PREDICCIÓ DE LA DEMANDA EN UNA CADENA HOTELERA AMB MACHINE LEARNING

Grau en estadística aplicada UAB

01/10/2019 – 15/01/2020

bismart
Analytics
Artificial Intelligence

Miquel Vallverdú Esteve
Tutora: Raquel García Blanco (Bismart)
Tutor: Albert Ruiz Cirera (UAB)

Agraïments

Abans de res, voldria expressar el meu agraïment a totes aquelles persones que m'han ajudat a realitzar aquest treball.

Per començar, a les persones de Bismart, l'empresa que m'ha donat l'oportunitat de desenvolupar el projecte. En especial a la meva tutora, la Raquel García Blanco, per totes les hores que ha dedicat a ajudar-me, i a les seves constants crítiques constructives que m'han portat a aprendre tant i a sentir-me orgullós del treball que he realitzat. A en Pau Maymó, per haver dedicat un munt d'hores de manera desinteressada i generosa en ajudar-me quan més ho he necessitat. I a en Joan Teixidó, per ajudar-me també en els aspectes on menys còmode m'he sentit.

En segon lloc, a les persones que he conegut a la Universitat Autònoma de Barcelona. A l'Albert Ruiz, el meu tutor, i a en Toni Lozano, per haver-me aconsellat quan els hi he demanat. Així com també a tots els professors i companys que, durant els quatre anys i mig d'estada a la universitat, m'han ajudat a aprendre els coneixements necessaris per a poder concloure els estudis amb aquest treball.

I per últim, als éssers més propers, per l'ànim que he rebut i per que m'han demostrat que podia comptar amb ells quan fes falta.

Abstracte

Aquest treball té la intenció de transmetre els passos a seguir per a construir un model d'aprenentatge automàtic capaç de predir la demanda d'una cadena hotelera, des del tractament de les dades en 'brut' fins a la posada en escena del model. Es construiran diferents models de regressió per a predir la variable d'interès, emprant una tècnica d'aprenentatge supervisat, i es compararan entre ells per a escollir-ne el que es consideri millor. Finalment, es prova l'eficàcia del model escollit amb dades reals.

Abstracto

Este trabajo tiene la intención de transmitir los pasos a seguir para construir un modelo de aprendizaje automático capaz de predecir la demanda de una cadena hotelera, desde el tratamiento de datos en 'sucio' hasta la puesta en escena del modelo. Se construirán diferentes modelos de regresión para predecir la variable de interés, usando una técnica de aprendizaje supervisado, y se compararán entre ellos para escoger el que se considere mejor. Finalmente, se prueba la eficacia del modelo escogido con datos reales.

Abstract

This work intends to transmit the steps to follow to build a machine learning model able to predict a hotel chain demand, from the handling with raw data to the staging of the model. Different regression models will be constructed to predict the variable of interest, using a supervised learning technique, and will be compared between them to choose the one that is best considered. Finally, the performance of the chosen model is tested with real data.

ÍNDEX

1 INTRODUCCIÓ

Contextualització

Aquest projecte s'ha desenvolupat a l'empresa *Bismart Business Intelligence Specialist Services SL*, a on he cursat les meves pràctiques, tant curriculars com extracurriculars. Bismart es defineix com una consultora tecnològica especialitzada en solucions i serveis de Data Management i Analytics. Es dediquen a oferir solucions de negoci i solucions tecnològiques. Les solucions tecnològiques s'ofereixen per a tres àrees diferents: per a l'àrea d'integració de dades, per a l'àrea de gestió i emmagatzematge de dades i per a l'àrea d'anàlisi de dades. En aquest sentit, l'empresa vol ampliar el ventall de serveis i començar a oferir solucions orientades en l'àmbit de la ciència de dades o *Data Science*. I aquí és on aquest projecte entra en escena.

L'objectiu principal d'aquest projecte serà desenvolupar un model predictiu amb *Machine Learning* per a la demanda d'una cadena hotelera client. Aquest projecte es durà a terme inicialment com a una prova de concepte, amb l'objectiu d'incorporar-lo al portfoli de productes de l'empresa si el resultat es interessant.

Com que diversos clients de Bismart són del sector hoteler, l'empresa vol convertir els serveis que els hi ofereix en una solució replicable, per tal de fer els mínims canvis possibles a l'hora de vendre la solució a un client o un altre, així que serà essencial documentar-lo de tal manera que qualsevol altre treballador de l'empresa pugui saber en tot moment quins passos s'han seguit, i que serveixi de guia per quan es vulgui replicar per a altres clients del mateix sector.

Sobre el Machine Learning i la seva aplicació en el problema

El *Machine Learning* o aprenentatge automàtic és una disciplina dins l'àmbit de la intel·ligència artificial capaç de desenvolupar sistemes que aprenguin automàticament. Aprendre, en aquest context, vol dir identificar patrons complexes en milions de dades. I la màquina que aprèn, realment és un algoritme que és capaç de predir comportaments futurs mitjançant les relacions i patrons que ofereixen les dades.

Perquè és tant important i està tant de moda el *Machine Learning* en l'actualitat? Doncs per que la quantitat de dades que es generen creix de manera exponencial, tot el que fem genera dades. Extreure informació valuosa sobre elles suposa un avantatge competitiu. Ja que com a resultat s'obtenen prediccions d'alt valor que ajuden a prendre millors decisions i a desenvolupar millors accions de negoci[1]. A més, s'ha reduït molt la dificultat de computar algoritmes de *Machine Learning*. Anys enrere, començar a moure's en aquest àmbit era complex i estava a l'abast de molt pocs, en l'actualitat, qualsevol que disposi d'un ordinador a casa pot començar a introduir-se en aquesta disciplina, ja que tot el codi necessari per a la computació es troba a internet. Lo únic que cal fer com a usuari és tractar aquest codi de manera específica per a les dades que es disposen.

Un algoritme de *Machine Learning* pot 'aprendre' de dues maneres diferents, de manera supervisada o de manera no supervisada. La diferència entre elles és que en l'aprenentatge supervisat, l'algoritme 's'entrena' fent servir resultats ja coneguts per a la variable resposta, mentre que en l'aprenentatge no supervisat no es coneixen resultats a priori.

En aquest cas s'aplicarà una tècnica de *Machine Learning* per a predir el nombre d'habitacions ocupades, agrupades per certes característiques com el tipus d'habitació, el tipus de règim i d'altres que més endavant es detallaran, per a una cadena hotelera, en funció de les dades que

tenen guardades. Per tant, com es coneixen les dades ja passades sobre el nombre d'habitacions ocupades, serà un **model d'aprenentatge supervisat**.

Es poden aplicar dues tècniques en l'aprenentatge supervisat: tècniques de classificació i tècniques de regressió. En aquest cas, **s'utilitzaran tècniques de regressió**, ja que la variable resposta és de caràcter numèric.

L'objectiu d'aplicar una tècnica de *Machine Learning* sobre aquestes dades, és el de poder oferir a la cadena hotelera unes prediccions prou acurades que els ajudin a prendre decisions de les quals en puguin obtenir un benefici.

Motivació

Les motivacions que m'han portat a realitzar aquest projecte són les següents:

- La importància creixent de l'aprenentatge automàtic i la ciència de dades en el món professional.
- Poder conèixer de primera mà la feina d'un científic de dades professional
- Aprendre a tractar amb tots els problemes que puguin sorgir a l'hora de solucionar un problema de *Machine Learning*.
- La gran varietat de conceptes i mètodes que es tracten en aquest treball, que puc relacionar amb els coneixements adquirits en el grau d'estadística.
- Aprendre nous llenguatges de programació com *SQL* i *Python*.

En resum, la gran motivació és aprendre tot el que pugui i guanyar una experiència que m'ajudi en el futur a ser un bon professional d'aquest àmbit.

2 PLANTEJAMENT I OBJECTIUS

En aquest treball es podran veure detalladament tots els passos que conformen el procés d'elaborar un model de *Machine Learning* adient per predir la demanda d'una cadena hotelera. Des de l'anàlisi i neteja de les dades en "brut", fins al desenvolupament del model predictiu.

Saber i conèixer la tipologia dels hotels pot ser molt útil a l'hora de prendre alguna decisió que no depengui únicament de les dades. Així que s'ha de saber que la cadena hotelera que proporciona les dades de l'estudi és una cadena amb 7 hotels diferents arreu de Catalunya. Dos per la Costa Brava, un per la costa de Barcelona i quatre per la Costa Daurada. Pel que fa la tipologia dels hotels, són tots hotels a prop de la platja i de quatre estrelles, bàsicament busquen clients disposats a passar-hi les vacances. Tots els hotels disposen de piscina, espais recreatius i de restauració.

Per a la realització d'aquest treball, es disposa d'una base de dades real on hi ha informació de cada reserva feta a un hotel determinat pertanyent a la cadena hotelera en qüestió. Hi ha informació de les reserves, de la producció i de l'ocupació en diferents taules, i s'ha de trobar la manera d'ajuntar aquestes taules de manera adequada per a obtenir el conjunt de dades adient, que reculli la màxima informació possible per al desenvolupament del model. Creant també, a partir de les dades existents, les variables necessàries.

L'**objectiu principal** del treball és desenvolupar, mitjançant la informació que pugui proporcionar aquesta base de dades, un model adient per predir la demanda d'habitacions, en el camp de l'aprenentatge automàtic. I mostrar tota la sèrie de passos a seguir per a desenvolupar-lo, de manera clara i entenedora.

Finalment, es provarà la precisió del model escollit amb dades reals i es representaran els resultats obtinguts, amb la intenció de ser capaços de poder oferir conclusions interessants per a la cadena hotelera client sobre les prediccions fetes. Com per exemple, si en un període determinat els hi convindria pujar els preus ja que ho tindran tot ple, o al contrari, si convindria baixar-los per aconseguir vendre més habitacions ja que amb els preus actuals no ho ompliran tot.

Per acabar, es redactaran totes aquelles conclusions i reflexions sobre el projecte, tant d'àmbit professional com d'àmbit acadèmic o personal.

3 DISSENY DEL PROJECTE: FULL DE RUTA

Per a la realització del projecte hem elaborat un full de ruta amb les següents tasques dividides en tres blocs, resumint el que s'hi podrà trobar en cada una d'elles:

ESTUDI I DEFINICIÓ DEL PROBLEMA

El primer pas és analitzar la base de dades, entendre la informació proporcionada de cada una de les variables i identificar i definir la variable objectiu de l'estudi i les candidates a predir-la.

DATA ENGINEERING: MANIPULACIÓ DE DADES

Un cop s'han analitzat les taules de la base de dades i es té clar quina informació ha de proporcionar el conjunt de dades final, es fan tots els passos possibles per a aconseguir-la. Bàsicament, en aquest bloc s'explicaran els passos que fan referència a tot el procés de manipulació de la base de dades per a obtenir el conjunt de dades esmentat, i deixar-lo llest per començar a desenvolupar el següent bloc.

DATA SCIENCE: A LA RECERCA DEL MILLOR MODEL

La ciència de dades és un camp interdisciplinari que involucra mètodes científics, processos i sistemes per extreure coneixement o un millor enteniment de les dades. Es pot definir també com un concepte per unificar estadístiques, anàlisis de dades, aprenentatge automàtic i els mètodes relacionats per a comprendre i analitzar els fenòmens reals[2]. En aquest cas, comprendre i analitzar el fenomen de la demanda en una cadena hotelera.

Pre-processament

Es mostra tot el que fa referència a la neteja i transformació de les dades com ara analitzar la presència de valors nuls i modificar-los o eliminar-los, canviar els formats de les variables que calgui canviar, eliminar variables que no aportin informació útil, etc, per tal de resoldre els possibles errors que aportin les dades.

Anàlisi descriptiu

En qualsevol estudi de caire estadístic és essencial elaborar un anàlisi descriptiu, cal estudiar la distribució de cada una de les variables, així com el seu comportament respecte la variable objectiu. Fent un anàlisi univariat de totes les variables, així com un anàlisi bivariat de cada una d'elles respecte la variable d'interès, per tal d'entendre millor la informació que proporcionen i detectar si hi ha diferències en els comportaments de cada nivell respecte la variable objectiu.

Feature Engineering

Per tal d'ajudar al futur model a desenvolupar unes prediccions acurades, cal analitzar quines variables seran les que aportaran informació per a fer-les. Aquest apartat engloba tant la creació com l'eliminació de variables. Per a seleccionar les variables explicatives que formaran part del model predictiu, s'analitzaran aspectes com les correlacions o la multicol·linealitat entre elles.

Construcció i entrenament del model

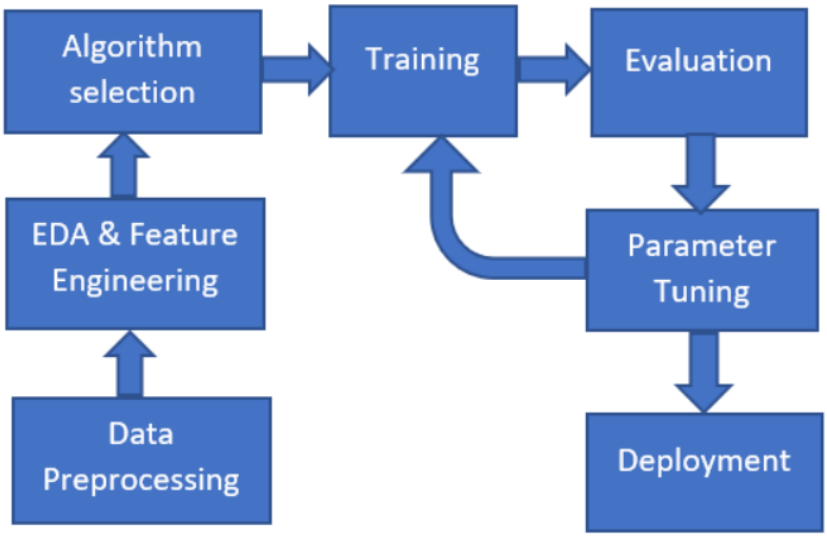
Un cop tenim les variables predictores de la variable resposta, es construiran diversos models. Es dividiran les dades en dues mostres, una d'entrenament i una altra de test, la mostra test no contindrà informació sobre la variable resposta. S'entrenaran els models buscant millorar les prediccions fetes per a la mostra test amb les dades d'entrenament.

Avaluació del model

Per avaluar la potència i precisió dels models es tindrà en compte el valor de l'R^2 i alguns valors d'error com l'error quadràtic mitjà (MSE) o l'error absolut mitjà (MAE). Es compararan els valors que ofereixen cada un dels models entre ells i s'escollirà el model que més s'adeqüi a l'objectiu del treball.

Posada a en escena

Es realitzarà una prova real per al model escollit per a comprovar que l'algoritme funciona correctament. El model haurà de ser capaç de predir dades noves que no s'han fet servir en cap moment, i s'avaluaran els resultats.



Il·lustració 1: Esquema resum del Bloc 3: Data Science

4 INDICACIONS GENERALS PER AL SEGUIMENT DEL PROJECTE

Per al seguiment del projecte teniu a la vostra disposició diversos arxius que complementen l'explicació feta en aquest document, tant *.csv* com *.ipynb*, a més de l'annex que s'hi troba en les pàgines finals.

Aquests arxius es troben en el repositori GitHub del següent enllaç:

https://github.com/KeloV10/TFG-MachineLearning

Les **indicacions generals** són les següents:

Per a l'estudi i definició del problema i per a l'extracció de dades, el codi emprat s'hi troba en l'annex, a més, en l'explicació dels passos s'hi detalla específicament on es pot consultar.

A partir del preprocessament fins a l'avaluació del model, es farà un resum/explicació del contingut desenvolupat en els *Jupyter Notebooks* (els arxius *.ipynb*), mostrant imatges que il·lustrin les explicacions quan sigui el cas.

En els *Notebooks* s'hi troba el codi *Python* necessari per a l'execució de cada pas, a més de comentaris i gràfics addicionals. La idea és que els *Notebooks* siguin una guia que ajudi, si cal, a seguir l'explicació dels passos seguits per a desenvolupar el projecte. Hi ha un *Notebook* per a cada un dels apartats, degudament indicats a l'inici de l'explicació de cada un d'ells.

Per últim, també s'indicarà en cada un dels passos, l'arxiu *.csv*, de la carpeta *datasets*, que s'utilitza per dur-los a terme.

4.1 EINES PER A DESENVOLUPAR EL PROJECTE

El software que s'utilitza per a fer el primer estudi de les dades i entendre la informació que proporciona cada una de les variables, definir la variable objectiu i construir i definir el conjunt de dades o dataset final, serà el *Microsoft SQL Server Management*. On s'hi treballa amb codi *SQL*.

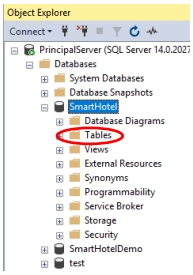
Un cop es tingui el conjunt de dades final, s'extreurà per fer tota la resta del projecte amb *Python* a través del *Jupyter Notebook*.

5 DESENVOLUPAMENT DEL PROJECTE

S'apliquen cada una de les tasques del full de ruta

5.1 ESTUDI I DEFINICIÓ DEL PROBLEMA

S'identifica quina és la base de dades en la que es treballarà. Ens connectem a un servidor remot de *Microsoft SQL Server* Management anomenat *PrincipalServer*. La base de dades que s'ha fet servir en aquest projecte s'anomena *SmartHotel*.



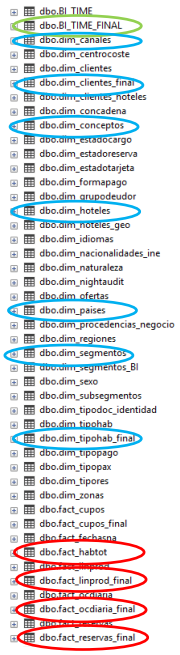
Il·lustració 2: Bases de dades

Si es despleguen les taules, veiem que tenim un total de 53 taules.

Com que no es faran servir totes, ara l'objectiu és analitzar cada una d'elles i veure quines poden ser les taules que proporcionin informació útil per al model.

Fent un primer cop d'ull es decideix que les taules encerclades en vermell en la *il·lustració 3* són les taules de fets que es faran servir per a desenvolupar el conjunt de dades final, i les encerclades en blau, són les taules de dimensions que proporcionen informació sobre les taules de fets.

En aquest cas, la **variable resposta**, que és el **nombre d'habitacions ocupades** en un dia determinat, i amb altres característiques que a continuació s'exposaran, s'haurà de crear a partir de les taules de fets. També és interessant saber el nombre d'habitacions previstes per ocupar, en aquesta mateixa data, en les setmanes anteriors fins a un any, és a dir, saber l'evolució per setmanes, durant un any, de la previsió de les habitacions que s'ocuparan en un dia determinat.



Il·lustració 3: Taules de la base de dades

El conjunt de dades final haurà de contenir informació de la variable resposta, és a dir, el **nombre d'habitacions ocupades** agrupades per **hotel**, **tipus d'habitació**, **tipus de règim**, **tipus de segment**, **client**, **país de l'agència client** i **tipus de canal**, per la **data d'ocupació**, amb el **preu mig** per a cada agrupació. A més del **nombre d'habitacions i el preu mig de cada setmana enrere fins a 52 setmanes** (un any). Conformant un total de 114 variables.

Les taules de fets contenen:

- *dbo.fact_linprod_final*
  - 10.114.440 files
  - 71 columnes
- *dbo.fact_reservas_final*
  - 687.936 files
  - 57 columnes
- *dbo.fact_habtot*
  - 307.094 files
  - 8 columnes
- *dbo.fact_ocdiaria_final*
  - 2.250.451 files
  - 44 columnes

La taula encerclada en verd, és una taula de temps auxiliar, on hi ha camps referents a l'any, la setmana, el dia, etc. Aquesta taula ajudarà a desenvolupar les variables per a la previsió d'ocupació.

S'analitzen les variables de cada una de les taules per veure quines seran útils per al model, hi ha variables que poden ser útils però no contenen informació, i d'altres que directament no proporcionen dades rellevants.

En les taules hi ha informació des de l'any 2011 fins l'any 2020. Per a fer aquest treball només s'han tingut en compte dades **a partir de 2018**. Per tant, el nombre de files es redueix considerablement. A més, com s'ha comentat en la secció anterior, al final del projecte es voldrà fer una prova real, amb dades que ja han passat, així que es decideix agafar les dades a partir de 2018 **fins a juliol del 2019 inclòs** per a construir el model. Les dades d'agost i setembre del 2019 es guardaran a part, simulant que encara no han passat, per a la realització de la prova amb dades reals.

Finalment, les variables, per taules, que formaran part del conjunt de dades final, o bé ajudaran a desenvolupar-lo[1], seran les següents:

- ***dbo.fact_reservas_final***: taula de reserves, cada fila correspon a una única reserva amb les característiques d'aquesta, hi ha tantes files com reserves.
  - *idhotel*: codi identificador de l'hotel, p.e. *D041*
  - *idreserva**: codi identificador de la reserva, p.e. *542765*
  - *idcliente*: codi identificador de l'agència client d'on prové la reserva, p.e. *0000002003*
  - *idtipohab*: codi identificador del tipus d'habitació que s'ha reservat, p.e. *DBL*
  - *idregimen*: codi identificador del tipus de règim que ha comprat la reserva, p.e. *MP*
  - *id_canal*: codi identificador del tipus de canal per on s'ha produït la reserva, p.e. *B2B*
  - *fechareserva**: dia de la formalització de la reserva, p.e. *2018-02-21*
  - *fechacancelacion**: dia de la cancel·lació de la reserva, en cas que no es cancel·li, NULL, p.e. *2018-04-23*
  - *fechaentrada**: dia que la reserva té previst entrar a l'hotel, p.e. *2018-08-08*
  - *fechasalida**: dia que la reserva té previst sortir de l'hotel, p.e. *2018-08-13*
  - *cantidad**: número corresponent a la quantitat d'habitacions reservades per reserva, p.e. *1*
  - *idpais_cliente*: codi identificador del país de l'agència client que ha venut la reserva, p.e. *ES*
  - *idsegmento*: codi identificador del segment de la reserva, només hi ha informació de si és per vacances o és desconegut, p.e. *VACAC*

- ***dbo.fact_linprod_final***: taula de producció, cada fila correspon a un càrrec fet a una reserva, pot haver-hi més d'un càrrec per reserva, hi ha tantes files com càrrecs.
  - *idreserva**: ídem que l'explicació anterior
  - *fechaprod**: dia on s'hi va produir un càrrec sobre un concepte per a una reserva en concret, p.e. *2018-08-30*

[1]Les variables marcades (*) no apareixeran al conjunt de dades final però ajuden a desenvolupar-lo

- *idconcepto**: codi identificador del tipus de concepte per el qual existeix un càrrec, p.e. *TI*
- *importessin**: quantitat corresponent al càrrec i d'on s'extreuran les variables referents al preu, p.e. *44.972727*

- ***dbo.fact_ocdiaria_final***: taula d'ocupació, cada fila correspon a un dia on una reserva estava ocupant un hotel. Per a cada reserva, existiran tantes files com nits hagi comprat dita reserva.
  - *idreserva**: ídem que l'explicació anterior
  - *fechaocup**: data corresponent a l'ocupació d'una reserva, p.e. *2019-09-16*

- ***dbo.BI_TIME_FINAL***: taula de temps, cada fila correspon a un dia pel qual tenim informació, hi ha tantes files com dies presents en la base de dades. En la resta de variables hi ha informació sobre les característiques del dia, com l'any, el semestre, el mes, la setmana, etc. És una taula creada expressament per ajudar a fer operacions en les que sigui necessari tenir en compte el temps.
  - *fecha**: data, p.e. *2019-02-04*
  - *AñoSemana**: any i setmana, separats per un punt, de la data en concret, p.e. *2018.30*

Les variables que podem relacionar amb les taules de dimensions encerclades en blau en la *il·lustració 2*, i que per tant, disposen d'informació complementaria que ajuda a entendre el significat dels seus valors, són les següents:

- ***dbo.fact_reservas_final***
  - *idhotel* → ***dbo.dim_hoteles***
  - *idcliente* → ***dbo.dim_clientes_final***
  - *idtipohab* → ***dbo.dim_tipohab_final***
  - *id_canal* → ***dbo.dim_canales***
  - *idpais_cliente* → ***dbo.dim_paises***
  - *idsegmento* → ***dbo.dim_segmento***

- ***dbo.fact_linprod_final***
  - *idconcepto* → ***dbo.dim_conceptos***

5.2 DATA ENGINEERING: MANIPULACIÓ DE DADES

A continuació es mostra una guia-resum de passos a seguir per a aconseguir el conjunt de dades final amb la informació que es demana. El codi emprat en cada pas es mostra i es comenta a l'annex.

1. Abans que res, s'ha d'efectuar una categorització o 'mapping' de les variables categòriques que faci falta, segons l'anàlisi primari que s'ha fet. En aquest cas les variables que es retocaran seran les següents: *idtipohab*, *idregimen*, *id_canal*, *idpais_cliente*, *idcliente*. L'opció triada per a fer-ho consisteix en duplicar la taula *dbo.fact_reservas_final* i en la nova hi haurà exactament la mateixa informació però amb les variables esmentades categoritzades com es pretén. Per tant, els passos posteriors enllaçaran amb la nova taula que s'anomena *dbo.fact_reservas_mapping*. L'objectiu d'aquest pas és el de reduir nivells de les variables per així també reduir dimensions en el conjunt de dades final [Annex, Queries 1-5 i Observació 1].

2. Crear una taula on hi hagi l'import agrupat per reserves i per data de producció. Per exemple, una reserva que entri el dia 1/1 i surt el 4/1, estarà 3 nits a l'hotel, amb la qual cosa, es mostraran 3 files en aquesta taula, una per cada pagament corresponent a la nit que s'allotja. Obtenint així, el cost per nit que li suposa l'habitació a cada reserva [Annex, Query 6].
3. En cas de trobar valors nuls, es substitueixen per l'import més proper tenint en compte els tipus d'habitació, de règim, de segment i de canal. Si no troba cap, fa una mitjana dels imports d'un mateix tipus d'habitació i règim.
   Les dades estan filtrades a partir de l'any 2018 per a la data d'entrada, s'eliminen els conceptes que no tinguin a veure amb el tipus de règim [Annex, Query 6]. Ja que hi ha molts conceptes que fan referència a menús pagats en el restaurant, per exemple, i no tenen res a veure amb el preu de l'habitació, i amb conseqüència, no ens interessen.
4. Per veure com va oscil·lant el nombre de reserves fetes per a un dia determinat al llarg del temps, es creen dues noves taules a partir de l'anterior. On hi hagi, en una taula, l'import i el nombre de reserves, agrupades per dia d'ocupació, per a les reserves que no es cancel·len, i en l'altra, el mateix però en resultat negatiu per a les reserves que es cancel·len, llavors es "fusionen" les taules amb un *union all*. S'obtindrà una taula amb el nombre d'habitacions ocupades i l'import mig agrupat tal i com s'explica en l'apartat anterior(estudi i definició del problema), així com també el recorregut durant un any [Annex, Query 6].
5. Per últim, s'han de despivotar les files per obtenir una columna per a cada setmana anterior a la data d'ocupació, i una fila per cada data d'ocupació amb les característiques concretades anteriorment [Annex, Query 7].

Finalment, el conjunt de dades final està llest per començar a desenvolupar el projecte en llenguatge *Python*.

Variables finals i objectiu de l'estudi

Abans de començar a tractar amb el conjunt de dades final, és convenient recordar i aclarir quines són les variables que formen part d'aquest conjunt, ja que algunes d'elles són variables creades en el procés de tractament de dades en *SQL Server Management* com és el cas de la variable resposta. Així com també, recordar i aclarir l'objectiu d'aquest projecte:

La variable resposta pren el nom de ***NHab***. I com s'ha esmentat en la secció de l'estudi i definició del problema, és una variable que proporciona informació del **nombre d'habitacions reservades(*NHab*) per a un dia(*fechaocupacion*), per a un hotel(*idhotel*), per a un tipus d'habitació(*idtipohab*), per a un tipus de règim(*idregimen*), per a un canal(*id_canal*), per a una agència client(*idcliente*), per a un país d'origen de l'agència client(*idpais_cliente*) i per a un segment(*idsegmento*) determinats**. Cada una de les files del conjunt de dades creat mostrarà informació del nombre d'habitacions que s'ocupen agrupades per a les característiques esmentades, a més de mostrar informació del nombre d'habitacions previstes per ocupar, amb aquestes característiques, en cada una de les setmanes anteriors fins a un any (***S1H, S2H, ..., S52H***), el preu mig que va tindre una habitació amb aquestes característiques (***PreuHab***) i el preu previst que tindria en cada una de les setmanes anteriors fins a un any (***S1, S2, ..., S52***). Convé

recordar també, que les dades mostren les observacions de **tot l'any 2018 fins el mes de juliol del 2019**. Amb un total de 52.688 files i 114 columnes.

	idhotel	fechaocupacion	idcliente	idtipohab	idregimen	idsegmento	id_canal	idpais_cliente	NHab	S1H	S2H	S3H
0	D021	2018-03-28	Huespedes	DBL	MP	ZZZ	ZZ	ES	0	0	0	0
1	D021	2018-03-29	Huespedes	DBL	AD	ZZZ	ZZ	ES	7	7	7	5
2	D021	2018-03-29	Huespedes	DBL	MP	ZZZ	ZZ	ES	6	5	4	3
3	D021	2018-03-31	Huespedes	DBL	AD	ZZZ	ZZ	ES	15	15	7	6
4	D021	2018-03-30	Huespedes	DBL	AD	ZZZ	ZZ	ES	23	18	9	6

	S51H	S52H	PreuHab	S1	S2	S3	S4		S50	S51	S52
	0	0	90.895386	90.895386	90.895386	90.895386	90.895386		90.895386	90.895386	90.895386
	0	0	133.241558	123.293506	123.293506	129.120000	129.120000		170.909090	170.909090	170.909090
[...]	0	0	156.207576	187.449091	196.447727	215.869697	215.869697	[...]	211.645454	211.645454	211.645454
	0	0	130.412624	129.869594	124.784195	125.410349	128.746965		170.909090	170.909090	170.909090
	0	0	138.602766	138.320706	144.462626	157.103030	157.103030		170.909090	170.909090	170.909090

Il·lustració 4: Capçalera del conjunt de dades

L'**objectiu** a assolir amb aquestes dades és el de desenvolupar un model d'aprenentatge automàtic, capaç de **predir la variable objectiu (*NHab*)**, mitjançant les variables explicatives esmentades. Però no hi seran totes, ja que algunes s'hauran d'eliminar si es considera que no aportaran informació útil per a predir la variable objectiu, a més, en el pas del *Feature Engineering*, se'n construiran de noves. Per tant, encara queden passos per acabar de definir quines seran les variables predictores del model.

5.3 DATA SCIENCE: A LA RECERCA DEL MILLOR MODEL

5.3.1 Preprocessament

***Arxius necessaris**: 1preproce.ipynb, dates.csv, dataset.csv*

"El propòsit fonamental del preprocessament de dades és manipular i transformar les dades primàries per a poder exposar el contingut de la informació, o fer-lo més accessible" [3].

En aquest apartat s'expliquen els següents passos: es revisa la tipologia de les dades, s'analitza com tractar els valors nuls, els casos duplicats i els outliers, si n'hi ha, així com etiquetar degudament cada una de les variables categòriques en els casos que així es requereixi, amb l'objectiu d'entendre les dades i detectar possibles errors a priori.

Revisió dels tipus de variables

En general, en el món de les dades es consideren dos tipus de variables: les variables numèriques i les variables categòriques, i les categòriques poden ser nominals o ordinals. I en *Python*, normalment les variables categòriques estan definides com a tipus *object* mentre que les numèriques poden ser de tipus *int* si són enteres o *float* si contenen decimals. En la *il·lustració 5* es poden veure els tipus de cada una de les variables de l'estudi.

Tipus	categòrica/object		numèrica	
	nominal	ordinal	int	float
Variables	*idhotel , idcliente , idtipohab , idregimen , idsegmento , id_canal , idpais_cliente*	*fechaocupacion (data)*	*NHab , S1H , S2H , ..., S52H*	*PreuHab , S1 , S2 , ..., S52*

Il·lustració 5: Variables de l'estudi i el seu tipus

La variable *fechaocupacion* és un cas especial, ja que al tractar-se d'una data hauria d'estar en format *datetime*, però es deixa en format *object* pel que s'explica més endavant.

Comprovar duplicats

Un conjunt de dades amb files duplicades pot conduir a fer anàlisis erronis, i amb conseqüència, extreure conclusions falses. En aquest cas es comprova que no hi ha cap duplicat en el conjunt de dades.

Tractament de nulls

Per evitar complicacions posteriors a l'hora de construir els models, cal comprovar si existeixen dades *nulls* i decidir com tractar-les. Es comprova que un 3.1% de les files contenen valors *nulls* i també es veu que tots aquests valors corresponen a cada una de les 53 variables de tipus *float*, les que fan referència al preu. I es veu que són les mateixes en cada una de les files.

Es decideix eliminar-los, ja que se sap que és degut a dades que no tenen informació en la base de dades de producció, ja siguin cancel·lacions o dades que per motius que no se saben, no s'hi ha guardat la informació, i per tant, no proporcionaran informació útil.

S'afegeix la variable dia de l'any

D'acord amb la norma ISO8601[4] el primer dia de l'any cau en dilluns i la primera setmana és la setmana que conté el primer dijous de l'any. Seguint aquesta lògica es crea la variable ***diaAño*** en l'arxiu *dates.csv*, que conté les dates dels anys que es fan servir en l'estudi, i s'afegeix al conjunt de dades. La idea és que aquesta nova variable substitueixi la variable *fechaocupacion* i que ajudi a l'algoritme que fa les prediccions a trobar patrons, per exemple, entre els dilluns, els dimarts, etc, de cada any. A més, a partir d'aquesta variable es creen variables noves en el pas del *feature engineering*.

Tractament d'outliers

Fent un primer cop d'ull als estadístics (*il·lustració 6*) que mostren les dades, es veuen casos estranys en les variables referents al preu de les habitacions.

	PreuHab
count	51043.000000
mean	125.322046
std	74.941721
min	-692.527278
25%	72.510410
50%	108.527272
75%	161.090910
max	1237.398542

Il·lustració 6: estadístics per a PreuHab

Se sap que els preus de les habitacions que ofereix aquesta cadena hotelera oscil·len aproximadament entre els 40 euros/nit per a les habitacions més barates i els 700 euros per a les més cares.

Els mínims d'aquestes variables són negatius, lo qual falseja la informació que es vol per als preus, ja que els hotels no regalen habitacions i menys les compren. En contraposició, també s'hi veuen uns màxims massa elevats, amb preus que no es corresponen a la realitat. Està clar que aquests casos són susceptibles d'analitzar i tractar.

Per coneixement de negoci, s'eliminen totes aquelles files que tenen imports negatius, nuls o 0, ja que falsejarien la informació sobre els preus mig de les habitacions.

Veient els mínims de les variables que informen sobre els imports, i després d'investigar els preus que ofereix la cadena hotelera en qüestió, es decideix agafar només aquelles files que els imports mínims superin els 40€, ja que hi havien preus molt baixos (menys de 10€) que creiem que són descomptes especials i, per tant, no corresponen al preu real i es poden eliminar.

En el cas dels màxims, s'analitza el percentil a mesura que es va augmentant el preu, i es veu que a partir d'uns 300 euros, el percentil 1 és molt petit. En un principi s'havia decidit tallar la mostra a partir del percentil 1. Però fent una mica d'investigació en la web de la mateixa cadena hotelera es veu com poden haver-hi preus de fins a 740 euros, en èpoques de temporada alta i amb les condicions màximes. Per tant, es decideix eliminar totes aquelles files on el *PreuHab* sigui superior a 750 euros.

Reetiquetatge

Abans de començar amb l'anàlisi descriptiu, es fa un reetiquetatge d'algunes variables categòriques per tal de treballar amb dades més entenedores.

- *idhotel*: Es canvia l'identificador per les sigles de l'hotel.
- *idsegmento*: Es canvia el nom del nivell 'ZZZ' per 'DESCONOCIDO'.
- *id_canal*: Ídem que l'anterior.

Finalment, les dimensions del conjunt de dades abans de començar l'anàlisi descriptiu són de 47.630 files i 115 columnes.

5.3.2 Anàlisi descriptiu

***Arxius necessaris**: 2descript.ipynb, dates.csv, dfFinal.csv*

L'anàlisi descriptiu proporciona una base de coneixement que si s'interpreta de forma correcta, les dades poden oferir perspectives útils que puguin servir d'orientació per al desenvolupament del model [5].

En general, els gràfics que s'han utilitzat han estat histogrames, per analitzar la distribució de les dades, i diagrames de caixa, per a detectar possibles outliers o diferents comportaments sobre la variable objectiu segons el nivell. A més, en l'arxiu *2descript.ipynb*, s'hi pot trobar una versió estesa d'aquest anàlisi amb gràfics i comentaris complementaris als que ja es mostra en aquest document. En alguns casos l'anàlisi bivariat es fa amb altres variables a part de la variable resposta.

En l'arxiu *21descriptReduit.ipynb* es fa el mateix anàlisi però restringint el valor de *NHab* per a valors més petits per veure si les dades es comporten d'una altra manera fent aquest canvi. La conclusió que es pot treure és que en general es comporten de la mateixa manera.

Amb el mateix propòsit s'han desenvolupat els arxius *22descrExtremsNhab.ipynb* i *22descrExtremsPreu.ipynb* agafant només les dades inferiors al percentil 5 del valor de *NHab* i el valor de *PreuHab* respectivament.

5.3.2.1 Variable resposta

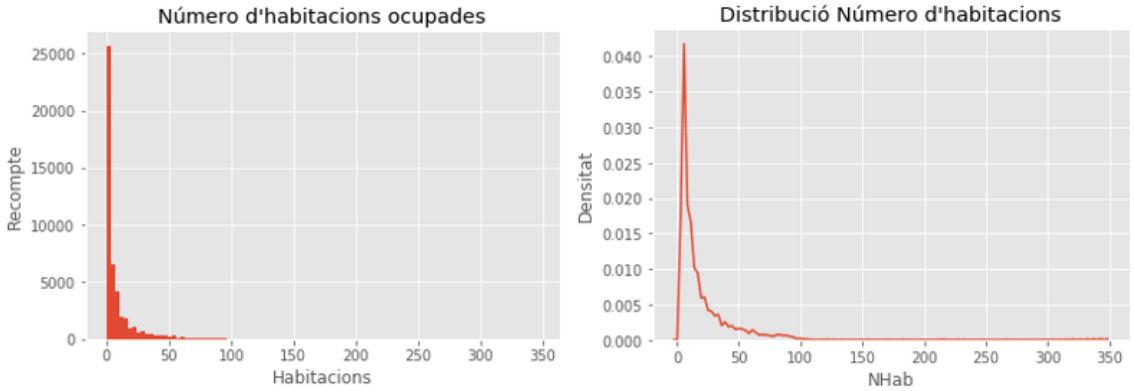
Nombre d'habitacions (NHab)

S'analitza la variable objectiu d'estudi, el nombre d'habitacions. Primer es mostra un resum dels estadístics descriptius més importants i es veu com hi ha una gran diferència entre el valor del percentil 75 i el valor màxim. Han d'haver-hi pocs valors alts per a *NHab*.

	NHab
count	47630.000000
mean	9.775667
std	18.072125
min	0.000000
25%	1.000000
50%	3.000000
75%	10.000000
max	346.000000

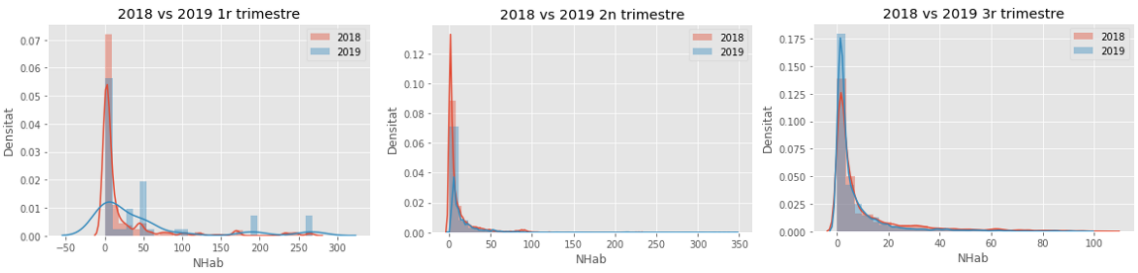
Il·lustració 7: Descriptius de NHab

En la *il·lustració 8* es veu un histograma i la distribució i de la variable i es veu clarament com hi ha molt pocs casos entre 50 i 300, i com n'hi ha molts casos amb valors petits.



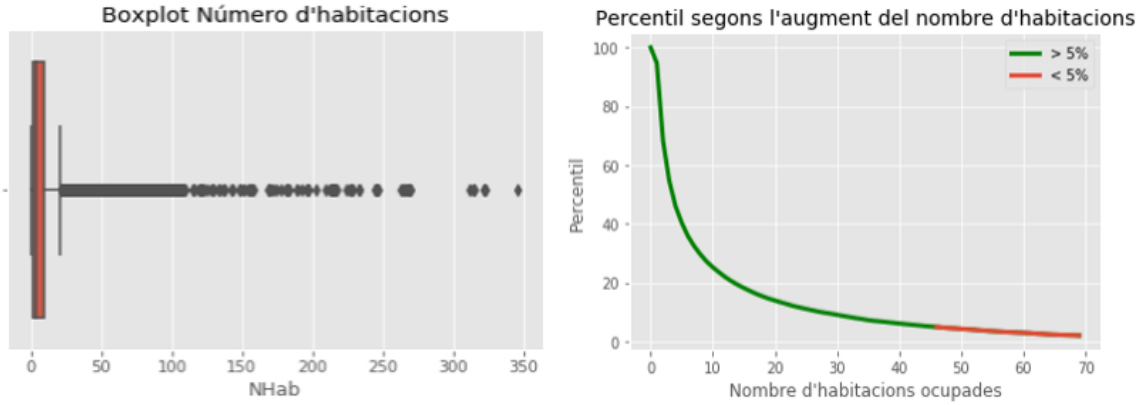
Il·lustració 8: Histograma i diagrama de distribució per a la variable NHab.

En la *il·lustració 9* es compara la distribució de la variable per trimestres, entre 2018 i 2019 (no es té en compte l'últim trimestre ja que no es tenen en compte les dades a partir d'octubre de 2019).



Il·lustració 9: Distribució per trimestres de la variable objectiu

S'observen petites diferencies en els dos primers trimestres, sembla que per a l'any 2018, la variable *NHab* pren valors més petits. Però es comporten de manera similar en tots dos anys, havent-hi majoria de valors petits. En el 3r trimestre, per al 2019 només s'agafen les dades de juliol.

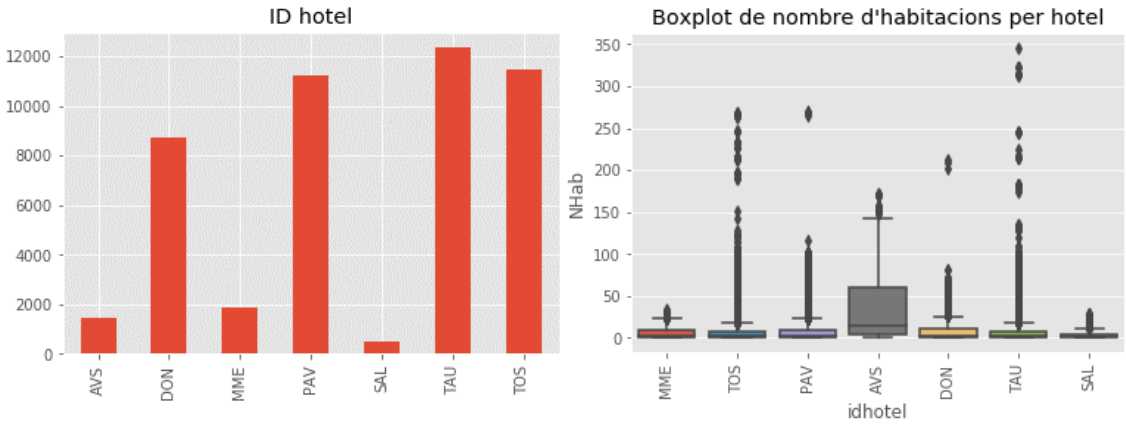


Il·lustració 10: Diagrama de caixa i comportament del percentil de NHab.

En el diagrama de caixa de la *il·lustració 10* es confirma que el gran gruix dels valors de *NHab* prenen un valor més petit que 10, aproximadament. En el gràfic del percentil es mostra el valor de *NHab* a partir del qual el percentil és inferior a 5. Concretament a partir del valor 46.

5.3.2.2 Variables categòriques

Hotel (idhotel)



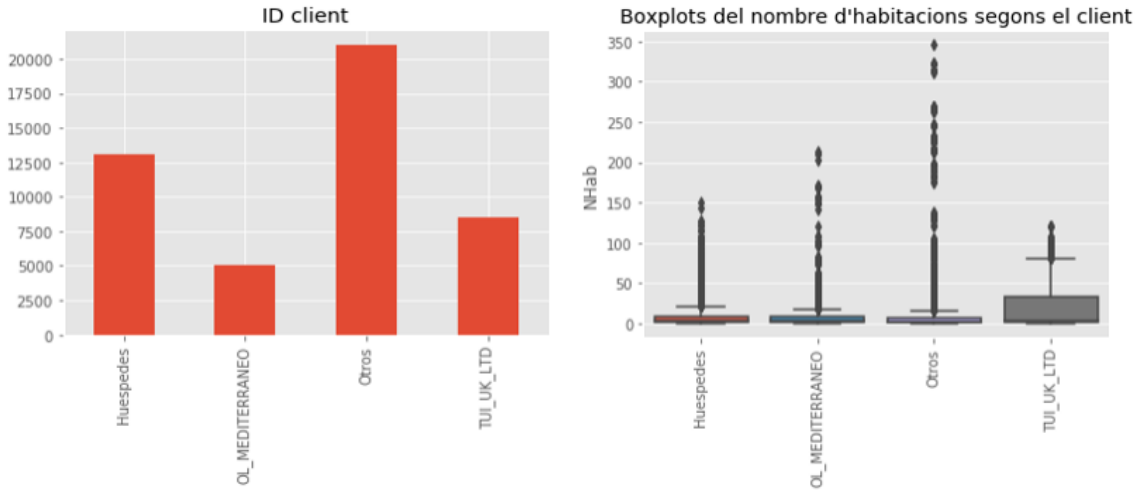
Il·lustració 11: Univariat: Histograma. Bivariat: diagrames de caixa respecte la variable resposta.

L'hotel que apareix en més casos és el *TAU*, hi ha una gran diferencia entre els 3 que menys apareixen i la resta.

L'hotel *AVS* es comporta molt diferent a la resta, mentre que el percentil 75 en la resta d'hotels són aproximadament 10 habitacions, en aquest cas es troba al voltant de 60.

El *MME* i el *SAL* gairebé no tenen casos per sobre de 50. La resta d'hotels es comporten d'una manera similar amb la variable resposta. Tenen molts casos força més grans respecte al percentil 75.

Client (idcliente)

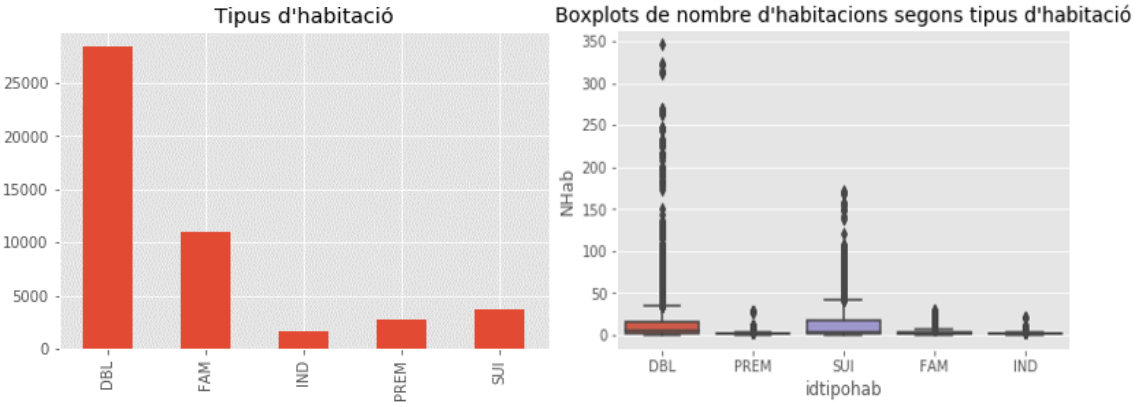


Il·lustració 12: Univariat: Histograma. Bivariat: diagrames de caixa respecte la variable resposta.

Les reserves fetes des de les agències *TUI_UK_LTD* i *OLIMPIA_MEDITERRANEO* suposen aproximadament un 27% de les dades, mentre que les fetes directament pels hostes suposen un altre 27%. La resta de casos són reserves fetes a partir d'altres agències clients.

En els diagrames de caixa de la *il·lustració 12* destaca el comportament per a l'agència *TUI_UK_LTD*, això sembla indicar que el valor de *NHab* tendeix a fer-se més gran quan l'agència client és aquesta empresa, no obstant, el valor màxim de *NHab* respecte a aquesta empresa no arriba a les 150 habitacions, mentre que la resta de nivells el superen.

Habitació (idtipohab)

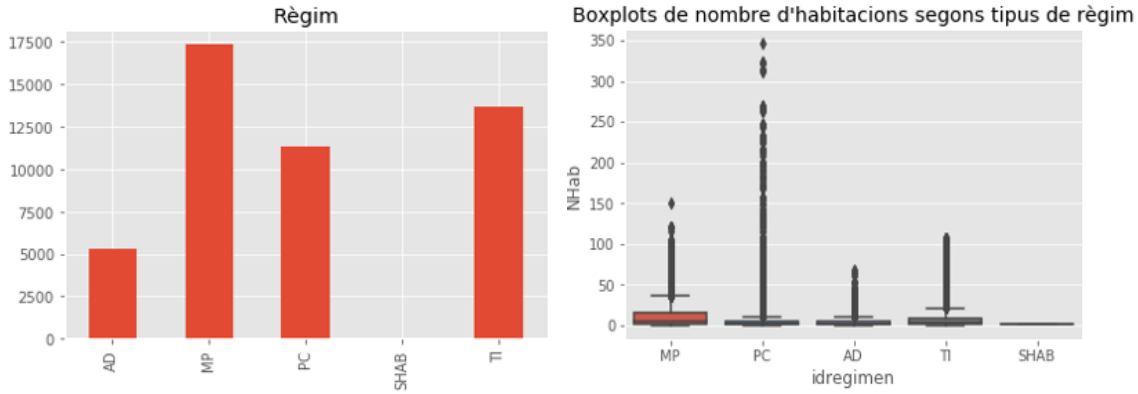


Il·lustració 13: Univariat: Histograma. Bivariat: diagrames de caixa respecte la variable resposta.

Es pot veure també com el tipus d'habitació doble és el més influent amb un gairebé 70% de les observacions.

En el diagrama de caixa bivariat destaca el comportament de les suites, que s'assembla al de l'hotel *AVS* vist en l'anàlisi bivariat de la variable *idhotel*. Aquesta relació té sentit quan un s'adona que en aquest hotel només s'hi venen suites.

Règim (idregimen)

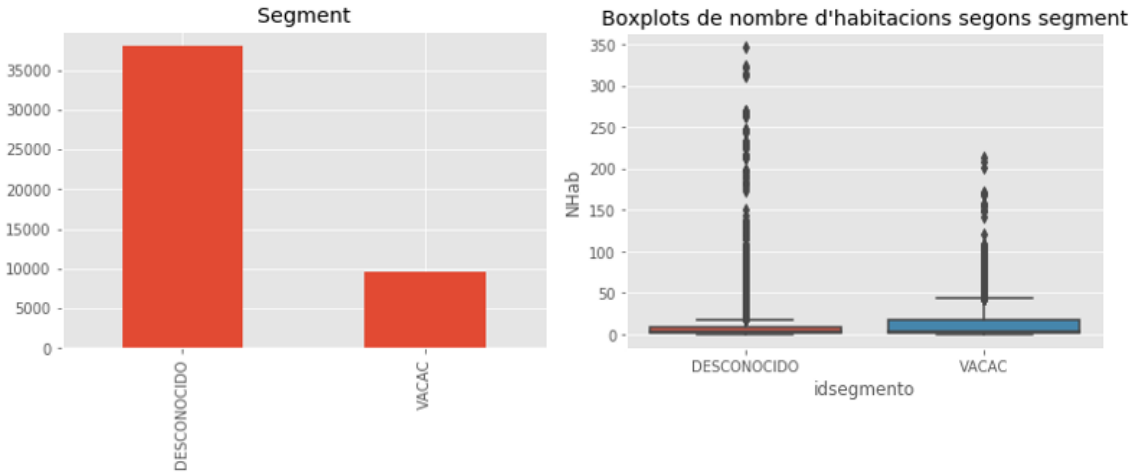


Il·lustració 14: Univariat: Histograma. Bivariat: diagrames de caixa respecte la variable resposta.

El règim més observat és el de mitja pensió, mentre que només s'hi veuen 10 casos on es demani només habitació.

Mentre que els règims de pensió completa són els que mostren un valor de *NHab* més elevat, tal i com es veu en la *il·lustració 14*.

Segment (idsegmento)



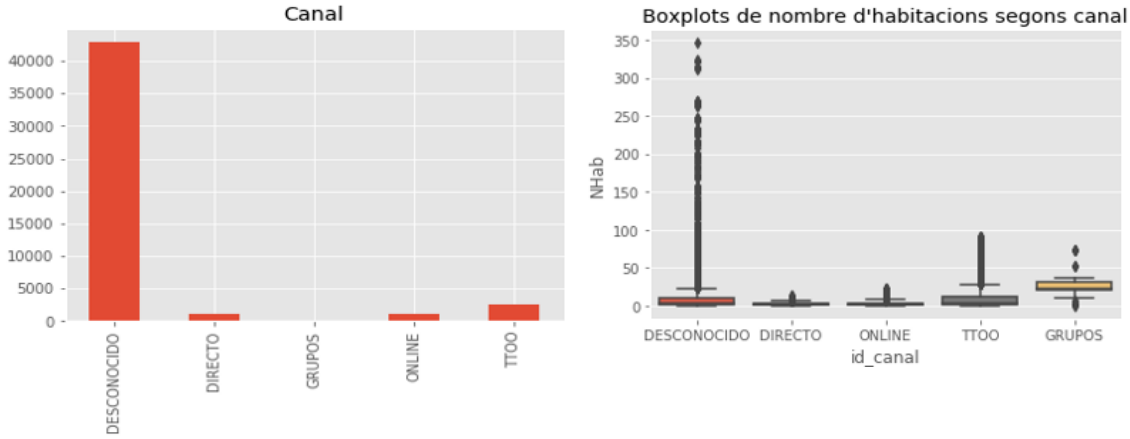
Il·lustració 15: Univariat: Histograma. Bivariat: diagrames de caixa respecte la variable resposta.

La gran majoria d'observacions tenen un segment desconegut, i només un 20% són provinents d'un segment vacacional.

Això segurament és degut a una mala recollida de dades, i pot ser que moltes de les que pertanyen a un segment desconegut siguin també vacacionals, ja que aquests tipus d'hotels estan especialment fets per a estades vacacionals.

És interessant veure en els diagrames de caixa que el valor de *NHab* tendeix a augmentar en les reserves que es vénen per a un segment vacacional.

Canal (id_canal)

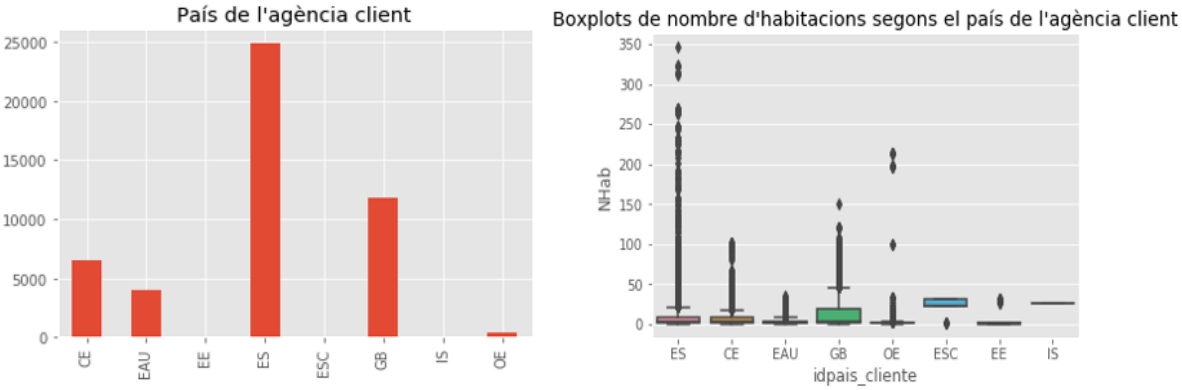


Il·lustració 16: Univariat: Histograma. Bivariat: diagrames de caixa respecte la variable resposta.

El cas d'aquesta variable és semblant al de la variable *idsegmento*, sembla que les dades no s'han recollit correctament, ja que gairebé un 86% de les observacions és de caràcter desconegut.

Però també és interessant veure com el comportament d'aquesta variable respecte a la objectiu és semblant per a tots els nivells excepte per als *GRUPOS*, on més del 75% de les dades d'aquest nivell corresponen a valors força alts de *NHab*.

País origen de l'agència client (idpais_cliente)

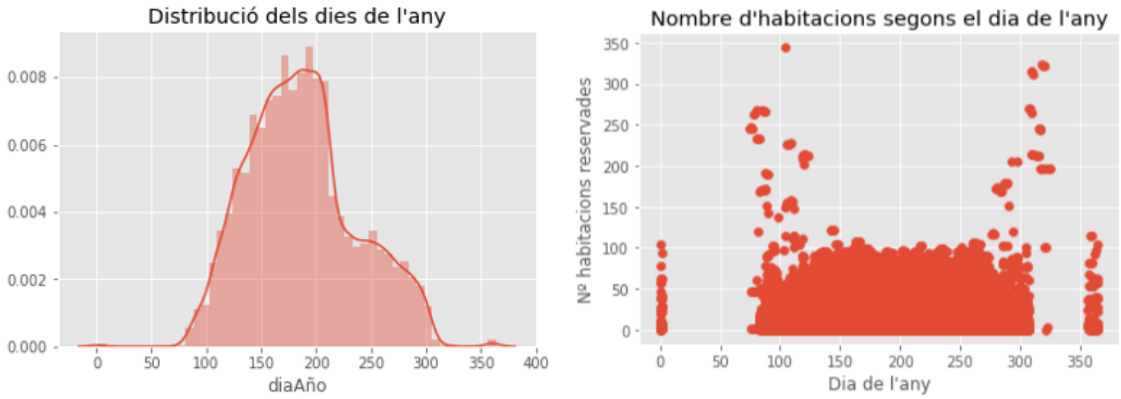


Il·lustració 17: Univariat: Histograma i recompte. Bivariat: diagrames de caixa respecte la variable resposta.

En més de la meitat de les observacions, el país d'origen de l'agència client és Espanya. Té certa lògica ja que els hotels es troben a Catalunya.

Es veuen comportaments diferents segons els països d'origen. Destaquen els casos de Gran Bretanya, Escandinàvia i Islàndia.

Dia de l'any de la data d'ocupació (diaAño)



Il·lustració 18: Univariat: Distribució. Bivariat: diagrama de dispersió respecte la variable resposta.

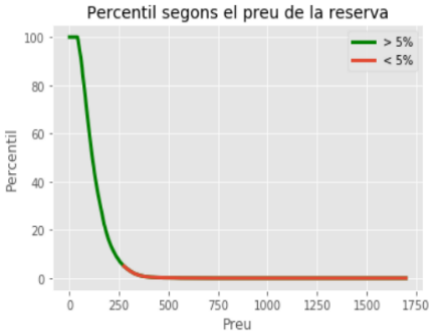
En la *il·lustració 18* es veu que, aproximadament, a partir de mitjans de març fins a principis de novembre, és la zona per on es distribueixen les observacions, amb una gran densitat sobretot per a l'època estiuenca, com es lògic, atenent al tipus d'hotels. Tot i que es nota l'absència del mes d'agost i setembre del 2019.

D'altra banda en el diagrama de dispersió es veu com, a part de lo que ja s'ha vist en la distribució, hi ha valors de *NHab* per a les festes de Nadal. També cal destacar que el valor de *NHab* pren valors alts en èpoques de inici de primavera i per la tardor.

5.3.2.3 Variables numèriques

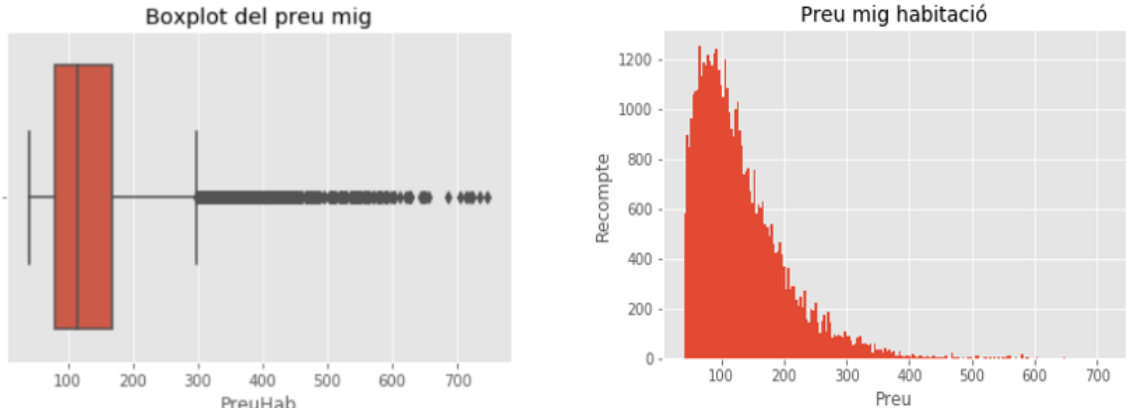
Preu mig de l'habitació (PreuHab)

	PreuHab
count	47630.000000
mean	131.448197
std	72.537042
min	40.000000
25%	79.038319
50%	113.953165
75%	165.453409
max	713.005287



Il·lustració19: Principals descriptius i percentil PreuHab.

El preu mig d'una habitació és de 131 euros, mentre que els mínims i màxims coincideixen amb el punt de tall establert en el preprocessament. Hi ha una gran distància entre el preu del percentil 75 i el màxim. Es veu com a mesura que augmenta el preu disminueixen el nombre d'observacions. Concretament, a partir de 272 euros, les dades restants només representen un 5% del total.



Il·lustració 20: Diagrama de Caixa i histograma i percentil.

Més enllà dels 300 euros són preus poc freqüents com es veu en el diagrama de caixa. La majoria de preus ronden entre els 80 i els 170 euros aproximadament.

Il·lustració 21: Diagrama de dispersió respecte a NHab

En l'anàlisi bivariat respecte la variable objectiu resulta una correlació de -0.148, lo qual indica que, lleugerament, com més alt és el número d'habitacions, més baix és el preu. Té cert sentit, ja que una reserva gran d'habitacions pot comportar descomptes.

5.3.3 Feature engineering

***Arxius necessaris**: 3featEng.ipynb, dfFE.csv, dates.*csv

El *feature engineering* engloba tot allò que fa referència a la manipulació de les variables amb l'objectiu de donar-li al model aquelles variables que li seran d'ajut per a desenvolupar una predicció eficaç. En aquest projecte, el feature engineering està dividit en dos passos: el feature construction i el feature selection.

En el *feature construction* s'hi construiran totes aquelles variables que en un principi s'ha considerat que poden aportar informació útil per a la predicció de la variable objectiu. A més d'un petit anàlisi descriptiu per a cada una d'elles per a comprovar que s'han construït i afegit correctament.

En el *feature selection* s'avaluarà la capacitat explicativa de cada una de les variables predictores respecte la variable objectiu, i en funció d'aquest valor, es seleccionaran unes o altres. Intentant reduir al màxim les dimensions, sense perdre cap variable que pugui aportar un cert valor d'informació útil.

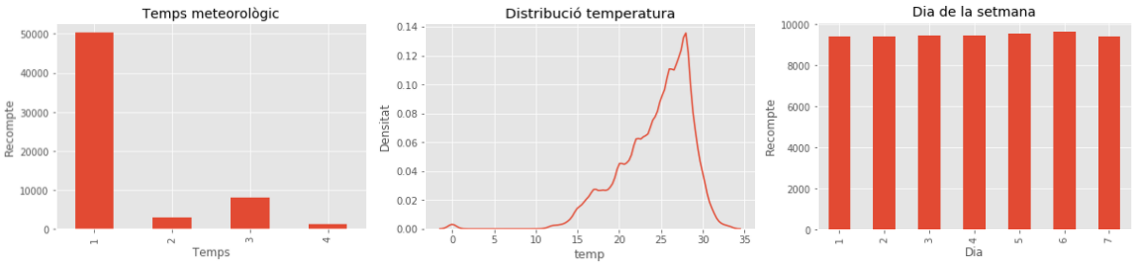
Feature Construction

En funció del que s'ha vist en l'anàlisi descriptiu i el coneixement de negoci que es tingui, es construiran unes variables o unes altres. En aquest projecte, s'ha cregut convenient afegir al conjunt de dades les següents variables:

- *diaSemana*: variable categòrica ordinal. Número corresponent al dia de la setmana. (De l'1 al 7)
- *findes*: variable categòrica ordinal. Pren el valor *'entreSemana'* de dilluns a dijous, *'visperaFinde'* els divendres i *'finde'* els dissabtes i diumenges.
- *temporada*: variable categòrica nominal. Pren el valor *'alta'* o *'baixa'* si el dia corresponent es troba en temporada alta o baixa, respectivament.
- *mes*: variable categòrica ordinal. Número corresponent al mes de l'any. (De l'1 al 12)
- *weather*: variable categòrica ordinal. Pren valors de l'1 al 4, de bon temps a mal temps.
- *temp*: variable numèrica. temperatura en graus centígrads.
- *holiday*: variable categòrica nominal. Pren el valor '*National holiday*' si es festiu i 0 si no ho és en el país d'origen del client.

La idea és que aquestes variables ajudin a l'algoritme a trobar patrons entre diferents espais temporals, com s'ha dit en el pas on s'afegeix la variable *diaAño.*

Per posar un exemple, amb les variables que indiquen els dies de la setmana, el mes i els festius, l'algoritme podrà detectar que per a l'època del març o abril de cada any hi ha uns dies que l'hotel passa d'estar buit a estar ple (Setmana Santa). Com que aquesta festivitat cau en dies diferents cada any, sense aquestes variables l'algoritme no podria detectar aquests tipus de patrons, es per això que és important poder crear variables a partir de la data.



Il·lustració 22: Distribució del temps, temperatura i mesos.

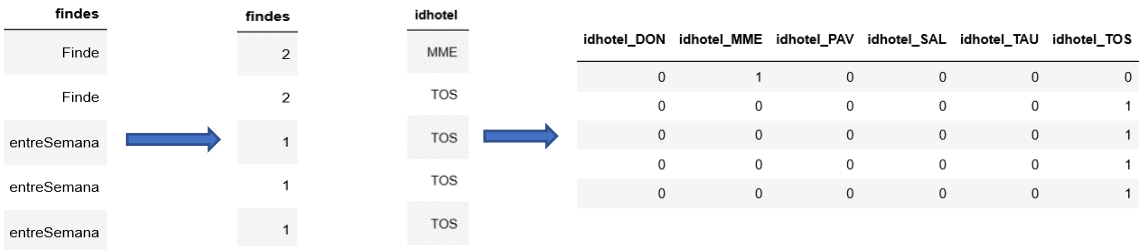
En la *il·lustració 22* es veuen els comportaments d'algunes de les variables creades. Gairebé sempre fa bon temps. Les temperatures solen ser a partir de 15 graus fins als 30 aproximadament. I que cada dia de la setmana aporta, aproximadament, el mateix nombre d'observacions.

A més, en l'arxiu *featEng.ipynb* es poden veure els comportaments per a la resta de variables creades. D'on es pot observar que la distribució segons el dia de la setmana és gairebé idèntic per a tots els dies, que hi ha més observacions en temporada baixa, que la gran majoria d'observacions són de dies no festius i que els mesos que aporten més informació són els mesos d'estiu, juny, juliol, agost i setembre, mentre que no hi ha cap dada per al febrer.

Un cop ja es tenen totes les dades considerades necessàries per a desenvolupar el model predictiu, queda un últim pas per arribar a la selecció de variables.

Totes les variables han d'estar en format numèric per a poder emprar les tècniques de *feature selection*.

Les variables categòriques ordinals es transformen a numèriques amb la funció *OrdinalEncoder* que les transforma amb variables de tipus *int*. Per a la resta de variables categòriques es crearan *dummies*, és a dir, es crearan tantes columnes com nivells tingui la variable, i cada columna contindrà valors 1 i 0 segons si l'observació pertany al nivell en qüestió o no.



Il·lustració 23: Exemple de transformació de variable categòrica ordinal i nominal a numèrica.

Degut a la creació de *dummies* el nombre de variables augmenta fins a 143, incloent la variable objectiu.

Feature Selection

En el camp de l'aprenentatge automàtic, el feature selection[6] o selecció de característiques (selecció de variables), és el procés de selecció d'un subconjunt rellevant de característiques(variables) que s'utilitzaran en la construcció del model.

Les tècniques per a arribar a fer la selecció de variables s'usen per diverses raons:

- ✓ Simplificar models per fer-los més fàcils d'interpretar
- ✓ Escurçar la durada del temps d'entrenament del model
- ✓ Reduir dimensions del conjunt de dades
- ✓ Reduir l'overfitting(sobre ajust), o formalment, reducció de la variància

En aquest projecte s'han desenvolupat dues tècniques basades en estadística (multicol·linealitat i selecció de les *k* millors) i quatre basades en models (regressió lineal, arbre de decisió, *random forest* i *xgboost*) per a determinar la selecció de variables. L'objectiu de desenvolupar aquestes tècniques és el de veure quina utilitat tenen les variables predictores per a l'algoritme a l'hora d'explicar la variable resposta, i ser capaços de seleccionar només aquelles variables que expliquin una part subjectivament significativa de la mateixa. Cada tècnica avalua les variables tenint en compte diferents aspectes, però la idea és trobar coincidències en totes les tècniques. Per exemple, una variable que es consideri poc important amb una tècnica, però amb una altra sigui important, no s'eliminarà. Al final de la secció s'exposen les variables que s'han decidit eliminar i perquè.

❖ Multicol·linealitat

Per començar s'analitza la presència de multicol·linealitat entre variables. Com és lògic, degut a com està feta la base de dades, ja s'intueix que moltes variables referents a les setmanes anteriors, tant en preu com en número d'habitacions, estaran fortament correlacionades, ja que en molts casos estan fetes a partir de les setmanes anteriors o posteriors a la setmana en qüestió. Per tenir una idea visual és representa una matriu de correlacions, en la *il·lustració 24*

es mostra una part d'aquesta matriu, per a les variables referents al nombre d'habitacions ocupades i als preus durant les setmanes anteriors.

	NHab	S1H	S2H	S3H	S4H	S5H	S6H	S7H	S8H	S9H
NHab	1	0.994	0.94	0.889	0.862	0.842	0.829	0.816	0.805	0.795
S1H	0.994	1	0.946	0.897	0.87	0.851	0.839	0.827	0.816	0.805
S2H	0.94	0.946	1	0.959	0.936	0.923	0.911	0.9	0.89	0.88
S3H	0.889	0.897	0.959	1	0.987	0.977	0.967	0.956	0.947	0.938
S4H	0.862	0.87	0.936	0.987	1	0.993	0.985	0.975	0.967	0.958
S5H	0.842	0.851	0.923	0.977	0.993	1	0.994	0.987	0.98	0.972
S6H	0.829	0.839	0.911	0.967	0.985	0.994	1	0.994	0.989	0.983
S7H	0.816	0.827	0.9	0.956	0.975	0.987	0.994	1	0.996	0.99
S8H	0.805	0.816	0.89	0.947	0.967	0.98	0.989	0.996	1	0.996
S9H	0.795	0.805	0.88	0.938	0.958	0.972	0.983	0.99	0.996	1

	PreuHab	S1	S2	S3	S4	S5	S6	S7	S8	S9
PreuHab	1	0.998	0.992	0.986	0.981	0.975	0.97	0.965	0.959	0.955
S1	0.998	1	0.995	0.989	0.983	0.977	0.972	0.967	0.961	0.957
S2	0.992	0.995	1	0.994	0.989	0.983	0.978	0.972	0.966	0.963
S3	0.986	0.989	0.994	1	0.994	0.988	0.983	0.978	0.972	0.968
S4	0.981	0.983	0.989	0.994	1	0.994	0.989	0.984	0.978	0.974
S5	0.975	0.977	0.983	0.988	0.994	1	0.994	0.989	0.984	0.98
S6	0.97	0.972	0.978	0.983	0.989	0.994	1	0.994	0.988	0.984
S7	0.965	0.967	0.972	0.978	0.984	0.989	0.994	1	0.994	0.99
S8	0.959	0.961	0.966	0.972	0.978	0.984	0.988	0.994	1	0.996
S9	0.955	0.957	0.963	0.968	0.974	0.98	0.984	0.99	0.996	1

Il·lustració 24: Mostra de correlacions entre els nombres d'habitacions i els preus de les setmanes anteriors

Tot i que en condicions normals, una correlació entre variables superior a 0.8 ja seria motiu per eliminar una d'elles, s'ha considerat aquest cas de les setmanes anteriors com un cas especial, ja que la pròpia cadena hotelera fa ús d'aquestes dades i es busca fer un model precís però també que reflecteixi bé el problema que s'intenta resoldre. Per això s'ha establert, en aquests casos, que s'eliminaran totes aquelles variables corresponents a la setmana més antiga que superin un 98% de correlació amb altres variables.

Aplicant aquest criteri, s'eliminen totes les variables referents a setmanes anteriors excepte les ***S2H***, ***S3H*** i ***S52H***. **Tots els preus excepte el preu de la setmana actual queden fora del model**. Es redueix a 41 el nombre de variables.

	NHab	S2H	S3H	S52H	PreuHab	diaAño	weather	temp	mes	diaSemana	findes	idhotel_DON	idhotel_MME
NHab	1	0.94	0.889	0.523	-0.15	0.0377	-0.0327	-0.0298	0.0384	0.00248	0.00473	-0.0194	-0.0389
S2H	0.94	1	0.959	0.584	-0.153	0.0383	-0.0308	-0.0177	0.0392	-0.00153	-3.31e-05	-0.0233	-0.0464
S3H	0.889	0.959	1	0.632	-0.154	0.0306	-0.0292	-0.0105	0.0315	-0.00356	-0.00261	-0.0239	-0.0506
S52H	0.523	0.584	0.632	1	-0.0628	0.0473	-0.0254	0.0661	0.046	-0.00457	-0.00643	-0.00061	-0.0386
PreuHab	-0.15	-0.153	-0.154	-0.0628	1	0.176	-0.0346	0.452	0.168	0.0021	-0.0028	-0.0342	0.134
diaAño	0.0377	0.0383	0.0306	0.0473	0.176	1	-0.151	0.471	0.979	-0.00412	-6.36e-05	0.0628	-0.0119
weather	-0.0327	-0.0308	-0.0292	-0.0254	-0.0346	-0.151	1	-0.218	-0.157	0.0299	0.0328	-0.0764	-0.00235
temp	-0.0298	-0.0177	-0.0105	0.0661	0.452	0.471	-0.218	1	0.46	-0.00709	0.00263	0.124	-0.00119
mes	0.0384	0.0392	0.0315	0.046	0.168	0.979	-0.157	0.46	1	-0.00954	-0.00489	0.0634	-0.0116
diaSemana	0.00248	-0.00153	-0.00356	-0.00457	0.0021	-0.00412	0.0299	-0.00709	-0.00954	1	0.689	0.00237	0.00192
findes	0.00473	-3.31e-05	-0.00261	-0.00643	-0.0028	-6.36e-05	0.0328	0.00263	-0.00489	0.689	1	0.00496	0.00212
idhotel_DON	-0.0194	-0.0233	-0.0239	-0.00061	-0.0342	0.0628	-0.0764	0.124	0.0634	0.00237	0.00496	1	-0.0965
idhotel_MME	-0.0389	-0.0464	-0.0506	-0.0386	0.134	-0.0119	-0.00235	-0.00119	-0.0116	0.00192	0.00212	-0.0965	1

Il·lustració 25: Mostra de correlacions entre les variables restants.

En la *il·lustració 25* es mostren les correlacions entre 13 variables. Com es veu, hi ha correlacions fortes entre les tres primeres variables, però es decideix mantenir-les pel que s'acaba d'explicar. Pel que fa a la resta de variables, només s'elimina la variable *diaAño*, ja que es correlaciona 0.98 punts amb la variable *mes*. I es creu que la variable *mes* pot oferir patrons més interessants a l'algoritme.

- ❖ Select K Bests

Aquest mètode retorna les *K* variables que més variància expliquen sobre la variable objectiu. La funció emprada en *Python* té una opció per retornar els p-valors de cada variable respecte la interacció lineal amb la variable objectiu.

La hipòtesi nul·la, en aquest cas, diu que no hi ha presència d'interaccions lineals entre la variable objectiu i l'explicativa, així que les variables interessants són aquelles les quals el seu p-valor sigui inferior a 0.05 i es pugui dir que hi ha prou evidències estadístiques com per refusar la hipòtesi nul·la i creure que existeix una interacció lineal amb la variable *NHab*. Però això no vol dir que no puguin tindre cap altre tipus de relació amb la variable objectiu, per a això es mira la importància de les variables en la secció de la *Feature Importance*.

Feature	pvalue
diaSemana	0.588209
findes	0.301843
idhotel_PAV	0.117128
idhotel_TOS	0.370770
idregimen_SHAB	0.124604
idpais_cliente_OE	0.508091

Il·lustració 26: p-valors superiors a 0.05

En la *il·lustració 26* s'observen totes aquelles variables predictores amb un p-valor superior a 0.05.

- Regressió lineal

L'objectiu de desenvolupar una regressió lineal és analitzar els coeficients que tindrà cada una de les variables. És possible que hi hagi variables que en el pas anterior mostrin un p-valor significatiu, però els seus coeficients siguin tant baixos que també es consideri l'opció d'eliminar-los degut a la poca influència que tenen. Amb influència es fa referència al fet que el coeficient en qüestió estigui llunyà a 0 (influent) o proper a 0 (poc influent) en valor absolut.

Features	Coefficients
diaAño	0.008465
diaSemana	-0.000123
PreuHab	-0.001139
idpais_cliente_IS	-0.001219

Il·lustració 27: coeficients més petits que 0.01 en valor absolut

Les variables de la *il·lustració 27* s'estudien en la *Feature Importance* per veure si s'eliminen del model.

- Feature importance

Finalment, és desenvolupa el mateix procediment per a tres algoritmes diferents: *Decision Tree*, *Random Forest* i *XGBoost*. Per a cada un dels algoritmes s'avalua la importància de les variables i es busquen coincidències per tal d'eliminar-ne.

En l'apartat corresponent de l'arxiu *3featEng.ipynb* es poden veure quines són les variables menys influents i, per tant, candidates a estudiar la seva eliminació del model.

Conclusions de la selecció de variables

La *il·lustració 28* mostra les variables que es podrien eliminar i els motius, que són els següents:

1. Correlació excessiva entre variables
2. p-valor > 0.05
3. Coeficient massa baix
4. Molt poca importància en la feature importance
5. Poca importància en la feature importance (com a complement a un dels 3 primers)

	Motius				
Variables	1	2	3	4	5
diaSemana		X	X		X
findes		X		X	
idhotel_DON				X	
idhotel_MME				X	
idhotel_SAL				X	
idhotel_TOS		X			X
idtipohab_IND				X	
idregimen_SHAB		X		X	
id_canal_DIRECTO				X	
id_canal_ONLINE				X	
idpais_cliente_EAU				X	
idpais_cliente_EE				X	
idpais_cliente_GB	X				
idpais_cliente_IS			X	X	
idpais_cliente_OE		X			X
holiday_No festiu				X	

Il·lustració 28: Variables candidates a ser eliminades

5.3.4 Construcció del model

***Arxius necessaris**: dfModel.csv, dates.csv, 4model.ipynb, 4modelV2.ipynb, 41VisualParam.ipynb, 41VisualParamV2.ipynb*

En aquest projecte s'ha decidit dissenyar dues estratègies per a construir el model. En la primera estratègia es construeix un model amb totes les variables excepte les que s'eliminen per multicol·linealitat, i en la segona estratègia es construeix un model eliminant les variables descrites en la *il·lustració 28*, a part de les variables eliminades per multicol·linealitat.

Per tal de trobar el millor model, per a totes dues estratègies s'ajusten 5 algoritmes que es mostren en la *il·lustració 29*: [7] [8] [9] [10] [11]

Models	Algoritme implementat a *python*
Regressió lineal múltiple	LinearRegression()
K-Nearest Neighbors (KNN)	KNeighborsRegressor()
Decision Tree	DecisionTreeRegressor()
Random Forest	RandomForestRegressor()
Gradient Boosting	XGBRegressor()

Il·lustració 29: Models i els corresponents algoritmes

S'han seleccionat aquests models perquè són alguns dels algoritmes típics de regressió aplicables a *Machine Learning*.

Com és un model d'aprenentatge supervisat, per a l'entrenament del model es divideix la mostra en dues parts. D'una banda, la mostra d'entrenament amb totes les variables (75% de les dades), que serveix per a que l'algoritme vagi reconeixent patrons entre les variables predictores i la variable objectiu. Per altra banda, la mostra test amb totes les variables però "amagant" la variable objectiu (25% de les dades), que és la que haurà d'endevinar l'algoritme en funció d'allò que ha après amb les dades d'entrenament. Un cop s'ha entrenat el model s'avalua l'encert de l'algoritme comparant les prediccions obtingudes amb els resultats reals.

Criteris d'avaluació

Per a avaluar la precisió dels models, tant per a la mostra d'entrenament com per a la mostra tet, s'utilitzen 4 estadístics diferents:

- ✓ R^2: representa la proporció de la variància que explica el model sobre la variància real. Pren valors entre 0 i 1 i com més elevat més variància explica el model sobre la real i, per tant, és més precís.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- ✓ MSE: *Mean Square Error* o error quadràtic mitjà. Mesura la mitjana dels errors al quadrat.

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- ✓ RMSE: *Root Mean Square Error* o arrel de l'error quadràtic mitjà. És l'arrel quadrada del MSE.

$$RMSE = \sqrt{MSE}$$

- ✓ MAE: *Mean Absolute Error* o error absolut mitjà. Representa, en mitjana, la diferencia absoluta entre les dades reals i les prediccions.

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Un aspecte a tindre en compte a l'hora d'avaluar un model és valorar la presència d'*overfitting* o sobre ajust. Tot i que no és un criteri que mesuri la precisió dels models, és una bona pràctica tenir-lo en compte a l'hora d'escollir un model o un altre en cas que els resultats obtinguts en els estadístics esmentats estiguin ajustats. Es diu que hi ha sobre ajust quan l'algoritme només s'ajusta a aprendre dels casos particulars que se li ensenyen, amb la qual cosa, serà incapaç de reconèixer noves dades d'entrada que es surtin una mica dels rangs establerts i amb tota probabilitat s'equivocarà a l'hora de fer la predicció [12]. Es podrà parlar de que un model està sobre ajustat si els resultats de la mostra d'entrenament són força millors que les de la mostra test.

Cross-validation

Per comprovar si la manera de dividir les dades en una mostra d'entrenament i una altra de test afecta en els resultats del model s'utilitza la tècnica del *cross-validation* o validació creuada.

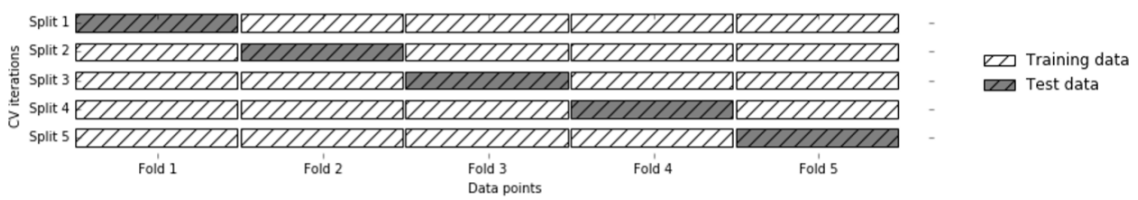
La validació creuada és un mètode estadístic per avaluar el rendiment generalitzat, és més estable i exhaustiu que utilitzar una sola divisió en un conjunt d'entrenament i un de test.

En *cross-validation*, les dades es divideixen repetidament i es formen diversos models. La versió més utilitzada de validació creuada és *k-fold cross-validation*, on k és un número especificat per l'usuari, normalment 5 o 10. En aquest cas, *k = 5*.

Quan es realitza la validació creuada de cinc vegades, les dades es divideixen en cinc parts de (aproximadament) la mateixa mida, anomenats *folds* o plecs. A continuació, s'entrena una seqüència de models.

El primer model s'entrena utilitzant el primer plec com a conjunt de prova, i els plecs restants (2-5) s'utilitzen com a conjunt d'entrenament. El model es construeix utilitzant les dades dels plecs 2-5, i després s'avalua la precisió en el plec 1. A continuació, es crea un altre model, aquesta vegada utilitzant el plec 2 com a conjunt de prova i les dades als plecs 1, 3, 4 i 5 com a conjunt d'entrenament.

Aquest procés es repeteix utilitzant els plecs 3, 4 i 5 com a conjunts de prova. Per a cadascun d'aquests cinc fraccionaments de les dades en conjunts d'entrenament i test es calcula la precisió. En la *il·lustració 30* s'il·lustra el procés [13].



Il·lustració 30: Cross-validation de 5 parts

Hyperparameter tuning

Els algoritmes de la *il·lustració 29* disposen d'una sèrie de paràmetres per defecte, anomenats híper paràmetres, que l'investigador pot modificar. El procés d'optimitzar aquests híper paràmetres per tal d'obtenir millors resultats predictius és el que es coneix com a *hyperparameter tuning*. Un exemple d'híper paràmetre seria el nombre d'arbres a construir en un random forest. En aquest projecte s'optimitzen els paràmetres utilitzant la tècnica del *Cross Validation* amb una funció *que s'anomena GridSearchCV*[14] i que s'explica a continuació.

❖ Aspectes a tindre en compte en la optimització d'híper paràmetres

Com s'ha dit, es fa servir la funció *GridSearchCV* per al procés de l'*Hyperparameter tuning*. Aquesta funció rep una sèrie de paràmetres per a un model determinat, i fa una combinació entre tots els paràmetres possibles. Per a cada combinació de paràmetres entrena 5 models diferents usant la tècnica del *Cross Validation*. I retorna els paràmetres d'aquell model amb una precisió més elevada de la mostra *train*, en funció del criteri que se li passi (en aquest estudi, el que ve per defecte, que és el MSE). Un cop es tenen els valors òptims s'entrena el model i s'avalua la seva precisió amb els estadístics comentats anteriorment.

Es pot deduir que el principal aspecte a tindre en compte a l'hora d'optimitzar els paràmetres amb *GridSearchCV* és el temps de computació. Si, per exemple, es volen optimitzar 9 paràmetres de l'algoritme *XGBRegressor()*, i per a cada un dels paràmetres se li passen a l'algoritme un rang de tant sols 5 valors, fer totes les combinacions possibles comporta entrenar 15.120 models, a més, com es fa servir la tècnica de *Cross Validation* el nombre de models augmentaria a 75.600, lo qual és totalment inviable ja que l'ordinador podria estar dies o setmanes compilant. Per tant, s'ha de buscar una alternativa amb la intenció de reduir el temps computacional. En aquest projecte s'ha decidit anar optimitzant d'un en un, o de dos en dos com a màxim els paràmetres. Els gràfics representats als arxius *41VisualParam.ipynb* i el *41VisualParamV2.ipynb*, corresponents a la primera i segona estratègia respectivament, són d'utilitat per saber quin rang de valors se li passen al *GridSearchCV* des de l'inici, ja que es mostra per a cada híper paràmetre, un gràfic on es pot veure per a quins valors la diferència entre l'R^2 de *train* i de *test* es fa més petita, per tal de controlar l'*overfitting*.

En els arxius *4model.ipynb* i *4modelV2.ipynb*, corresponents a la primera i segona estratègia respectivament, es mostra com s'han optimitzat els híper paràmetres i quins són els híper paràmetres definitius per a cada un dels models.

5.3.5 Avaluació del model

***Arxius necessaris**: dfModel.csv, dates.csv, 4model.ipynb, 4modelV2.ipynb*

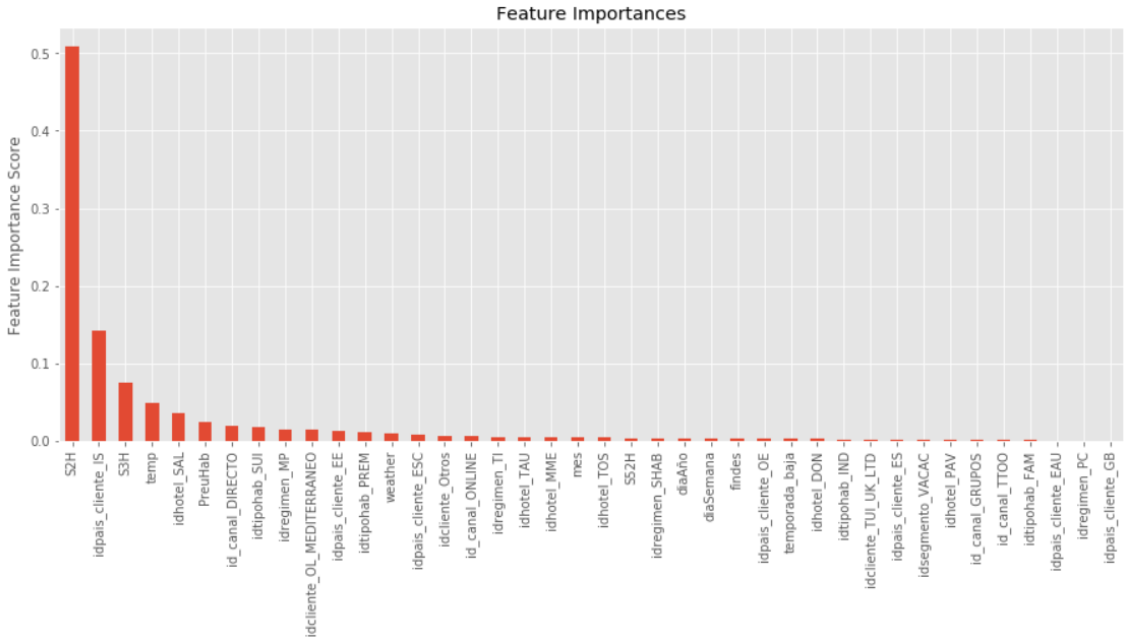
Amb els híper paràmetres definits s'avaluen els models, a continuació es mostren els resultats obtinguts en les dues estratègies i les conclusions que s'extreuen.

Estratègia 1, conjunt de 40 variables predictores

Estadístic:	R^2		MSE		RMSE		MAE	
Mostra:	Train	Test	Train	Test	Train	Test	Train	Test
Regressió Lineal	0,8842	0,8989	37,4227	34,3760	6,1174	5,8631	1,5308	1,5747
K-Nearest Neighbors	0,9639	0,9477	11,7058	17,7865	3,4214	4,2174	1,0849	1,4133
Decision Tree	0,9729	0,9439	8,7759	19,1096	2,9624	4,3715	1,1784	1,3071
Random Forest	0,9864	0,9617	4,4499	13,0338	2,1095	3,6102	0,7834	1,0224
XGBoost	0,9949	0,9723	1,6880	9,4449	1,2992	3,0733	0,6166	0,9540

Il·lustració 31: Resultats obtinguts per als models de l'estratègia 1

Mirant únicament els resultats per a la mostra test, hi ha un clar guanyador, el model amb *XGBoost*, ja que millora en tots els resultats respecte a la resta de models. Tenint en compte l'*overfitting*, el model que menys diferències mostra entre els resultats obtinguts per a les dades d'entrenament i test és el *KNN*. Com que les diferències en aquest sentit no són gaire elevades, i tant l'R^2 com els errors de la mostra test de l'*XGBoost* són força millors que els del *KNN*, es decideix escollir l'*XGBoost* com a millor model entre els presentats.



Il·lustració 32: Importància de cada una de les variables en el model XGBoost

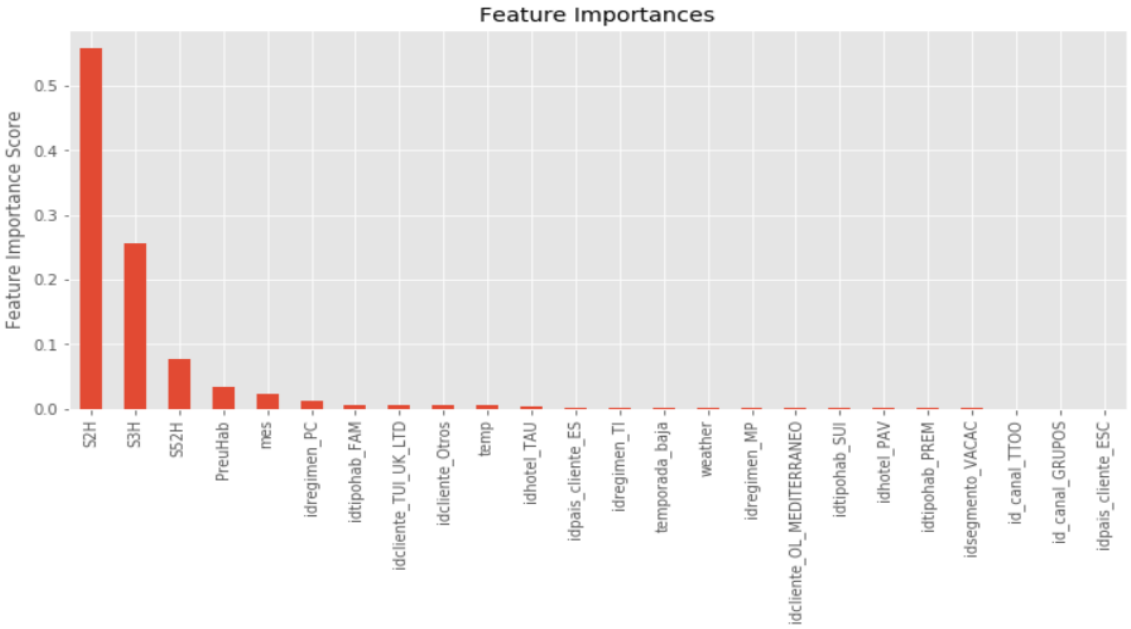
En la *il·lustració 32* es mostra la importància de cada una de les variables per al model guanyador. La variable que més importància té és *S2H* lo qual té sentit ja que s'ha de recordar que es correlaciona en 0.93 punts amb la variable a predir. També pren molta importància el fet que el país de l'agència client sigui d'Islàndia, veient el descriptiu es dedueix que l'algoritme ha detectat que quan això passa, el valor de NHab se situa al voltant de 30.

Estratègia 2, conjunt de 24 variables predictores

Estadístic:	R^2		MSE		RMSE		MAE	
Mostra:	Train	Test	Train	Test	Train	Test	Train	Test
Regressió Lineal	0,8838	0,8989	37,5551	34,3750	6,1282	5,8630	1,5095	1,5551
K-Nearest Neighbors	0,9859	0,9645	4,6008	12,1317	2,1449	3,4831	0,6932	1,1481
Decision Tree	0,9745	0,9583	8,2827	14,1701	2,8780	3,7643	1,1916	1,3194
Random Forest	0,9791	0,9609	6,7696	13,3386	2,6018	3,6522	0,9644	1,1179
XGBoost	0,9939	0,9702	2,0250	10,1622	1,4230	3,1878	0,6646	0,9939

Il·lustració 33: Resultats obtinguts per als models de l'estratègia 2

Els resultats per a les dades *test* dels models *KNN* i *Decision Tree*, en general, milloren respecte l'anterior estratègia. Mentre que el *Random Forest* i l'*XGBoost* empitjoren suaument. Tenint en compte l'*overfitting* milloren els resultats del *Decision Tree* i del *Random Forest* i empitjoren els del *KNN* i l'*XGBoost*. Com els resultats test són força semblants (excepte per a la regressió lineal que és força pitjor que la resta), s'ha decidit el millor model entre aquells que la diferència dels resultats *train* i *test* per a l'R^2 és inferior a 0.02 i aquest model és el *Random Forest*.



Il·lustració 34: Importància de cada una de les variables en el model Random Forest

La variable *S2H* segueix sent la que pren una importància més elevada amb gran diferència. A diferència del model de la primera estratègia, en aquest les variables més importants són les que presenten una correlació més elevada amb la variable objectiu (*S2H*, *S3H*, *S52H*, *PreuHab*).

5.3.6 Posada en escena: *Random Forest*

***Arxius necessaris**: dates.csv, df2020.csv, dfModel.csv, dfReals.csv, 5DadesReals.ipynb, 6ModelReal.ipynb*

Per acabar, es posa en marxa un dels models guanyadors. Per decidir quin d'ells s'agafa, una opció seria parlar amb els responsables de la cadena hotelera i ensenyar-los els models que s'han entrenat, amb les variables que tenen en compte, la importància que li donen a les variables i els resultats obtinguts dels estadístics, i deixar que ells siguin els que prenguin la decisió.

Una altra opció és decidir en funció del propi criteri de l'investigador, que és la que s'ha dut a terme, ja que aquest projecte és una prova de concepte i no era viable la primera opció.

Per a decidir amb quin model es fa la posada en escena amb dades reals, com els resultats són força semblants, s'ha optat per seguir un criteri conservador, és a dir, la prioritat ha estat evitar l'*overfitting*, així que el model escollit és el *Random Forest* de la segona estratègia.

Per a que l'algoritme funcioni correctament, les dades noves han de seguir el mateix procés pel que han passat les dades amb les que s'ha treballat, això es pot veure detalladament en l'arxiu *5DadesReals.ipynb*. Finalment, en l'arxiu *6ModelReal.ipynb* es mostra com s'han fet les prediccions. Un dels aspectes a mencionar és que s'han corregit forçadament les prediccions per tal de que com a màxim, el valor de la predicció sigui igual a la capacitat màxima de l'hotel en qüestió.

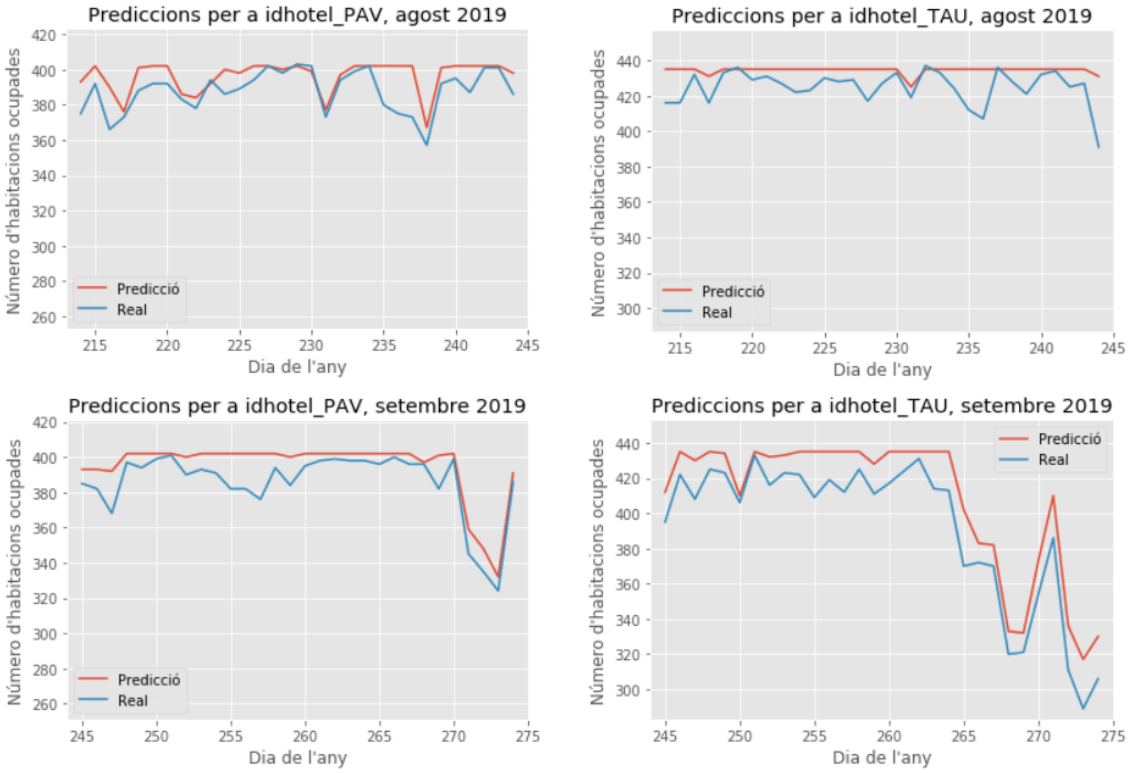
A continuació es mostren els resultats obtinguts per a les dades reals d'agost i setembre de l'any 2019, fent les prediccions amb aquest model.

Estratègia 2: Random Forest				
Estadístic:	R^2	MSE	RMSE	MAE
Agost	0,9674	5,4364	2,3316	1,0101
Setembre	0,9732	4,6733	2,1618	0,912

Il·lustració 35: Resultats Random Forest en dades reals

Es milloren els resultats aconseguits amb les dades test. Segurament, el fet d'haver corregit les prediccions com s'ha comentat, ha fet augmentar la precisió i reduir els errors.

En general, es pot parlar de que s'han obtingut uns bons resultats. Per veure-ho un a mica més en detall, en la *il·lustració 37* es representen les prediccions i el que realment va passar per a dos hotels (*PAV* i *TAU*) i per a tots dos mesos.



Il·lustració 37: Representació gràfica de les prediccions i la realitat

Una conclusió que es podria extreure veient els gràfics és que la predicció sempre està per sobre del que passa realment. A conseqüència d'això, es podria dir que sempre que en un dia determinat la predicció doni un valor inferior al màxim de capacitat d'un hotel es complirà. Així que quan això passa podria ser una bona oportunitat per a la cadena hotelera per baixar els preus per intentar arribar al màxim. Però tot això seria més consistent amb més dades i més representacions.

6 CONCLUSIONS

En aquest apartat mostro les conclusions i reflexions que extrec del projecte un cop acabat, tant del compliment dels objectius proposats i possibles aspectes a millorar com de l'experiència que m'ha aportat.

Objectius i aspectes a millorar

Crec que he pogut assolir l'objectiu principal que em proposava a l'inici del treball, penso que he construït un model adient per a predir la variable d'interès i que aquest document podrà servir com una guia orientativa dels passos a seguir per a la construcció d'un model de *Machine Learning*.

Per què considero que és un bon model? Doncs bé, la precisió R^2 del model escollit per a les dades test és de 0.9609 i per a les dades reals de 0.97 aproximadament. Però focalitzar-nos únicament en aquests valors per veure si és un bon model seria un error, encara que siguin valors tan alts. Cal recordar que aquest és un problema especial, on hem fet servir variables explicatives molt correlacionades entre elles, i una d'elles es correlaciona 0.939 punts amb la variable objectiu. Per tant, cal que el model construït tingui una precisió superior a aquest 0.939 per ser considerat un model adient.

Ara bé, el fet d'haver aconseguit construir un model adient, no vol dir que els resultats obtinguts i els passos que s'han seguit siguin els ideals. Una de les coses que he aprés és que en moltes situacions no hi ha una estratègia definida. Per exemple, en la selecció de variables el fet de quedar-nos amb una variable o una altre no és cap ciència certa. Un altre exemple el trobem en l'elecció del model que es fa servir per predir les dades, seria igual de correcte haver escollit l'*XGBoost* ja que els resultats són molt semblants. El mateix passa en passos com la *feature construction*, l'*hyperparameter tuning*, etc.

Per altra banda, alguns dels aspectes a millorar de cara al futur a l'hora de realitzar un projecte semblant serien els següents:

- Establir un màxim per al valor de la predicció depenent de la capacitat màxima de cada hotel tal i com s'ha fet en la posada en escena. Potser així els resultats *test* haurien estat millors i es reduiria la presència d'*overfitting*.
- Entrenar un model amb més dades. A l'inici del projecte es va decidir fer servir les dades a partir de 2018, ja que al tractar-se d'una prova de concepte es va prioritzar intentar reduir el temps de computació. La conseqüència d'això és que el model no troba tots aquells patrons que podria trobar amb més dades d'entrenament, i que podrien fer que obtinguéssim millors prediccions. També ajudaria a reduir la presència d'*overfitting*.
- Tenir més mesos per a provar la precisió del model, per veure a partir de quan comença a fer prediccions allunyades de la realitat. Amb dos mesos de prova no es poden treure conclusions gaire sòlides.
- La separació entre les dades d'entrenament i de test s'ha fet de manera aleatòria, una opció per a canviar-ho seria tindre en compte la temporalitat, agafar com a mostra test dos mesos seguits, per exemple, i comparar els resultats obtinguts.

Experiència personal

Entrant en l'àmbit personal, voldria fer menció a aquells aspectes que més dificultat m'han causat i, per acabar, l'experiència que en trec de tot plegat:

- El primer i el que més temps m'ha portat, ha estat el fet d'haver de manipular la base de dades en brut per a obtenir el conjunt de dades que ens interessava utilitzant codi *SQL*, em va resultar molt complex la computació d'algunes *queries* i vaig haver de demanar ajuda en moltes ocasions.
- Un altre aspecte que m'ha portat temps és l'optimització d'híper paràmetres. En aquest cas, no és per no saber com programar-ho sinó perquè li vaig donar molta importància al fet de reduir l'*overfitting*, quan realment no en tenia tanta, ja que les diferències entre les mostres *train* i *test* eren força semblants. A més, molts dels paràmetres a optimitzar no els coneixia, i alguns encara em costa saber què volen dir.
- També dir que el fet de no saber programar ni en *SQL* ni en *Python* ha fet que el temps de dedicació a fer certes coses hagi estat molt superior al que hauria dedicat algú amb uns coneixements mínims d'aquests llenguatges.

Tot i que sóc conscient que em queda molt per aprendre, gràcies a haver treballat amb aquests llenguatges considero que he obtingut unes bases sòlides d'*SQL* i sobretot de *Python* que m'ajudaran en la meva carrera professional.

En general, em sento molt satisfet d'haver desenvolupat aquest projecte de *Data Science* que, tot i ser una prova de concepte dins l'empresa, podria ser un problema real perfectament, on apareixen molts dels conceptes treballats en el grau, que m'han permès relacionar els coneixements adquirits en els estudis amb el món professional.

7 BIBLIOGRAFIA

[1] https://cleverdata.io/que-es-machine-learning-big-data/

[2] https://es.wikipedia.org/wiki/Ciencia_de_datos, 07/01/2020

[3] Dorian Pyle, Data Preparation for Data Mining, Morgan Kaufmann Publishers, 1999

[4] http://www.numerosemana.es/iso8601.html

[5] https://www.questionpro.com/blog/es/analisis-descriptivo/

[6] Feature selection – Wikipedia, 25/11/19

[7] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

[8] https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html

[9] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html

[10] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[11] https://xgboost.readthedocs.io/en/latest/python/python_api.html

[12] https://www.aprendemachinelearning.com/que-es-overfitting-y-underfitting-y-como-solucionarlo/

[13] Introduction to Machine Learning with Python. Andreas C.Müller & Sarah Guido

[14] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

8 ANNEX

8.1 QUERIES

- ***Query 1***:

```
CASE WHEN [idtipohab] IN('CUA', 'DBL', 'DBL 02', 'DBL 03', 'DBL 03 VM',
       'DBL 04', 'DBL 05', 'DBL 07', 'DBL 2/4', 'DBL M+IND', 'DBL SB', 'DBL VM',
       'DBL/4 01', 'DBL2', 'DBL2/3', 'DBL3', 'DBL3P', 'DBL4', 'DBLP', 'DBLP 2/4',
        'DBLP2/3', 'DBLP4', 'DBLVP', 'DLB1', 'DUO') THEN 'DBL'
             WHEN [idtipohab] IN('APART 2/4', 'APART 4/6', 'APART DUP', 'APART')
THEN 'APART'
             WHEN [idtipohab] IN('SUI' ,'SUI 2/4', 'SUI 2/6', 'SUI 4/6') THEN
'SUI'
             WHEN [idtipohab] IN('FAM', 'FAM2/4', 'FAMP2/4') THEN 'FAM'
             WHEN [idtipohab] IN('IND', 'INDSB') THEN 'IND'
             WHEN [idtipohab] IN('PREM', 'PREMP', 'PREMP4') THEN 'PREM'
             ELSE [idtipohab]
             END AS [idtipohab]
```

Comentari: *S'agrupen els tipus d'habitacions entre les que tenen característiques similars, i resulten 7 nivells:*

- *DBL = Doble*
- *SUI = Suite*
- *PREM = Premium*
- *FAM = Familiar*
- *IND = Individual*
- *APART = Apartament*
- *VIRTUAL*

- ***Query 2***:

```
CASE WHEN [idregimen] IN('MP', 'MPAL', 'MPAV') THEN 'MP'
             WHEN [idregimen] IN('PC', 'PCAV') THEN 'PC'
             ELSE [idregimen]
             END AS [idregimen]
```

Comentari: *S'agrupen els tipus de règim entre els que tenen característiques similars, i resulten 5 nivells:*

- *MP = Mitja pensió*
- *PC = Pensió Completa*
- *TI = Tot Inclós*
- *AD = Habitació i esmorzar*
- *SHAB = Només habitación*

- ***Query 3***:

```
CASE WHEN [id_canal] IN('B2B', 'B2C') THEN 'ONLINE'
             WHEN [id_canal] IN('TTOO-DI', 'TTOO-GA', 'TTOO-RE') THEN 'TTOO'
             WHEN [id_canal] IN('WEB PROPIA', 'DIRECTO') THEN 'DIRECTO'
             ELSE [id_canal]
             END as [id_canal]
```

Comentari: *S'agrupen els tipus de canal entre els que tenen característiques similars, i resulten 5 nivells:*

- *TTOO = Tour Operador*
- *DIRECTO*
- *ONLINE*
- *GRUPOS*
- *ZZ = Desconegut*

- ***Query 4***:

```
CASE WHEN [idpais_cliente] = 'IE' THEN 'GB'
            WHEN [idpais_cliente] IN('CH', 'NL', 'BE', 'PL', 'DE', 'IT', 'AT',
'CZ') THEN 'CE'
            WHEN [idpais_cliente] IN('NO', 'SE') THEN 'ESC'
            WHEN [idpais_cliente] IN('PT', 'AD', 'FR') THEN 'OE'
            WHEN [idpais_cliente] IN('EE', 'RO', 'RU', 'LT', 'MD', 'UA') THEN
'EE'
            WHEN [idpais_cliente] = 'AE' THEN 'EAU'
            ELSE [idpais_cliente]
            END AS [idpais_cliente]
```

Comentari: *S'agrupen els països segons la localització en el mapa i la influència que tenen en les dades, i resulten 8 nivells:*

- *ES = Espanya*
- *GB = Gran Bretanya*
- *CE = Centre Europa*
- *ESC = Escandinàvia*
- *OE = Oest Europa*
- *EE = Est Europa*
- *EAU = Emirats Àrabs Units*
- *IS = Islàndia*

- ***Query 5***:

```
CASE WHEN [idcliente] = '0000002001' THEN 'TUI_UK_LTD'
            WHEN [idcliente] = '0000002003' THEN 'OL_MEDITERRANEO'
          WHEN [idcliente] = 'Huespedes' THEN 'Huespedes'
            ELSE 'Otros'
            END as [idcliente]
```

Comentari: *S'agrupen les agències clients segons la influència que tenen en les dades, i resulten 4 nivells:*

- *Huespedes = Hostes*
- *TUI_UK_LTD = Agència més influent*
- *OL_MEDITERRANEO = Segona agencia més influent*
- *Otros = La resta de nivells*

❖ ***Query 6***:

```
CREATE VIEW [dbo].[importes] AS

SELECT  idreserva,
              idhotel,
              idcliente,
              idtipohab,
              idregimen,
              idsegmento,
              id_canal,
              fecha,
              fechaentrada,
              fechasalida,
              fechareserva,
              fechacancelacion,
              importe,
              CASE WHEN importe IS NULL THEN
                     CASE WHEN LAG(importe) OVER(PARTITION BY idtipohab,
idregimen, idsegmento, id_canal ORDER BY fecha, fechareserva DESC) IS NULL THEN
                           CASE WHEN LAG(importe, 2) OVER(PARTITION BY idtipohab,
idregimen, idsegmento, id_canal ORDER BY fecha, fechareserva DESC) IS NULL THEN
                                  CASE WHEN LAG(importe, 3) OVER(PARTITION BY
idtipohab, idregimen, idsegmento, id_canal ORDER BY fecha, fechareserva DESC) IS
NULL THEN
                                                AVG(importe) OVER(PARTITION BY
idtipohab, idregimen)
                                  ELSE LAG(importe, 3) OVER(PARTITION BY
idtipohab, idregimen, idsegmento, id_canal ORDER BY fecha, fechareserva DESC) END
                           ELSE LAG(importe,2) OVER(PARTITION BY idtipohab,
idregimen, idsegmento, id_canal ORDER BY fecha, fechareserva DESC) END
                     ELSE LAG(importe) OVER(PARTITION BY idtipohab, idregimen,
idsegmento, id_canal ORDER BY fecha, fechareserva DESC) END
              ELSE importe END AS NEWimporte
FROM (
              SELECT  r.idreserva,
                           r.idhotel,
                           r.idcliente,
                           r.idtipohab,
                           r.idregimen,
                           r.idsegmento,
                           r.id_canal,
                           f.fecha,
                           r.fechaentrada,
                           r.fechasalida,
                           r.fechareserva,
                           r.fechacancelacion,
                           SUM(p.importesin) AS importe
              FROM [dbo].[fact_reservas_mapping] r
              INNER JOIN [SmartHotel].[dbo].[BI_TIME_FINAL] AS f ON
              (
                     f.fecha >= r.fechaentrada
                     AND f.fecha < r.fechasalida
                     AND YEAR(fecha) > 2017
                     AND YEAR(fechareserva) > 2016
              )
              LEFT JOIN [SmartHotel].[dbo].[fact_linprod_final] p ON
              (
                     r.idreserva = p.idreserva
                     AND
                     f.fecha = p.fechaprod
                     AND
```

```
                fechacancelacion IS NULL
            )
            WHERE (idconcepto IN ('AD', 'HAB', 'MP', 'MPAL', 'TI', 'TI AD', 'TI
AD HAB', 'TI JU', 'TI NI', 'TI NI HAB', 'ACF', 'MPAV', 'PC', 'PCAV', 'SHAB') OR
idconcepto IS NULL)
                AND YEAR(fecha) > 2017 AND (idtipohab <> 'ZZZ' OR idregimen
<> 'ZZZ')
            GROUP BY r.idreserva,
                     r.idhotel,
                     r.idcliente,
                     r.idtipohab,
                     r.idregimen,
                     r.idsegmento,
                     r.id_canal,
                     f.fecha,
                     r.fechaentrada,
                     r.fechasalida,
                     r.fechareserva,
                     r.fechacancelacion
            HAVING SUM(p.importesin) >= 1 OR SUM(p.importesin) IS NULL
      ) a

GO
```

Comentari: *S'ajunten les taules* ***dbo.fact_reservas_final*** *i* ***dbo.BI_TIME_FINAL*** *en una taula nova per obtenir una fila per a cada nit de la reserva a l'hotel. Aquesta taula s'ajunta amb la taula* ***dbo.fact_linprod_final*** *segons reserva i segons data per obtenir els imports per nit i per reserva. La taula s'anomena* ***dbo.importes****.*

- ***Query 7***:

```
CREATE VIEW dbo.datasetwithnulls

AS

select
      idhotel,
      fechaocupacion,
      idcliente,
      idtipohab,
      idregimen,
      idsegmento,
      id_canal,
      idpais_cliente,
      SUM(ISNULL([S1H], 0)) AS  [NHab],
      SUM(ISNULL([S2H], 0)) AS  [S1H],
      SUM(ISNULL([S3H], 0)) AS  [S2H],
      SUM(ISNULL([S4H], 0)) AS  [S3H],
      SUM(ISNULL([S5H], 0)) AS  [S4H],
      SUM(ISNULL([S6H], 0)) AS  [S5H],
      SUM(ISNULL([S7H], 0)) AS  [S6H],
      SUM(ISNULL([S8H], 0)) AS  [S7H],
      SUM(ISNULL([S9H], 0)) AS  [S8H],
      SUM(ISNULL([S10H], 0)) AS [S9H],
      SUM(ISNULL([S11H], 0)) AS [S10H],
      SUM(ISNULL([S12H], 0)) AS [S11H],
      SUM(ISNULL([S13H], 0)) AS [S12H],
      SUM(ISNULL([S14H], 0)) AS [S13H],
      SUM(ISNULL([S15H], 0)) AS [S14H],
      SUM(ISNULL([S16H], 0)) AS [S15H],
      SUM(ISNULL([S17H], 0)) AS [S16H],
      SUM(ISNULL([S18H], 0)) AS [S17H],
```

```
SUM(ISNULL([S19H], 0)) AS [S18H],
SUM(ISNULL([S20H], 0)) AS [S19H],
SUM(ISNULL([S21H], 0)) AS [S20H],
SUM(ISNULL([S22H], 0)) AS [S21H],
SUM(ISNULL([S23H], 0)) AS [S22H],
SUM(ISNULL([S24H], 0)) AS [S23H],
SUM(ISNULL([S25H], 0)) AS [S24H],
SUM(ISNULL([S26H], 0)) AS [S25H],
SUM(ISNULL([S27H], 0)) AS [S26H],
SUM(ISNULL([S28H], 0)) AS [S27H],
SUM(ISNULL([S29H], 0)) AS [S28H],
SUM(ISNULL([S30H], 0)) AS [S29H],
SUM(ISNULL([S31H], 0)) AS [S30H],
SUM(ISNULL([S32H], 0)) AS [S31H],
SUM(ISNULL([S33H], 0)) AS [S32H],
SUM(ISNULL([S34H], 0)) AS [S33H],
SUM(ISNULL([S35H], 0)) AS [S34H],
SUM(ISNULL([S36H], 0)) AS [S35H],
SUM(ISNULL([S37H], 0)) AS [S36H],
SUM(ISNULL([S38H], 0)) AS [S37H],
SUM(ISNULL([S39H], 0)) AS [S38H],
SUM(ISNULL([S40H], 0)) AS [S39H],
SUM(ISNULL([S41H], 0)) AS [S40H],
SUM(ISNULL([S42H], 0)) AS [S41H],
SUM(ISNULL([S43H], 0)) AS [S42H],
SUM(ISNULL([S44H], 0)) AS [S43H],
SUM(ISNULL([S45H], 0)) AS [S44H],
SUM(ISNULL([S46H], 0)) AS [S45H],
SUM(ISNULL([S47H], 0)) AS [S46H],
SUM(ISNULL([S48H], 0)) AS [S47H],
SUM(ISNULL([S49H], 0)) AS [S48H],
SUM(ISNULL([S50H], 0)) AS [S49H],
SUM(ISNULL([S51H], 0)) AS [S50H],
SUM(ISNULL([S52H], 0)) AS [S51H],
SUM(ISNULL([S53H], 0)) AS [S52H],
SUM(ISNULL([S1], 0)) / NULLIF(SUM(ISNULL([S1H], 0)), 0) AS   [PreuHab],
SUM(ISNULL([S2], 0)) / NULLIF(SUM(ISNULL([S2H], 0)), 0) AS   [S1],
SUM(ISNULL([S3], 0)) / NULLIF(SUM(ISNULL([S3H], 0)), 0) AS   [S2],
SUM(ISNULL([S4], 0)) / NULLIF(SUM(ISNULL([S4H], 0)), 0) AS   [S3],
SUM(ISNULL([S5], 0)) / NULLIF(SUM(ISNULL([S5H], 0)), 0) AS   [S4],
SUM(ISNULL([S6], 0)) / NULLIF(SUM(ISNULL([S6H], 0)), 0) AS   [S5],
SUM(ISNULL([S7], 0)) / NULLIF(SUM(ISNULL([S7H], 0)), 0) AS   [S6],
SUM(ISNULL([S8], 0)) / NULLIF(SUM(ISNULL([S8H], 0)), 0) AS   [S7],
SUM(ISNULL([S9], 0)) / NULLIF(SUM(ISNULL([S9H], 0)), 0) AS   [S8],
SUM(ISNULL([S10], 0)) / NULLIF(SUM(ISNULL([S10H], 0)), 0) AS [S9],
SUM(ISNULL([S11], 0)) / NULLIF(SUM(ISNULL([S11H], 0)), 0) AS [S10],
SUM(ISNULL([S12], 0)) / NULLIF(SUM(ISNULL([S12H], 0)), 0) AS [S11],
SUM(ISNULL([S13], 0)) / NULLIF(SUM(ISNULL([S13H], 0)), 0) AS [S12],
SUM(ISNULL([S14], 0)) / NULLIF(SUM(ISNULL([S14H], 0)), 0) AS [S13],
SUM(ISNULL([S15], 0)) / NULLIF(SUM(ISNULL([S15H], 0)), 0) AS [S14],
SUM(ISNULL([S16], 0)) / NULLIF(SUM(ISNULL([S16H], 0)), 0) AS [S15],
SUM(ISNULL([S17], 0)) / NULLIF(SUM(ISNULL([S17H], 0)), 0) AS [S16],
SUM(ISNULL([S18], 0)) / NULLIF(SUM(ISNULL([S18H], 0)), 0) AS [S17],
SUM(ISNULL([S19], 0)) / NULLIF(SUM(ISNULL([S19H], 0)), 0) AS [S18],
SUM(ISNULL([S20], 0)) / NULLIF(SUM(ISNULL([S20H], 0)), 0) AS [S19],
SUM(ISNULL([S21], 0)) / NULLIF(SUM(ISNULL([S21H], 0)), 0) AS [S20],
SUM(ISNULL([S22], 0)) / NULLIF(SUM(ISNULL([S22H], 0)), 0) AS [S21],
SUM(ISNULL([S23], 0)) / NULLIF(SUM(ISNULL([S23H], 0)), 0) AS [S22],
SUM(ISNULL([S24], 0)) / NULLIF(SUM(ISNULL([S24H], 0)), 0) AS [S23],
SUM(ISNULL([S25], 0)) / NULLIF(SUM(ISNULL([S25H], 0)), 0) AS [S24],
SUM(ISNULL([S26], 0)) / NULLIF(SUM(ISNULL([S26H], 0)), 0) AS [S25],
SUM(ISNULL([S27], 0)) / NULLIF(SUM(ISNULL([S27H], 0)), 0) AS [S26],
```

```
	SUM(ISNULL([S28], 0)) / NULLIF(SUM(ISNULL([S28H], 0)), 0) AS [S27],
	SUM(ISNULL([S29], 0)) / NULLIF(SUM(ISNULL([S29H], 0)), 0) AS [S28],
	SUM(ISNULL([S30], 0)) / NULLIF(SUM(ISNULL([S30H], 0)), 0) AS [S29],
	SUM(ISNULL([S31], 0)) / NULLIF(SUM(ISNULL([S31H], 0)), 0) AS [S30],
	SUM(ISNULL([S32], 0)) / NULLIF(SUM(ISNULL([S32H], 0)), 0) AS [S31],
	SUM(ISNULL([S33], 0)) / NULLIF(SUM(ISNULL([S33H], 0)), 0) AS [S32],
	SUM(ISNULL([S34], 0)) / NULLIF(SUM(ISNULL([S34H], 0)), 0) AS [S33],
	SUM(ISNULL([S35], 0)) / NULLIF(SUM(ISNULL([S35H], 0)), 0) AS [S34],
	SUM(ISNULL([S36], 0)) / NULLIF(SUM(ISNULL([S36H], 0)), 0) AS [S35],
	SUM(ISNULL([S37], 0)) / NULLIF(SUM(ISNULL([S37H], 0)), 0) AS [S36],
	SUM(ISNULL([S38], 0)) / NULLIF(SUM(ISNULL([S38H], 0)), 0) AS [S37],
	SUM(ISNULL([S39], 0)) / NULLIF(SUM(ISNULL([S39H], 0)), 0) AS [S38],
	SUM(ISNULL([S40], 0)) / NULLIF(SUM(ISNULL([S40H], 0)), 0) AS [S39],
	SUM(ISNULL([S41], 0)) / NULLIF(SUM(ISNULL([S41H], 0)), 0) AS [S40],
	SUM(ISNULL([S42], 0)) / NULLIF(SUM(ISNULL([S42H], 0)), 0) AS [S41],
	SUM(ISNULL([S43], 0)) / NULLIF(SUM(ISNULL([S43H], 0)), 0) AS [S42],
	SUM(ISNULL([S44], 0)) / NULLIF(SUM(ISNULL([S44H], 0)), 0) AS [S43],
	SUM(ISNULL([S45], 0)) / NULLIF(SUM(ISNULL([S45H], 0)), 0) AS [S44],
	SUM(ISNULL([S46], 0)) / NULLIF(SUM(ISNULL([S46H], 0)), 0) AS [S45],
	SUM(ISNULL([S47], 0)) / NULLIF(SUM(ISNULL([S47H], 0)), 0) AS [S46],
	SUM(ISNULL([S48], 0)) / NULLIF(SUM(ISNULL([S48H], 0)), 0) AS [S47],
	SUM(ISNULL([S49], 0)) / NULLIF(SUM(ISNULL([S49H], 0)), 0) AS [S48],
	SUM(ISNULL([S50], 0)) / NULLIF(SUM(ISNULL([S50H], 0)), 0) AS [S49],
	SUM(ISNULL([S51], 0)) / NULLIF(SUM(ISNULL([S51H], 0)), 0) AS [S50],
	SUM(ISNULL([S52], 0)) / NULLIF(SUM(ISNULL([S52H], 0)), 0) AS [S51],
	SUM(ISNULL([S53], 0)) / NULLIF(SUM(ISNULL([S53H], 0)), 0) AS [S52]
from
(
	SELECT
	a.idhotel
	,fechaocupacion
	,a.idcliente
	,a.idtipohab
	,a.idregimen
	,a.idsegmento
	,a.id_canal
	,idpais_cliente
	,SUM(importe) AS importe
	,'S' + cast(RN as varchar) + 'H' as SH
	,'S' + cast(RN as varchar) as S
	,SUM(cantidad) as cantidad
	FROM (
		SELECT r.[idhotel]
			,r.[idreserva]
			,f.fecha as fechaocupacion
			,s.AñoSemana
			,ROW_NUMBER() OVER (PARTITION BY r.idhotel, r.idreserva,
f.fecha ORDER BY s.AñoSemana DESC) AS RN
			,cantidad
			,NEWimporte as importe
			,r.idcliente
			,r.idtipohab
			,r.idregimen
			,r.idsegmento
			,r.id_canal
			,r.idpais_cliente

		FROM [SmartHotel].[dbo].[fact_reservas_mapping] r
		inner join [SmartHotel].[dbo].[BI_TIME_FINAL] as f on
		(
		f.fecha >= r.fechaentrada
		and
```

```
                f.fecha < r.fechasalida
                and
                YEAR(fecha) > 2017
                )
                left join [dbo].[importes] o
                on r.idreserva = o.idreserva AND f.fecha = o.fecha and r.idhotel
= o.idhotel
                inner join (SELECT distinct AñoSemana FROM
[SmartHotel].[dbo].[BI_TIME_FINAL]) as s
                on s.AñoSemana >= concat(year(r.fechareserva), '.', right('0'+
cast(DATEPART(wk, r.fechareserva) as varchar),2))
                and s.AñoSemana <= concat(year(f.fecha), '.', right('0'+
cast(DATEPART(wk, f.fecha) as varchar),2))
              union all
              SELECT r.[idhotel]
                      ,r.[idreserva]
                      ,f.fecha as fechaocupacion
                      ,s.AñoSemana
                      ,ROW_NUMBER() OVER (PARTITION BY r.idhotel, r.idreserva,
f.fecha ORDER BY s.AñoSemana DESC) AS RN
                      ,-cantidad
                      ,-NEWimporte as importe
                      ,r.idcliente
                      ,r.idtipohab
                      ,r.idregimen
                      ,r.idsegmento
                      ,r.id_canal
                      ,r.idpais_cliente
                FROM [SmartHotel].[dbo].[fact_reservas_mapping] r
                inner join [SmartHotel].[dbo].[BI_TIME_FINAL] as f
                on f.fecha >= r.fechaentrada
                and f.fecha < r.fechasalida
                and YEAR(fecha) > 2017
                left join [dbo].[importes] o
                on r.idreserva = o.idreserva AND f.fecha = o.fecha and r.idhotel
= o.idhotel
                inner join (SELECT distinct AñoSemana FROM
[SmartHotel].[dbo].[BI_TIME_FINAL]) as s
                on s.AñoSemana >= concat(year(r.fechacancelacion), '.',
right('0'+ cast(DATEPART(wk, r.fechacancelacion) as varchar),2)) and
                s.AñoSemana <= concat(year(f.fecha), '.', right('0'+
cast(DATEPART(wk, f.fecha) as varchar),2))
                WHERE r.fechacancelacion is not null) a
                WHERE a.idtipohab <> 'ZZZ'
                AND a.idregimen <> 'ZZZ'
        GROUP BY
        a.idhotel, fechaocupacion, RN, a.idcliente, a.idtipohab, a.idregimen,
a.idsegmento, a.idpais_cliente, a.id_canal
) pre
        pivot
        (
              avg(cantidad)
              for [SH] in ([S1H], [S2H], [S3H], [S4H], [S5H], [S6H], [S7H],
[S8H], [S9H],[S10H],

        [S11H],[S12H],[S13H],[S14H],[S15H],[S16H],[S17H],[S18H],[S19H],[S20H],

        [S21H],[S22H],[S23H],[S24H],[S25H],[S26H],[S27H],[S28H],[S29H],[S30H],

        [S31H],[S32H],[S33H],[S34H],[S35H],[S36H],[S37H],[S38H],[S39H],[S40H],

        [S41H],[S42H],[S43H],[S44H],[S45H],[S46H],[S47H],[S48H],[S49H],[S50H],
```

```
                                [S51H],[S52H],[S53H]
                                )
        ) pivHab
        pivot
        (
                avg(importe)
                for [S] in ( [S1], [S2], [S3], [S4], [S5], [S6], [S7], [S8],
[S9],[S10],

        [S11],[S12],[S13],[S14],[S15],[S16],[S17],[S18],[S19],[S20],

        [S21],[S22],[S23],[S24],[S25],[S26],[S27],[S28],[S29],[S30],

        [S31],[S32],[S33],[S34],[S35],[S36],[S37],[S38],[S39],[S40],

        [S41],[S42],[S43],[S44],[S45],[S46],[S47],[S48],[S49],[S50],
                                [S51],[S52],[S53]
                                )
        ) piv

group by      idhotel, fechaocupacion, idcliente, idtipohab, idregimen,
idsegmento, idpais_cliente, id_canal

GO
```

Comentari: *S'ajunten les taules* ***dbo.importes*** *amb* ***dbo.fact_reservas_maping*** *i* ***dbo.BI_TIME_FINAL****. Segons reserva i segons data. Es filtra per any més gran a 2017, i es treuen els tipus d'habitació o règims sense informació, ja que les reserves per a aquests casos tenen informació falsa.*

8.2 OBSERVACIONS

Observació 1: *Les 5 primeres queries formen part d'una sola, que té per objectiu proporcionar exactament la mateixa informació que la taula 'dbo.fact_reservas_final' però amb les variables en qüestió recategoritzades, creant la taula 'dbo.fact_reservas_mapping' tal i com s'explica en el document. Es fa d'aquesta manera per no haver de modificar la taula original, ja que és una taula que es fa servir per a altres projectes i podria causar problemes modificar-la sense avisar.*