A partir del preprocessament fins a les conclusions, es farà un resum/explicació del contingut desenvolupat en el *Jupyter Notebook* (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 dels comentaris oportuns.

La idea és que els *Notebooks* siguin una guia que ajudi 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.

Preprocessament

*preproce.ipynb*

Un cop ja hi ha les dades guardades en un arxiu *.csv* es pot començar el preprocessament.

En aquest apartat s’expliquen els passos aplicats i es mostra en cada un d’ells com es van modificant les dimensions de la base de dades, si es que es modifiquen, per tal de tenir un control a simple vista les dades que es van eliminant:

Inici: 122413 files i 114 columnes

* Filtrar per dades entre l’1 de gener de 2018 i el 30 de setembre de 2019

Se sap que la base de dades conté informació des de gener de 2018 fins a octubre de 2020. Però només interessa treballar amb aquelles dates que ja han passat per a assolir l'objectiu d'aquest estudi, per tant, com l'última actualització de les dades consta a octubre de 2019, s'agafen totes les dades des de gener de 2018 fins a setembre del 2019 inclós. La resta de dades es guarden en un arxiu i es podran utilitzar a la part final del projecte per comprovar l'eficàcia del model.

Actual: 107160 files i 114 columnes

* Primer anàlisi de les variables i el seu tipus

Anàlisi de cada tipus de variable i transformació a un altre tipus si cal. Hi ha 8 variables categòriques (*idhotel*, *fechaocupacion*, *idcliente*, *idtipohab*, *idregimen*, *idsegmento*, *id\_canal*, *idpais\_cliente*. I 106 variables numèriques, de les quals 53 corresponen al nombre d'habitacions reservades i altres 53 al preu mig per habitació.

En ambdós casos, una variable correspon a la setmana actual, *NHab*, en el cas de les habitacions reservades, i *PreuHab*, en el cas del preu mig de les habitacions, mentre que les altres 52 són les dades de cada una de les setmanes enrere fins a completar un any.

Fent un primer cop d'ull als estadístics que mostren les dades, es veuen casos extranys en les variables referents al preu de les habitacions.

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 és totalment fals, 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, ho veiem més endavant.

En aquest pas no ha calgut modificar la base de dades.

* Comprovar duplicats

No hi ha cap duplicat en el conjunt de dades

* Tractament de nulls

Es comprova que hi ha 3534 nulls en 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.

Actual: 103626 files i 114 columnes.

* S’afegeix la variable data

A partir de l’arxiu *dates.csv* s’afegeix la variable *diaAño* corresponent al dia de l’any. Per tant, hi haurà informació del dia de l’any de la variable *fechaocupacion*.

La variable nova ha de ser de tipus categòrica.

Actual: 103626 files i 115 columnes

* Filtrar segons el preu mig

En aquest pas es tractaran els casos estranys que s’han vist en el primer anàlisi.

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.

Actual: 102377 files i 115 columnes

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.

Actual: 97875 files i 115 columnes

En el cas dels màxims, s’analitza el percentatge de dades sobre el total de la mostra que va quedant a mesura que es va augmentant el preu, i es veu que a partir d’uns 300 euros, el percentatge de dades sobre el total és molt petit. En un principi s’havia decidit tallar la mostra a partir d’aquell preu que, les dades que el superin, conformessin menys d’un 1%. Però veient 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.

Actual: 97864 files i 115 columnes

### Categorització

Es fa una categorització de totes les variables categòriques excepte la variable *idcliente*.

* *idhotel*: Es canvia l’identificador per el nom de l’hotel.
* *idtipohab*: S’agrupen les habitacions en 7 nivells diferents.
* *idregimen*: S’agrupen els règims en 7 nivells diferents.
* *idsegmento*: Es canvia el nom d’un nivell.
* *idcanal*: S’agrupen els canals en 5 nivells diferents.
* *idpais\_cliente*: S’agrupen els països segons la influència de cada un i la localització en el mapa, quedant agrupats en 8 nivells diferents.