 3 arima_man Málaga       Test   96.5 104.   96.5  57.9  57.9   NaN   NaN 0.241
## 4 SNaive    Málaga       Test   48.3  57.7  48.3  27.9  27.9   NaN   NaN 0.412
```

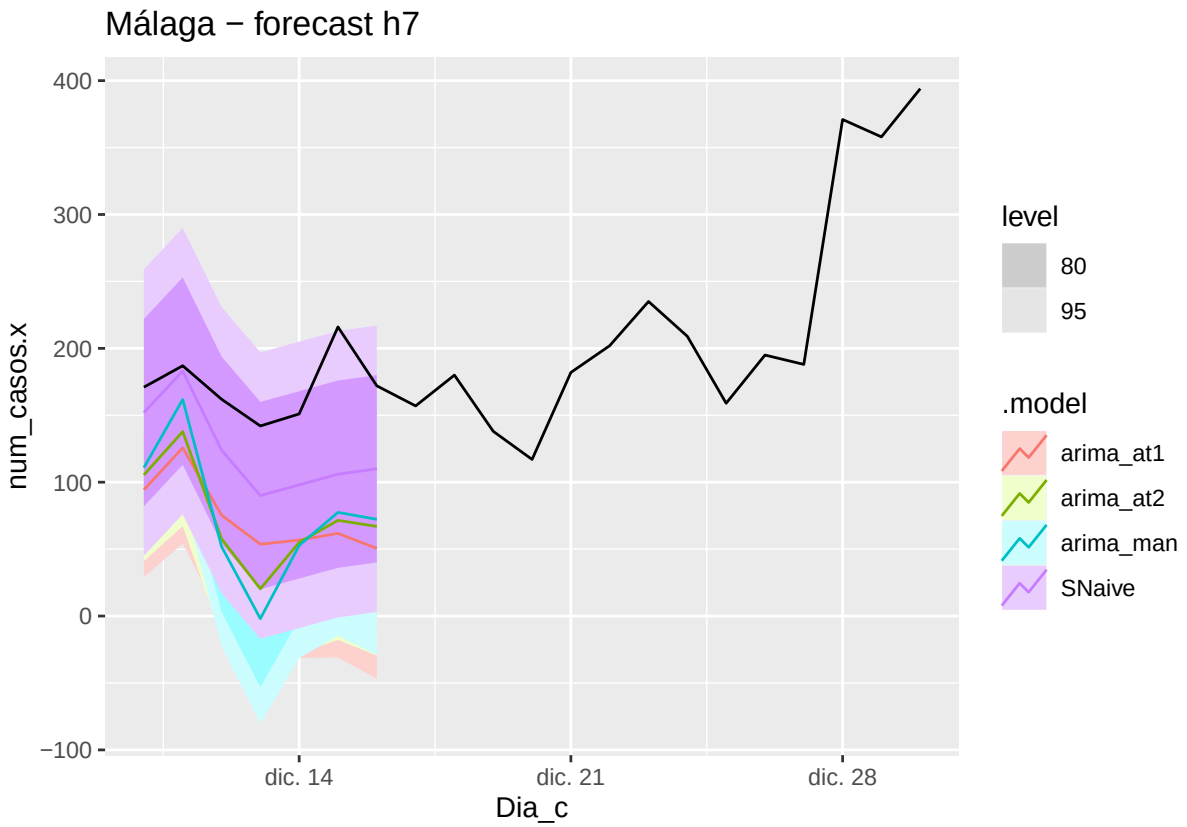
```
fabletools::accuracy(Mal_fc_h14, Mal_N_cases_tt)
```

```
## # A tibble: 4 x 11
##   .model    sub_region_2 .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE  ACF1
##   <chr>     <chr>        <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Málaga       Test   109. 115.  109.   62.2  62.2   NaN   NaN 0.471
## 2 arima_at2 Málaga       Test   121. 128.  121.   71.2  71.2   NaN   NaN 0.532
## 3 arima_man Málaga       Test   124. 134.  124.   74.3  74.3   NaN   NaN 0.522
## 4 SNaive    Málaga       Test    49   63.3  49.4  26.9  27.2   NaN   NaN 0.516
```

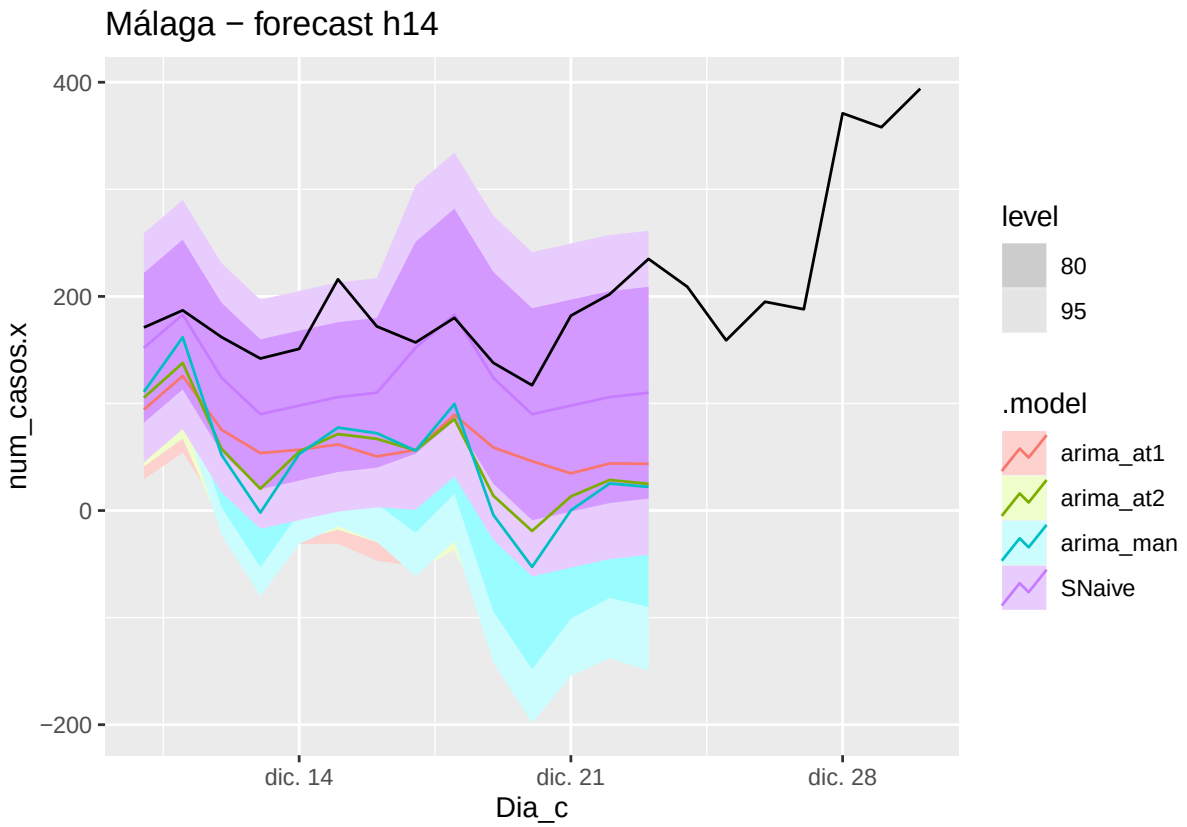
```
fabletools::accuracy(Mal_fc_h21, Mal_N_cases_tt)
```

```
## # A tibble: 4 x 11
##   .model    sub_region_2 .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE  ACF1
##   <chr>     <chr>        <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Málaga       Test  143.   165. 143.   66.7  66.7   NaN   NaN 0.625
## 2 arima_at2 Málaga       Test  173.   200. 173.   81.7  81.7   NaN   NaN 0.718
## 3 arima_man Málaga       Test  181.   212. 181.   86.2  86.2   NaN   NaN 0.720
## 4 SNaive    Málaga       Test   80.8  118.  83.4  33.0  34.6   NaN   NaN 0.637
```

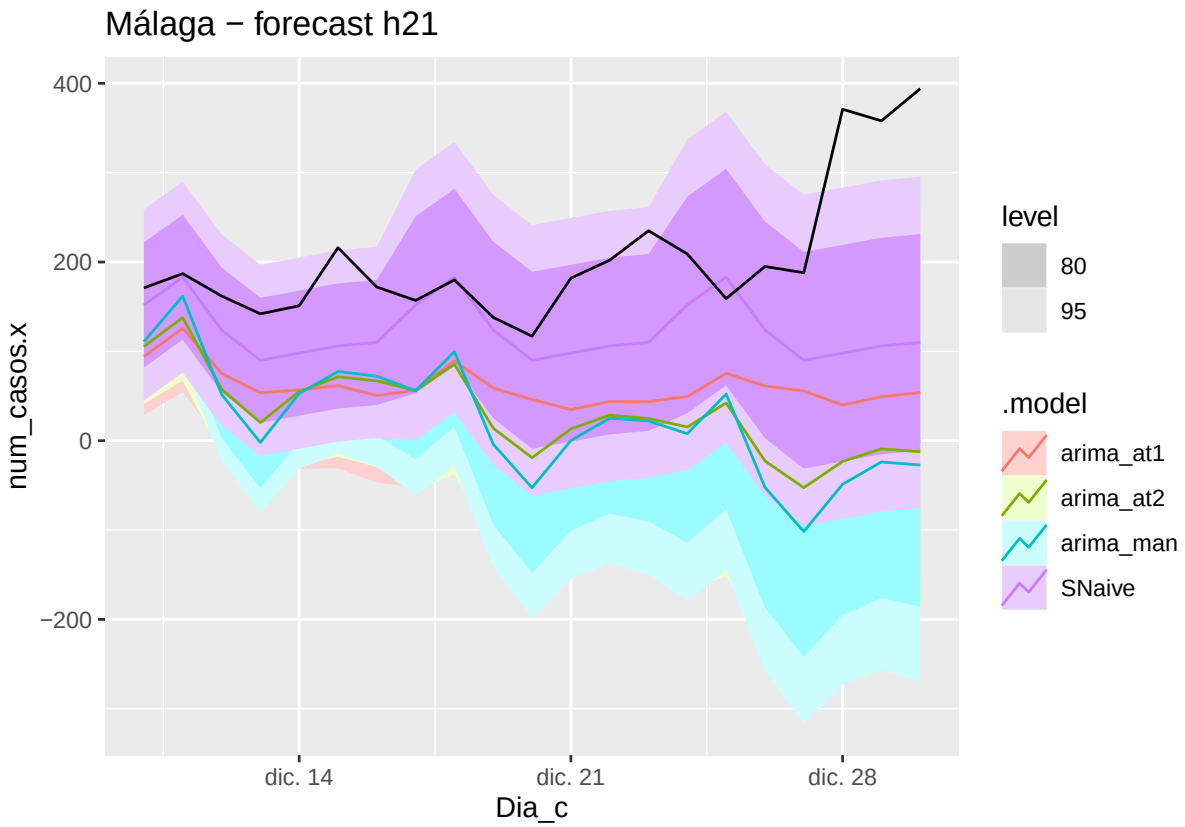
```
# Plots
Mal_fc_h7 %>%
  autoplot(Mal_N_cases_tt) +
  labs(title="Málaga - forecast h7")
```




```
Mal_fc_h14 %>%
  autoplot(Mal_N_cases_tt) +
  labs(title="Málaga - forecast h14")
```




```
Mal_fc_h21 %>%
  autoplot(Mal_N_cases_tt) +
  labs(title="Málaga - forecast h21")
```




```
####### Cádiz #######
# Model train
Cad_fit_model <- Cad_N_cases_tr %>%
  model(
    SNaive = SNAIVE(num_casos.x),
    arima_man = ARIMA(num_casos.x ~ pdq(2,1,2) + PDQ(1,1,1)),
    arima_at1 = ARIMA(num_casos.x),
    arima_at2 = ARIMA(num_casos.x, stepwise = FALSE,approx = FALSE))

Cad_fit_model %>% pivot_longer(!sub_region_2,
                           names_to = "Model name",
                           values_to = "Orders")
```

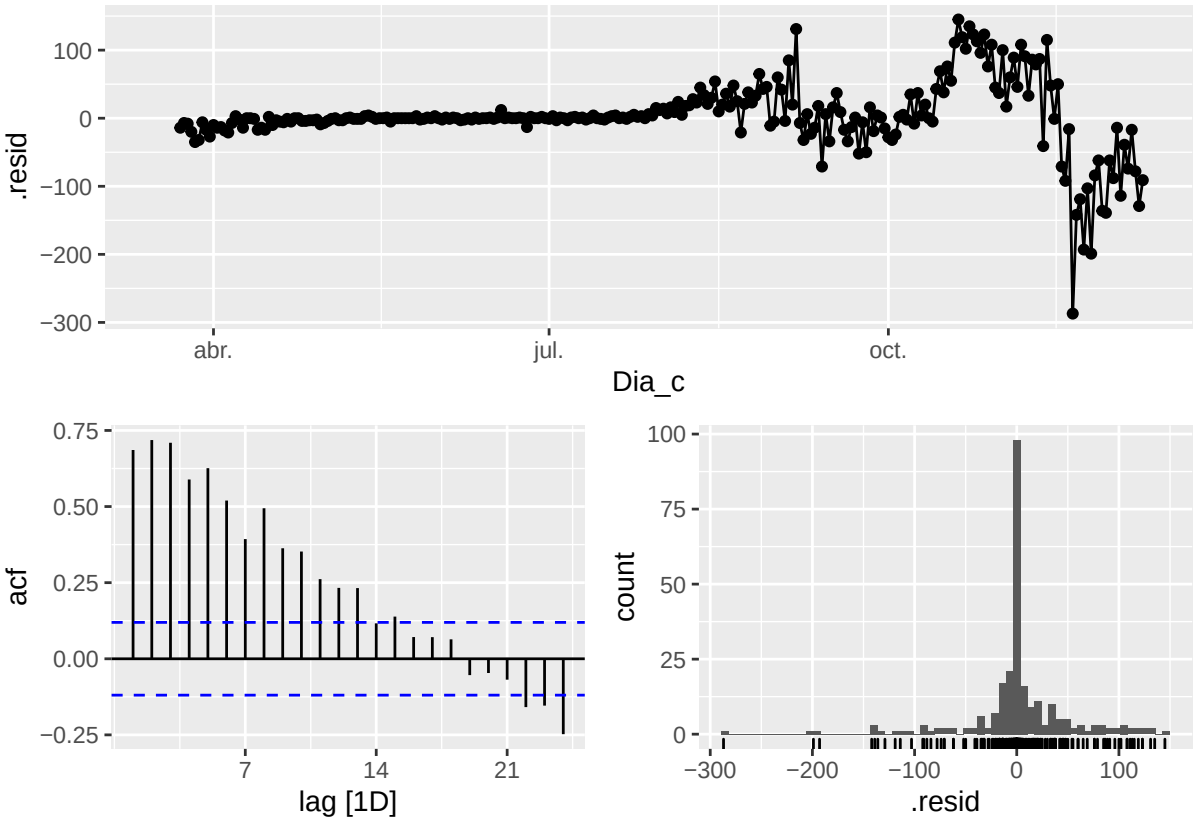
```
## # A mable: 4 x 3
## # Key:     sub_region_2, Model name [4]
##   sub_region_2 `Model name`                    Orders
##   <chr>        <chr>                          <model>
## 1 Cádiz        SNaive                        <SNAIVE>
## 2 Cádiz        arima_man    <ARIMA(2,1,2)(1,1,1)[7]>
## 3 Cádiz        arima_at1    <ARIMA(2,1,1)(2,0,0)[7]>
## 4 Cádiz        arima_at2    <ARIMA(2,1,1)(1,0,1)[7]>
```

```
# Good model >> Less Sigma / More BIC or AIC
glance(Cad_fit_model) %>% arrange(AICc) %>% select(.model:BIC)
```

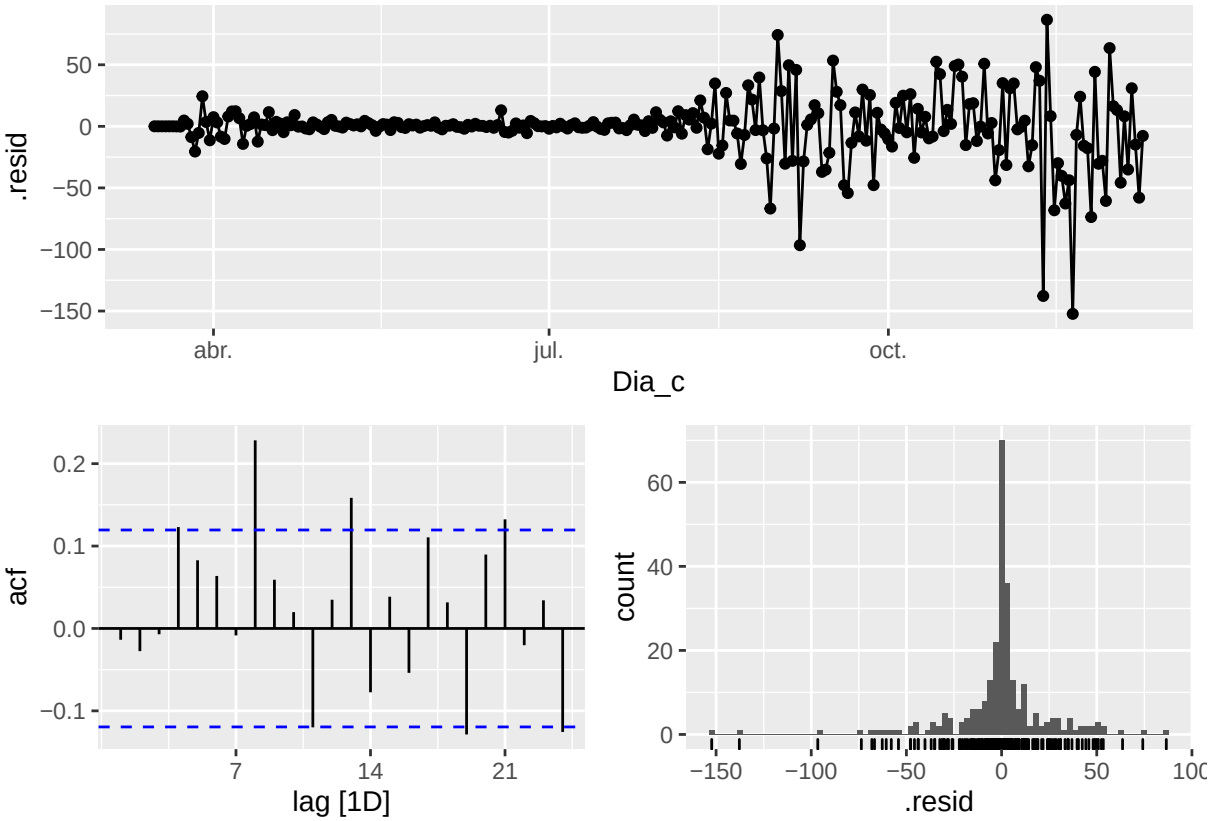
```
## # A tibble: 4 x 6
##   .model    sigma2 log_lik   AIC  AICc   BIC
```

```
##   <chr>      <dbl>   <dbl> <dbl> <dbl> <dbl>
## 1 arima_man   689.  -1223. 2459. 2459. 2484.
## 2 arima_at2   680.  -1255. 2522. 2523. 2544.
## 3 arima_at1   725.  -1263. 2538. 2538. 2560.
## 4 SNaive     2562.     NA    NA    NA    NA
```

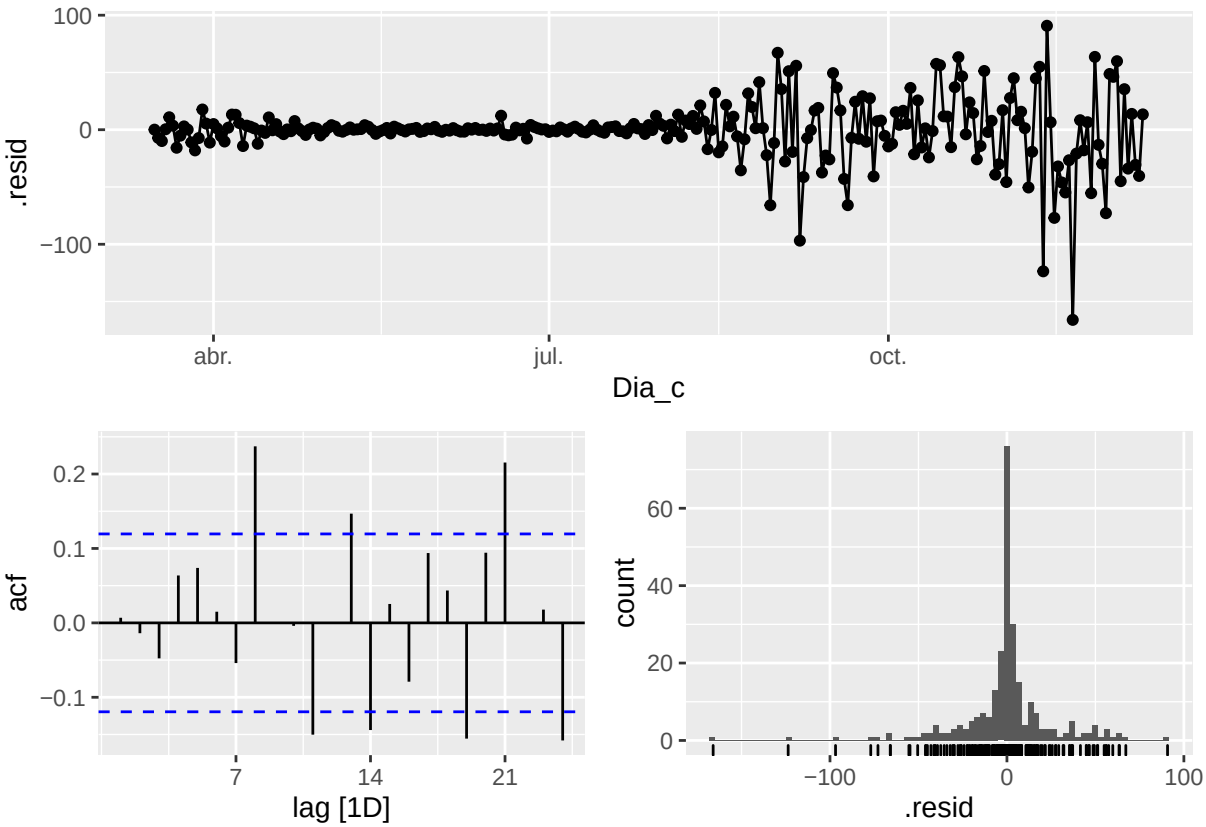
```
Cad_fit_model %>% select(SNaive) %>% gg_tsresiduals()
```



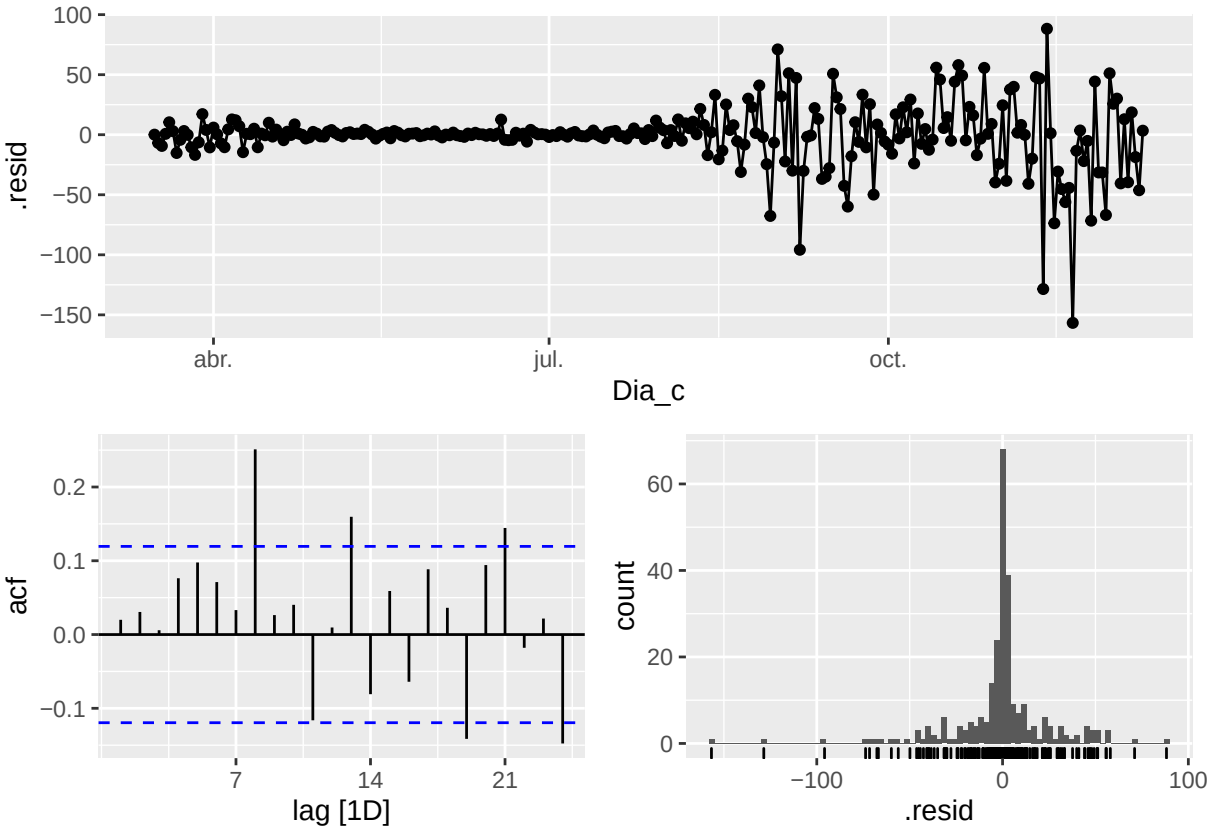

```
Cad_fit_model %>% select(arima_man) %>% gg_tsresiduals()
```




```
Cad_fit_model %>% select(arima_at1) %>% gg_tsresiduals()
```




```
Cad_fit_model %>% select(arima_at2) %>% gg_tsresiduals()
```




```
augment(Cad_fit_model) %>%
  features(.innov, ljung_box, lag=7)
```

```
## # A tibble: 4 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>       <dbl>     <dbl>
## 1 Cádiz        arima_at1    4.18     0.758
## 2 Cádiz        arima_at2    6.32     0.503
## 3 Cádiz        arima_man    7.47     0.382
## 4 Cádiz        SNaive     710.       0
```

```
augment(Cad_fit_model) %>%
  features(.innov, ljung_box, lag=14)
```

```
## # A tibble: 4 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>       <dbl>     <dbl>
## 1 Cádiz        arima_at1    38.3  0.000461
## 2 Cádiz        arima_at2    37.5  0.000611
## 3 Cádiz        arima_man    36.4  0.000910
## 4 Cádiz        SNaive      899.   0
```

```
augment(Cad_fit_model) %>%
  features(.innov, ljung_box, lag=21)
```

```
## # A tibble: 4 x 4
##   sub_region_2 .model    lb_stat  lb_pvalue
##   <chr>        <chr>       <dbl>      <dbl>
```

```
## 1 Cádiz        arima_at1    66.7 0.00000117
## 2 Cádiz        arima_at2    56.9 0.0000369
## 3 Cádiz        arima_man    53.8 0.000105
## 4 Cádiz        SNaive      911.  0
```

```
# Forecast
Cad_fc_h7<-fabletools::forecast(Cad_fit_model, h=7)
Cad_fc_h14<-fabletools::forecast(Cad_fit_model, h=14)
Cad_fc_h21<-fabletools::forecast(Cad_fit_model, h=21)
# Accuracy
fabletools::accuracy(Cad_fc_h7, Cad_N_cases_tt)
```

```
## # A tibble: 4 x 11
##   .model    sub_region_2 .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE    ACF1
##   <chr>     <chr>        <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl>
## 1 arima_at1 Cádiz        Test  78.8   79.3  78.8 50.7   50.7   NaN   NaN -0.368
## 2 arima_at2 Cádiz        Test  82.2   82.7  82.2 53.3   53.3   NaN   NaN  0.0954
## 3 arima_man Cádiz        Test  86.6   87.6  86.6 56.6   56.6   NaN   NaN  0.165
## 4 SNaive    Cádiz        Test   3.57  20.6  17.6  2.68  10.8   NaN   NaN -0.166
```

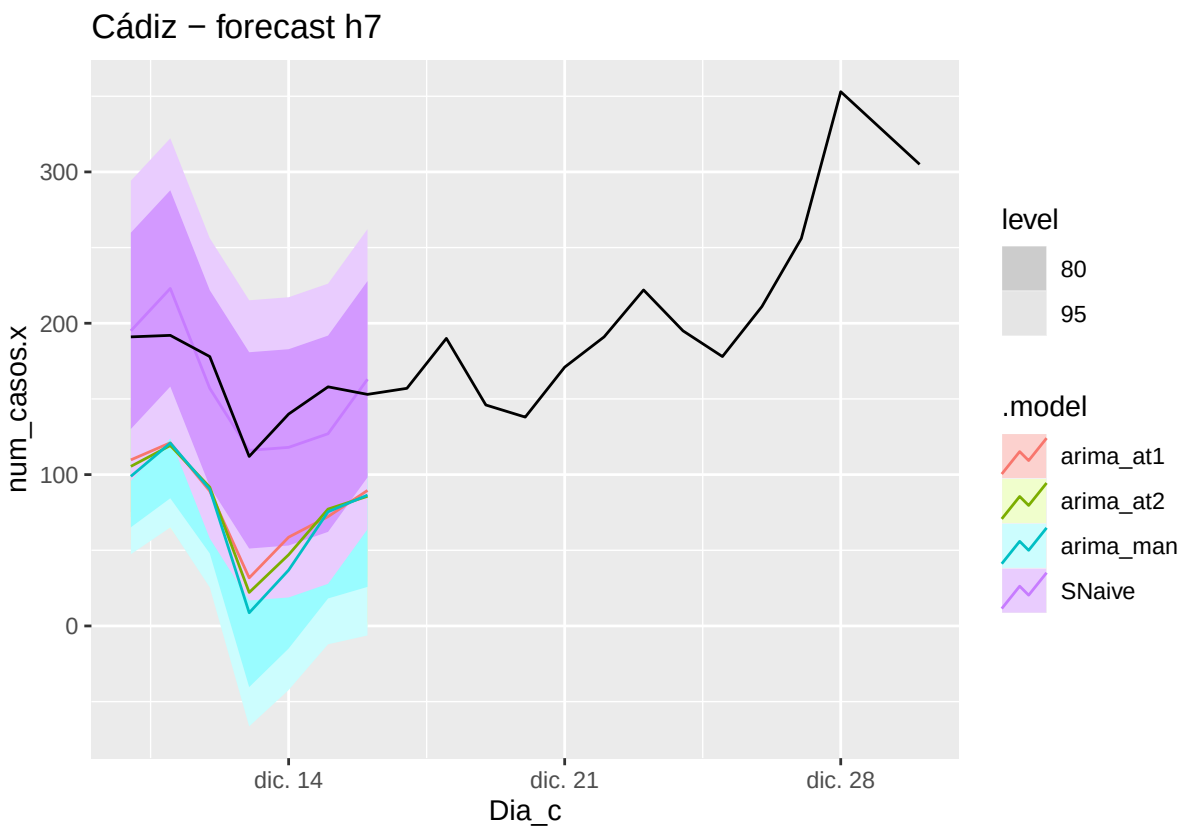
```
fabletools::accuracy(Cad_fc_h14, Cad_N_cases_tt)
```

```
## # A tibble: 4 x 11
##   .model    sub_region_2 .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE  ACF1
##   <chr>     <chr>        <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Cádiz        Test  113.  119.  113.  67.6   67.6   NaN   NaN 0.730
## 2 arima_at2 Cádiz        Test  119.  126.  119.  71.8   71.8   NaN   NaN 0.731
## 3 arima_man Cádiz        Test  123.  131.  123.  75.1   75.1   NaN   NaN 0.712
## 4 SNaive    Cádiz        Test   10.1  34.3  28.8  5.48  16.5   NaN   NaN 0.583
```

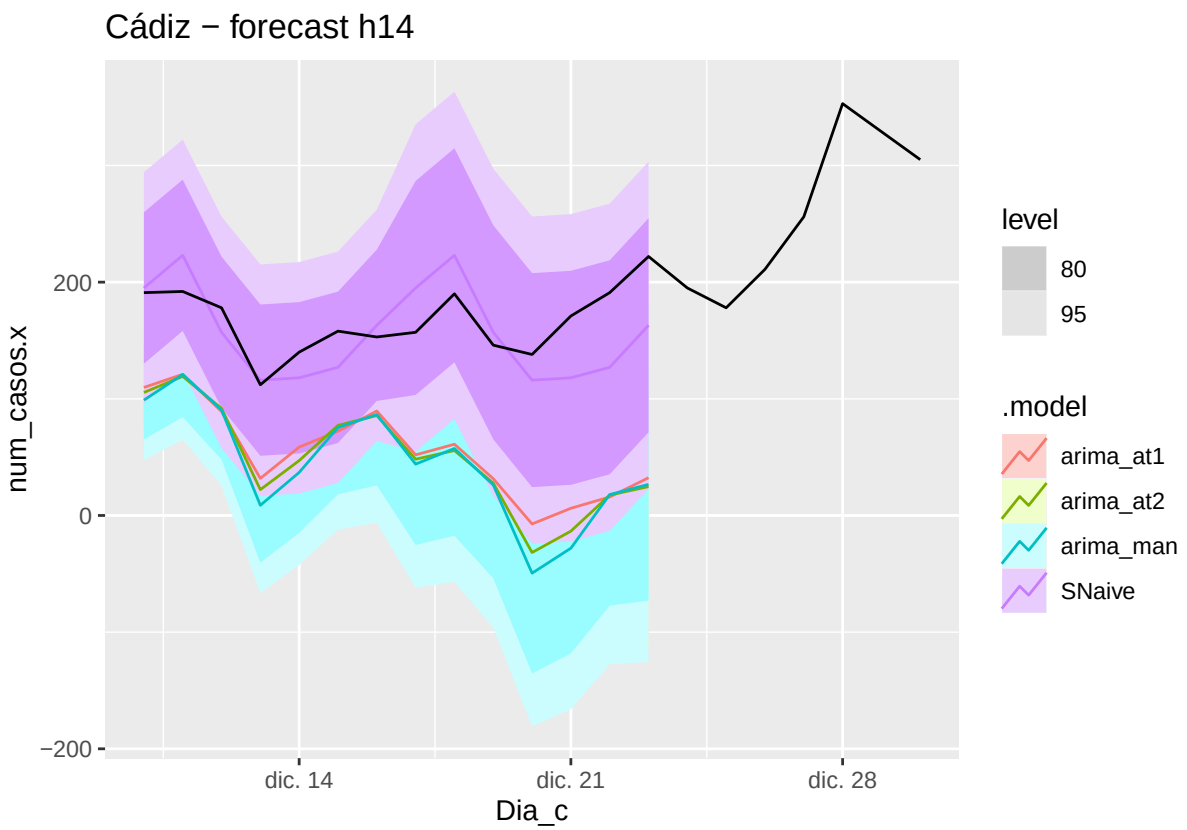
```
fabletools::accuracy(Cad_fc_h21, Cad_N_cases_tt)
```

```
## # A tibble: 4 x 11
##   .model    sub_region_2 .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE  ACF1
##   <chr>     <chr>        <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Cádiz        Test  168.  193.  168.   80.4  80.4   NaN   NaN 0.859
## 2 arima_at2 Cádiz        Test  179.  207.  179.   85.7  85.7   NaN   NaN 0.857
## 3 arima_man Cádiz        Test  185.  213.  185.   88.9  88.9   NaN   NaN 0.847
## 4 SNaive    Cádiz        Test   41.4  86.5  58.1  14.6  24.4   NaN   NaN 0.758
```

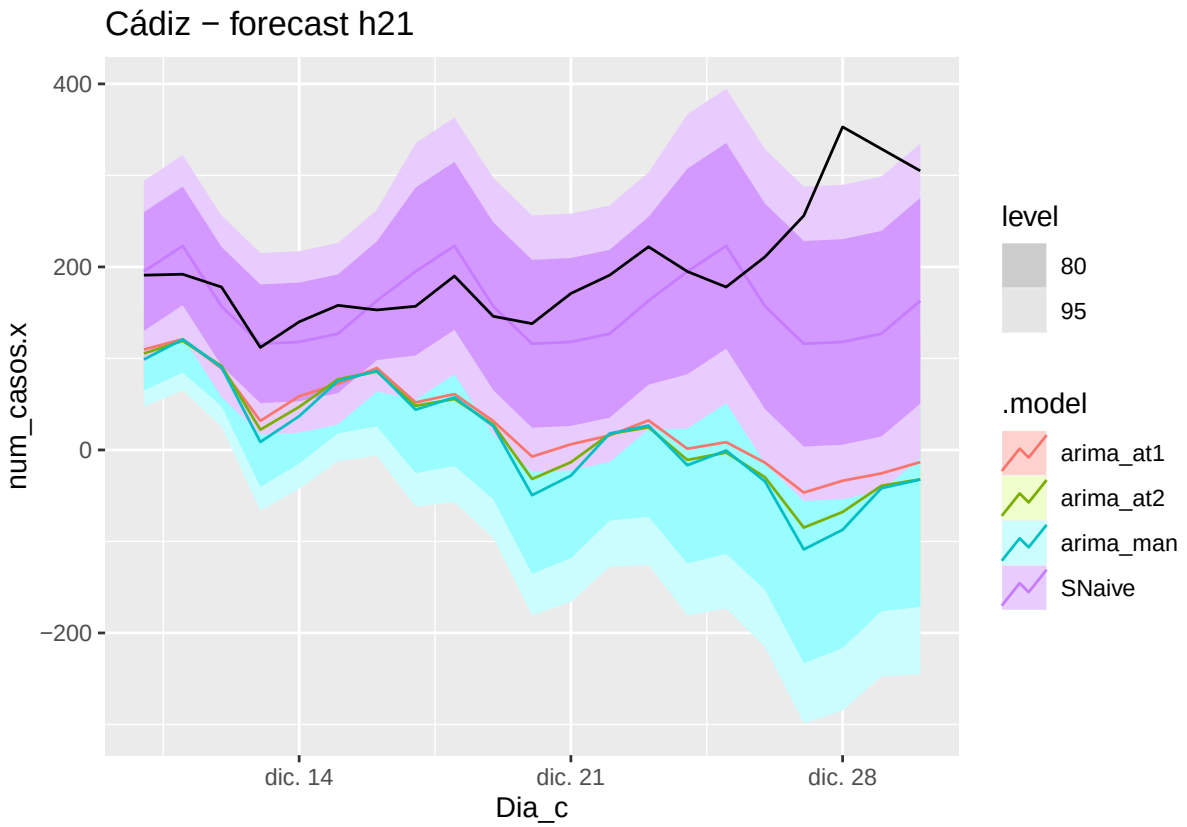
```
# Plots
Cad_fc_h7 %>%
  autoplot(Cad_N_cases_tt) +
  labs(title="Cádiz - forecast h7")
```




```
Cad_fc_h14 %>%
  autoplot(Cad_N_cases_tt) +
  labs(title="Cádiz - forecast h14")
```



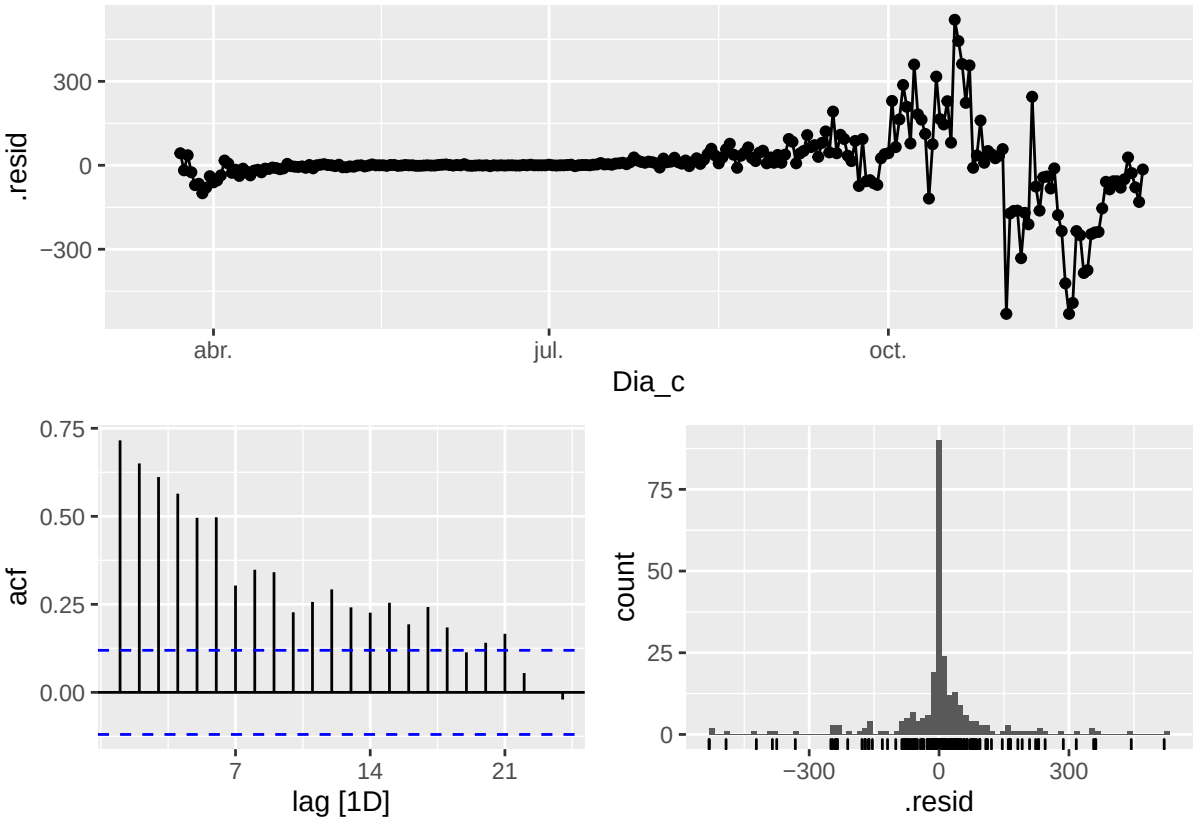

```
Cad_fc_h21 %>%
  autoplot(Cad_N_cases_tt) +
  labs(title="Cádiz - forecast h21")
```



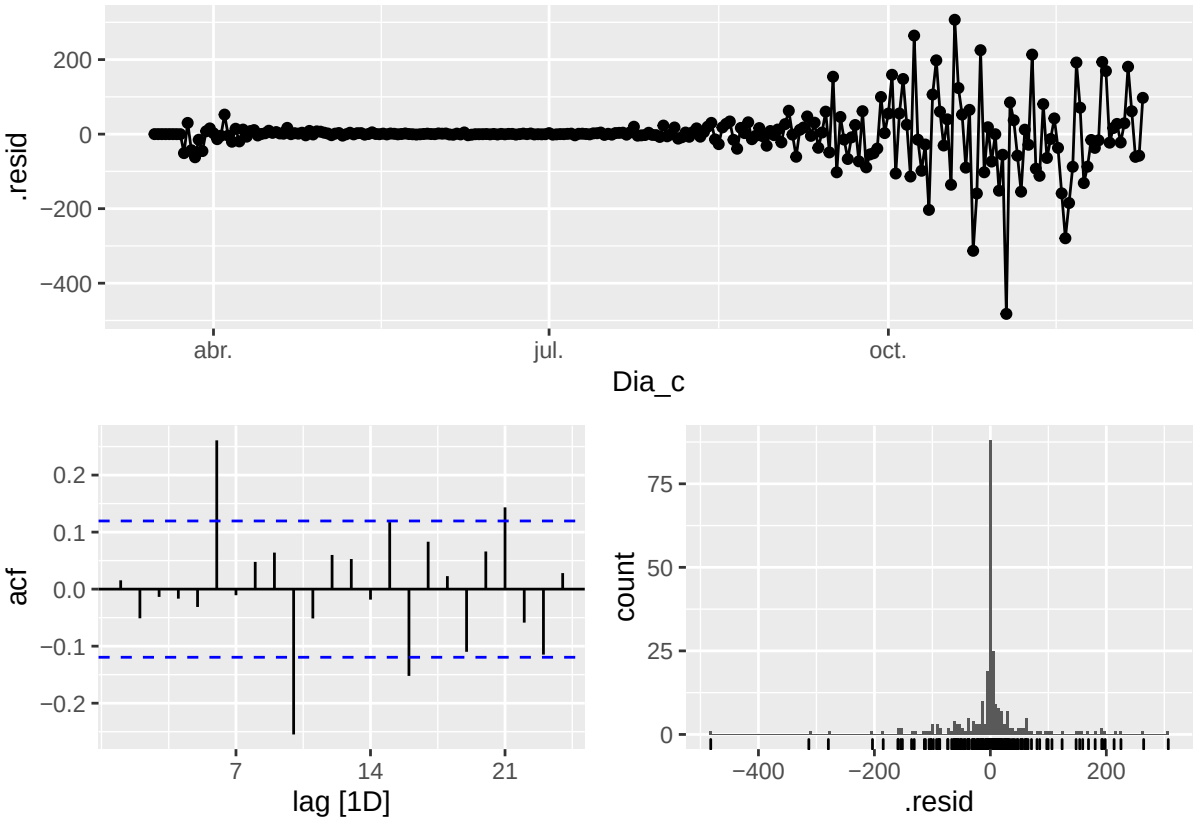

```
####### Sevilla #######
# Model train
Sev_fit_model <- Sev_N_cases_tr %>%
  model(
    SNaive = SNAIVE(num_casos.x),
    arima_man = ARIMA(num_casos.x ~ pdq(2,1,2) + PDQ(1,1,1)),
    arima_at1 = ARIMA(num_casos.x),
    arima_at2 = ARIMA(num_casos.x, stepwise = FALSE,approx = FALSE))

# Good model >> Less Sigma / More BIC or AIC
glance(Sev_fit_model) %>% arrange(AICc) %>% select(.model:BIC)
```

```
## # A tibble: 4 x 6
##   .model    sigma2 log_lik   AIC  AICc   BIC
##   <chr>      <dbl>   <dbl> <dbl> <dbl> <dbl>
## 1 arima_at2  5249.  -1493. 2998. 2998. 3019.
## 2 arima_man  5592.  -1494. 3003. 3003. 3028.
## 3 arima_at1  5485.  -1499. 3008. 3008. 3025.
## 4 SNaive    14589.     NA    NA    NA    NA
```

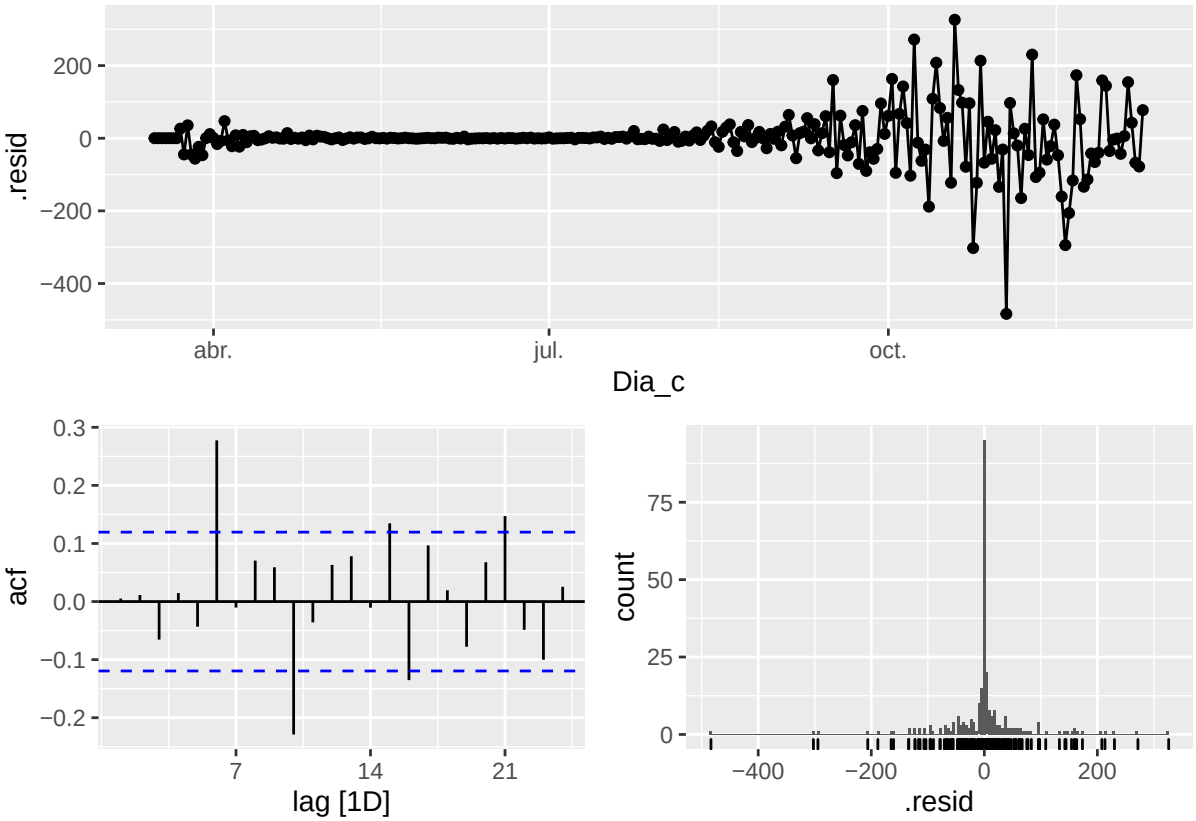
```
Sev_fit_model %>% select(SNaive) %>% gg_tsresiduals()
```



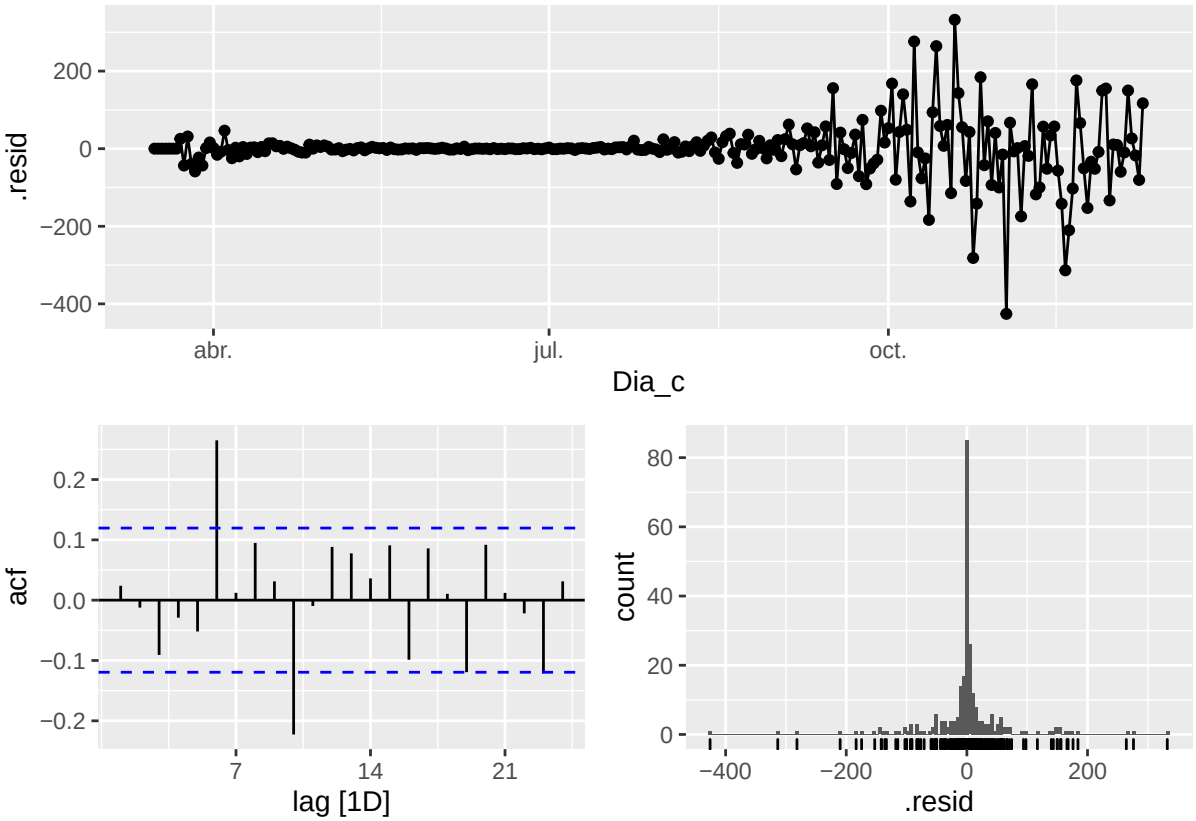

```
Sev_fit_model %>% select(arima_man) %>% gg_tsresiduals()
```




```
Sev_fit_model %>% select(arima_at1) %>% gg_tsresiduals()
```




```
Sev_fit_model %>% select(arima_at2) %>% gg_tsresiduals()
```




```
augment(Sev_fit_model) %>%
  features(.innov, ljung_box, lag=7, dof=3)
```

```
## # A tibble: 4 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>       <dbl>     <dbl>
## 1 Sevilla      arima_at1    23.2  0.000117
## 2 Sevilla      arima_at2    22.9  0.000131
## 3 Sevilla      arima_man    20.1  0.000482
## 4 Sevilla      SNaive      591.   0
```

```
augment(Sev_fit_model) %>%
  features(.innov, ljung_box, lag=14, dof=3)
```

```
## # A tibble: 4 x 4
##   sub_region_2 .model    lb_stat  lb_pvalue
##   <chr>        <chr>       <dbl>      <dbl>
## 1 Sevilla      arima_at1    43.5 0.00000879
## 2 Sevilla      arima_at2    44.0 0.00000730
## 3 Sevilla      arima_man    42.8 0.0000118
## 4 Sevilla      SNaive      743.  0
```

```
augment(Sev_fit_model) %>%
  features(.innov, ljung_box, lag=21, dof=3)
```

```
## # A tibble: 4 x 4
##   sub_region_2 .model    lb_stat   lb_pvalue
##   <chr>        <chr>       <dbl>       <dbl>
```

```
## 1 Sevilla      arima_at1    66.3 0.000000190
## 2 Sevilla      arima_at2    58.0 0.00000430
## 3 Sevilla      arima_man    66.4 0.000000181
## 4 Sevilla      SNaive      815.  0
```

```
# Forecast
Sev_fc_h7<-fabletools::forecast(Sev_fit_model, h=7)
Sev_fc_h14<-fabletools::forecast(Sev_fit_model, h=14)
Sev_fc_h21<-fabletools::forecast(Sev_fit_model, h=21)
# Accuracy
fabletools::accuracy(Sev_fc_h7, Sev_N_cases_tt)
```

```
## # A tibble: 4 x 11
##   .model    sub_region_2 .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE    ACF1
##   <chr>     <chr>        <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl>
## 1 arima_at1 Sevilla      Test   88.2  93.7  88.2  61.1  61.1   NaN   NaN -0.0150
## 2 arima_at2 Sevilla      Test  101.  106.  101.   67.4  67.4   NaN   NaN -0.373
## 3 arima_man Sevilla      Test  123.  128.  123.   85.2  85.2   NaN   NaN  0.156
## 4 SNaive    Sevilla      Test   35.1  48.5  39.1  21.1  23.8   NaN   NaN -0.360
```

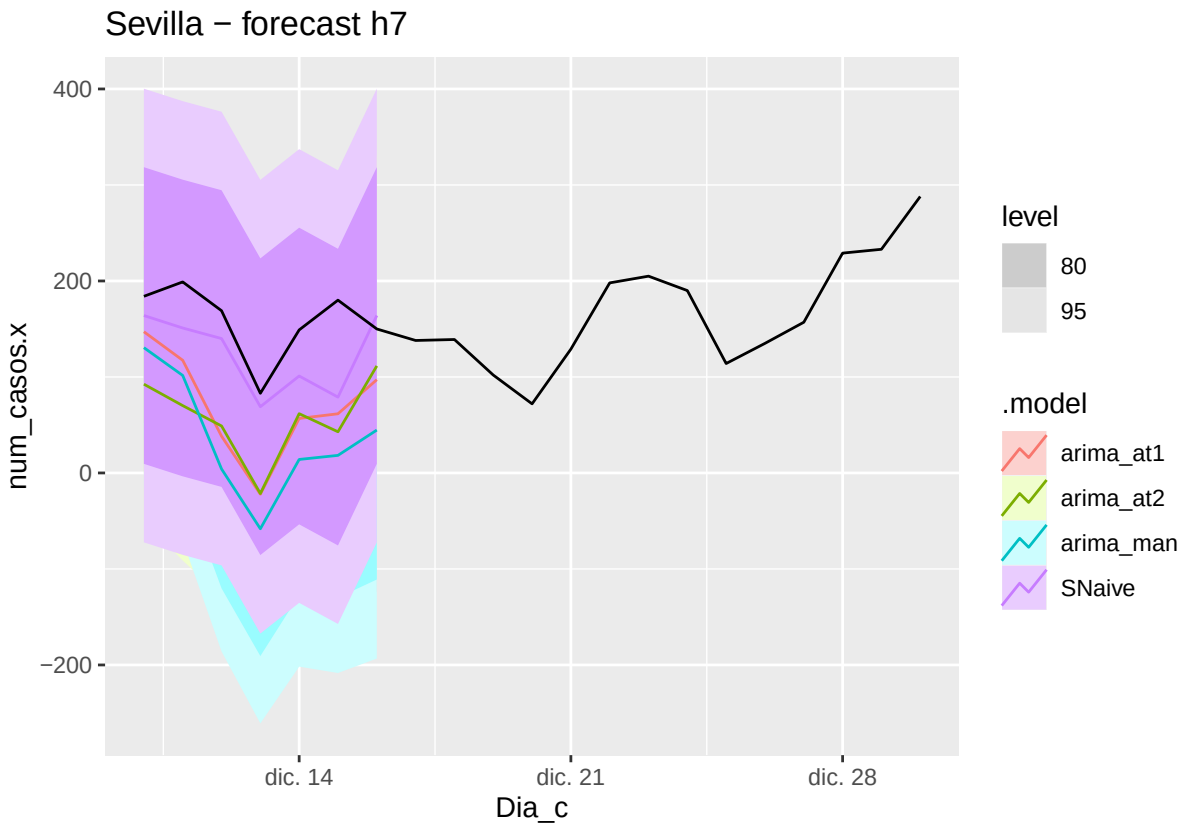
```
fabletools::accuracy(Sev_fc_h14, Sev_N_cases_tt)
```

```
## # A tibble: 4 x 11
##   .model    sub_region_2 .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE  ACF1
##   <chr>     <chr>        <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Sevilla      Test  102.  110.  102.   76.5  76.5   NaN   NaN 0.436
## 2 arima_at2 Sevilla      Test   98.6 103.   98.6  70.2  70.2   NaN   NaN 0.186
## 3 arima_man Sevilla      Test  163.  175.  163.  123.  123.    NaN   NaN 0.614
## 4 SNaive    Sevilla      Test   25.8  50.2  38.6  13.5  24.1   NaN   NaN 0.299
```

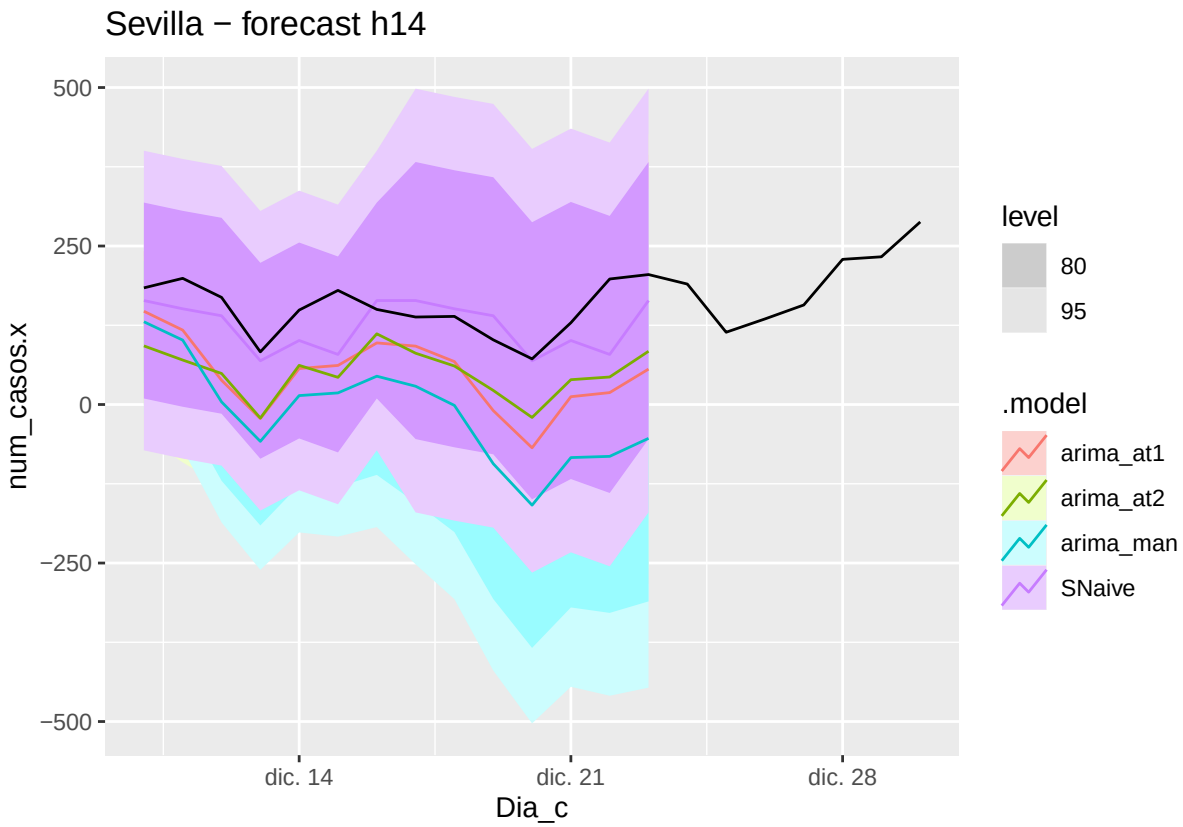
```
fabletools::accuracy(Sev_fc_h21, Sev_N_cases_tt)
```

```
## # A tibble: 4 x 11
##   .model    sub_region_2 .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE  ACF1
##   <chr>     <chr>        <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Sevilla      Test  136.  153.  136.   86.9  86.9   NaN   NaN 0.682
## 2 arima_at2 Sevilla      Test  125.  136.  125.   77.8  77.8   NaN   NaN 0.639
## 3 arima_man Sevilla      Test  227.  254.  227.  146.  146.    NaN   NaN 0.779
## 4 SNaive    Sevilla      Test   40.0  69.2  52.5  18.5  29.0   NaN   NaN 0.547
```

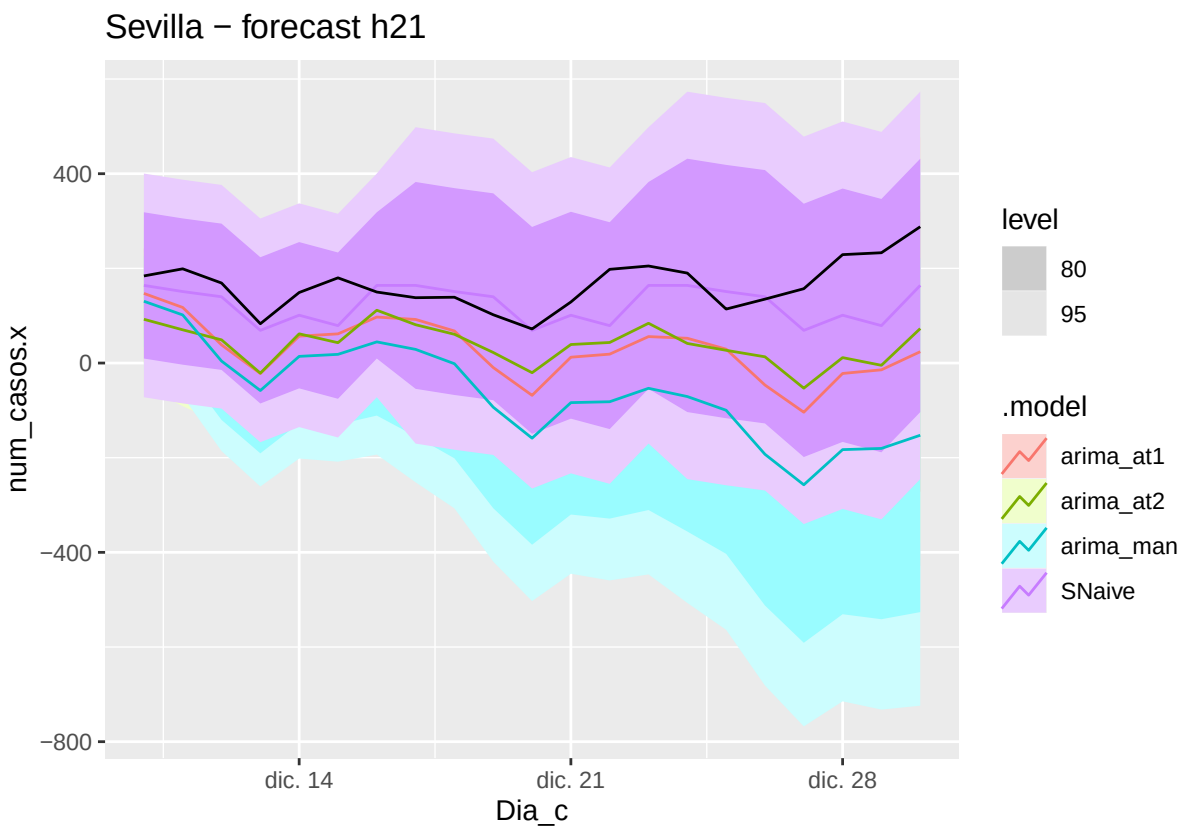
```
# Plots
Sev_fc_h7 %>%
  autoplot(Sev_N_cases_tt) +
  labs(title="Sevilla - forecast h7")
```




```
Sev_fc_h14 %>%
  autoplot(Sev_N_cases_tt) +
  labs(title="Sevilla - forecast h14")
```




```
Sev_fc_h21 %>%
  autoplot(Sev_N_cases_tt) +
  labs(title="Sevilla - forecast h21")
```



3.3.4 Multivariate (7, 14, 21 days) Málaga

```
# Model train
# We have added mobility variables to models
fit_model <- Mal_N_cases_tr %>%
  model(
    SNaive = SNAIVE(num_casos.x),
    arima_man = ARIMA(num_casos.x ~ retail_and_recreation_percent_change_from_baseline +
                        grocery_and_pharmacy_percent_change_from_baseline +
                        parks_percent_change_from_baseline +
                        transit_stations_percent_change_from_baseline +
                        workplaces_percent_change_from_baseline +
                        residential_percent_change_from_baseline + Total + pdq(2,1,2) +
                        PDQ(1,1,1)),
    arima_at1 = ARIMA(num_casos.x~ retail_and_recreation_percent_change_from_baseline +
                        grocery_and_pharmacy_percent_change_from_baseline +
                        parks_percent_change_from_baseline +
                        transit_stations_percent_change_from_baseline +
                        workplaces_percent_change_from_baseline +
                        residential_percent_change_from_baseline + Total),
    arima_at2 = ARIMA(num_casos.x~ retail_and_recreation_percent_change_from_baseline +
                        grocery_and_pharmacy_percent_change_from_baseline +
                        parks_percent_change_from_baseline +
                        transit_stations_percent_change_from_baseline +
                        workplaces_percent_change_from_baseline +
```

```
                        residential_percent_change_from_baseline + Total,
                      stepwise = FALSE,approx = FALSE))

# Show and report model
fit_model
```

```
## # A mable: 1 x 5
## # Key:     sub_region_2 [1]
##   sub_region_2   SNaive                                arima_man
##   <chr>         <model>                                  <model>
## 1 Málaga       <SNAIVE> <LM w/ ARIMA(2,1,2)(1,1,1)[7] errors>
## # ... with 2 more variables: arima_at1 <model>, arima_at2 <model>
```

```
report(fit_model)
```

```
## # A tibble: 4 x 9
##   sub_region_2 .model    sigma2 log_lik   AIC  AICc   BIC ar_roots   ma_roots
##   <chr>        <chr>      <dbl>   <dbl> <dbl> <dbl> <dbl> <list>     <list>
## 1 Málaga       SNaive     2991.      NA    NA    NA    NA  <NULL>     <NULL>
## 2 Málaga       arima_man   989.  -1266. 2560. 2562. 2610. <cpl [9]>  <cpl [9]>
## 3 Málaga       arima_at1   967.  -1298. 2620. 2621. 2663. <cpl [14]> <cpl [8]>
## 4 Málaga       arima_at2   901.  -1288. 2604. 2605. 2654. <cpl [8]>  <cpl [10]>
```

```
# Good model >> Less Sigma - More BIC or AIC
fit_model %>% pivot_longer(!sub_region_2,
                           names_to = "Model name",
                           values_to = "Orders")
```

```
## # A mable: 4 x 3
## # Key:     sub_region_2, Model name [4]
##   sub_region_2 `Model name`                                  Orders
##   <chr>        <chr>                                        <model>
## 1 Málaga       SNaive                                      <SNAIVE>
## 2 Málaga       arima_man    <LM w/ ARIMA(2,1,2)(1,1,1)[7] errors>
## 3 Málaga       arima_at1    <LM w/ ARIMA(0,1,1)(2,0,1)[7] errors>
## 4 Málaga       arima_at2    <LM w/ ARIMA(1,1,3)(1,0,1)[7] errors>
```

```
glance(fit_model) %>% arrange(AICc) %>% select(.model:BIC)
```

```
## # A tibble: 4 x 6
##   .model    sigma2 log_lik   AIC  AICc   BIC
##   <chr>      <dbl>   <dbl> <dbl> <dbl> <dbl>
## 1 arima_man   989.  -1266. 2560. 2562. 2610.
## 2 arima_at2   901.  -1288. 2604. 2605. 2654.
## 3 arima_at1   967.  -1298. 2620. 2621. 2663.
## 4 SNaive     2991.     NA    NA    NA    NA
```

```
# We use a Ljung-Box test >> large p-value, confirmins residuals are similar to white noise.
augment(fit_model) %>%
  features(.innov, ljung_box, lag=7)
```

```
## # A tibble: 4 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>       <dbl>     <dbl>
## 1 Málaga       arima_at1   26.8   0.000360
## 2 Málaga       arima_at2    8.77  0.269
## 3 Málaga       arima_man   22.6   0.00202
```

```
## 4 Málaga       SNaive     521.    0
```

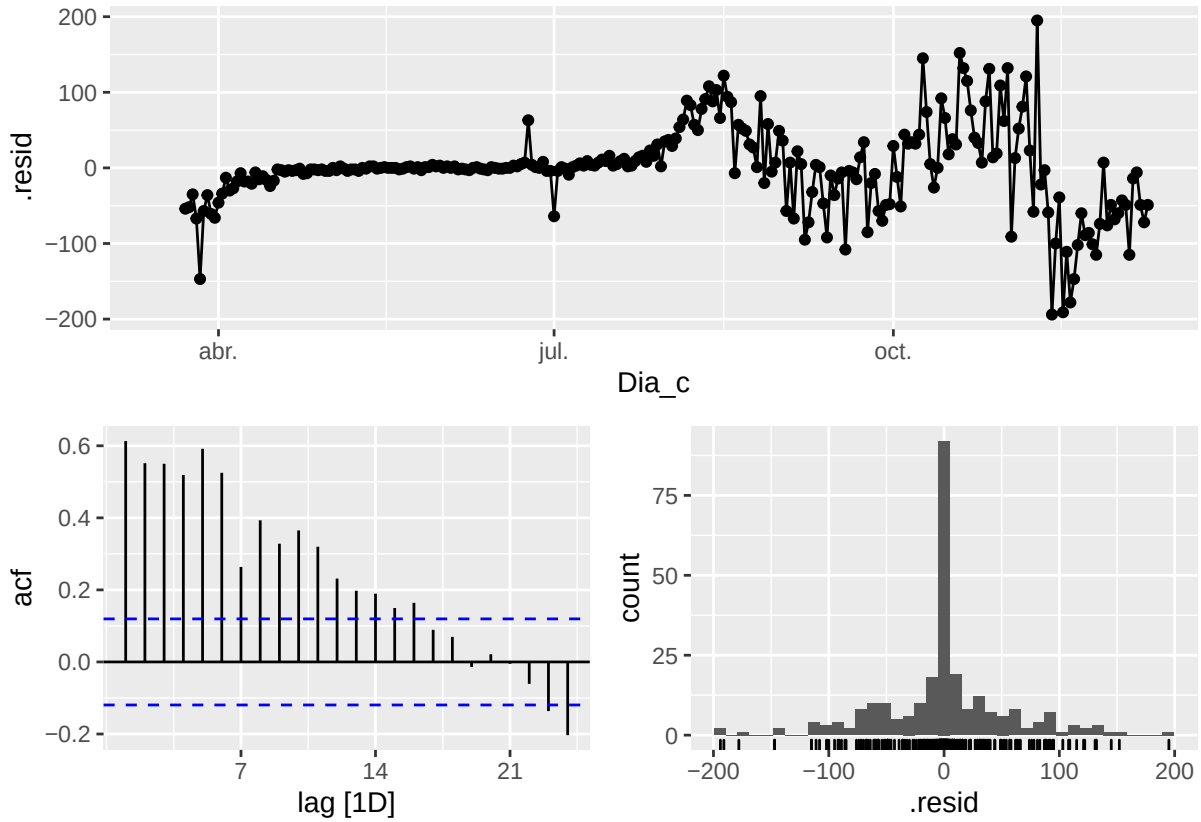
```
augment(fit_model) %>%
  features(.innov, ljung_box, lag=14)
```

```
## # A tibble: 4 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>       <dbl>     <dbl>
## 1 Málaga       arima_at1    34.5   0.00174
## 2 Málaga       arima_at2    16.7   0.272
## 3 Málaga       arima_man    31.1   0.00535
## 4 Málaga       SNaive      693.    0
```

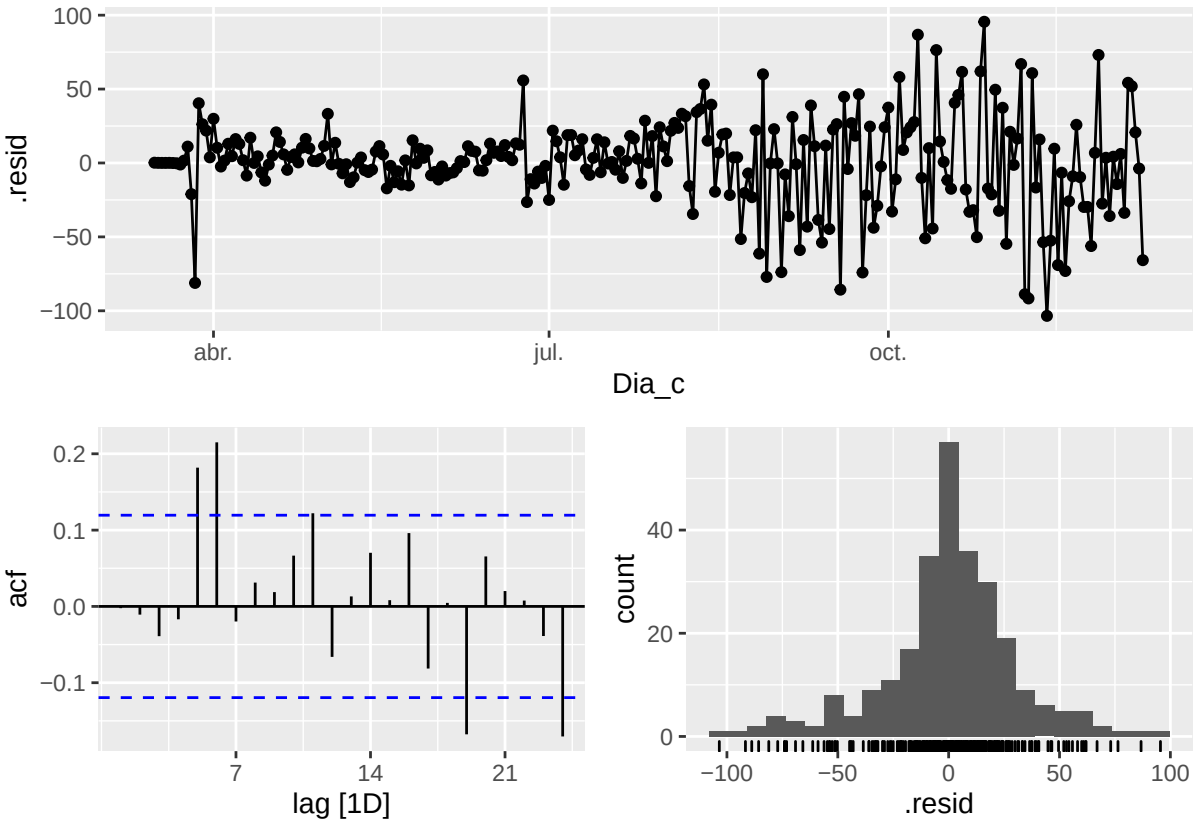
```
augment(fit_model) %>%
  features(.innov, ljung_box, lag=21)
```

```
## # A tibble: 4 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>       <dbl>     <dbl>
## 1 Málaga       arima_at1    48.3  0.000623
## 2 Málaga       arima_at2    31.1  0.0723
## 3 Málaga       arima_man    45.3  0.00159
## 4 Málaga       SNaive      711.   0
```

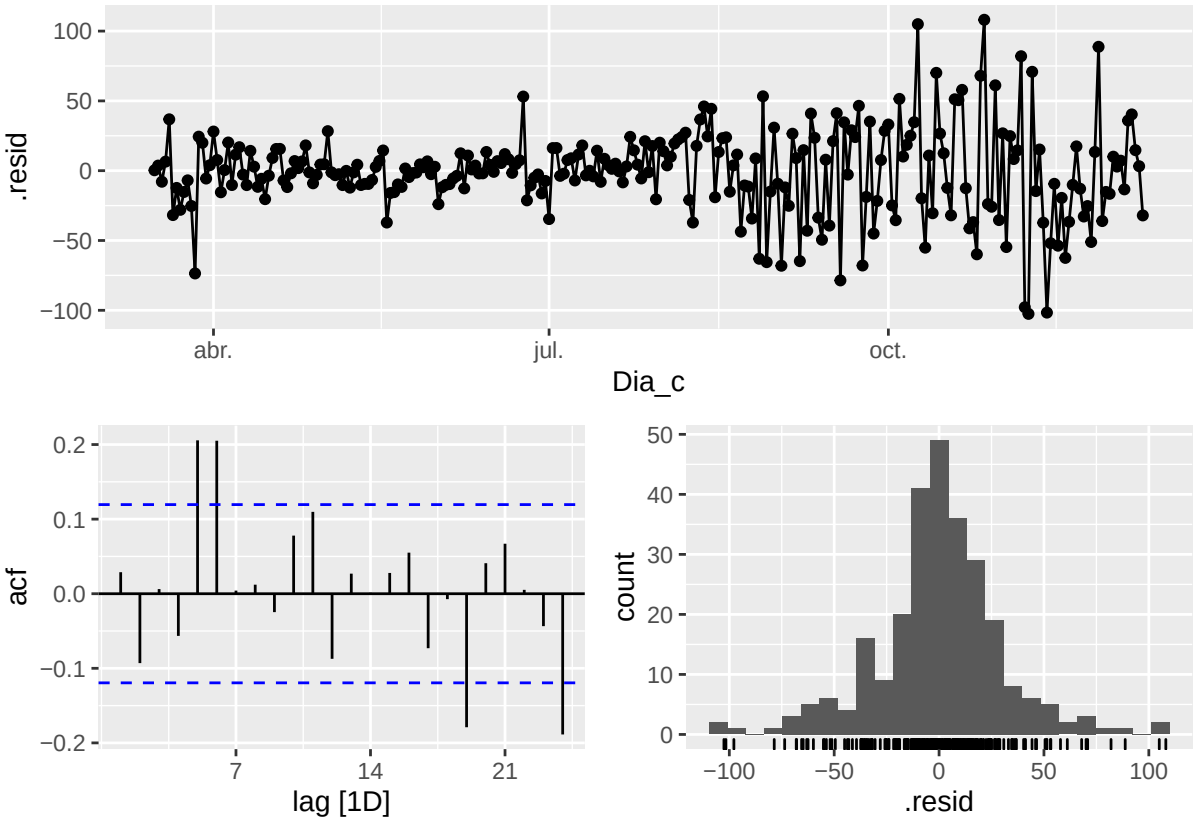
```
fit_model %>% select(SNaive) %>% gg_tsresiduals()
```



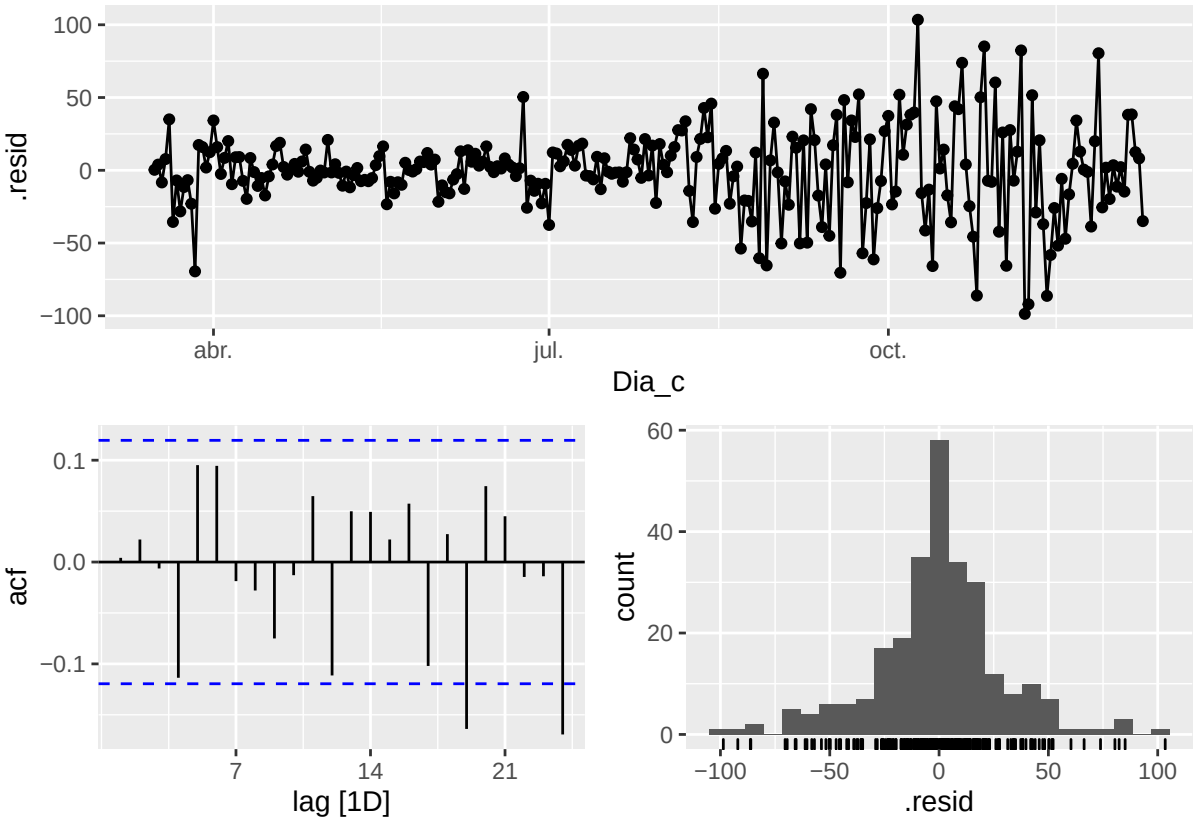

```
fit_model %>% select(arima_man) %>% gg_tsresiduals()
```




```
fit_model %>% select(arima_at1) %>% gg_tsresiduals()
```




```
fit_model %>% select(arima_at2) %>% gg_tsresiduals()
```




```
# Significant spikes out of 30 is still consistent with white noise.
# To be sure, use a Ljung-Box test, which has a large p-value, confirming that the
# residuals are similar to white noise.
# Note that the alternative models also pass this test.

# New data (dynamic regression)
# Here it is needed generate future values for the exogenous variables
# Foir simplicity we select a rand number included into the 2nd and 3rd quantile for
# the variable
# h7
Mal_N_cases_fr7 <- new_data(Mal_N_cases_tr, 7) %>%
  mutate(retail_and_recreation_percent_change_from_baseline =
           runif(7,quantile(Mal_N_cases_tt$retail_and_recreation_percent_change_from_baseline,
                            0.25),
                 quantile(Mal_N_cases_tt$retail_and_recreation_percent_change_from_baseline,
                          0.75)),
         grocery_and_pharmacy_percent_change_from_baseline =
           runif(7,quantile(Mal_N_cases_tt$grocery_and_pharmacy_percent_change_from_baseline,
                            0.25),
                 quantile(Mal_N_cases_tt$grocery_and_pharmacy_percent_change_from_baseline,
                          0.75)),
         parks_percent_change_from_baseline =
           runif(7,quantile(Mal_N_cases_tr$parks_percent_change_from_baseline,
                            0.25),
                 quantile(Mal_N_cases_tt$parks_percent_change_from_baseline,
                          0.75)),
```

```
          transit_stations_percent_change_from_baseline =
            runif(7,quantile(Mal_N_cases_tt$transit_stations_percent_change_from_baseline,
                             0.25),
                    quantile(Mal_N_cases_tt$transit_stations_percent_change_from_baseline,
                             0.75)),
          workplaces_percent_change_from_baseline =
            runif(7,quantile(Mal_N_cases_tt$workplaces_percent_change_from_baseline,
                             0.25),
                    quantile(Mal_N_cases_tt$workplaces_percent_change_from_baseline,
                             0.75)),
          residential_percent_change_from_baseline =
            runif(7,quantile(Mal_N_cases_tt$residential_percent_change_from_baseline,
                             0.25),
                    quantile(Mal_N_cases_tt$residential_percent_change_from_baseline,
                             0.75)),
          Total = runif(7,quantile(Mal_N_cases_tt$Total,0.25),
                        quantile(Mal_N_cases_tt$Total,0.75)))

# h14
Mal_N_cases_fr14 <- new_data(Mal_N_cases_tr, 14) %>%
  mutate(retail_and_recreation_percent_change_from_baseline =
            runif(14,quantile(Mal_N_cases_tt$retail_and_recreation_percent_change_from_baseline,
                              0.25),
                    quantile(Mal_N_cases_tt$retail_and_recreation_percent_change_from_baseline,
                             0.75)),
          grocery_and_pharmacy_percent_change_from_baseline =
            runif(14,quantile(Mal_N_cases_tt$grocery_and_pharmacy_percent_change_from_baseline,
                              0.25),
                    quantile(Mal_N_cases_tt$grocery_and_pharmacy_percent_change_from_baseline,
                             0.75)),
          parks_percent_change_from_baseline =
            runif(14,quantile(Mal_N_cases_tt$parks_percent_change_from_baseline,
                              0.25),
                    quantile(Mal_N_cases_tt$parks_percent_change_from_baseline,
                             0.75)),
          transit_stations_percent_change_from_baseline =
            runif(14,quantile(Mal_N_cases_tt$transit_stations_percent_change_from_baseline,
                              0.25),
                    quantile(Mal_N_cases_tt$transit_stations_percent_change_from_baseline,
                             0.75)),
          workplaces_percent_change_from_baseline =
            runif(14,quantile(Mal_N_cases_tt$workplaces_percent_change_from_baseline,
                              0.25),
                    quantile(Mal_N_cases_tt$workplaces_percent_change_from_baseline,
                             0.75)),
          residential_percent_change_from_baseline =
            runif(14,quantile(Mal_N_cases_tt$residential_percent_change_from_baseline,
                              0.25),
                    quantile(Mal_N_cases_tt$residential_percent_change_from_baseline,
                             0.75)),
          Total = runif(14,quantile(Mal_N_cases_tt$Total,0.25),
                        quantile(Mal_N_cases_tt$Total,0.75)))
```

```r
# h21
Mal_N_cases_fr21 <- new_data(Mal_N_cases_tr, 21) %>%
  mutate(retail_and_recreation_percent_change_from_baseline =
           runif(21,quantile(Mal_N_cases_tt$retail_and_recreation_percent_change_from_baseline,
                             0.25),
                 quantile(Mal_N_cases_tt$retail_and_recreation_percent_change_from_baseline,
                          0.75)),
         grocery_and_pharmacy_percent_change_from_baseline =
           runif(21,quantile(Mal_N_cases_tt$grocery_and_pharmacy_percent_change_from_baseline,
                             0.25),
                 quantile(Mal_N_cases_tt$grocery_and_pharmacy_percent_change_from_baseline,
                          0.75)),
         parks_percent_change_from_baseline =
           runif(21,quantile(Mal_N_cases_tt$parks_percent_change_from_baseline,
                             0.25),
                 quantile(Mal_N_cases_tt$parks_percent_change_from_baseline,
                          0.75)),
         transit_stations_percent_change_from_baseline =
           runif(21,quantile(Mal_N_cases_tt$transit_stations_percent_change_from_baseline,
                             0.25),
                 quantile(Mal_N_cases_tt$transit_stations_percent_change_from_baseline,
                          0.75)),
         workplaces_percent_change_from_baseline =
           runif(21,quantile(Mal_N_cases_tt$workplaces_percent_change_from_baseline,
                             0.25),
                 quantile(Mal_N_cases_tt$workplaces_percent_change_from_baseline,
                          0.75)),
         residential_percent_change_from_baseline =
           runif(21,quantile(Mal_N_cases_tt$residential_percent_change_from_baseline,
                             0.25),
                 quantile(Mal_N_cases_tt$residential_percent_change_from_baseline,
                          0.75)),
         Total = runif(21,quantile(Mal_N_cases_tt$Total,0.25),
                       quantile(Mal_N_cases_tt$Total,0.75)))

# Forecast
fc_fh7<-fabletools::forecast(fit_model, new_data = Mal_N_cases_fr7)
fc_fh14<-fabletools::forecast(fit_model, new_data = Mal_N_cases_fr14)
fc_fh21<-fabletools::forecast(fit_model, new_data = Mal_N_cases_fr21)

# Accuracy
fabletools::accuracy(fc_fh7, Mal_N_cases)
```

```
## # A tibble: 4 x 11
##   .model    sub_region_2 .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE    ACF1
##   <chr>     <chr>        <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl>
## 1 arima_at1 Málaga       Test   47.6  53.6  47.6  26.8  26.8  1.35 0.981  0.0240
## 2 arima_at2 Málaga       Test   37.7  43.8  37.7  21.1  21.1  1.07 0.802 -0.0677
## 3 arima_man Málaga       Test   57.7  63.1  57.7  34.4  34.4  1.64 1.16  -0.157
## 4 SNaive    Málaga       Test   48.3  57.7  48.3  27.9  27.9  1.37 1.06   0.412
```

```r
fabletools::accuracy(fc_fh14, Mal_N_cases)
```

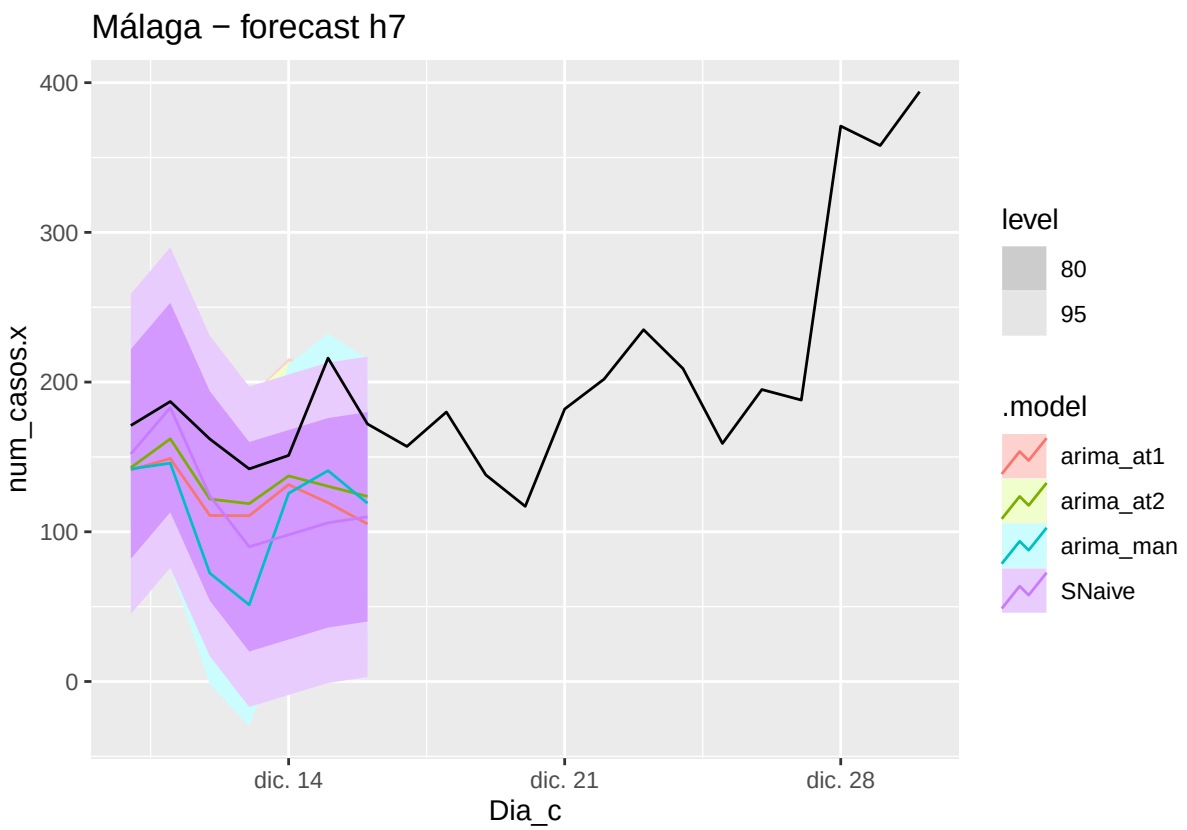
```
## # A tibble: 4 x 11
```

```
##   .model    sub_region_2 .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE  ACF1
##   <chr>     <chr>        <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Málaga       Test   70.9  83.7  73.2  39.6  40.8  2.08  1.53 0.221
## 2 arima_at2 Málaga       Test   59.1  70.1  61.1  33.0  34.1  1.74  1.28 0.196
## 3 arima_man Málaga       Test   80.6  91.0  82.7  47.8  49.0  2.35  1.67 0.136
## 4 SNaive    Málaga       Test   49    63.3  49.4  26.9  27.2  1.40  1.16 0.516
```

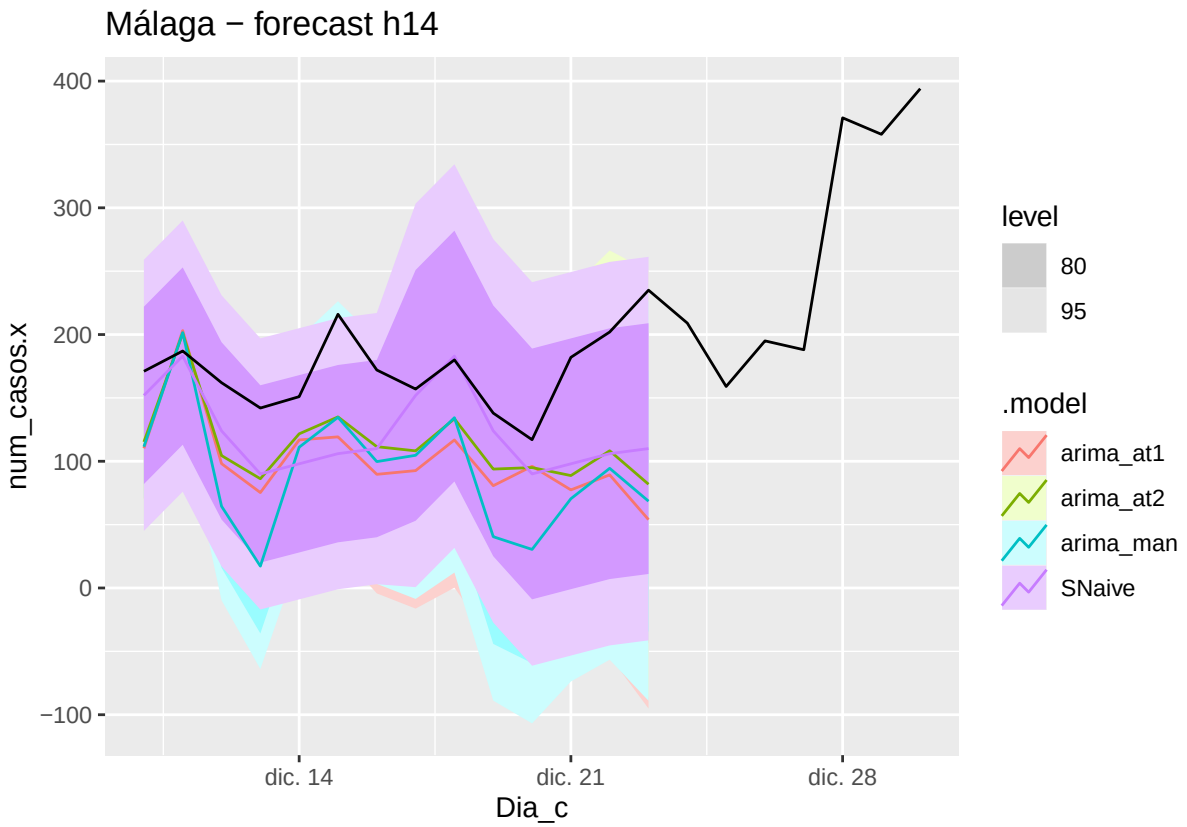
```
fabletools::accuracy(fc_fh21, Mal_N_cases)
```

```
## # A tibble: 4 x 11
##   .model    sub_region_2 .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE  ACF1
##   <chr>     <chr>        <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Málaga       Test  110.   142. 111.   48.7  49.1  3.16  2.60 0.625
## 2 arima_at2 Málaga       Test   97.7  130.  98.7  42.2  42.8  2.81  2.39 0.638
## 3 arima_man Málaga       Test  124.   153. 125.   57.6  58.2  3.55  2.81 0.636
## 4 SNaive    Málaga       Test   80.8  118.  83.4  33.0  34.6  2.37  2.16 0.637
```

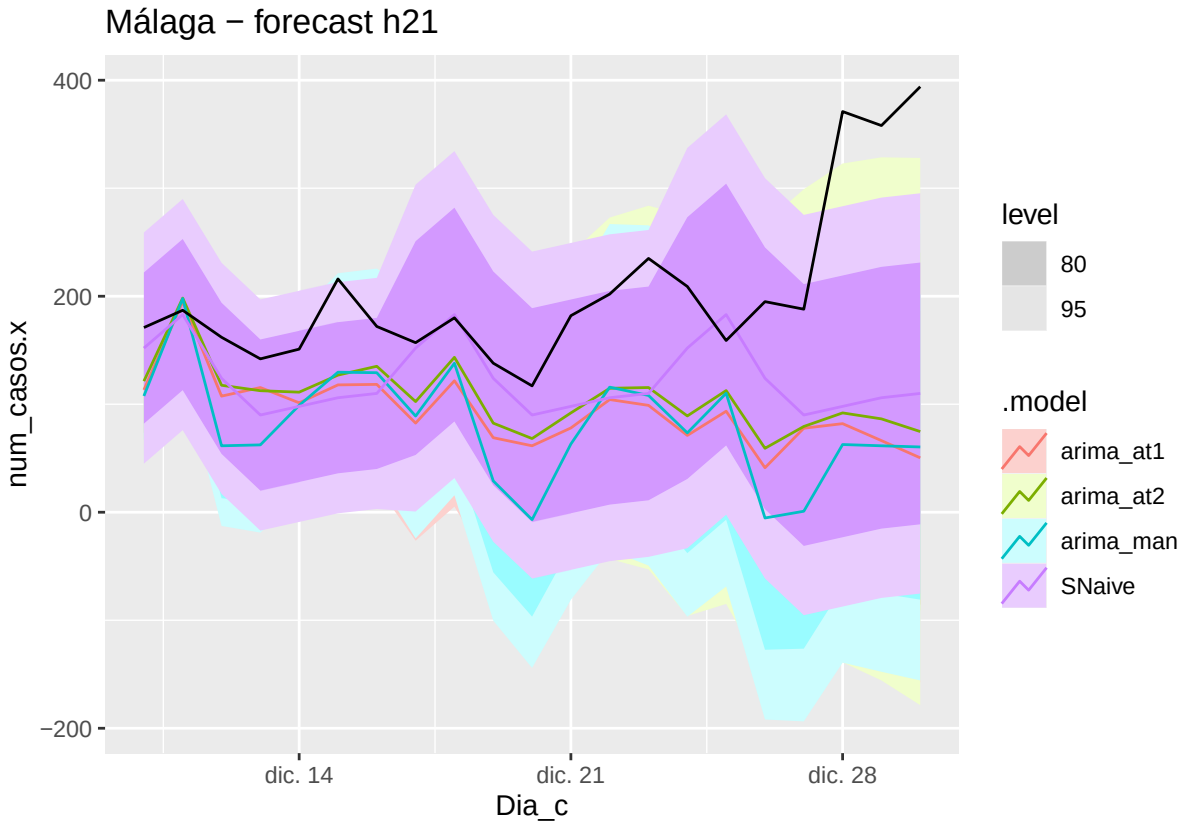
```
# Plots
fc_fh7 %>%
  #filter(.model=='arima_at1'|.model=='arima_at2'|.model=='SNaive') %>%
  autoplot(Mal_N_cases_tt) +
  labs(title="Málaga - forecast h7")
```




```
fc_fh14 %>%
  #filter(.model=='arima_at1'|.model=='arima_at2'|.model=='SNaive') %>%
  autoplot(Mal_N_cases_tt) +
  labs(title="Málaga - forecast h14")
```




```
fc_fh21 %>%
  #filter(.model=='arima_at1'|.model=='arima_at2'|.model=='SNaive') %>%
  autoplot(Mal_N_cases_tt) +
  labs(title="Málaga - forecast h21")
```



3.4 Till here 16-Apr-2021

Bibliography

Baayen, R Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press.

Hothorn, Torsten, and Brian S Everitt. 2014. *A Handbook of Statistical Analyses Using R*. CRC press.

Hyndman, Rob J., and George Athanasopoulos. 2021. "Forecasting: Principles and Practice, 3rd." OTexts: Melbourne, Australia. OTexts.com.

Liviano Solas, Daniel, and Maria Pujol Jover. n.d. *Analisis de Datos Y Estadastica Descriptiva Con R Y R-Commander*. UOC.

Teetor, Paul. 2011. *R Cookbook: Proven Recipes for Data Analysis, Statistics, and Graphics*. O'Reilly Media, Inc.

Vegas Lozano, Esteban. n.d. *Preprocesamiento de Los Datos*. UOC.