Exploratory Data Analysis

Hoteles en Portugal

1. Introducción.

“El análisis exploratorio de datos es una forma de analizar datos definido por John W. Tukey (E.D.A: Exploratory data analysis) es el tratamiento estadístico al que se someten las muestras recogidas durante un proceso de investigación en cualquier campo científico” [1] . Esta memoria presenta los diferentes pasos y aspectos que se han seguido durante la realización de un EDA sobre datos de hoteles en Portugal. La motivación para la realización del mismo parte de la necesidad de consolidar la formación obtenida durante la primera parte del bootcamp de The Bridge: Data Science, poniendo en práctica los conocimientos adquiridos mostrando así la comprensión y adquisición de los mismos.

1. Temática

La temática escogida es información respectiva a dos hoteles situados en Portugal, el primero de ellos situado en una ciudad y el segundo situado en el campo. Inicialmente no fue la temática escogida. La primera temática escogida fue información sobre partidos de fútbol en La Liga pero la dificultad al acceso de datos con información relevante y suficiente provocó que se cambiase a la temática actual.

La elección de la temática final, los hoteles en Portugal, estuvo causada por los siguientes puntos positivos: la gran cantidad de datos, tanto en distintas instancias como en variables distintas, lo que nos permite obtener resultados más fiables al tener una gran cantidad de muestras; la temática en sí, es interesante y accesible para gente no experta su campo de interés; la reutilizabilidad de los datos, la cantidad de datos y el tipo de datos nos van a permitir trabajar en el futuro con estos mismos datos en el proyecto de Machine Learning.

1. Obtención de los datos

Los datos utilizados han sido obtenidos de la plataforma Kaggle [3], pertenecientes al artículo publicado en [2]. Los datos se corresponden a dos hoteles, uno es un hotel urbano y el otro hotel es un resort en el campo, incluye información de cuándo se realizó la reserva, duración de la estadía, el número de adultos, niños y/o bebés, y el número de espacios de estacionamiento disponibles entre otras cosas. Los datos van desde el 1 de julio de 2015 al 31 de agosto de 2017. Los datos se componen de las siguientes variables:

**hotel** : Hotel (H1 = Resort Hotel or H2 = City Hotel)

**is\_canceled** : Value indicating if the booking was canceled (1) or not (0)

**lead\_time** : Number of days that elapsed between the entering date of the booking into the PMS and the arrival date

**arrival\_date\_year** : Year of arrival date

**arrival\_date\_month** : Month of arrival date

**arrival\_date\_week\_number** : Week number of year for arrival date

**arrival\_date\_day\_of\_month** : Day of arrival date

**stays\_in\_weekend\_nights** : Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel

**stays\_in\_week\_nights** : Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel

**adults** : Number of adults

**children** : Number of children

**babies** : Number of babies

**meal** : Categories are presented in standard hospitality meal packages:

Undefined/SC – no meal package;

BB – Bed & Breakfast;

HB – Half board (breakfast and one other meal – usually dinner);

FB – Full board (breakfast, lunch and dinner)

**country** : Country of origin. Categories are represented in the ISO 3155–3:2013 format

**market\_segment** : Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”

**distribution\_channel** : Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”

**is\_repeated\_guest** : Value indicating if the booking name was from a repeated guest (1) or not (0)

**previous\_cancellations** : Number of previous bookings that were cancelled by the customer prior to the current booking

**previous\_bookings\_not\_canceled** : Number of previous bookings not cancelled by the customer prior to the current booking

**reserved\_room\_type** : Code of room type reserved. Code is presented instead of designation for anonymity reasons.

**assigned\_room\_type** : Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due

**booking\_changes** : Number of changes/amendments made to the booking from the moment the booking was entered on the PMS

**deposit\_type** : Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories:

No Deposit – no deposit was made;

Non Refund – a deposit was made in the value of the total stay cost;

Refundable – a deposit was made with a value under the total cost of stay

**agent** : ID of the travel agency that made the booking

**company** : ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons

**days\_in\_waiting\_list** : Number of days the booking was in the waiting list before it was confirmed to the customer

**customer\_type**: Contract - when the booking has an allotment or other type of contract associated to it;

Group – when the booking is associated to a group;

Transient – when the booking is not part of a group or contract, and is not associated to other transient booking;

Transient-party – when the booking is transient, but is associated to at least other transient booking

**adr** : Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights

**required\_car\_parking\_spaces** : Number of car parking spaces required by the customer

**total\_of\_special\_requests** : Number of special requests made by the customer (e.g. twin bed or high floor)

**reservation\_status** : Reservation last status, assuming one of three categories:

Canceled – booking was canceled by the customer;

Check-Out – customer has checked in but already departed;

No-Show – customer did not check-in and did inform the hotel of the reason why

**reservation\_status\_date** : Date at which the last status was set. This variable can be used in conjunction with the ReservationStatus to understand when was the booking canceled or when did the customer checked-out of the hotel

1. Hipótesis

ddddddddddd

1. Preprocesado y limpia

ddddddddddddddddddddd

1. exploratorio

ddddddddddddddddddddddddd

1. Análisis

ddddddddddddddddddddddddddddd

1. Conclusiones

dddddddddddddddddddddddddddddddddddddddddddddddddddddddddddd

Referencias

[1] *Análisis exploratorio de datos*. (n.d.). Wikipedia. Retrieved January 12, 2022, from <https://es.wikipedia.org/wiki/An%C3%A1lisis_exploratorio_de_datos>

[2] Antonio, N., de Almeida, A., & Nunes, L. (2019). Hotel booking demand datasets. *Data in Brief*, *22*, 41–49. https://doi.org/10.1016/j.dib.2018.11.126

[3] Kaggle: Your Machine Learning and Data Science Community. Retrieved January 12, 2022, from http://kaggle.com