Exemple d'aplicació de tècniques de mineria de dades a la indústria hotelera

Components del grup:

Subgrup 1

Guillem Querol Llaveria
Pablo Morante López
Antoni Ramoneda Montoya
Carles Requena Sánchez
Aleix Salvador Barrera

Subgrup 2

Laura Julià Melis
Victor Miranda Hernández
Marta Piñol Palau
Sofía Touceda Suárez

Data d'entrega: **27.09.2018**

Definició del projecte i assignació.

Aquest treball es desenvoluparà amb l'objectiu de millorar les experiències de viatges en l'àmbit de la industria hotelera. Per a assolir aquest objectiu, s'utilitzaran una sèrie de tècniques de mineria de dades.

- **Font i informació sobre la base de dades.**

Com a materia prima per a l'anàlisi, s'han utilitzat dades importades a través d'una API des de 'Booking'. Aquestes dades són propietat de *Booking* però un usuari de kaggle les ha fet públiques i en permet l'ús amb finalitats acadèmiques. A la web trobem dues versions de les dades: un primer *'dataset'* amb la informació obtinguda de *Booking* (https://www.kaggle.com/jiashenliu/ 515k-hotel-reviews-data-in-europe), i un de segon, en el qual la informació del primer ha estat enriquida amb noves variables. Aquest últim és el que s'utilitza en aquest treball https://www.kaggle.com/ycalisar/hotel-reviews-dataset-enriched.

La base de dades conté, aproximadament, 515.000 opinions de clients, i les puntuacions otorgades pels mateixos, que han contractat estades a un total de 1.493 hotels d'Europa. *Booking* també proporciona les coordenades de cada hotel per a realitzar geolocalitzacions.

- **Estructura de la base de dades.**

A grans trets, la matriu de dades resultant conté 41 variables, 21 de les quals són numèriques i 20 categòriques, i 515.738 observacions. Tanmateix, per a ajustar-nos a la dimensionalitat de les dades proposada a les bases del treball, s'ha seleccionat aleatòriament un subconjunt de 5.000 observacions i s'han eliminat de la base de dades les variables **[Hotel_Address, Hotel_State, Room_Type, Tags, Day_of_Week, Day_of_Year, Bed_Type, Week_of_Month, Week_of_Year, Quarter_of_Year,Reviewer_Country]**, que es consideren poc rellevants en relació als objectius del treball. D'aquesta manera, assegurem que tots els procediments requerits podran ser implementats de manera eficient i satisfactòria amb les nostres dades.

Respecte als valors **missing**, la base de dades revela un total de 3.350, que representen un 2, 23% de la matriu de dades completa ($m \cdot n$). En aquest sentit, presentem la Taula 1 que mostra com es distribuieixen els valors missing entre les diferents variables, així com la Figura 1.1 on es representa un histograma que resumeix la taula anterior.

variable	nre.Missings	freq.Missings
Hotel_lat	23	0.0153 %
Hotel_lng	23	0.0153 %
Businesses_100m	23	0.0153 %
Businesses_1km	23	0.0153 %
Businesses_5km	23	0.0153 %
Room_Type_Level	3209	2.1393 %
Trip_Type	14	0.0093 %
Reviewer_Nationality	10	0.0067 %
Negative_Review	2	0.0013 %

Table 1: Taula de valors missing

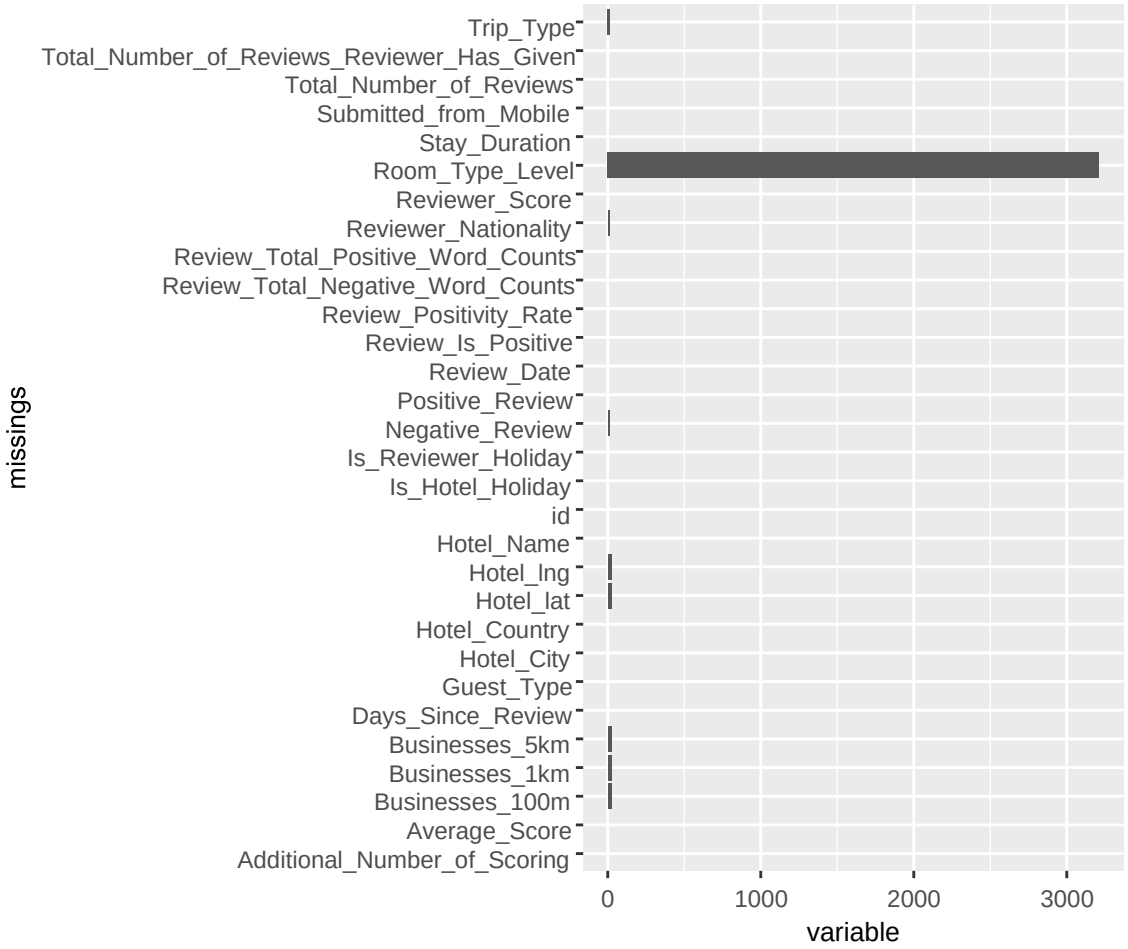

Figure 1: Histograma dels valors missing