UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

Escola Tècnica Superior d'Enginyeria Informàtica

# Hotel Occupancy

**Team 8**

Data Science Degree, Project III

**Delivery Date**:
15/05/2021

**Authors**:
Andrea García Pastor
Cándido García Rodríguez
Júlia Gregori Torres
Lluna Pérez Pérez
Clara Salelles Roman
Arturo Serrano Moliner

# Table of contents

-poner las bases de datos que buscamos
-poner lo de latitudes y longitudes
-¿? una mejor explicación del ajuste de los modelos a los datos
-revisar imagenes
-evaluación del valor --> valor real de los mapas/ proyecto
-poner legado mejor en memoria, hablar de las lecciones aprendidas, y crear repositorio github con código diciendo que se podría aplicar para otros tipos de negocios
-revisar anexos y borrar lo no necesario y referenciarlos
-RESUMEN al inicio
-reestructurar todo/ escribir algunas cosas mejor

# 1. Introduction

## 1.1 Relevance of project

An investor who wants to build a hotel in Spain in 2023, decides to contact a group of young data analysts to help him decide where to install the establishment in order to have the largest possible number of customers and, therefore, obtain the maximum amount of profit.

In order to carry out the question posed by the investor, both territorial factors (autonomous communities, provinces...) and demographic factors (travelers in the corresponding area) are taken into account. Once the community and the province have been chosen according to the increase in the number of total travellers (both national and foreign), it is necessary to delimit the number of possible tourist areas and points in that province.

To do this, a prediction will be made of the most interesting variables that we will study in depth during the project and the area or point that best suits the client's interests will be obtained. Finally, a report will be elaborated with different aspects that will be useful, such as the possible competition.

## 1.2 Project goals

As previously mentioned, the objective of the study is to be able to predict where it is most appropriate to install a hotel establishment. However, it is necessary to study the behavior of the different locations in previous years in order to determine the anomalies of hotel occupancy over time.

**The main objective of the project is:** to predict a future hotel occupancy that is able to satisfy the needs and maximize the profits of the investor. To this end, we will identify anomalies and trends in hotel occupancy along with other useful variables, in order to forecast a possible optimal strategy for opening a new hotel.

However, it should be noted that with the development and evolution of the project, the initially proposed objectives were not maintained. The different objectives that the group has been

able to contemplate during the progress of the study can be found in more detail in *Annex 1: Goals and values*.

# 2. Data and methodology

## 2.1 Data collection procedure

Firstly, as we had issues with our data as we hadn't valuable data related with the economy of the hotels, or about other data related to transport

The data with which the project will be developed come from the web repository of the National Institute of Statistics (INE), especially those related to the hotel occupancy survey. To achieve our objective we decided to create 4 new datasets:

**Dataset 1: Types of accommodation**

To create this database we have put together the following datasets:

- **2.1**: Travellers, overnight stays by type of accommodation by Autonomous Communities and Cities.
- **2.2**: Average stay, by type of accommodation, by Autonomous Communities and Autonomous Cities.
- **2.3**: Estimated establishments, estimated bedplaces and staff employed by type of accommodation and by Autonomous Communities and Cities.

This dataset allows us to differentiate 4 types of tourism: rural tourism, Airbnb apartments, Camping and hotels. Thanks to this dataset we will discover the occupancy differences between each type of accommodation and we will be able to analyze how the number of travelers varies. In this way, we will be able to create a model that fits the influence of the type of accommodation that will be studied later. This won't be a main objective for us, but an additional information that could help the investor decide where to invest.

Next, we create 3 datasets that present the same information but are collected with greater granularity. We go from Autonomous Communities and Provinces, Zones and Tourists Spots. When obtaining the data from INE we did not have the option of a pre-aggregated base in which the hierarchy was defined with which to work later. Therefore, we downloaded 3 different ones:

**Dataset 2: Zones**

The resulting database of zones, explained below, has been formed with the union of the following datasets:

- **1.3**: Travellers and overnight stays by tourist areas
- **1.9**: Average stay by tourist areas
- **1.13**: Establishments, estimated bedplaces, occupancy rates and staff employed by tourist areas

**Dataset 3: Tourist spots**

To create this database we have put together the following datasets:

- **1.4**: Travellers and overnight stays by tourist spots

- **1.10**: Average stay by tourist spots
- **1.14**: Establishments, estimated bedplaces, occupancy rates and staff employed by tourist spots

## Dataset 4: Communities and provinces

This database is made up of the following databases extracted from the INE:

- **1.2**: Travellers and overnight stays by autonomous communities and provinces
- **1.8**: Average stay by autonomous community and province.
- **1.12**: Establishments, bedplaces, occupancy rates and staff employed by Autonomous Community and province.

In addition, within these three data sets there are different aggregations of categorical variables related to geographic location. For a more detailed definition of the above variables, both categorical and numerical, please refer to *Annex 2:Data*.

| DATASET | ROWS | VARIABLES |
|---|---|---|
| **Types of accommodation** | 5472 | 11 |
| **Zones** | 2220 | 17 |
| **Tourist spots** | 6300 | 17 |
| **Communities and provinces** | 3744 | 15 |

*Figure 2.1 Table of number of rows and variables of datasets*

It's important to choose a good methodology that fits the needs of our project. In our case we believe that the most convenient is CRISP-DM, which has six stages:

- **Business Understanding**: this phase focuses on understanding the client's needs. It focuses mainly on defining clear objectives and then defining a data mining problem and a preliminary plan designed to achieve the defined objectives. In *Annex 3: Domain, knowledge, context and innovation* we can find in more depth the team's perception of the proposed problem.

- **Data Understanding:** prior to modelling, it is necessary to understand the data with which we are going to work. In general terms, in this phase, data related to the aspects already mentioned for areas, tourist spots and communities and provinces are collected. These will allow the development of the project with the aim of fulfilling the main objective, which is to help the investor to locate the opening of a hotel with the idea of maximising its benefits. Likewise, the same information is also collected by types of accommodation, information that may be of interest to the investor.

- **Data Preparation:** once the data is understood, it is important to make the necessary modifications in order to prepare the data sets for the realization of the model. In our case, and in general, we have proceeded to extract the variables that interest us and to eliminate the *'total _'* variables in order to convert the classes belonging to the *'total _'* characteristic into new and different attributes. This process can be seen in more detail in *Annex 4:Data preparation*.

The remaining modeling, evaluation and deployment phases will be discussed in detail throughout the report. However, it should be noted that the development of the methodology used can be found in depth in *Annex 5: Methodology*.

## 2.2 Mineable View

As mentioned in the previous section, it's very important that the data layout is adequate in order to carry out a good modeling of the problem and, consequently, a good predictive model.

In addition, commented in a general way above, we have mainly transformed the categorical columns that we consider important, into new variables that encompass the numerical data of each category. If we focus individually on the transformed data, the modifications have been similar and repetitive, and all of them end with a general adjustment.

After the creation of each dataframe, we proceeded to clean them, in order to train the models and create maps available to the client to better transmit the information. As the data was provided by the INE, the presence of outliers was null, focusing our cleaning mainly on the omission of years prior to 2015 and after 2019 (not including both).

We will introduce the necessary variables into the time series models in order to cluster the results and to be able to make more complete forecasts. The desired model is in particular a gruped time series.

Before getting to know the respective input and output variables for each dataset, it is important to note that, in order to be able to work with the time series, the variable "Periodo" has been split into "Año" and "Mes". Furthermore, the ordering of the data is based on the increasing order of the years followed by the months, i.e. the first observation of the dataset will be January 2015, consecutively it will be February 2015 and so on until December 2019.

Once the transformations are known, we proceed to determine the corresponding minable view. The input variables we have used are the period and the zones and the output, the occupancy rate. This prediction will illustrate whether the ratio of overnight stays to the number of places offered (number of nights/places) in each area or point has increased or decreased. This is interesting because the higher the number of overnight stays, the higher the percentage. This means that the number of hotel rooms and nights spent by the traveller(s) is, on average, higher and therefore the benefits will be higher.

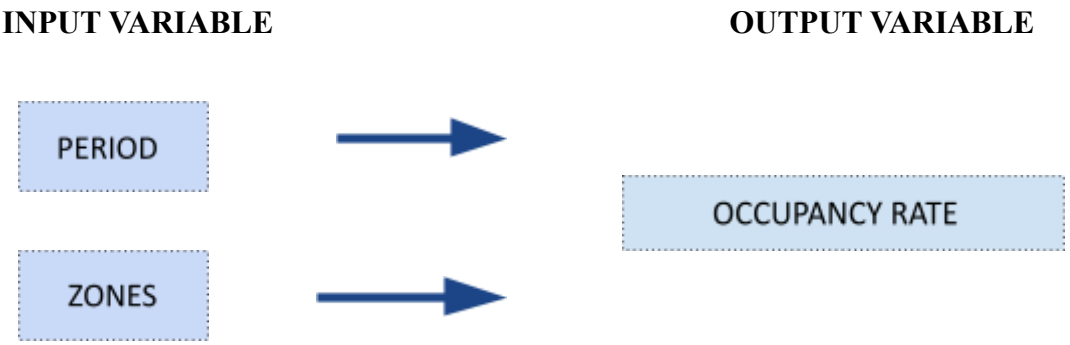


*Figure 2.2 Model Schema*

For information purposes, we will also have as output variables the number of spanish and foreign travellers, since it may be of interest to the client. However, this is not our main objective, so its explanation is developed in the annexes section.

# 3. Models and results

As the main objective is to find a model to predict the occupancy rate (a numerical variable) for all zones and touristic spots depending on the period, it has been decided to make a model of grouped time series. For this purpose, several types with different parameters have been considered as possible answers.

For training purposes, the data has been divided into two parts:

- **Training**: periods from 2015 to 2018. An imputation of missing data has been made with seasonal splits.
- **Test**: periods from 2019.

After several tests, the models that best fit the data were:

- **ETS** with seasonality and trend.
- **ARIMA** with seasonality and trend.

Anyway we have also tried other models such as TSLM and SNAIVE, but in general the best predictions are made with ETS and ARIMA models.

The following metrics will be used to assess them:

| | |
|---|---|
| **MAE** | Average of prediction errors |
| **MAPE** | Mean absolute percent error for each time period. |
| **RMSE** | Root mean square error. The scale of the errors is equal to the scale of the targets. |

## 3.1 Zones

As the model has to be properly adjusted for all areas of Spain, we will look at it in aggregate. In this case, the **ETS** model is the best.

Evaluating the models, ETS and ARIMA, by means of different metrics, we observed that the mean absolute percentage error (MAPE) is higher in the case of ARIMA, since it is 3.29%, while in ETS it is 2.5%. Therefore, we can state that in the first model indicated, the predictions are 0.8% less far from reality with respect to the second model. Analyzing the root mean square error (RMSE), we reaffirm that the ETS model gives better results, since it has a value of 1.41%, the predicted observations are the indicated percentage farther from the total observed observations, using as a reference the occupancy rate per vacancy, while in ARIMA it is 2%, so that the predicted and observed values are farther apart.

| .model | Comunidad | Zonas | MAE | MAPE | RMSE | ACF1 |
|---|---|---|---|---|---|---|
| ets | *<aggregated>* | *<aggregated>* | 1.248750 | 2.504207 | 1.408170 | 0.02830193 |
| arima | *<aggregated>* | *<aggregated>* | 1.647672 | 3.286925 | 1.998388 | -0.06356689 |

*Figure 3.1 Global performance of each model*

Now we have generated the metrics for all communities, as we would like to see how well the model adjusts to them and how much we should trust the predictions for a specific community.

We can see that the mean percentage error goes from 1.26% in Canarias with the ETS model up to 15.31% in Murcia using ARIMA.

| .model | Comunidad | Zonas | MAE | MAPE | RMSE | ACF1 |
|---|---|---|---|---|---|---|
| ets | *Canarias* | *<aggregated>* | 1.261027 | 1.775008 | 1.654378 | 0.135850673 |
| ets | *Cataluña* | *<aggregated>* | 1.497481 | 3.490683 | 1.776836 | -0.120093056 |
| arima | *Andalucía* | *<aggregated>* | 1.375325 | 2.883594 | 1.789114 | 0.177001215 |
| ets | *Andalucía* | *<aggregated>* | 1.439168 | 3.020872 | 1.799974 | 0.112721628 |
| ets | *Galicia* | *<aggregated>* | 1.641690 | 5.413395 | 1.890885 | -0.031477704 |
| arima | *Cataluña* | *<aggregated>* | 1.770027 | 4.013787 | 1.965815 | 0.010073675 |
| arima | *Galicia* | *<aggregated>* | 1.575821 | 5.775815 | 2.080269 | 0.082232086 |
| ets | *País Vasco* | *<aggregated>* | 1.704569 | 3.648491 | 2.267772 | -0.119270336 |
| arima | *Extremadura* | *<aggregated>* | 2.124940 | 6.898470 | 2.384955 | -0.062351044 |
| arima | *Comunitat Valenciana* | *<aggregated>* | 1.998786 | 3.227736 | 2.746559 | 0.465105326 |
| ets | *Pirineos* | *<aggregated>* | 2.757096 | 10.362880 | 2.969278 | -0.315629036 |
| ets | *Comunitat Valenciana* | *<aggregated>* | 2.271839 | 3.666016 | 3.011133 | 0.436956956 |
| arima | *Aragón* | *<aggregated>* | 2.447432 | 7.019791 | 3.017041 | -0.005686549 |
| arima | *Canarias* | *<aggregated>* | 2.435702 | 3.446285 | 3.045649 | 0.553829663 |
| arima | *Pirineos* | *<aggregated>* | 2.772409 | 11.094953 | 3.258615 | -0.362296823 |
| ets | *Balears, Illes* | *<aggregated>* | 2.617014 | 5.492946 | 3.458123 | -0.280509659 |
| ets | *Aragón* | *<aggregated>* | 2.592334 | 7.164609 | 3.506568 | 0.066738699 |
| ets | *Extremadura* | *<aggregated>* | 3.274537 | 9.099026 | 3.732859 | -0.011687204 |
| arima | *País Vasco* | *<aggregated>* | 3.279305 | 7.549176 | 3.784753 | -0.213758718 |
| ets | *Asturias, Principado de* | *<aggregated>* | 3.295772 | 8.654836 | 3.852580 | -0.157013300 |
| ets | *Navarra, Comunidad Foral de* | *<aggregated>* | 3.558879 | 14.367188 | 4.154257 | 0.328054066 |
| arima | *Asturias, Principado de* | *<aggregated>* | 3.569760 | 9.961515 | 4.226420 | 0.056689935 |
| arima | *Balears, Illes* | *<aggregated>* | 3.567415 | 7.437996 | 4.651234 | -0.378344734 |
| arima | *Navarra, Comunidad Foral de* | *<aggregated>* | 4.229476 | 16.916366 | 4.863481 | 0.292228356 |
| ets | *Murcia, Región de* | *<aggregated>* | 4.523881 | 9.204708 | 5.747375 | -0.090455254 |
| arima | *Murcia, Región de* | *<aggregated>* | 7.378671 | 15.310300 | 8.205418 | -0.141400381 |

*Figure 3.2 Performance of the model for each community*

As the **ETS** model is the one that performs best in general for predicting occupancy rates for the Zones dataset, we'll use it as our predictive model. Then we calculate the prediction until the year 2023 for all the zones in Spain. In this case we will introduce to our model all of our data (2015-2019) and we will impute all the missing values as we have done before with the training set, but now with all the data set.

In *figure 3.3* we can see the model predictions with 80% and 95% confidence intervals for each zone. The first plot corresponds to the aggregated level for all the times series, so we recommend that the reader to ignore it.

We can observe that the predictions, as expected, are very seasonal. They should be useful if we observe a trend on them. We could say, for example, that the predictions for Rias Baixas estimate that the occupancy rate has a positive linear trend. This means that occupancy rate will be greater for 2023 than for 2019. The opposite happens in the North of Extremadura, where the linear trend is negative.

We can also see that there are zones whose predictions remain the same and are estational (Costa de la Luz de Cádiz). For those zones, there won't be a change in the occupation rate.

A large confidence interval in our predictions will indicate us to be less confident of our predictions. The results could increase, decrease or stay the same in the interval shown, so it won't be as accurate as we would like them to be.

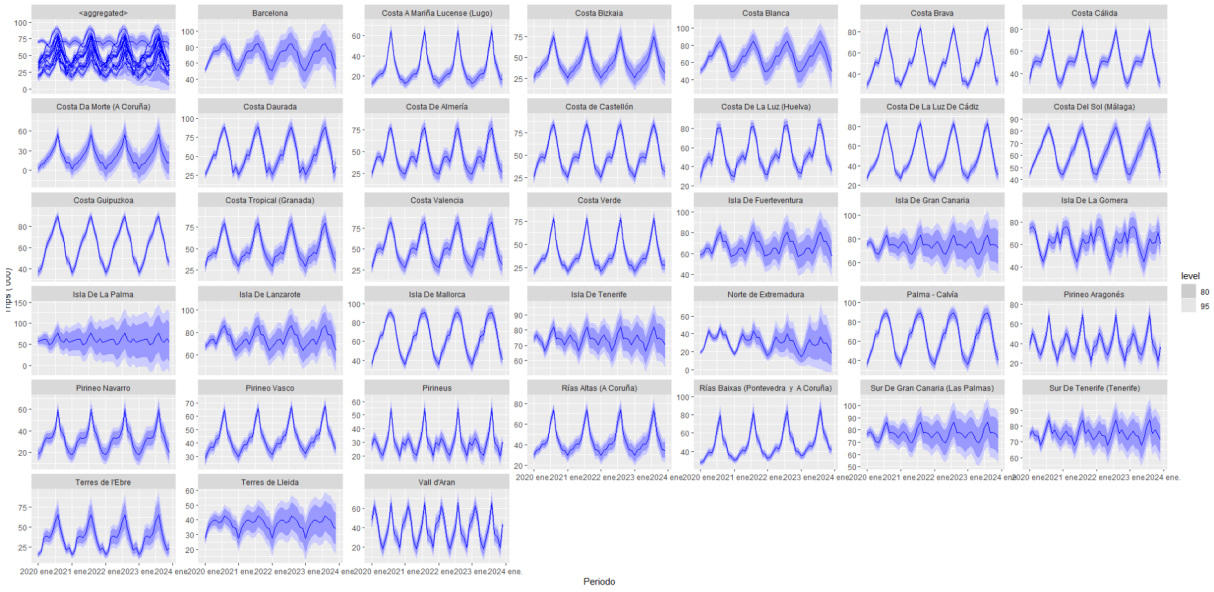


*Figure 3.3 Prediction of Occupancy rate*

To see the results of the model carried out for the dataset of zones choosing as a variable the values of both foreign and Spanish travellers, see Annex 7:Foreing travelers.

## 3.2 Touristic Spots

The procedure is the same as for the zones. Looking at the metrics for Spain, the model with the lowest error is again **ETS**. In all metrics (excluding AFC1 because it is not a precision metric), ARIMA has a superior error.

| | .model | Comunidad | Puntos turísticos | MAE | MAPE | RMSE | ACF1 |
|---|---|---|---|---|---|---|---|
| 1 | ets | *<aggregated>* | *<aggregated>* | 0.9468466 | 1.558375 | 1.215849 | -0.2285406 |
| 2 | arima | *<aggregated>* | *<aggregated>* | 2.1791028 | 3.349737 | 2.584953 | -0.1096294 |

*Figure 3.4 Global performance of each model*

If we focus on the models for each community, the result does not vary with respect to that obtained with the zones. Although the error range (taking into account the two models) is high (from 1.6% to 11.34%), the model that best fits each Spanish community is still **ETS**.

| | .model | Comunidad | Puntos turísticos | MAE | MAPE | RMSE | ACF1 |
|---|---|---|---|---|---|---|---|
| 1 | ets | Andalucía | <aggregated> | 0.9857893 | 1.630958 | 1.197566 | 0.063046825 |
| 2 | ets | Cataluña | <aggregated> | 0.9269452 | 1.594817 | 1.224576 | 0.220457900 |
| 3 | arima | Madrid, Comunidad de | <aggregated> | 1.2638244 | 1.911614 | 1.565103 | 0.130094570 |
| 4 | ets | Castilla-La Mancha | <aggregated> | 1.6280159 | 3.943348 | 1.863236 | 0.421133709 |
| 5 | ets | Canarias | <aggregated> | 1.5135510 | 2.125067 | 1.973678 | 0.423203273 |
| 6 | arima | Andalucía | <aggregated> | 1.6304077 | 2.674644 | 2.055221 | 0.203627945 |
| 7 | ets | Extremadura | <aggregated> | 1.6300506 | 3.608208 | 2.124642 | 0.314794612 |
| 8 | ets | Madrid, Comunidad de | <aggregated> | 1.6971065 | 2.643640 | 2.140897 | 0.094847755 |
| 9 | ets | Comunitat Valenciana | <aggregated> | 2.0617042 | 3.179731 | 2.227401 | 0.573332153 |
| 10 | ets | Galicia | <aggregated> | 1.8575297 | 3.914009 | 2.238643 | 0.210113052 |
| 11 | arima | Castilla-La Mancha | <aggregated> | 1.6628930 | 4.385052 | 2.259325 | 0.324861589 |
| 12 | arima | País Vasco | <aggregated> | 1.9142824 | 3.093560 | 2.333090 | 0.454097624 |
| 13 | ets | País Vasco | <aggregated> | 1.9590213 | 3.344026 | 2.481802 | 0.386562726 |
| 14 | arima | Rioja, La | <aggregated> | 2.3192312 | 4.381403 | 2.862149 | 0.183231885 |
| 15 | arima | Cataluña | <aggregated> | 2.3671732 | 3.350515 | 2.863129 | 0.145693530 |
| 16 | ets | Cantabria | <aggregated> | 2.5019088 | 5.223680 | 2.942527 | -0.378664849 |
| 17 | ets | Rioja, La | <aggregated> | 2.1191255 | 4.066870 | 3.089482 | 0.315216450 |
| 18 | ets | Aragón | <aggregated> | 2.5377786 | 5.789636 | 3.112504 | -0.003272021 |
| 19 | arima | Castilla y León | <aggregated> | 2.3380449 | 4.300455 | 3.202235 | 0.209733655 |
| 20 | arima | Asturias, Principado de | <aggregated> | 2.8457746 | 5.775361 | 3.379576 | -0.168454506 |
| 20 | arima | Asturias, Principado de | <aggregated> | 2.8457746 | 5.775361 | 3.379576 | -0.168454506 |
| 21 | arima | Galicia | <aggregated> | 2.7198501 | 5.731957 | 3.392739 | 0.170974900 |
| 22 | arima | Canarias | <aggregated> | 2.7945289 | 3.918609 | 3.475518 | 0.584519052 |
| 23 | ets | Navarra, Comunidad Foral de | <aggregated> | 2.6745474 | 4.851692 | 3.693794 | 0.180061599 |
| 24 | ets | Balears, Illes | <aggregated> | 3.1324050 | 4.887270 | 3.815897 | -0.622769647 |
| 25 | ets | Asturias, Principado de | <aggregated> | 3.5256115 | 6.961124 | 3.961183 | -0.279023288 |
| 26 | arima | Extremadura | <aggregated> | 3.4999289 | 7.715629 | 4.119905 | 0.377414177 |
| 27 | arima | Murcia, Región de | <aggregated> | 3.2541319 | 6.464858 | 4.216986 | -0.046832856 |
| 28 | arima | Navarra, Comunidad Foral de | <aggregated> | 3.3103694 | 5.655671 | 4.248612 | 0.152970992 |
| 29 | ets | Castilla y León | <aggregated> | 3.6136767 | 6.782323 | 4.350906 | 0.160022010 |
| 30 | arima | Comunitat Valenciana | <aggregated> | 3.4578842 | 5.249309 | 4.550145 | 0.573150355 |
| 31 | ets | Murcia, Región de | <aggregated> | 3.4865626 | 6.837722 | 4.611196 | -0.053739612 |
| 32 | arima | Cantabria | <aggregated> | 4.6581519 | 9.149852 | 5.040885 | -0.447002287 |
| 33 | arima | Balears, Illes | <aggregated> | 5.1153760 | 7.682369 | 5.518347 | -0.655636420 |
| 34 | arima | Aragón | <aggregated> | 5.1935836 | 11.343363 | 5.715775 | -0.093471572 |

*Figure 3.5 Performance of the model for each touristic spot*

We then calculated the predictions up to 2023, with data from 2015 to 2019. At this stage we have imputed the 2019 data for the missing data with a seasonal split, as they are part of the train set, not the test set, for the final forecasts.

As there are more tourist spots than zones, the graphs look smaller, but the interpretation is the same. The first sub-graph, which corresponds to the set of predictions for all the points, will not be used. We are more interested in seeing the predictions for each individual point.

As with the zones, the occupancy rate predictions are shown with confidence intervals of 80% and 95%, which means that the predictions can oscillate in that range with a very small error. Predictions with an increasing trend (Avila), decreasing (Santa Cruz de Tenerife), with a wide interval (Palencia) and predictions that remain stationary (Palma) are observed.

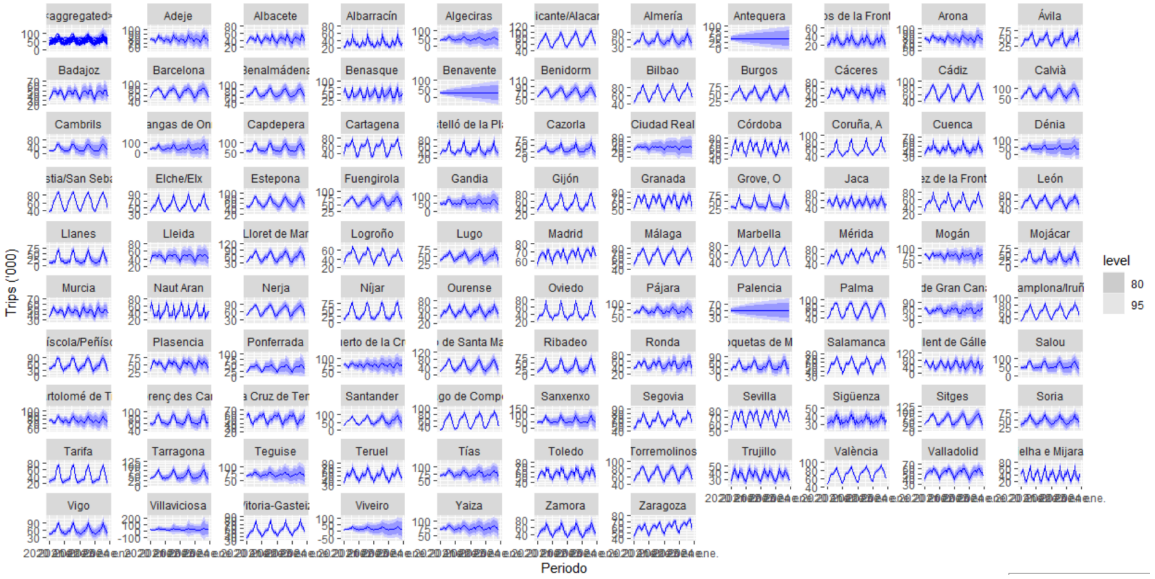


*Figure 3.6 Predictions of Occupancy rate*

# 4. Project Use

The two possible scenarios that we could have are:

- **The prediction of occupancy rate is higher than in reality**, which can lead to benefit loss.
- **The prediction of occupancy rate is lower than in reality**, which can trigger staff workload saturation, short supply of products or infrastructures (for instance with hotels, there can be less rooms prepared than needed, not enough food to feed the customers, not enough cleaning personnel). This can start off client insatisfaction and also benefit loss.

We are aware of the limitations of our modeling, and we will consider any possible errors to ensure the least errors as possible. Although we must take into account that due to the coronavirus pandemic, the hotel industry has undergone changes that couldn't have been predicted by anyone, so we will inform any potential client of our services about this.

Moreover, we have to take into account that if there is a lot of occupancy in an area and we advise all investors to open a hotel in that area, occupancy may decrease because there will be more supply and customers will be divided, so we will take this into account, so that our actions interfere as little as possible to the interests of the investors.

## 4. 1 Legacy

After carrying out this project we thought of creating an application that would respond to the interests of investors.

The scope of development of the project does not know issues related to economic factors with respect to the implantation of hotels. For future iterations and mineable views, we think that adding variables of this type could be very interesting because it would take the study problem to a much more complete and complex environment, but closer to a real problem issue. We will be able to compare not only the occupancy, but also the benefits that would be obtained, being substantially different from an investment in luxury hotels than in hotels with one or two stars.

# 5. Conclusions

Let's put it in context: an investor who wants to build a hotel in Spain for 2023 contacts us to help him choose where he should establish the premises in order to have as many customers as possible.

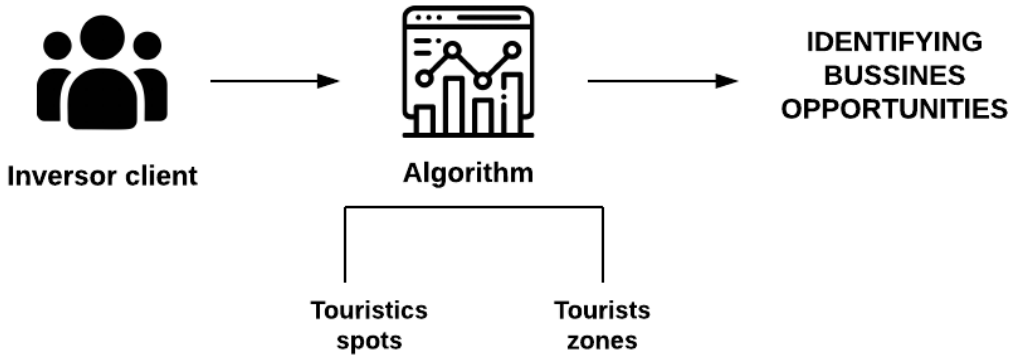


*Figure 5.1 Project diagram*

First we will create heat maps of the Spanish provinces based on the growth or decrease in the number of tourists, both spanish and foreign. From there we will choose the one with the highest increase because it is assumed that if there has been growth, the business in that particular province will do well, as long as the choice takes into account the client's preferences. For example, it could be the case that the province with the highest growth is Madrid but the person who hires our services prefers a coastal area. In this situation we will look at those provinces near the sea with the highest growth in the number of tourists, such as Canarias or Baleares.

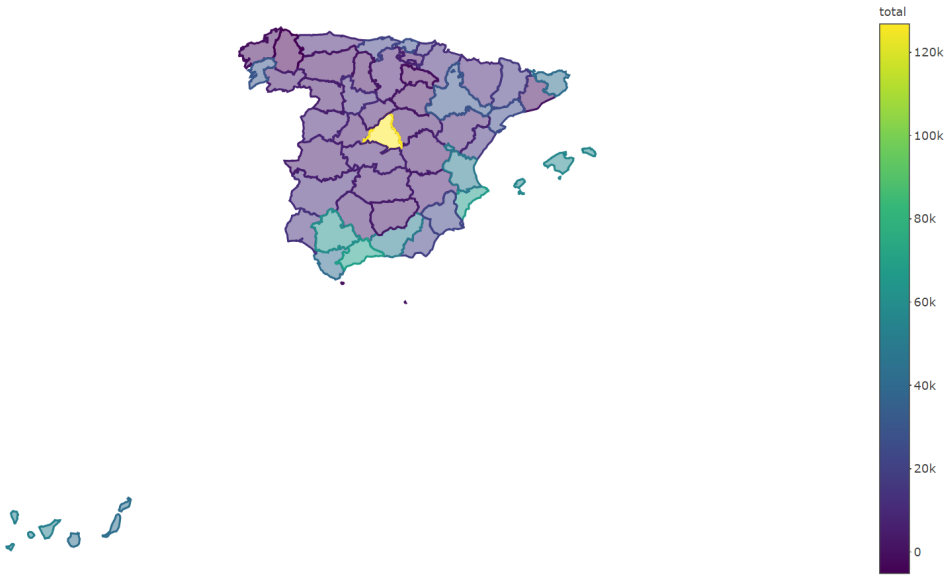


*Figure 5.2 Tourist growth density map*

With the following bubble maps by zones and points, we seek to provide intuitive information prior to the model, so that any of our potential investors can have an idea of the geographical distribution of occupancy in Spain and can consult any period of the years studied. In this way, apart from intuiting at a glance that the coastal areas have a higher occupancy rate than the interior of the peninsula, we can also denote points and/or areas with the highest occupancy, such as the case of Arona for the year 2019 (annual average) or the island of Mallorca for the month of August (inter-annual average).

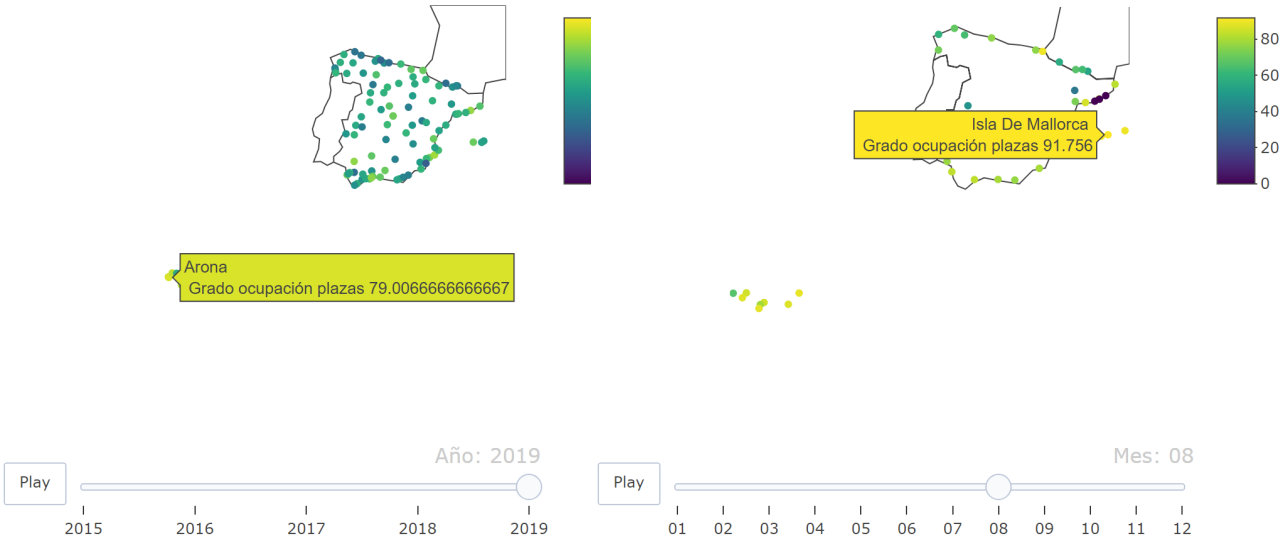


*Figure 5.3 Bubble maps by zones and points*

The objective of this is to give a more detailed service to the investor, giving him extra information apart from our predictions that gives him a retrospective vision of how the occupancy has been distributed in the last years.

As we have seen in the maps, Canarias and Baleares have seen an increase in number of travellers and have the highest occupancy rate, so we proceed to predict the occupancy rate and travelers for zones, and occupancy rate for touristic points.

For Arona's occupancy rate prediction , we can see a slight positive trend, and we observe how the confidence interval increases as predicted years are further apart from the training data. The forecast in 2023 could vary by 20 % positive or negative.

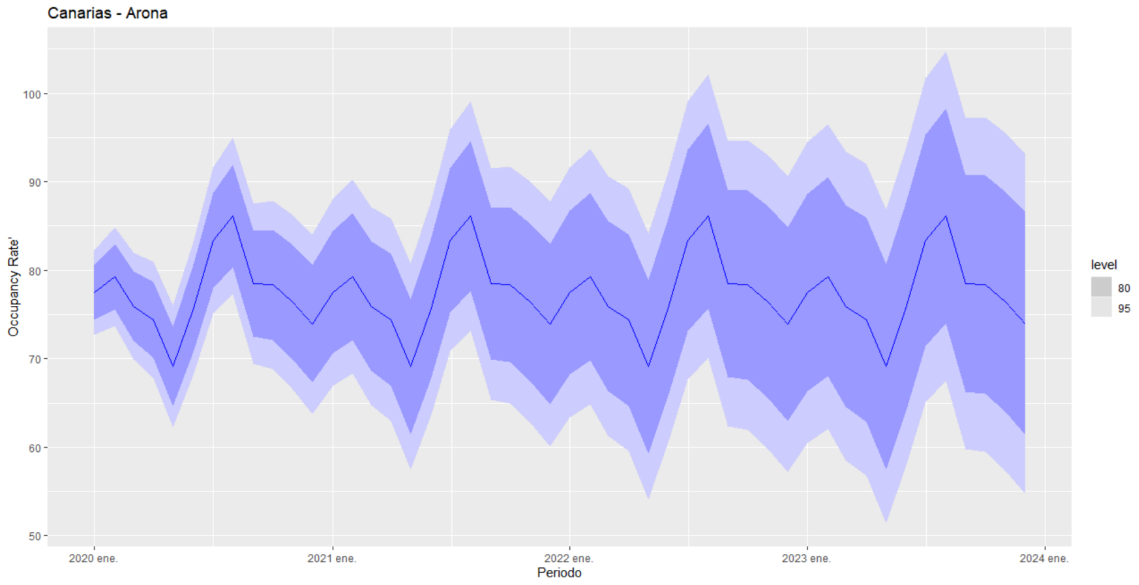


*Figure 5.4 Occupancy rate predictions for Arona*

On the other hand, the predictions for the island of Mallorca are shown in the following figure. We can see that they are quite seasonal, increasing in the summer months (up to 90%) and decreasing in winter (up to 40%). Moreover, there is a slight positive trend, so that the occupancy rate in the year of the opening of our client's hotel will be higher than it is.

However, we cannot be 100% sure, because the values could oscillate within a confidence interval. The optimum would be an increase in occupancy in July and August (although 100% would not be realistic), but it may be the case that it decreases to 80%.

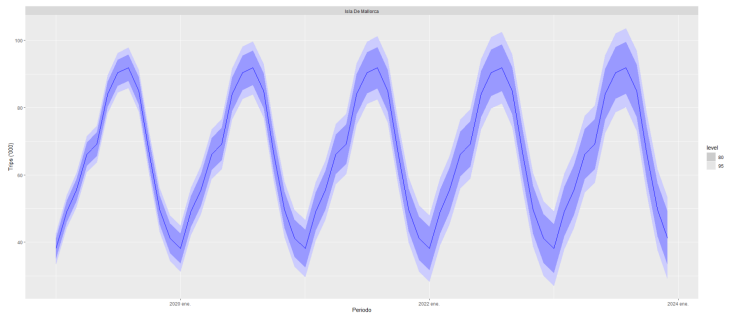

*Figure 5.5 Occupancy rate predictions for island of Mallorca*

Having looked at the two graphs, we can say that Arona is more stable than Mallorca (despite having a large confidence interval), if the client prefers a hotel that is active all year round.

On the other hand, if the investor prefesr a hotel with high profits in the summer months, Mallorca is the best area. In fact, the confidence interval is small, so the prediction will not be too far from reality.

As we said before, the main objective of our study is to be able to predict where in Spain it is convenient to establish a hotel establishment so that the investor client can obtain the greatest possible number of customers and, therefore, obtain maximum profits. In conclusion, the choice of the area or touristic point will depend on the client's preferences, so our job is to give them the most viable options to meet them.

# 6. Bibliography

## Databases:

Instituto Nacional de Estadística. (National Statistics Institute). (n.d.). Retrieved May 15, 2021, from Ine.es website: https://www.ine.es/dynt3/inebase/index.htm?padre=238&capsel=238

Expedia Hotel Recommendations. (n.d.). Kaggle.Com. Retrieved May 15, 2021, from https://www.kaggle.com/c/expedia-hotel-recommendations/data

Database - Eurostat. (n.d.). Europa.Eu. Retrieved May 15, 2021, from https://ec.europa.eu/eurostat/web/main/data/database

Meteomatics. (n.d.). Meteomatics.Com. Retrieved May 15, 2021, from https://www.meteomatics.com/en/?gclid=Cj0KCQiA4L2BBhCvARIsAO0SBdZs_5NV6TREwBcnyhy2zlkmREW5Fvzp7ywJj7PDf11pIvm2ATQrxtsaAoHIEALw_wcB

Bases de datos - Transporte - Estadísticas regionales y mundiales, datos nacionales, mapas, clasificaciones. (n.d.). Knoema.Es. Retrieved May 15, 2021, from https://knoema.es/atlas/topics/Transporte/datasets

No title. (n.d.). Retrieved May 15, 2021, from Europa.eu website: https://ec.europa.eu/eurostat/databrowser/view/icw_res_02/default/table?lang=en

## Articles:

Hosteltur. (n.d.). ACAV. Retrieved May 15, 2021, from Hosteltur.com website: https://www.hosteltur.com/tag/acav

Retrieved May 15, 2021, from Sharepoint.com website: https://upvedues-my.sharepoint.com/personal/arsermo1_upv_edu_es/Documents/Archivos%20de%20chat%20de%20Microsoft%20Teams/Effects%20of%20covid%20on%20hotel%20management.pdf

## Others:

Quinteros, O. E., Funes, A. and Ahumada, H. C. (2016). Construcción de una vista minable para aplicar minería de datos secuenciales temporales. XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016).

Retrieved May 9, 2021, from Regio7.cat website: https://www.regio7.cat/catalunya/municipis.html

3 Visualización de datos. (n.d.). Retrieved May 9, 2021, from Hadley.nz website: https://es.r4ds.hadley.nz/visualizaci%C3%B3n-de-datos.html

Sancho, R. S. (n.d.). Prefacio. Retrieved May 9, 2021, from Gitbooks.io website: https://rsanchezs.gitbooks.io/rprogramming/content/index.html

Forecast reconciliation. (n.d.). Retrieved May 9, 2021, from Tidyverts.org website: https://fabletools.tidyverts.org/reference/reconcile.html

Retrieved May 9, 2021, from Geodatos.net website: https://www.geodatos.net/coordenadas

2.2 Time plots. (n.d.). Retrieved May 9, 2021, from Otexts.com website: https://otexts.com/fpp3/time-plots.html

Zuur, M. (2019, August 29). Time-series forecasting using R with fable. Retrieved May 9, 2021, from Thinkwisesoftware.com website: https://community.thinkwisesoftware.com/news-updates-21/time-series-forecasting-using-r-with-fable-651

Zuur, M. (2019, August 29). Time-series forecasting using R with fable. Retrieved May 9, 2021, from Thinkwisesoftware.com website: https://community.thinkwisesoftware.com/news-updates-21/time-series-forecasting-using-r-with-fable-651

Linear Regression Example in R using lm() Function – Learn by Marketing. (n.d.). Retrieved May 9, 2021, from Learnbymarketing.com website: http://www.learnbymarketing.com/tutorials/linear-regression-in-r/

Parra, F. (n.d.). 8 Series Temporales. Retrieved May 9, 2021, from Bookdown.org website: https://bookdown.org/content/2274/series-temporales.html

RStudio Community. (n.d.). Retrieved May 9, 2021, from Rstudio.com website: https://community.rstudio.com/

Vega, J. B. M. (n.d.). R para principiantes. Retrieved May 9, 2021, from Bookdown.org website: https://bookdown.org/jboscomendoza/r-principiantes4/condicionales.html

Confusing about using cross validation tsibble. (2020, December 22). Retrieved May 9, 2021, from Rstudio.com website: https://community.rstudio.com/t/confusing-about-using-cross-validation-tsibble/91463

Home - RDocumentation. (n.d.). Retrieved May 9, 2021, from Rdocumentation.org website: https://www.rdocumentation.org/

Introduction to feasts. (n.d.). Retrieved May 9, 2021, from R-project.org website: https://cran.r-project.org/web/packages/feasts/vignettes/feasts.html

# 7. Annexes

## 7.1 Annex 1: Goals and values

First of all, we considered several ideas of objectives that we could take for the realisation of the project, for some of them we needed to study other data. Among the objectives, we found the following:

- Optimisation of the hotel's profit through simulations with the main variables.

- Obtaining the most visited airports and tourist areas, in order to sell the hotel's information to maximise profit.

- Creation of customer profiles, country of origin, average spending, length of stay, etc., with the idea of knowing their travel trends and preferences.
    - Value: creation of a website to inform tourists about average occupancy, average prices...

- Obtaining the most suitable tourist areas in Spain for the opening or expansion of hotels, taking into account different characteristics related to the length of stay and accessibility.

- Study of the most frequented airports in Europe and the characteristics related to tourists, in order to maximise flight sales.

- Study of incentives for tourists to come and stay as long as it is in the hotels' interest.
    - Value: include offers and vouchers for the area, celebration of events...

After having studied and discussed one by one the previous objectives, we decided to go ahead with the fourth one, and make some modifications to achieve a more suitable objective, taking into account the data we can obtain.

In this way, the main objective, already discussed in the main report, is to predict a future hotel occupancy that is able to satisfy the needs and maximize the profits of the investor. To this end, we will identify anomalies and trends in hotel occupancy along with other useful variables, such as average stay by category, in order to forecast a possible optimal strategy for opening a new hotel.

Once the general objective has been defined, we can position ourselves as: a company that offers assistance to the tourism sector e.g. restaurants, leisure activities, hotels, travel agencies, etc., because if we correctly predict the occupancy rate for a certain period, they can plan in the most optimal way possible their work shifts, the necessary staff, new hires, the use of infrastructures, marketing strategies and with this, be able to increase revenues thanks to our forecast.

## 7.2 Annex 2: Data

The variables that appear in the different dataset used to build the model are defined below:.

- Period (**Periodo**): variable of categorical type, which will be converted into date type when modelling. It corresponds to the time at which the data were collected. It consists of 72 values (6 years times 12 months), i.e. the database contains only data collected in the interval (2015, 2020) inclusive. It was decided not to work with the year 2021, as there was only one data, that of January.

- Average stay (**est_med**): numeric type variable that describes the average number of nights a tourist stays in a hotel of a specific category per month.

- Spanish travellers (**Viajeros_esp**): numeric variable that describes how many Spanish travellers stay one or more nights at a hotel of a specific category per month.

- Foreigners travellers (**Viajeros_ext**): numeric variable that describes how many foreigners travellers stay one or more nights at a hotel of a specific category per month.

- Spanish overnight stays (**Pernocataciones_esp**): numeric variable that describes how many nights Spanish tourists stay in a hotel of a specific category per month.

- Foreign overnight stays (**Pernoctaciones_ext**): numeric variable that describes the sum of how many nights the foreign tourists stays in a hotel of a specific category per month.

- Level of occupancy by bed places (**Grado_ocup_plazas**): numeric variable that describes the percentage relation between the total overnight stays during a month and the number of beds that can be used.

- Occupancy rate by bed places at weekends (**Grado_ocup_plazas_finsemana**): numeric variable that describes the percentage relation between the total overnight stays during friday's and saturday's of a month and the number of beds that can be used during the weekends.

- Occupancy rate by rooms (**Grado_ocup_habitaciones**): numeric variable that describes the percentage relation between the number of occupied rooms and the total number of rooms that can be occupied.

- Number of rooms (**Num_habitaciones**): numeric variable that describes the total number of rooms that can be occupied.

- Number of open establishments (**Num_establecimientos_abiertos**): numeric variable that describes the total number of open hotel establishments.

- Number of estimated places (**Num_plazas_estim**): numeric variable that describes the total number of estimated places.

- Staff employed (**Personal_emp**): numeric variable that indicates the total number of people working in hotels of the same category.

- Operation (**Operación**): categorical variable with four possible values that describes the type of accommodation. It's values are: Tourist Apartments, Camping, Hotel and Rural Tourism.

- Community (**Comunidad**): categorical variable with 19 possible values that indicates the autonomous community where the establishments are located. It's values are the different communities of Spain.

- Province (**Provincia**): categorical variable with 31 values corresponding to the provinces in which each zone is located, including the exceptions.

- Zones (**Zonas**): categorical variable which designates, within a given province, a specific area of tourist interest within that province. It has 72 different values, including exceptions.

- Tourist points (**Puntos turísticos**): categorical variable that stores all the names of the tourist points, a total of 106.

It is worth mentioning that all databases are **time series**, i.e. they have a time variable running from January 1999 to January 2021 (in total 265 monthly data), although we will focus on the years from 2015 to 2019. In addition, they all show the total number of tourists (numerical variable) according to country of residence, autonomous community, tourist area, etc.

We start an exploratory analysis to evaluate which targets fit the best according to our data. Although we have too many initial databases to summarise all our analysis in this work, we have done our best to show a representation with the analysis of three of our most important one.

**Dataset X1_13: <u>Establishments, estimated bedplaces, occupancy rates and staff employed by tourist areas</u>**

In this data set we can study several values for variables that describe the hotel management, such as the occupancy rates, the hired personal, and number of predicted vacancies.

First we have explored for missing data, and we found 3% missing in the Total column, some of it is caused by the effect of the pandemic but we have studied ways to replicate the lacking values with.

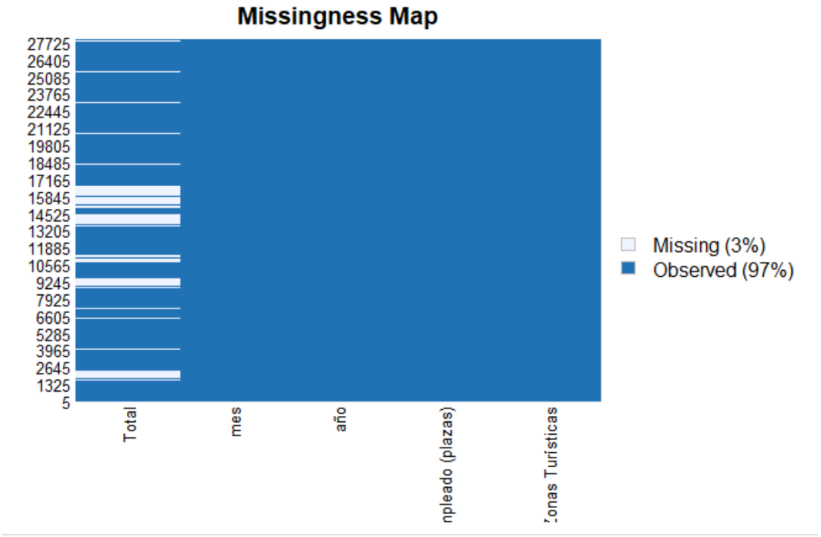


*Figure 7.1 Missingness Map of X1_13 dataset*

After having studied the missing data, we examined the density plot, where we can see that various variables follow the same tendency, the number of predicted vacancies, rooms and hired personal have a descending trend, having more frequency the least values, on the other hand the data for occupation degree escalate, being the highest values more frequent.

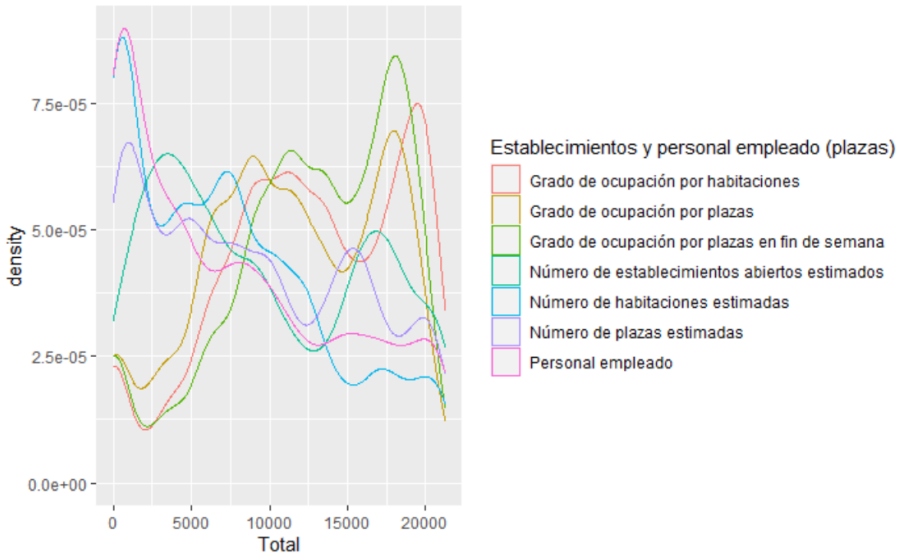


*Figure 7.2 Density plot of X1_13 dataset*

We can see in the box and whisker diagram that the data distribution for the different categories that has been homogeneous until 2020 besides the median value for occupation degree which have been increasing gradually since 2014, from 2020 onwards the median values and the distribution variation have decreased, although some outlier values can be seen.

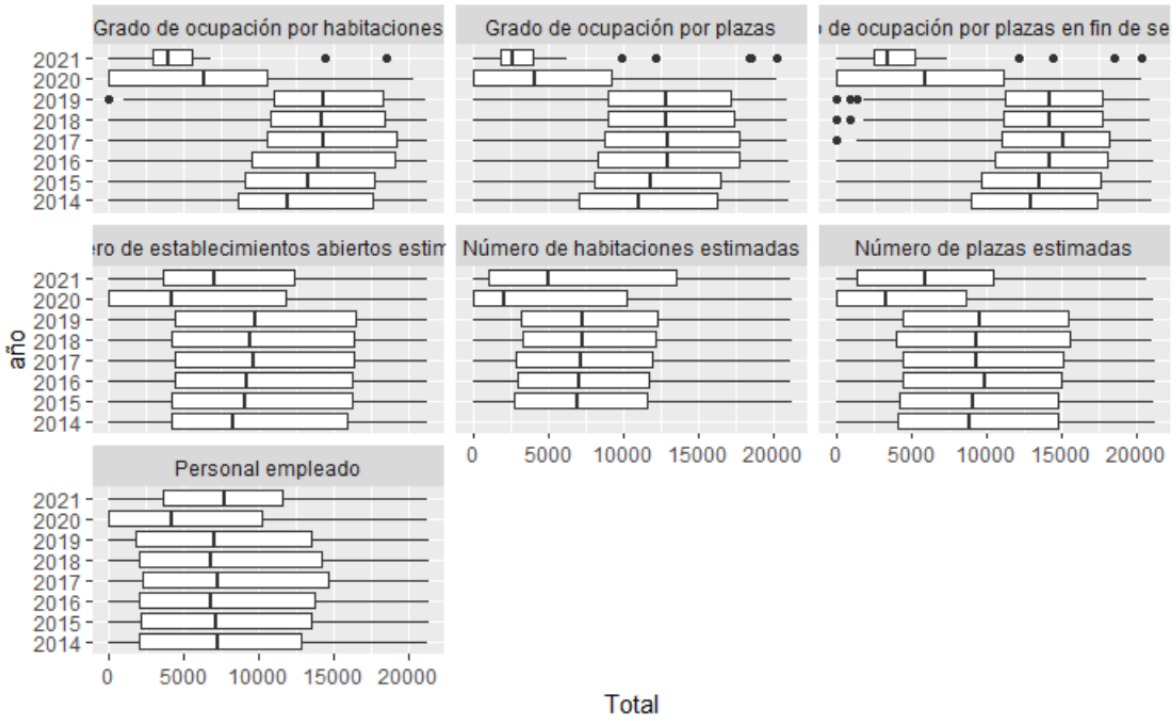


*Figure 7.3 Box and whiskers diagram of X1_13 dataset*

## Dataset X1_5 : Travelers and overnight stays by country of residence

Here we intended to observe the monthly density, we chose two years so that it could be better appreciated. As expected, many months follow the same distribution. We will study it in detail later.

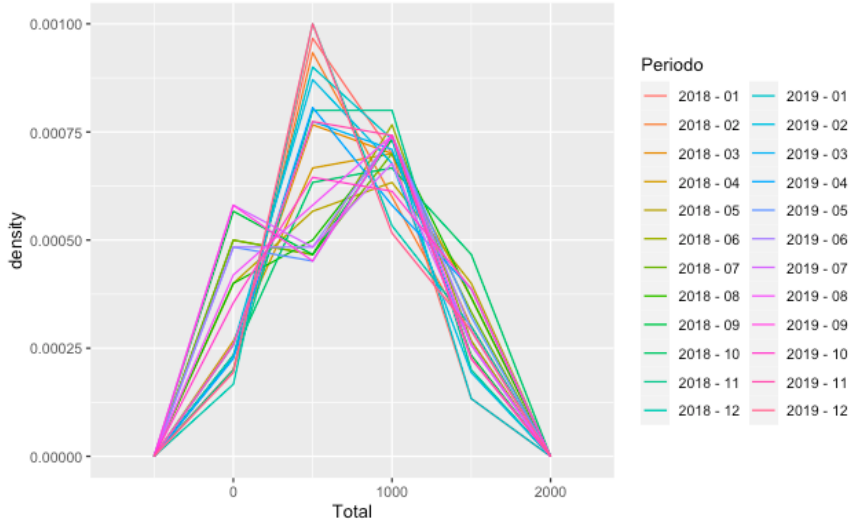


*Figure 7.4 Density plot of X1_5 dataset*

## Dataset X1_2 : Travellers and overnight stays by autonomous communities and provinces

As we can see, the distribution by month is practically the same in all years, with the exception of March, April and May 2020. This is due to the emergence of covid, the declaration of the state of alarm and mandatory confinement. In April, the highest occupancy is very low compared to the others. March and May occupancy is lower than in other years, although one reason for the occurrence of the outliers could be that the measures only affect the first or last fortnight of the month. It should also be noted that for 2021 there are only data for January.

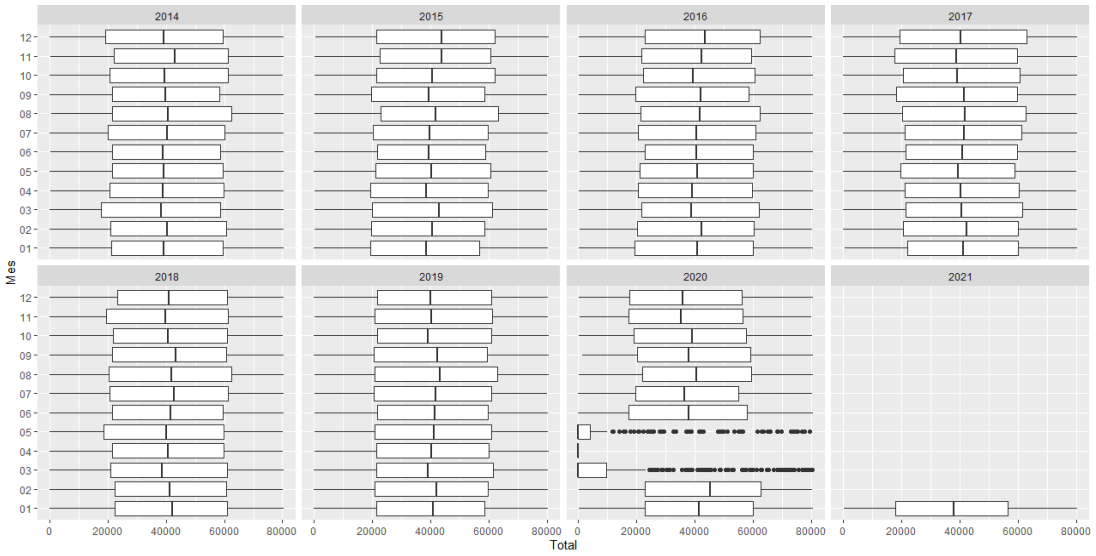


*Figure 7.5 Box and whiskers diagram of X1_2 dataset*

At first sight, in this graphic we note the presence of missing values (e.g. the occupancy of Andalucía in 2017); in total there are only 10, and all belonging to the variable "Total". Leaving this aside, the values do not vary much over the years in the same community, although there are some exceptions such as Madrid. We must also take into account that the values belonging to 2020 and 2021 (January) are determined by the presence of Covid. Finally, there are some outliers in Andalusia, Extremadura and Cantabria, but especially in Madrid and the Canary Islands in the most distant years.

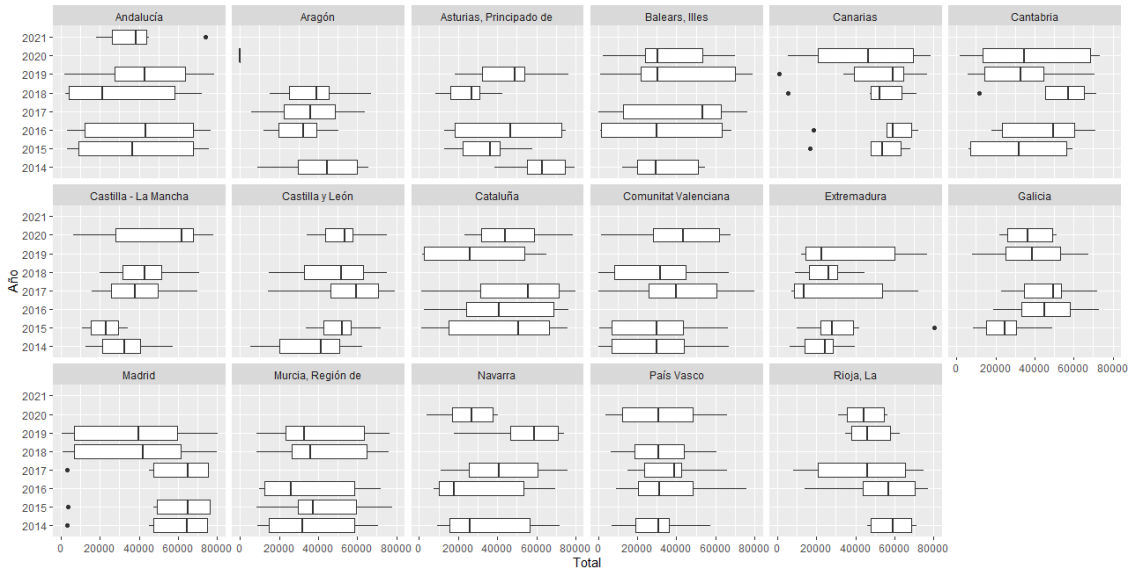


*Figure 7.6 Box and whiskers diagram of X1_2 dataset*

As mentioned above, attempts have been made to bring together different databases. But due to the nature of the given data, this task has been somewhat complicated. Even so, some of the themes searched were: air, sea and rail transport; economics and climatology.

## 7.3 Annex 3: Domain, knowledge, context and innovation

As briefly discussed in the previous section, the allocated data are for hotel occupancy in Spain. Let us put this in context.

The team was provided with a URL[1] to download the data, stored at the INE (Instituto Nacional de Estadística). Exactly the data refers to "Viajeros y pernoctaciones según país de residencia del viajero".

Once we had seen the data, the team started to think about what we could do with it. Being time series data, a first idea, as proposed along with the project, was to try to predict future hotel occupancy in Spain based on past data, i.e. the allocated data (e.g. before and after Covid). However, we found this procedure a bit poor for the scope of the project.

Firstly we studied this paper published last year which is centered on the tourism demand forecast. It is a very similar study to what we intend to do.

Gaussian processes for daily demand prediction in tourism planning - Tsang - 2020 - Journal of Forecasting - Wiley Online Library

The most remarkable points that we have extracted from this study are the following.

As in our project, their objective was to predict hotel occupancy rate at a region level (cities), they use gaussian processes, and compare them to different approaches such as linear regression, ARIMA models and other machine learning models like for example random forest.

They initially extract the most important predictive features to reduce dimensionality, which is what we thought of doing, so this means we are in the right working direction. Also they state that tourism forecasting for larger regions are more robust to local fluctuation, although they lose relevance in prediction for local tourism businesses, as we have our data separated by Autonomous communities, then provinces and lastly by touristic zones /points, this is very useful information as we can adequately proceed in our study knowing that predictions for communities or provinces will be more robust.

We also researched other models we could implement apart from regressions, principal component analysis for dimensionality reduction, and ARIMA/ machine learning models for predictions, and we found some interesting papers, in which they also did hotel occupancy forecasting but using neural networks, we are reluctant to use this type of models as we haven't yet implemented none although we've studied it on a theoric level, and we know the time cost of training this type of processes, although we will contemplate more thoroughly the benefits of using them and maybe try to implement one.

Tourism Room Occupancy Rate Prediction Based on Neural Network | SpringerLink

Room occupancy rate forecasting: a neural network approach | Emerald Insight

[1] https://www.ine.es/jaxiT3/Tabla.htm?t=2038&L=0

Moreover we considered integrating our databases with other datasets that could be interesting, such as air, maritime and rail transport data, climate, economic data or even INE tourism data but with additional information, such as hotel stars. At this point, different websites were searched for downloadable data (eurostat, aena, kaggle, meteomatics, etc.).

This was thought of as a way to innovate and to be able to find relationships that we may not have known existed and that, in reality, have a great influence on hotel occupancy.

Although the data provided are too general to be able to establish a concrete objective(s) and it is rather complicated to integrate them with other databases already created, as will be explained later [6.2 Data], we have thought of extracting information from Booking in order to be able to integrate data from the different accommodations with each corresponding tourist area or province. However, due to the nature of the data, we have focused the project [6.1 Goals and values] on the study of hotel occupancy for profit-maximising decisions for a company (e.g. travel agencies or investor groups).

## 7.4 Annex 4: Data Preparation

### Dataset 1: Types of accommodation

To create this database we have put together the following datasets:

- **2.1**: Travellers, overnight stays by type of accommodation by Autonomous Communities and Cities.
- **2.2**: Average stay, by type of accommodation, by Autonomous Communities and Autonomous Cities.
- **2.3**: Estimated establishments, estimated bedplaces and staff employed by type of accommodation and by Autonomous Communities and Cities.

On the basis of the modifications explained above and in order not to repeat operations again, we will comment on the codes above but focusing on the corresponding dataset.

The first to be modified has been base 2.3, whose variables are: *Operación* (the type of survey), *Comunidades y Ciudades Autónomas*, *Periodo, Establecimiento y personal empleado (plazas)* y *Total*. The steps to be taken were as follows:

1. Elimination of the *Total Nacional* value of the second variable.
2. Create 3 new columns, one for each value of the categorical variable *Establecimientos y personal empleado (plazas)*, thus obtaining the number of establishments open and staff employed..
3. Join the three selections in a new dataset taking as reference the operation, the period and the name of the autonomous communities and cities..

In the second base, far fewer operations have been performed, providing the new dataset with a single variable.:

1. Elimination of the *Total Nacional* value of the second variable.
2. Join the column *Estancia media* to the new base taking into account the operation, the period and the name of the autonomous communities and cities..

Unlike the others, 2.1 dataset has required a greater number of modifications, since instead of obtaining the values of a single categorical variable, the values of two categorical variables have been obtained.

1. Crate the columns *Viajeros residentes en España, Pernoctaciones residentes en España, Residentes en el Extranjero, Residentes en el Extranjero*.
2. Add the 4 new columns to the base created according to the values of *Operación*, *Comunidades y Ciudades Autónomas* and associated *Periodo*.

Finally, the years less than 2015 and greater than 2020 have been eliminated, the names of the communities and provinces have been modified so that no numbers appear (from *01 Andalucía* to *Andalucía*), the punctuation system has been modified from Spanish to English to avoid calculation and execution problems in R and missing values of numeric variables have been replaced by zeros.

Observing figure 7.7, it can be seen that the average number of Spanish travellers in hotels is clearly higher. The remaining group is headed by campsites, followed by flats and rural tourism. It also shows the negative effect, in terms of tourism, of the policy measures during the pandemic in 2020.

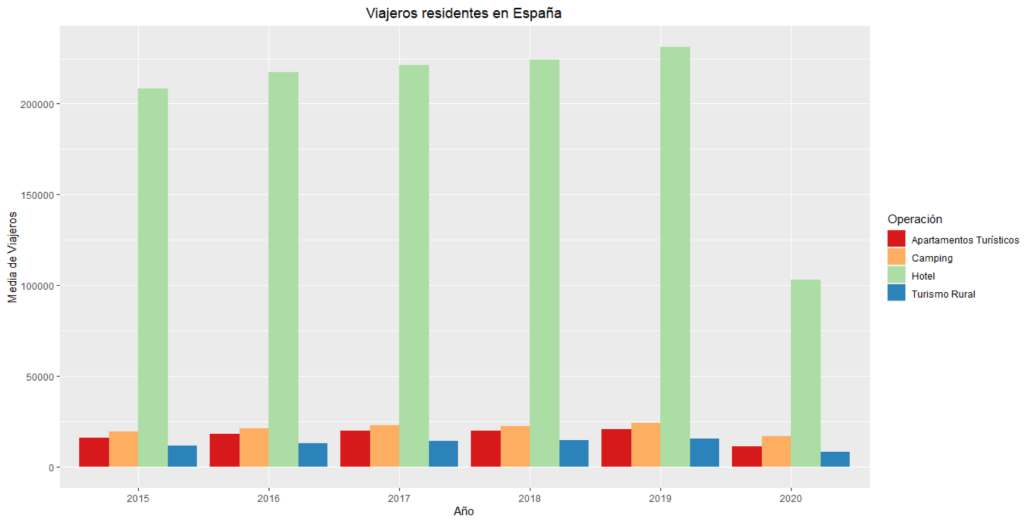


*Figure 7.7 National travellers's bars plot of Types of accommodation*

If we compare figure 7.7 with foreign travellers in figure 7.8, the order in the data is exactly the same as before, i.e. both types of travellers behave in the same way. However, there is a slight difference and that is that foreign people, excluding hotels, tend to travel on average more to flats rather than to campsites or to rural tourism. This makes sense since, generally speaking, when people visit another country they tend to go to the most important and central areas of the city and not to the outskirts.

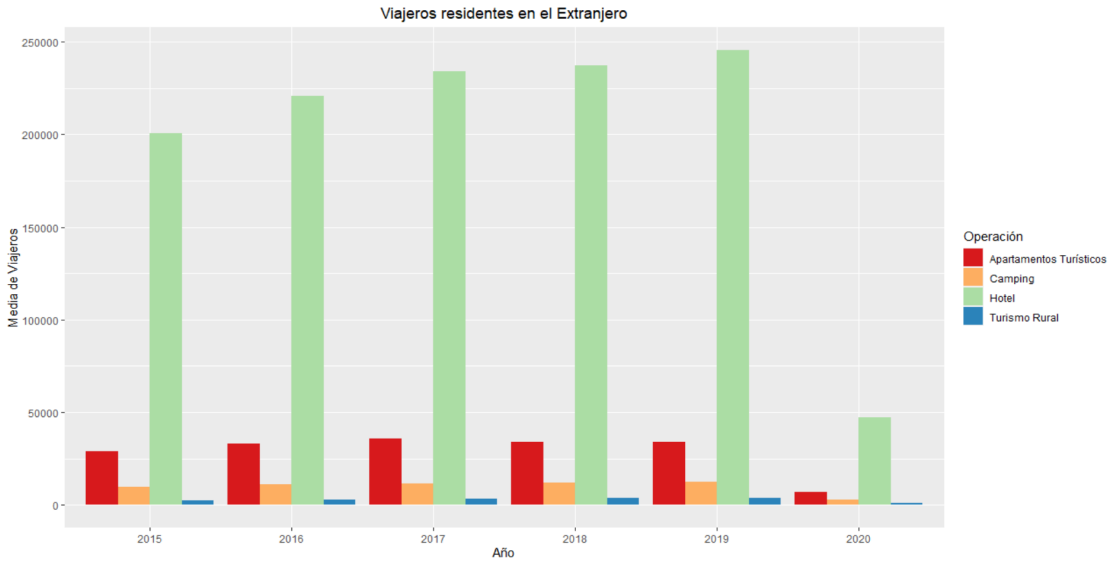


*Figure 7.8 Foreing travellers's bar plot of Types of accommodation*

This dataset allows us to differentiate 4 types of tourism: rural tourism, Airbnb apartments, Camping and hotels. Thanks to this dataset we will discover the occupancy differences between each type of accommodation and we will be able to analyze how the number of travelers varies. In this

way, we will be able to create a model that fits the influence of the type of accommodation that will be studied later.

```
  Operación          Comunidad           Periodo          Num_establecimientos_abiertos Num_plazas_estim  Personal_emp      est_med
 Length:5472        Length:5472        Length:5472        Min.   :   0.0                Min.   :     0   Min.   :    0   Min.   : 0.000
 Class :character   Class :character   Class :character   1st Qu.:   0.0                1st Qu.:  3468   1st Qu.:  147   1st Qu.: 1.910
 Mode  :character   Mode  :character   Mode  :character   Median :  48.0                Median :  8686   Median :  538   Median : 2.450
                                                          Mean   : 394.9                Mean   : 32063   Mean   : 3171   Mean   : 3.063
                                                          3rd Qu.: 535.0                3rd Qu.: 27920   3rd Qu.: 2257   3rd Qu.: 3.640
                                                          Max.   :3721.0                Max.   :360788   Max.   :62964   Max.   :31.000
 Viajeros_ext      Viajeros_esp     Pernoctaciones_ext Pernoctaciones_esp
 Min.   :      0   Min.   :     0   Min.   :       0   Min.   :      0
 1st Qu.:    448   1st Qu.:  3148   1st Qu.:    1296   1st Qu.:   8285
 Median :   2514   Median :  11400   Median :    7074   Median :  29498
 Mean   :  59766   Mean   :  63250   Mean   :  278855   Mean   : 162356
 3rd Qu.:  21821   3rd Qu.:  48885   3rd Qu.:   66077   3rd Qu.: 136634
 Max.   :1692776   Max.   :1346828   Max.   :10103202   Max.   :4374388
```

*Figure 7.9 Summary of Types of accommodation*

## Dataset 2: Zones

The resulting database of zones, explained below, has been formed with the union of the following datasets:

- **1.3**: Travellers and overnight stays by tourist areas
- **1.9**: Average stay by tourist areas
- **1.13**: Establishments, estimated bedplaces, occupancy rates and staff employed by tourist areas

As in the previous sub-sections, the aim of the data preparation has been to convert the values of the categorical variables into new variables, apart from further data cleaning.

For dataset 1.9, all the variables present have been retained but with some modifications:

1. Separate the column *Zonas Turísticas* into *Comunidad* and *Zona* as it was in the form "Andalucía: Costa De Almería". However, not all values had the same format, i.e. there are zones that do not belong to a specific community. To fix this, in these cases it has been assumed that the zone is the same as the community (such as Pirineos).

In contrast to the previous version, some variables have been omitted in 1.3, although many modifications have been similar:

1. Separate community and zone in the same way as in dataset 1.9.
2. Create the columns *Viajeros residentes en España, Pernoctaciones residentes en España, Residentes en el Extranjero* and *Residentes en el Extranjero* in the same way as in the previous datasets.
3. Add the 4 new columns to the base created according to the associated *Comunidad*, *Zona* and *Periodo* values.

The operations for base 1.13 are similar to those above:

1. Separate community and area in the same way as in dataset 1.9 and 1.3.
2. Create 7 new columns, one for each value of the categorical variable *Establecimientos y personal empleado (plazas)*, thus obtaining the number of establishments open, estimated

bedplaces and rooms, the degree of occupancy by bedplaces, weekend bedplaces and rooms, and the staff employed.

1. Add the created columns to the new dataset according to the associated *Comunidad*, *Zona* and *Periodo* values.

Afterwards, we create two new variables in the resulting database referring to the coordinates of a corresponding area, with the purpose of using them for a map visualisation, which will help the investor to locate the hotel establishment for economic convenience.

- Latitude (**Latitud**): numerical variable indicating the geographical latitude of the corresponding tourist spot.
- Longitude (**Longitud**): numerical variable indicating the geographical longitude of the corresponding tourist spot.

Finally, the years less than 2015 and greater than 2020 have been eliminated, and the spaces at the beginning of the name of the tourist areas have been eliminated (" Barcelona" - "Barcelona") and created the variable *Provincia* to store the province to which each zone belongs. There are some exceptions, such as Pirineos, where the community and the province inherit the name of the tourist area.

We made a box and whiskers graph to observe the *Est_media* variable by community and year as an example. As we can see, the northern or mountainous areas have a very specific seasonality range (between 2 and 2.5), leaving 2020 aside. The coastal areas, however, have a much wider range, with the Canary Islands standing out above all (exceeding 7.5), followed by the Islas Baleares, Andalucía, the Valencian Community and Catalonia. As for last year, the values start at zero and increase to where they are used to in other years. This is due to the fact that during the state of alarm, travel was prohibited and therefore seasonality was null. In addition, mandatory confinement was lifted in spring, so people took the opportunity to go on holiday, especially in summer. This explains the increase in values.

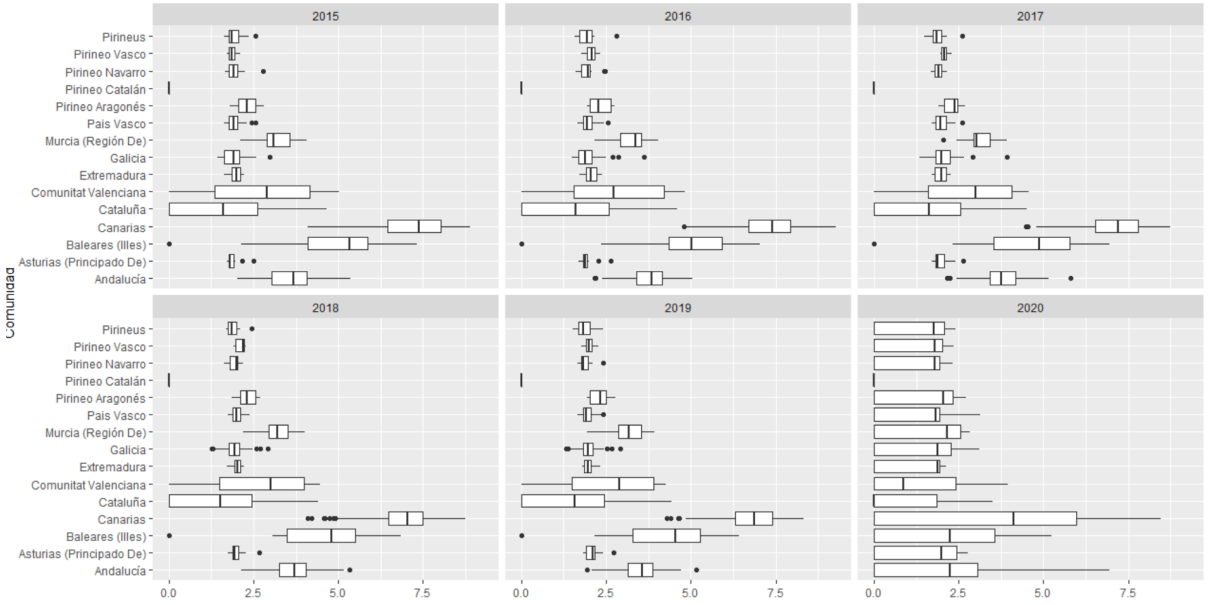


*Figure 7.10 Box and whiskers diagram of Zones*

It should be noted that there are anomalous values, especially in mountainous areas, such as Asturias, Galicia and the Basque Country, but also in the Canary Islands and Andalusia.

In this dataset we will be able to study the different tourist areas in Spain and analyze if there is a different behavior between inland and coastal areas. It will help us when creating our prediction model to know which are the areas with more occupancy and which are not. In addition, by having the period we will also be able to know which dates are the most frequented in each area.

```
   Zonas              Periodo          Grado_ocup_habitaciones Grado_ocup_plazas Grado_ocup_plazas_finsemana Num_habitaciones
Length:3384        Length:3384        Min.   : 0.00           Min.   : 0.00     Min.   : 0.00               Min.   :     0
Class :character   Class :character   1st Qu.:22.09           1st Qu.:17.93     1st Qu.:21.47               1st Qu.:   891
Mode  :character   Mode  :character   Median :44.60           Median :38.01     Median :47.46               Median :  3677
                                      Mean   :44.32           Mean   :39.48     Mean   :44.14               Mean   : 10614
                                      3rd Qu.:71.26           3rd Qu.:64.44     3rd Qu.:69.66               3rd Qu.: 15235
                                      Max.   :95.73           Max.   :95.68     Max.   :93.62               Max.   :126980
Num_plazas_estim  Personal_emp   Num_establecimientos_abiertos  Viajeros_ext      Viajeros_esp    Pernoctaciones_ext Pernoctaciones_esp
Min.   :     0   Min.   :    0   Min.   :  0.0                 Min.   :     0.0   Min.   :     0   Min.   :     0     Min.   :     0
1st Qu.:  1763   1st Qu.:  198   1st Qu.: 21.0                 1st Qu.:   966.2   1st Qu.:  3861   1st Qu.:  1112     1st Qu.:  7927
Median :  7726   Median :  898   Median : 82.0                 Median :  7513.0   Median : 20728   Median :  6843     Median : 54808
Mean   : 22802   Mean   : 3523   Mean   :140.4                 Mean   : 60713.4   Mean   : 39360   Mean   : 88259     Mean   :111368
3rd Qu.: 35316   3rd Qu.: 4710   3rd Qu.:198.0                 3rd Qu.: 68953.2   3rd Qu.: 52454   3rd Qu.: 44377     3rd Qu.:152198
Max.   :268031   Max.   :46913   Max.   :949.0                 Max.   :998846.0   Max.   :340894   Max.   :995280     Max.   :998425
 Provincia          Comunidad          est_med
Length:3384        Length:3384        Min.   :0.000
Class :character   Class :character   1st Qu.:1.670
Mode  :character   Mode  :character   Median :2.325
                                      Mean   :2.887
                                      3rd Qu.:4.110
                                      Max.   :9.260
```

*Figure 7.11 Summary of Zones*

## Dataset 4: Tourist spots

To create this database we have put together the following datasets:

- **1.4**: Travellers and overnight stays by tourist spots
- **1.10**: Average stay by tourist spots
- **1.14**: Establishments, estimated bedplaces, occupancy rates and staff employed by tourist spots

We have followed the same working scheme for the creation of this dataset, i.e. the values of the categorical variables have been converted into new variables. In addition to having carried out a cleaning of the base:

1. Separate the period variable into year and month and delete the years less than 2015 and greater than 2021 for puntos_viajeros_pernoctaciones.
2. We create a dataset with a merge of dataset 1.14 and 1.10 according to provinces and period.
3. Add to the created dataset the base 1.4 from the tourist spots.

As in the zones database, in the tourist spots dataset, we create two variables that correspond to the geographical coordinates: *Latitud* and *Longitud,* that will be used to show in a map the touristic points, which will help the investor to take a decision.

In this case we are going to perform the same analysis explained in dataset 3 but applied to Tourists spots.

```
Puntos turísticos     Latitud         Longitud         Periodo          Grado_ocup_habitaciones Grado_ocup_plazas
Length:7632        Min.   :27.88   Min.   :-16.726   Length:7632        Min.   : 0.00           Min.   :  0.00
Class :character   1st Qu.:36.84   1st Qu.: -6.136   Class :character   1st Qu.:32.36           1st Qu.: 27.66
Mode  :character   Median :39.78   Median : -4.460   Mode  :character   Median :53.43           Median : 45.77
                   Mean   :38.92   Mean   : -4.520                      Mean   :49.17           Mean   : 43.32
                   3rd Qu.:42.34   3rd Qu.: -1.106                      3rd Qu.:71.11           3rd Qu.: 62.80
                   Max.   :43.66   Max.   :  3.435                      Max.   :96.54           Max.   :100.35
Grado_ocup_plazas_finsemana Num_habitaciones Num_plazas_estim  Personal_emp     Num_establecimientos_abiertos    est_med
Min.   : 0.00               Min.   :    0    Min.   :    0    Min.   :    0.0   Min.   :  0.00                Min.   : 0.000
1st Qu.:35.38               1st Qu.:  533    1st Qu.: 1014    1st Qu.:  127.0   1st Qu.: 17.00                1st Qu.: 1.550
Median :58.41               Median : 1338    Median : 2674    Median :  333.0   Median : 29.00                Median : 1.860
Mean   :50.56               Mean   : 3513    Mean   : 7290    Mean   : 1075.5   Mean   : 50.97                Mean   : 2.437
3rd Qu.:72.39               3rd Qu.: 3516    3rd Qu.: 7270    3rd Qu.:  906.2   3rd Qu.: 55.00                3rd Qu.: 2.900
Max.   :96.07               Max.   :46110    Max.   :89514    Max.   :15563.0   Max.   :914.00                Max.   :10.770
 Viajeros_ext       Viajeros_esp    Pernoctaciones_ext Pernoctaciones_esp  Provincia          Comunidad
Min.   :     0.0   Min.   :     0   Min.   :     0     Min.   :     0     Length:7632        Length:7632
1st Qu.:   678.2   1st Qu.:  3747   1st Qu.:  1136     1st Qu.:  8265     Class :character   Class :character
Median :  4179.0   Median : 11386   Median :  8062     Median : 24412     Mode  :character   Mode  :character
Mean   : 24434.3   Mean   : 20127   Mean   : 77317     Mean   : 46307
3rd Qu.: 18707.5   3rd Qu.: 23242   3rd Qu.: 41635     3rd Qu.: 51678
Max.   :703145.0   Max.   :444984   Max.   :996338     Max.   :792947
```

*Figure 7.12 Summary of Touristic spots*

## Dataset 5: Communities and provinces

This database is made up of the following databases extracted from the INE:

- **1.2**: Travellers and overnight stays by autonomous communities and provinces
- **1.8**: Average stay by autonomous community and province.
- **1.12**: Establishments, bedplaces, occupancy rates and staff employed by Autonomous Community and province.

As in the previous sub-sections, the aim of the data preparation was to convert the values of the categorical variables into new variables, apart from further cleaning.

The initial objective was, as in the previous databases, to convert the categories of the categorical variables into new variables, in order to be able to join the files, as they had different numbers of categories but the same number of provinces.

After this, and taking into account that there were some differences between the provinces and communities stored in base 1.2 and those stored in 1.8 and 1.12, we had to:

1. Remove the communities from the variable where the provinces were also stored (the communities added up to the same amount as their provinces.
2. Ensure that the province variable was the same for all bases, and if not, change the values to match.
3. Add a variable *Comunidad* to be repeated for each province within that community.

Then we merge the databases with the merge function through the variables period, community and province. The last thing left is to replace the missing values (also . / ...) and to transform the relevant variables to numeric type.

We would like to recall that there are no missing data in the Period variable in any of the datasets, although there are some missing periods.

## MISSING DATA:

Initially we had some missing data in our datasets, nothing too extreme and never surpassing 12% of unknown values. Because of this, we decided to replace them with 0, however we have now understood that this might lead to errors in our predictions, so we have been investigating ways to impute these values with a reliable method.

We decided to use ImputeTS function from caret package, which offers several methodologies to impute the missing values, we have tried:

- na_kalman: missing value imputation by Kalman Smoothing
- na_seadec: seasonally decomposed missing value imputation
- na_seasplit: seasonally splitted missing value imputation

Because of the strong seasonality we see in our data, we decided to do a split seasonality interpolation which is straightforward. If there are say 12 months, then the time series is treated as 12 separate time series, one for each season. Imputation for each of these separated time series is then performed based on the selected imputation approach.

As we have different zones/spots/communities , in order to do a correct imputation for each different time series and use the seasonal split for a same area, instead of using the mean of all, we needed to separate the original data into one dataset per zone/spot … and then do the seasonal split imputation, finally merging again the data frames.

```
138
139 fills<- function(zon){
140   zonas = (levels(as.factor(zon$Zonas)))
141   listofdfs <- list()
142   for (i in 1:length(unique(as.factor(zon$Zonas)))){
143     json_data<- filter(zon, Zonas == zonas[i] & year(yearmonth(Periodo)) != '2020')
144     filled <- na_seasplit(json_data, algorithm = "interpolation", find_frequency=TRUE)
145     df<- data.frame(filled)
146     names(df)<-names(zon)
147     listofdfs[[i]] <- df # save your dataframes into the list
148   }
149   zonas_final <- do.call(rbind, listofdfs)
150 }
151
```

*Figure 7.12 Code to do seasonal interpolation after separating original data frame in zones*

Here is an example of the zone dataset before replacing the missing values and after, *figure 7.13* and *7.14*. We are aware of the limitations of replacing the missing data using *ImputeTS* with seasonal split, as for zones where only some months are missing the imputation can be more reliable than for zones where there is a trend in missing values and there is a regular desinformation for the same area, but in that situation we have judged better to do a prediction than replacing directly the missing values with 0, as the algorithm will fill up the gaps with a low prediction as no data is available, and that error can only cause that if an investor decides to build a hotel there, there will be a higher value than expected in those months.

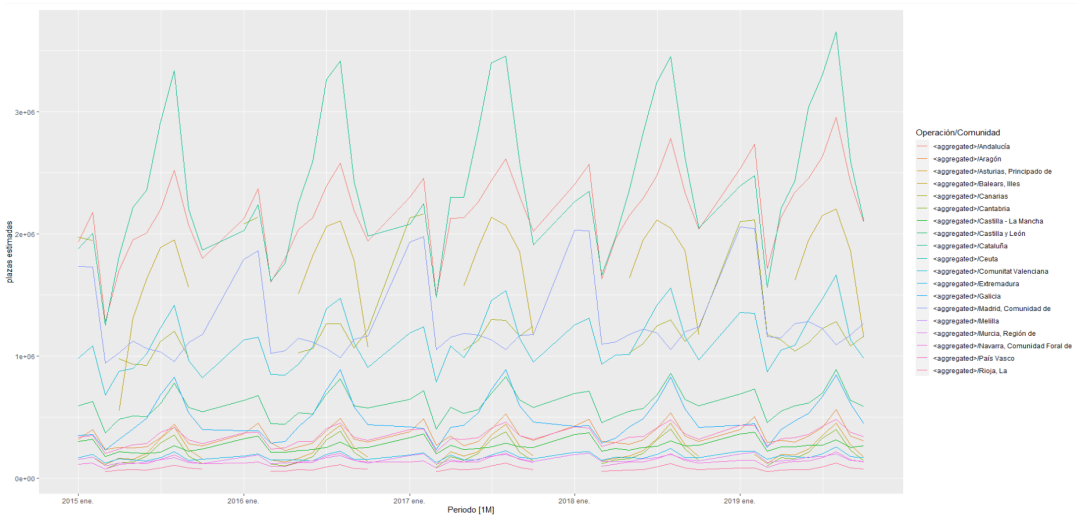


*Figure 7.13 Before replacing the missing values*

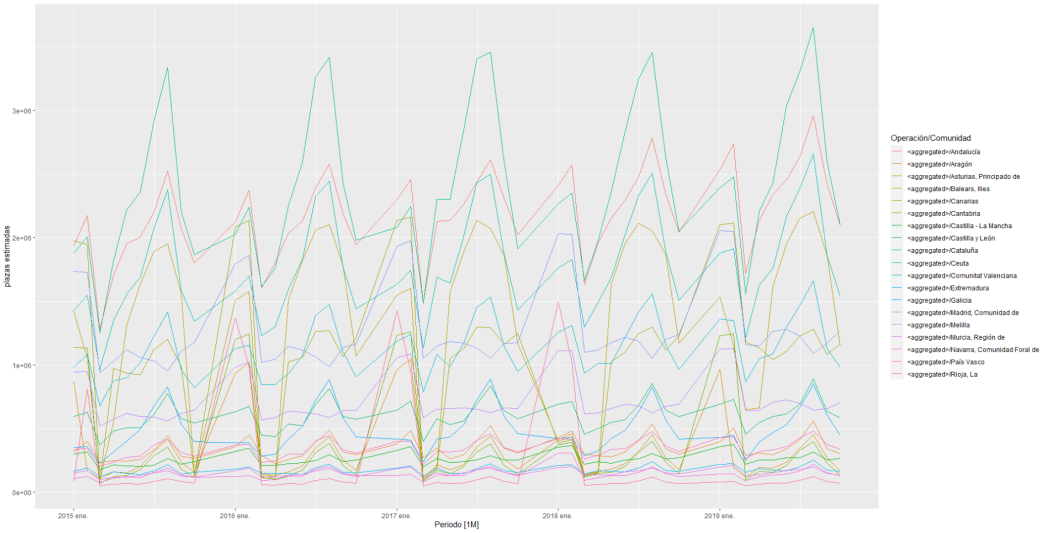


*Figure 7.14 After replacing missing values*

## 7.5 Annex 5: Methodology

**Crisp-DM** methodology is characterised by being the most popular. It is made up of six processes, which we will explain by adapting them to our project.

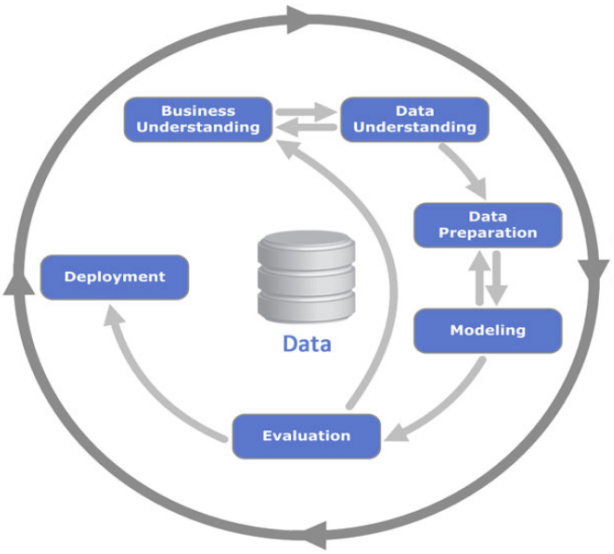


*Figure 7.15 Crisp-DM diagram*

- **Business Understanding**: the objective(s) of the team have been formulated from a business perspective in order to optimise the potential of the database. We start from a hotel occupancy database. With this initial context, a variety of objectives have been proposed, although it will be the process of data understanding that will help us to focus on them and make an optimal choice.

We understand the importance of the tourism business, especially in our country. In 2017 tourists produced 172.900 million euros four our economy, nearly 15 % of our PIB, so we understand how crucial this sector is for us, and the improvements that benefit maximization can produce.

Initially to we positioned as a hotel client point of view and tried to understand how could we increase client satisfaction and tourist arrivals to our country, nonetheless we understood that we couldn't do this because of the data we had, so we reconsidered our point of view and thought which business could benefit from hotel occupancy prediction, as having an insight about the amount of potential customers there can be in a certain period of time can help organizations prepare to suffice demand.

We ultimately decided on one hand, to help travel agencies/ other leisure businesses to improve marketing strategies and infrastructure optimisation, and on the other, help investment groups to decide the type of establishment (stars) and the region in which to build the hotels,in order to minimize the risk and maximize the benefits.

However we are aware of the problems caused by our data, as we don't have access to the economical data for the hotels, and we face very aggregated data, with poor specificity.

We are also aware that we can make erroneous predictions with our models but our concerns regarding this topic will be more extensively discussed in the impact assessment.

- **Data Understanding**: we have been working these days on this process, along with the previous one. However, it's in data understanding where we are spending most of our time, because once we have analysed the database properly, we will be able to formulate the objectives more thoroughly as we can understand what we can accomplish with our data.

- **Data Preparation**, **Modelling**, **Evaluation** and **Deployment** are the remaining processes. Here we'll specify the steps to be performed during the final part of the project, in order to achieve successfully data mining, we will need to merge our datasets to be able to do bottom up exploration, after we've achieved this we intend to do a regression model, or an exploratory time series model, that can allow us to interpret the importance of our variables. Afterwards as expected with time series data, we will model a time forecasting model to predict hotel occupation, mean stay and maybe other interesting variables, then we will evaluate the model and upgrade it to get better predictions, maybe playing with improvements with arima models.

The other methodology we have identified with has been **Scrum**, although we feel more comfortable with a slightly more flexible version called **SKI** (Structured Kanban Iteration).

This simplification has the same roles and meetings as Scrum, but is based on the working capacity of the members, which is useful because not everyone has the same amount of time available every day.

In addition, the cycles will be used to develop the ideas that have been proposed for the project (e.g. exploratory analysis) and thus, working on developing them and visualising the results. From this point, create new proposals and make more cycles, and always make planned meetings to follow carefully the development of the project.

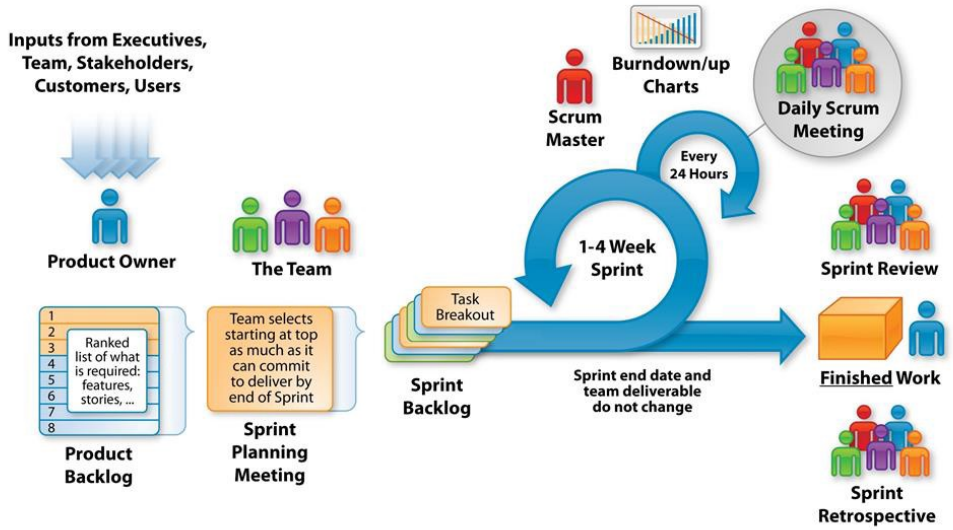


*Figure 7.16 SKI diagram*

Finally we decided to follow the Crisp-DM methodology, as its phases describe better our preferred line of work.

About monitoring our progress we are using trello to keep track of what we are planning to do, the things that we are currently doing and furthermore the activities we are planning to do. In addition we would like to recall that if we feel that is necessary we plan extra meetings to keep up with the work, although this meetings are not strict and we agree to meet the days in which all group members are available, this follows more the KSI methodology but we will implement this along with Crisp-DM because of convenience.

## 7.6 Annex 6: Foreign travelers

Evaluating the Naive Bayes, ARIMA, TSLM and ETS models, based on the values of foreign travelers, we can observe in the MAPE metric that the Naive Bayes model has the lowest value, since the predicted values deviate 2.35% from the real values. Since this model predicts in a simple way by repeating the values of previous periods, we decided to study the other possible models. The second model with the lowest predictive error is TSLM, with a value of 3.45% in the MAPE metric. In third place is the ETS model, whose percentage error, MAPE, only increases by 0.3%. Therefore, we decided to study it by analyzing each of the communities for these last two models.

| .model | Comunidad | Zonas | MAE | MAPE | RMSE | ACF1 |
|---|---|---|---|---|---|---|
| snaive | <aggregated> | <aggregated> | 71700.96 | 2.358095 | 100979.0 | -0.319012427 |
| arima | <aggregated> | <aggregated> | 132622.15 | 4.077851 | 168501.8 | -0.185418984 |
| tslm1 | <aggregated> | <aggregated> | 105850.33 | 3.446674 | 170851.5 | -0.119484027 |
| ets | <aggregated> | <aggregated> | 116059.29 | 3.719966 | 193343.2 | 0.004996718 |

*Figure 7.18 Global performance of each model*

We observe that the percentage error of the predicted values, which we measure using the MAPE metric, obtains in the case of the TSLM model too high values in some of the Autonomous Communities, as is the case of the Community of Navarre, with 52%, and the Balearic Islands, with 45%. For this reason, we decided to use the ETS model, as it has more constant errors.

| .model | Comunidad | Zonas | MAE | MAPE | RMSE | ACF1 |
|---|---|---|---|---|---|---|
| tslm1 | Extremadura | <aggregated> | 513.3125 | 21.711562 | 629.6334 | -0.16903321 |
| tslm1 | Navarra, Comunidad Foral de | <aggregated> | 625.7708 | 52.146658 | 797.0058 | -0.07179112 |
| tslm1 | Aragón | <aggregated> | 882.7917 | 11.669542 | 1011.3743 | -0.13323970 |
| tslm1 | Pirineos | <aggregated> | 986.0347 | 16.033708 | 1169.8730 | -0.09417775 |
| tslm1 | Asturias, Principado de | <aggregated> | 1089.1250 | 8.831588 | 1642.8225 | -0.25669895 |
| tslm1 | Murcia, Región de | <aggregated> | 1478.5417 | 15.112028 | 1673.0282 | 0.58680094 |
| tslm1 | País Vasco | <aggregated> | 2337.7083 | 6.983835 | 2757.7774 | 0.21528643 |
| tslm1 | Galicia | <aggregated> | 7254.5417 | 23.303742 | 8094.0692 | 0.40161231 |
| tslm1 | Comunitat Valenciana | <aggregated> | 22136.4583 | 12.518750 | 22512.2838 | 0.12124652 |
| tslm1 | Cataluña | <aggregated> | 18634.3333 | 1.870298 | 25546.6642 | -0.32796140 |
| tslm1 | Andalucía | <aggregated> | 25977.8611 | 8.736720 | 28428.5019 | 0.08946190 |
| tslm1 | Canarias | <aggregated> | 87954.5833 | 9.131441 | 92779.8530 | 0.62422928 |
| tslm1 | Balears, Illes | <aggregated> | 138908.1758 | 45.214951 | 161200.3148 | -0.28970293 |

| .model | Comunidad | Zonas | MAE | MAPE | RMSE | ACF1 |
|---|---|---|---|---|---|---|
| ets | Extremadura | <aggregated> | 522.6534 | 17.571586 | 625.3817 | -0.23122123 |
| ets | Navarra, Comunidad Foral de | <aggregated> | 585.3277 | 25.276799 | 830.7624 | 0.25098466 |
| ets | Aragón | <aggregated> | 899.0745 | 11.630360 | 1069.8359 | -0.16795484 |
| ets | Pirineos | <aggregated> | 968.1827 | 15.549050 | 1233.2739 | -0.07236654 |
| ets | Murcia, Región de | <aggregated> | 1316.9929 | 13.287065 | 1504.2806 | 0.52229389 |
| ets | Asturias, Principado de | <aggregated> | 1034.1193 | 8.597428 | 1624.6402 | -0.23356831 |
| ets | País Vasco | <aggregated> | 1922.9077 | 4.474031 | 2451.3425 | -0.52932300 |
| ets | Galicia | <aggregated> | 4113.4319 | 8.772318 | 5390.1032 | 0.17478339 |
| ets | Comunitat Valenciana | <aggregated> | 17760.0237 | 9.885535 | 18627.7913 | 0.27947468 |
| ets | Cataluña | <aggregated> | 18768.4793 | 2.178623 | 21709.3790 | -0.40879403 |
| ets | Andalucía | <aggregated> | 21145.4598 | 4.811532 | 27671.2396 | 0.49421749 |
| ets | Canarias | <aggregated> | 41533.8744 | 4.336531 | 49310.9361 | 0.57117791 |
| ets | Balears, Illes | <aggregated> | 89957.3301 | 21.666214 | 171977.8175 | -0.15433472 |

*Figure 7.18 Community performance of each model*

We can see now the prediction for number of travelers for the different zones, we can observe that there are zones as Costa de Valencia, Sur the Gran Canaria, Sur de Tenerife with a very wide confidence interval, we assume it is as these zones where several different areas are included, and this is what creates the poor fiability of this prediction.

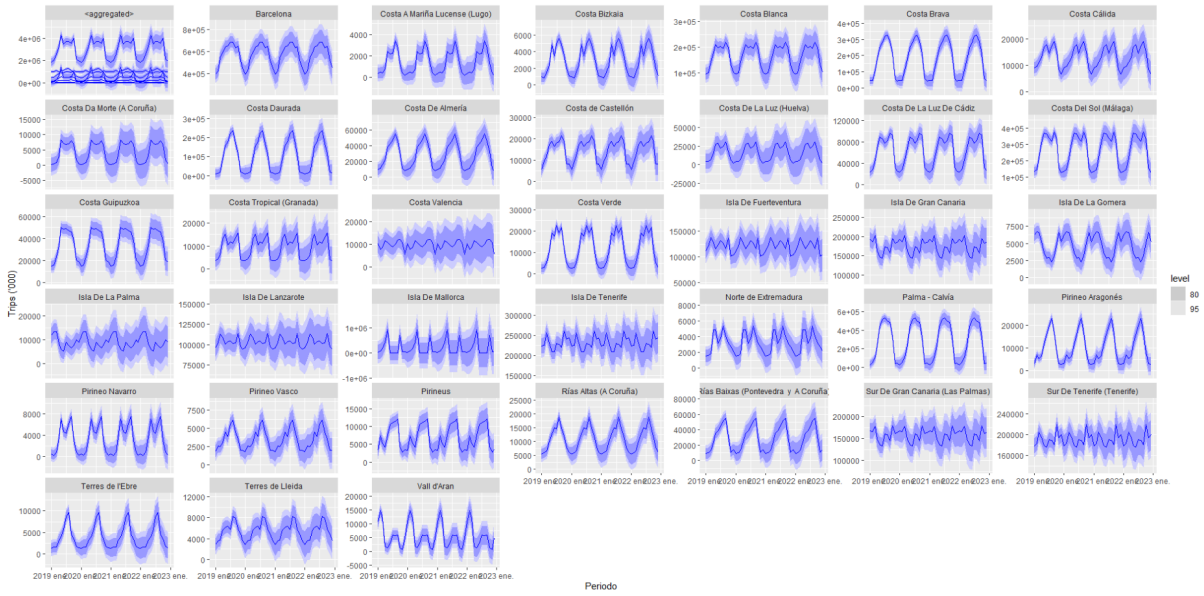


*Figure 7.19 Prediction for foreign travelers*

We will reproduce the same as with foreign travelers but with national travelers.

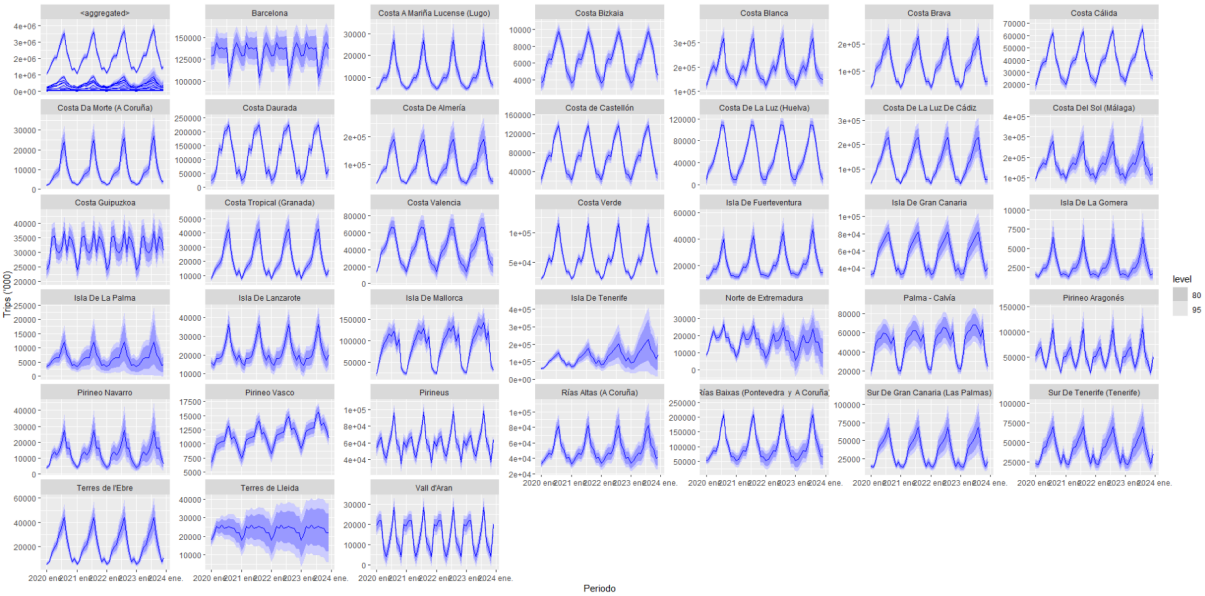


*Figure 7.20 Prediction for national travelers*

## 7.7 Annex 7: Use of technology

The project has been carried out largely with the Rstudio program, because we feel comfortable after years of practice with this language. In addition, by working with it during this project we have been able to learn more knowledge, libraries and new methods.

First of all, we have worked mainly with the **dyplr** library as it provides a simple grammar for the study of the different data frames and it is worth noting that the functions of this package are very fast, as they are implemented with the C++ language.

For data cleaning we have used the **tidyr** library, which allows us to sort the "dirty" data to obtain data objects in R in sorted format. The idea of the sorted data is that it is organized in such a way that each variable is in a column and each observation is in a row.

To separate the variables and the observations, we have to split the column into its two components. To do this tidyr has the separate function that takes the following arguments:

```
function (data, col, into, sep = "[^[[:alnum:]]+", remove = TRUE, convert = FALSE,
          extra = "warn", fill = "warn", ...)
```

In this way, we can divide the period variable into year and month:

```
Tipos_Aloj <- Tipos_Aloj %>% separate(Periodo, c("año", "mes"), "M")
```

After obtaining the different databases and having the mineable view explained before, we proceeded to make the model. Since we were dealing with periods we thought of studying time series, the reference we had from subjects studied as Digital Economics were the ARIMA models. We began to analyze these types of models in depth and even consulted with the professor of the subject where we learned this knowledge to be able to carry out our ideas. It helped us a lot.

Since we were dealing with time series to work with dates, we installed **lubridate**, a package with which we were already familiar from projects in other subjects. We reviewed the *Descriptive modeling* and *Visualization* practices to remind ourselves of its usefulness.

Also, thanks to the book 'Forecasting: Principles and Practice' we discovered new forecasting models that could be useful to us. Our data are hierarchical since they range from Autonomous Communities to Tourist Spots through Provinces and Tourist Zones. Therefore, after researching and reading the book we decided to study different models such as clustered time series among others. It was the most complicated part because of the difficulty of understanding the operation of the different libraries that were completely new to us.

We learned about the **tsibble** library that aims at managing temporal data and performing analysis in a smooth workflow. This package preserves temporal indexes as an essential data column and makes heterogeneous data structures possible. In addition to the tibble representation, a key composed of one or more variables is introduced to identify the observation units over time.

```
zon$Periodo <- yearmonth(as.character(zon$Periodo))
zonas <- zon %>% as_tsibble(key = c(i),index = Periodo, .drop = TRUE)
```

Following this library, we investigated and found the **feasts** package that facilitates the understanding of this type of representation, since it has a set of tools to analyze temporal data ordered in this format, tsibble.

Working with this format meant many execution errors that we were able to solve through Stack overflow, a public platform where we found a large collection of questions and answers about coding. It allowed us to see how other users had managed to solve the same errors. In this way, we applied the different tricks to our project until we reached the correct execution of the code.

One of the packages to highlight is **fable**, which provides us with a collection of univariate and multivariate time series models, including automatic ARIMA modeling. With fable we have the possibility to create, evaluate, visualize, combine models and make predictions.

```
zonas_gts <- zonas %>% aggregate_key(Zone, viajeros = sum(Viajeros_ext) + sum(Viajeros_esp),
                       est_med = sum(est_med), grado = sum(Grado_ocup_plazas),
                       establ = sum(Num_establecimientos_abiertos))

zonas_models <- train %>%
                model(ets = ETS(viajeros ~ season()), arima = ARIMA(viajeros ~ season())) %>%
                mutate(combination = (ets + arima ) / 2)

zonas_fc <- zonas_models %>% forecast(h = 12)
```

Finally, as for the visualization of the different communities, areas and tourist spots for the development of the mockup, we have used the **leaflet** library, that allows the creation of interactive web maps. We also use the **ggplot2** package for visualisations in both data preparation and model display.

## 7.8 Annex 8: Linear regression model

The objective of the linear regression model is to relate a dependent variable to several independent variables. Before running the time series, we ran, using the R language, linear regression models for the different databases available to us.

We use the **lm** function to run the model and create a new variable called *Viajeros_total*, which is the result of the sum of *Viajeros_ext* and *Viajeros_esp*, which we use as the dependent variable, since it is the one that gives us more general information on the number of tourists who stay overnight and, therefore, who travel to the corresponding place. We proceed to analyse the results of each of the datasets.

After making the model, we ran a summary of the model, in order to be able to interpret the results obtained. The last value, Pr(>|t|), is the one that indicates whether the variable is significant, in the case where the value is less than 0.05, or not significant, in the case where the value is greater than 0.05. In addition, star-shaped indications appear next to each value; the more star-shaped indications, the more significant the variable is with respect to the dependent variable, whereas if there aren't, the variable is not considered significant.

**Accommodation types**

In the database referring to the accommodation types, we can observe that all the variables, including the number of establishments open, the estimated number of bedplaces, the number of staff employed, the average stay and the overnight stays of foreigners and Spanish, are significant with respect to the total number of travellers.

```
Coefficients:
                                  Estimate Std. Error t value
(Intercept)                      9.809e+03  2.063e+03   4.755
tipos$Num_establecimientos_abiertos  2.498e+01  2.241e+00  11.145
tipos$Num_plazas_estim            1.699e-01  5.549e-02   3.062
tipos$Personal_emp                1.258e+01  5.509e-01  22.841
tipos$est_med                    -8.311e+03  5.198e+02 -15.989
tipos$Pernoctaciones_ext          5.216e-02  4.374e-03  11.925
tipos$Pernoctaciones_esp          4.244e-01  6.403e-03  66.280
                                 Pr(>|t|)
(Intercept)                      2.03e-06 ***
tipos$Num_establecimientos_abiertos  < 2e-16 ***
tipos$Num_plazas_estim            0.00221 **
tipos$Personal_emp                < 2e-16 ***
tipos$est_med                     < 2e-16 ***
tipos$Pernoctaciones_ext          < 2e-16 ***
tipos$Pernoctaciones_esp          < 2e-16 ***
```

*Figure 7.21 Linear Regression's Summary of Accommodation Types*

## Touristic Zones

```
Coefficients:
                                  Estimate Std. Error t value
(Intercept)                     -1.315e+04  2.498e+03  -5.262
zonas$Grado_ocup_habitaciones    3.150e+03  3.472e+02   9.072
zonas$Grado_ocup_plazas          1.706e+03  4.497e+02   3.793
zonas$Grado_ocup_plazas_finsemana -3.302e+03  2.621e+02 -12.600
zonas$Num_establecimientos_abiertos  2.968e+02  1.897e+01  15.642
zonas$Num_habitaciones           6.872e+00  1.648e+00   4.171
zonas$Num_plazas_estim          -3.068e+00  8.070e-01  -3.802
zonas$Personal_emp               8.687e+00  1.351e+00   6.428
zonas$Pernoctaciones_ext         1.871e-02  7.013e-03   2.668
zonas$Pernoctaciones_esp         1.750e-01  1.419e-02  12.335
zonas$est_med                   -1.536e+04  1.447e+03 -10.617

                                Pr(>|t|)
(Intercept)                     1.51e-07 ***
zonas$Grado_ocup_habitaciones    < 2e-16 ***
zonas$Grado_ocup_plazas         0.000152 ***
zonas$Grado_ocup_plazas_finsemana  < 2e-16 ***
zonas$Num_establecimientos_abiertos  < 2e-16 ***
zonas$Num_habitaciones          3.11e-05 ***
zonas$Num_plazas_estim          0.000146 ***
zonas$Personal_emp              1.47e-10 ***
zonas$Pernoctaciones_ext        0.007664 **
zonas$Pernoctaciones_esp         < 2e-16 ***
zonas$est_med                    < 2e-16 ***
```

*Figure 7.22 Linear Regression's Summary of Touristic Zones*

## Touristic Spots

```
                                Pr(>|t|)
(Intercept)                      < 2e-16 ***
puntos$Grado_ocup_habitaciones    0.00355 **
puntos$Grado_ocup_plazas         < 2e-16 ***
puntos$Grado_ocup_plazas_finsemana  < 2e-16 ***
puntos$Num_establecimientos_abiertos  < 2e-16 ***
puntos$Num_habitaciones         7.77e-13 ***
puntos$Num_plazas_estim          < 2e-16 ***
puntos$Personal_emp              < 2e-16 ***
puntos$Pernoctaciones_ext        < 2e-16 ***
puntos$Pernoctaciones_esp        < 2e-16 ***
puntos$est_med                   < 2e-16 ***
```

*Figure 7.23 Linear Regression's Summary of Touristic Spots*

Both on the basis of zones and tourist sites, all the variables, among which we find the degree of occupancy of rooms, places and weekend places, the number of establishments open, the number of rooms, the number of estimated places, the number of staff employed, the average stay and the overnight stays of foreigners and Spanish, are significant with respect to the total number of travellers.

## Provinces and communities

Finally, in the provinces and communities dataset, all the variables, including the occupancy rate of rooms, bedplaces and weekend bedplaces, the number of rooms, the number of estimated bedplaces, the average stay and overnight stays by foreigners and Spanish, are significant with respect to the total number of travellers, except for the number of establishments open and the number of staff employed, which are not.

```
Coefficients:
                                  Estimate Std. Error t value
(Intercept)                     -4.242e+04  6.294e+03  -6.739
prov$Grado_ocup_habitaciones     8.829e+03  4.791e+02  18.427
prov$Grado_ocup_plazas          -5.132e+03  5.834e+02  -8.797
prov$Grado_ocup_plazas_finsemana -2.270e+03  3.363e+02  -6.751
prov$Num_establecimientos_abiertos -2.634e+00  1.241e+01  -0.212
prov$Num_habitaciones            2.567e+01  1.765e+00  14.543
prov$Num_plazas_estim           -1.060e+01  9.628e-01 -11.005
prov$Personal_emp                7.797e-02  2.043e+00   0.038
prov$Pernoctaciones_ext         -2.654e-02  1.317e-02  -2.016
prov$Pernoctaciones_esp          5.264e-01  1.631e-02  32.273
prov$est_med                    -1.836e+04  2.881e+03  -6.372

                                Pr(>|t|)
(Intercept)                     1.84e-11 ***
prov$Grado_ocup_habitaciones     < 2e-16 ***
prov$Grado_ocup_plazas           < 2e-16 ***
prov$Grado_ocup_plazas_finsemana 1.70e-11 ***
prov$Num_establecimientos_abiertos   0.8319
prov$Num_habitaciones            < 2e-16 ***
prov$Num_plazas_estim            < 2e-16 ***
prov$Personal_emp                 0.9696
prov$Pernoctaciones_ext           0.0439 *
prov$Pernoctaciones_esp          < 2e-16 ***
prov$est_med                    2.10e-10 ***
```

*Figure 7.24 Linear Regression's Summary of Provinces and Communities*

## 7.9 Annex 9: Models and Code

### 7.9.1 Canarias - Arona's predictions

Looking at the graph, the prediction for the number of foreign travellers in 2023 remains more or less constant, without major changes throughout the months of the year. However, there is a small seasonality, as the value increases in the summer months and decreases in the winter months. In addition, the confidence in the year of hotel deployment is quite large, so the prediction could vary (up to a maximum error of 100,000 travellers) and not be as accurate.

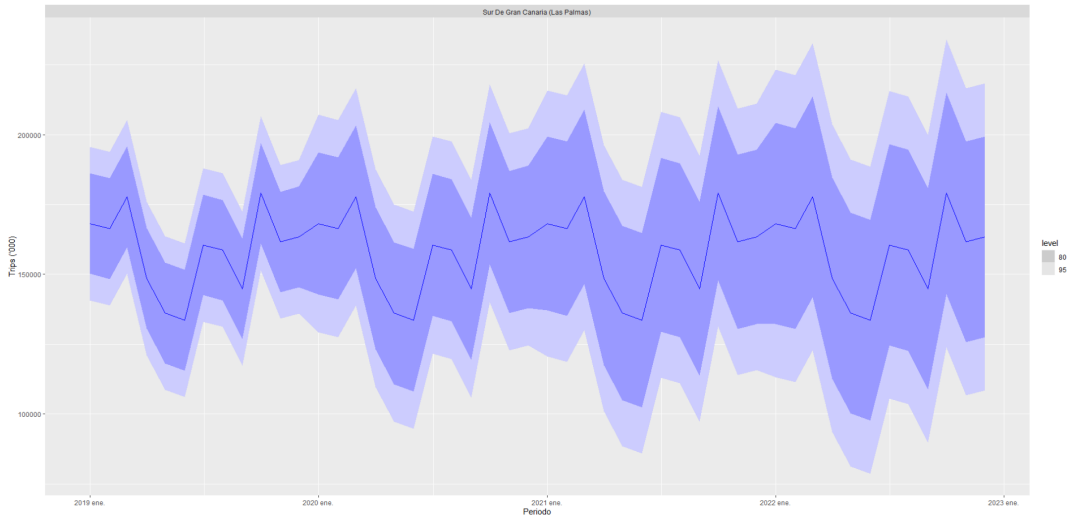


*Figure 7.25 Foreign travellers's predictions for Arona*

The predictions of national travellers are similar to those of foreign travellers. However, confidence is lower, which allows us to offer a more accurate prediction to the client. Note that although the scale is different, it refers to the same thing (100 means 100,000).

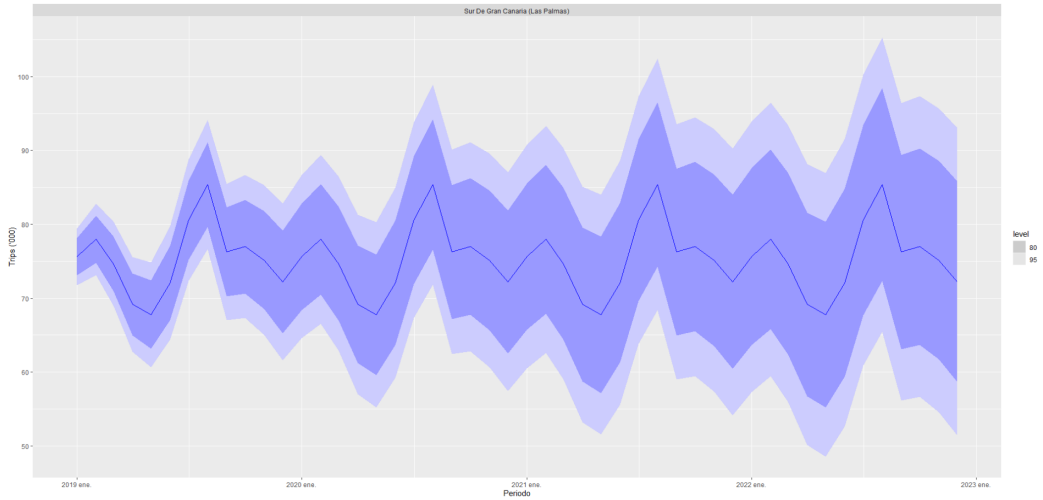


*Figure 7.26 National travellers's predictions for Arona*

### 7.9.2 Island of Mallorca's predictions

Looking at the graph, the predicted number of foreign travellers increases over the years, although it has a small drop in 2023. Moreover, it is also seasonal, as the values increase in summer and decrease in winter. In fact, the seasonal component is higher in Mallorca than in Arona. It should be noted that confidence is lower.

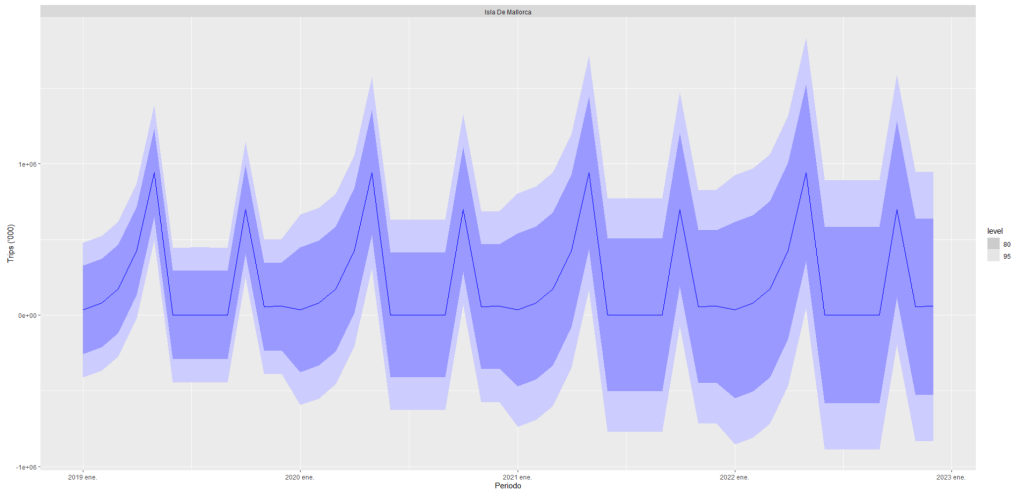


*Figure 7.27 Foreign travellers's predictions for island of Mallorca*

The predictions for domestic travellers are totally different as the confidence intervals are very small. This means that the values obtained for those years will be quite close to reality. Moreover, the number decreases sharply in the winter months, so the customer will have to take into account that at Christmas there will not be many Spaniards travelling to Mallorca. This is in contrast to the summer, when the number increases drastically. It should be noted that as the years go by, the values tend to increase.

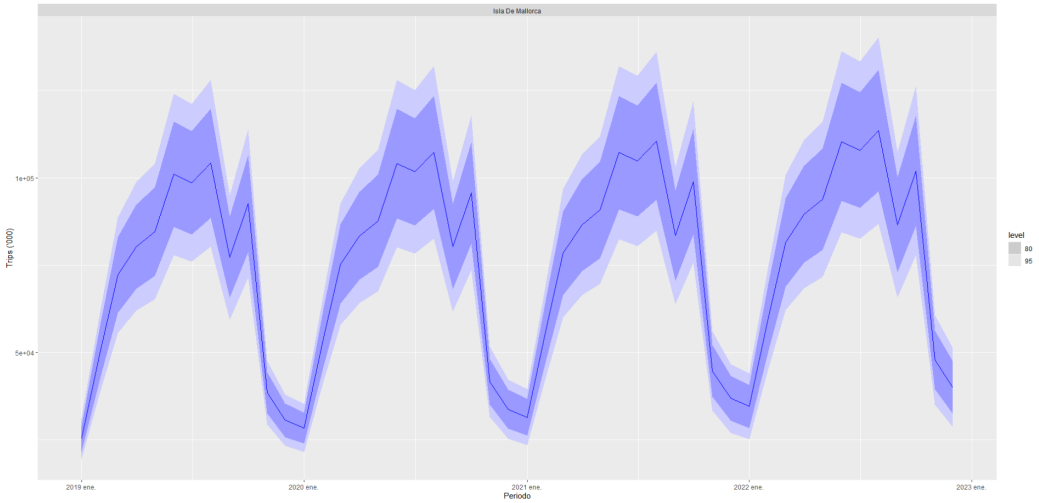


*Figure 7.28 National travellers's predictions for island of Mallorca*

### 7.9.3 Examples of the code used

```
residuals1<- train %>% model(ets = ETS(grado ~ season()+ trend())) %>% residuals()
residuals1 %>% autoplot()
residuals1 %>% autoplot() +
  facet_wrap(vars(Zonas), scales = "free_y", ncol = 3) +
  theme(legend.position = "none")

#arima
residuals <- train %>% model(arima = ARIMA(grado ~ season() + trend())) %>% residuals() %>%
  filter(!is_aggregated(Zonas))
autoplot(residuals, size = 1) +
  labs(y = "", x = "Periodo", title = "Residuos de grado de ocupación por Zona por Periodo (ARIMA)")+
  theme(plot.title = element_text(hjust=0.5, size = 25), legend.text = element_text(size =10),
        axis.title = element_text (size=15))

#accuray
accuracy <- fc %>% accuracy(test) %>%  arrange(RMSE)
accuracy <- select(accuracy,.model, Zonas, MAE, MAPE, RMSE)
res<-accuracy %>% filter(is_aggregated(Zonas))
```

*Figure 7.29 ARIMAand ETS's model validation*

```
#entrenamos los modelos
zonas_grad_models <- train %>%
  model(tslm1 = TSLM(grado),
        ets = ETS(grado ~ error()+trend() + season()),
        arima = ARIMA(grado ~ season()),
        snaive = SNAIVE(grado))

zonas_grad_models <- zonas_gts2 %>%
  model(
        ets = ETS(grado ~ error()+trend() + season()))
fc1 <- zonas_grad_models %>% forecast(h = 12)

fc2 <- zonas_grad_models %>% forecast(h = 48)

#autoplot de evaluación
fc1 %>% filter(!is_aggregated(Zonas)) %>%
  autoplot(test,level = NULL) +
  labs(y = "Trips ('000)") +
  facet_wrap(vars(Zonas), scales = "free_y")

#autoplot de la prediccion
fc2 %>% filter(!is_aggregated(Zonas)) %>%
  autoplot() +
  labs(y = "Trips ('000)") +
  facet_wrap(vars(Zonas), scales = "free_y")
```

*Figure 7.30 Prediction plots*