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.

2. 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.

3. 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. Para ver todos los datos dirigirse al apéndice A.

4. Hipótesis

Los dos hoteles tienen ubicaciones muy diferenciadas como son la ciudad y el campo, por lo tanto las características de las reservas es posible que varíen entre un hotel y otro. Como resultado es posible que la gestión de ambos hoteles no deba de realizarse de una manera centralizada, sino que será mejor gestionar parte de manera diferente. En este trabajo vamos a buscar áreas diferenciadas en las que haya que gestionar de manera diferente en ambos hoteles y cuales de manera conjunta.

5. Preprocesado y limpia

El paso previo a la exploración de los datos es el preprocesado y la limpieza de los mismos. En este paso vamos a preparar los datos para que sea fácil trabajar con ellos y eliminar datos corruptos que puedan provocar errores. También puede ser interesante reducir los datos mediante unas condiciones impuestas por el analista.

En el caso de los hoteles se realizan las siguientes operaciones de preproceso o limpieza:

- Se eliminan las entradas que no tienen clientes asignados, es decir que no tienen adultos, niños o bebés, ya que no tiene sentido que existan reservas vacías.
- Se sustituyen los valores sin definir en el régimen de estancia por los valores de “sin régimen” .
- Se eliminan las reservas canceladas, ya que para hallar las diferencias entre los dos hoteles vamos a centrarnos en los clientes que sí se alojan una vez llega el momento.
- Se ajusta el formato del mes de reserva para facilitar su uso durante el exploratorio.
- Nos quedamos únicamente con las reservas de particulares o reservas múltiples de particulares excluyendo los grupos y agencias. Esto es debido a que representan una parte muy pequeña de los datos.
- Por último dividimos los datos según al hotel al que pertenecen.

6. Análisis

El objetivo de este proyecto, como ya se ha explicado anteriormente, es analizar los distintos datos de los dos hoteles para buscar diferencias entre ellos y establecer o no una gestión única o individual para los hoteles. Nos vamos a enfocar en diversas características presentes de una forma u otra en los datos. Estas características son : cuándo se produce la estancia, qué tipo de grupos vienen (solo adultos, familias…), qué tipo de habitación se solicita más, longitud de la estancia y su distribución en fin de semana o no, país de procedencia, uso del aparcamiento y régimen de alojamiento. Todas estas características se

han analizado para los dos hoteles en conjunto y para cada hotel en particular, lo cual nos permite ver de una manera más clara las diferencias y si debemos gestionar esa parte de manera individual por hotel o de manera conjunta.

- Momento de la estancia: para obtener información sobre este punto miramos en los datos, mediante gráficas de barras, la distribución de las estancias atendiendo a cuando se producen teniendo en cuenta el año, mes y semana. Las gráficas de año nos aportan muy poca información debido a cómo se han recogido los datos, ya que no se trata de tres años completos si no que los datos del primer y último año son parciales. En cuanto a cómo se distribuyen las estancias de manera mensual y semanal se observa que durante los meses y semanas correspondientes a verano se produce un incremento de las estancias así como de los niños que se alojan en los hoteles. En el hotel de ciudad este cambio es menos pronunciado ya que su gráfica es más constante a lo largo del tiempo.

- Tipo de cliente: con el fin de analizar el tipo de cliente que viene representamos de manera gráfica cómo se distribuye el tipo de público que tenemos por habitación, esto lo representamos de manera gráfica con la cantidad de adultos que hacen cada reserva de habitación para solo adultos o para adultos con algún niño o bebé. En estas gráficas observamos que la mayor parte de nuestra base de clientes son parejas de adultos sin niños en la habitación, aunque en el hotel de campo esta tendencia es ligeramente menor.

- Tipo de habitación: en este punto analizamos si el reparto de habitaciones que tenemos en nuestro hotel se ajusta a la demanda de los clientes, para ello nos fijamos en la habitación deseada por el cliente en el momento de la reserva y la habitación asignada en el momento de la estancia. Generamos un *treemap* que nos muestra la distribución de habitación pedido / habitación asignada, en estas gráficas se observa que la mayor parte de nuestra base de clientes solicitan la opción A de habitación pero en una cantidad de casos significativa obtienen la habitación de tipo D. Para solucionar este problema se podría reducir el número de habitaciones de los tipos menos solicitados y aumentar las de tipo A para ajustarse mejor a las demandas de los clientes.

- Longitud de la estancia y distribución en la semana: para ver la duración de la estancia y cómo se distribuye en días entre semana o fin de semana generamos unas gráficas de barras apiladas de la suma de los días y de la media de los días.

En la gráfica de la suma de los días podemos observar que, al igual que en el primer punto, en verano incrementa el volumen de clientes, la diferencia entre semana y fin de semana se ajusta a la diferencia de 5 días para el primero enfrente de 2 días para el segundo. En cuánto a la media de días ofrece unos resultados muy interesantes, ya que en la ciudad se mantiene muy estable todo el año pero en el hotel de campo se incrementa durante el verano y en algunas semanas concretas, posiblemente coincidentes con puentes vacacionales. Otro punto a tener en cuenta es que la media de duración de estancia es significativamente superior en el hotel de campo.

- País de procedencia: al analizar la distribución de procedencia de los clientes de los hoteles podemos observar que la mayoría son de origen nacional (Portugal) pero teniendo también una importante representación de otros clientes de origen europeo. Además existe una diferencia entre la procedencia según el hotel de destino, en el caso de la ciudad hay mayor variedad de países de origen y no está tan marcado el origen nacional del cliente, en el caso del hotel de campo hay una menor variedad de países y la mayor parte de los clientes son nacionales aunque sin dejar de haber representación europea.

- Aparcamiento: la información obtenida al analizar el uso de los aparcamientos por los clientes es muy clara, la mayoría de los clientes no lo utiliza, sobre todo en el caso de la ciudad, en el caso del hotel de campo es poco utilizado pero su uso tiene algo más de relevancia para algunos clientes.

- Régimen de alojamiento: en este punto nos interesa ver cómo se distribuye la elección del cliente a la hora de elegir el régimen de alojamiento. La mayoría opta por la opción de desayuno, una pequeña parte por desayuno y una comida, otra parte de tamaño similar por ningún tipo de comida, y prácticamente nadie como régimen completo de comidas. En el caso de la ciudad es menor la gente que coge algo fuera del régimen de desayuno en comparación al hotel de campo.

7. Conclusiones

En este punto vamos a resumir la información obtenida por el análisis, primero exponiendo las diferencias de cada hotel y a continuación las conclusiones generales.

El hotel de la ciudad tiene un mayor número de clientes que se distribuyen de una manera constante a lo largo del año, con una duración similar de estancia. El origen del cliente es más diverso geográficamente y el uso del parking es muy reducido. El hotel de campo tiene menor volumen de clientes que se distribuyen de manera no uniforme a lo largo del tiempo siendo el verano el periodo de mayor afluencia. La duración de la estancia es superior al hotel de la ciudad y sus clientes no son tan geográficamente diversos como en el hotel de ciudad. El uso del parking es bajo pero necesario para un porcentaje de los clientes.

Las conclusiones generales son las siguientes: la mayoría de los clientes vienen en grupos de dos adultos sin niños a su cargo, las habitaciones más demandadas son las de tipo A pero no hay suficientes para abastecer la demanda y muchos clientes se acaban alojando en habitaciones de tipo D, la mayoría de los clientes se alojan con régimen de sólo desayuno.

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

Apéndice A

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