Tarea Académica 1 - TA1 semana 7

Profesora: Patricia Reyes Silva

Integrantes del grupo: Natalia Maury Castañeda (u201816996)
Oscar Flores Palermo (u201716498)
Carlos Eduardo Iparraguirre Marujo (u201810601)

Fecha de entrega: 5 de mayo del 2021
Ciclo: 2021-1

Índice

Contenido

Objetivo:

El objetivo de esta tarea académica es realizar un análisis exploratorio de un conjunto de datos, creando visualizaciones y obteniendo inferencias básicas utilizando R/RStudio como herramienta de software. Para ello se utilizará el dataset de Hotel Booking Demand.

Caso de análisis

Sobre el Dataset:

El dataset de Hotel Booking Demand contiene la información de reserva de hoteles, ya sean en una ciudad o un resort, que hacen los clientes. Incluye los detalles de cuándo se hizo la reserva, por cuánto tiempo se quedó, cantidad de personas que reservaron y el tipo de persona (adulto, niño, bebé, etc.), etc. El Dataset fue creado por Nuno Antonio, Ana Almeida, y Luís Nunes. Lo publicaron en el artículo "*Hotel Booking Demand Dataset*" para *Data in Brief*, perteneciendo al volumen 22 de esta página en el 2019 (Mostipak; 2020). Debido a que todos los autores pertenecen a Portugal y han trabajado en algún momento en las mismas instituciones al mismo tiempo, se asume que el artículo se publicó en Portugal pero no se especifica dicha información.

Los datos fueron recolectados a través de la extracción de las bases de datos SQL del sistema de gestión de propiedades (SGM) de hoteles. Para ello, se utilizaron consultas TSQL directamente ejecutadas en la base de datos del SGM de los hoteles. Para el análisis de los datos, se utilizaron comandos en r. Ambos hoteles examinados se encuentran en Portugal, específicamente en H1 en la región turística de Algarve está el hotel resort y H2 en la ciudad de Lisboa está el hotel de ciudad. La administración de los hoteles mencionados dio su permiso para que la información sea pública.

Sobre los autores:

Data in Brief es un *journal* que provee datasets. Esto les permite a los investigadores utilizar los datos de otros investigadores y también publicar sus propios datasets en artículos. El objetivo de la compañía es incrementar el tráfico en investigaciones de datos, facilitar la reproducción de datos al estar bien explicada en artículos, hacer que las investigaciones sean más fáciles de encontrar y accesibles a las personas (como un google de datasets), crear posibilidades para colaboraciones. El motivo de los objetivos de *Data in Brief* se basa en la premisa de que uno no sabe qué datos pueden ser útil para una persona, por lo que desean crear una plataforma que permita encontrar dichos datos (Elsevier B.V, ScienceDirect; S/F).

Ana Maria de Almeida es doctora (no se especifica en qué tiene su doctorado) e investigadora que actualmente pertenece al Instituto Universitario de Lisboa en el departamento de ciencias de la computación como profesora asistente. Tiene experiencia como investigadora senior en la Universidad de Coimbra y en ISTAR-IUL en el departamento de ciencias de la computación donde hizo una colaboración como investigadora principal. Sus habilidades están enfocadas

en matemática discreta, programación matemática, modelado y optimización de modelado, resolución de problemas, heurística, reconocimiento de patrones y matemática aplicada. Ha hecho 58 publicaciones relacionadas a datos utilizando matemática discreta, algoritmos, etc., ha realizado 3 proyectos relacionados a analítica predictiva y seguridad a lo largo de su carrera (ResearchGate, S/F).

Luís Nunes es un ingeniero de software que actualmente pertenece al Instituto Universitario de Lisboa en el departamento de ciencias de la computación como profesor asistente. En cuanto a habilidades y experiencia, sabe programación orientada a objetos, machine learning avanzado, reconocimiento de patrones, robótica, redes neuronales e inteligencia artificial, y visión computacional. Ha realizado 47 publicaciones relacionadas a los datos mediante machine learning, analítica predictiva, etc., ha realizado 4 proyectos relacionados al análisis, analítica predictiva, seguridad y *urban sensing* a lo largo de su carrera (ResearchGate, S/F).

Nuno Antonio es un ingeniero de datos que actualmente pertenece a la Universidade NOVA de Lisboa como profesor asistente. Tiene experiencia como profesor asistente en el Instituto Universitario de Lisboa y ha hecho un doctorado en el Instituto Universitario de Lisboa en el departamento de información científica y tecnológica. Sus habilidades se enfocan principalmente en la ciencia de datos[1], project management, Data mining, y machine learning avanzado. Ha realizado 36 publicaciones relacionadas a la analítica predictiva, machine learning, entre otros y 2 proyectos sobre analítica predictiva a lo largo de su carrera (ResearchGate, S/F).

Casos de uso:

Este dataset tiene distintos casos de uso dependiendo de la persona o compañía que vaya a usar los datos.Un caso de uso sería para las cadenas de hoteles saber cuál es el tiempo promedio que la gente se queda en el hotel, quienes son ese tipo de personas (adultos, familias, etc) y en base a ello hacer ofertas. Otra de ellas es que se puede ver en qué fechas la gente se queda más en los hoteles para realizar paquetes, promociones, etc. Otro caso de uso puede ser para las tiendas que están cerca a los hoteles para saber cuándo vendrá más gente y saber cuándo pueden vender más, hacer ofertas, hacia quienes dirigir sus productos, etc. Dependiendo del sujeto puede variar el caso de uso de este dataset.

La importancia de este análisis del dataset se basa en el caso de uso que le de ya que tendrá valor para dichas personas. Un ejemplo serían los dueños de hoteles, vendedores, páginas web como Trivago.com, etc. Todos los sujetos mencionados anteriormente, los datos de este dataset tienen un valor para ellos por lo que este análisis es importante e interesante para ellos. Mientras que si este mismo análisis se le muestra a un veterinario no le será de importancia, todo depende del sujeto que vea el análisis y el caso de uso que le dará.

[1] En cuanto a la ciencia de datos, las habilidades específicas son: extracción de la información, inteligencia del negocio, analítica del negocio y la analítica predictiva

Conjunto de datos

Variable	Descripción
hotel	Nombre del hotel (H1 = Resort Hotel or H2 = City Hotel)
is_canceled	Indica si la reservación es cancelada (1) o no (0)
lead_time	Número de días que transcurrieron entre la fecha de entrada de la reserva en el PMS y la fecha de llegada
arrival_date_year	Año (fecha) de llegada
arrval_date_month	Mes (fecha) de llegada
arrival_date_week_number	Número de la semana (fecha) de llegada en el año
arrival_date_day_of_month	El dia (fecha) de llegada
stays_in_weekend_nights	Número de noches que la persona se hospedó en dia de fin de semana (Sábado y Domingo)
stays_in_week_nights	Número de noches que la persona se hospedó en dia de semana (Lunes a Viernes)
adults	Número de adultos
children	Número de niños
babies	Numero de bebes
meal	Tipo de comida reservada. Se presenta por paquetes (comida cama y desayuno = BB, media pensión = HB y otros)
country	País de origen. Representado en el formato ISO

market_segment	Designación de segmento de mercado.
distribution_channel	Canal de distribución de reservas.
is_repeated_guest	Valor que indica si el nombre de la reserva fue de un huésped repetido (1) o no (0)
previous_cancellations	Número de reservas anteriores que fueron canceladas por el cliente antes de la reserva actual
previous_bookings_not_canceled	Número de reservas anteriores no canceladas por el cliente antes de la reserva actual
reserved_room_type	Código del tipo de habitación reservado.
assigned_room_type	Código del tipo de habitación asignada a la reserva
booking_changes	Número de cambios/modificaciones realizados a la reserva desde el momento en que se ingresó la reserva
deposit_type	indicación sobre si el cliente realizó un depósito para garantizar la reserva
agent	ID de la agencia de viajes de donde realizo la reserva
company	ID de la compañía o entidad donde realizo la reserva o el responsable pago la reserva.
days_in_waiting_list	Número de días que la reserva estuvo en lista de espera antes de que fuera confirmada al cliente
customer_type	Tipo de reserva de una de las cuatro categorías
adr	Tarifa diaria promedio según se define dividiendo la suma de todas las transacciones de alojamientos
required_car_parking_spaces	Número de plazas de aparcamiento requeridas por el cliente

total_of_special_requests	Número de solicitudes especiales realizadas por el cliente
reservation_status	Último estado de la reserva
reservation_status_date	Fecha en la que se estableció el último estado de la reserva.

Análisis de datos exploratorio

Para poder analizar los datos, primero se deben de cargar, inspeccionar y luego visualizar

Cargar datos

Para cargar los datos, se descargó el archivo csv desde Kaggle. Este viene en un archivo zip, por lo que antes deberá extraerse el archivo csv. Una vez extraído, este se pone en la misma carpeta que el proyecto de R designado a leer los datos. De ahí solo se realizan los comandos de lectura de archivos, en este caso read.csv debido al tipo de archivo. Se realiza el comando View para confirmar de que el data frame fue cargado correctamente.

Inspeccionar datos

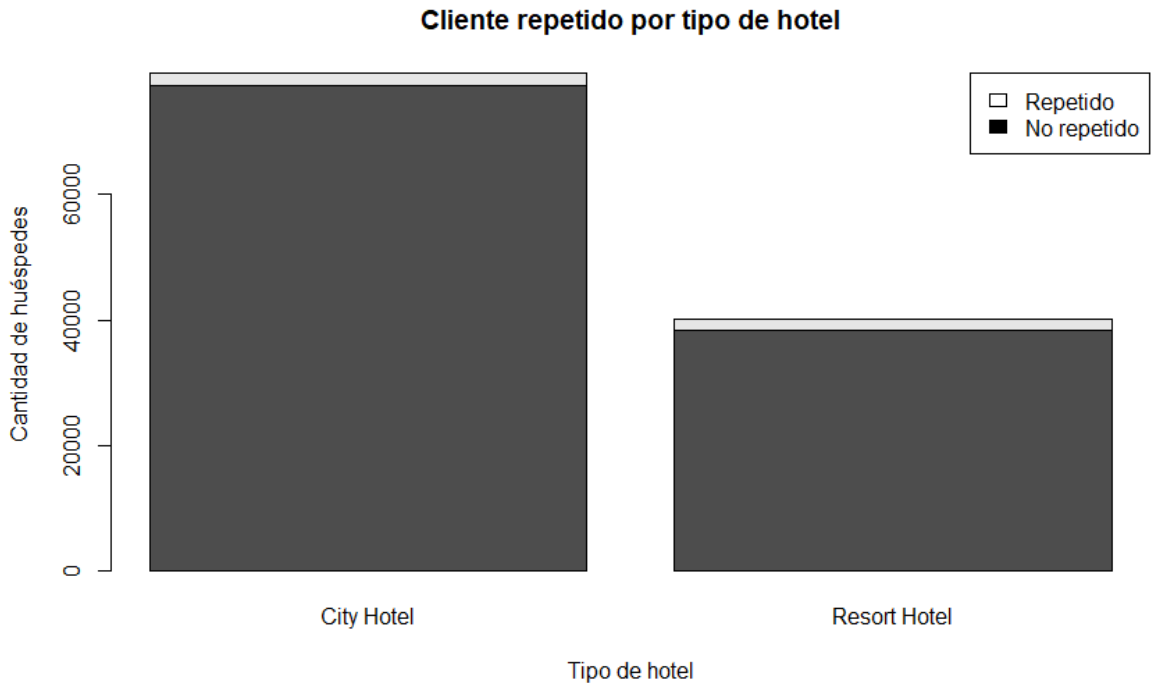
Primero usamos los comandos nrow y ncol para conocer las dimensiones del data frame adquirido, dando como resultado un data frame con 119390 filas y 32 columnas. Después se usa colnames para revisar el nombre de cada columna y ver que tipo de datos podrían hallarse en cada una. Los nombres de las columnas coinciden con las que están detalladas en la sección “Conjunto de datos”. Con str se puede ver el tipo de dato perteneciente a cada columna. En la mayoría de los casos el tipo de dato era int o char, con la única excepción en la columna “adr”, que tenía la estructura “num”. Utilizando summary, se puede ver un resumen de la información que contiene cada columna, y si una columna es de interés particular, ésta es revisada con table.

De la inspección de datos se puede detectar que los datos no están completamente preprocesados. Algunas tareas que deberían realizarse en el data frame serían cambiar variables binarias como “is_canceled” a tipo “logical” y cambiar los valores “NULL” a “NA” en “agent” y “company”.

Visualización de datos

A continuación, se harán diversos gráficos con los cuales se sacarán conclusiones posteriormente.

Clientes repetidos:



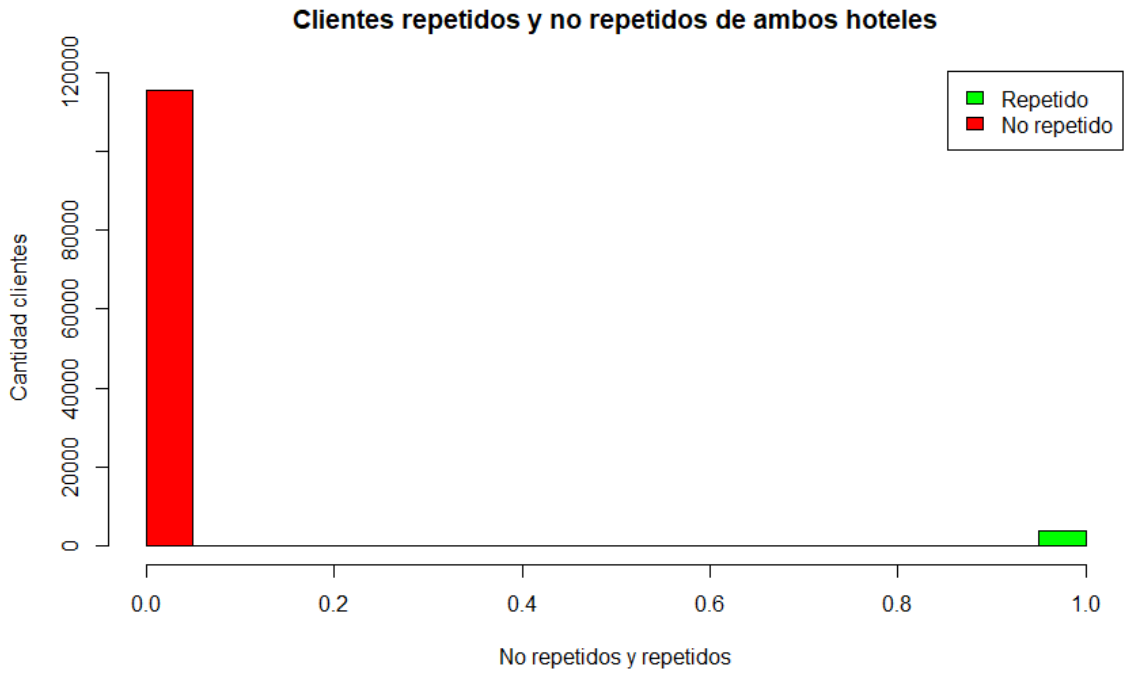
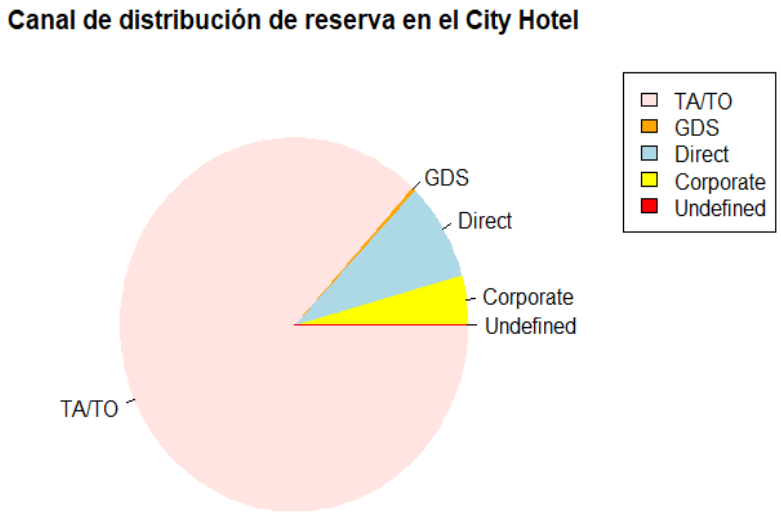
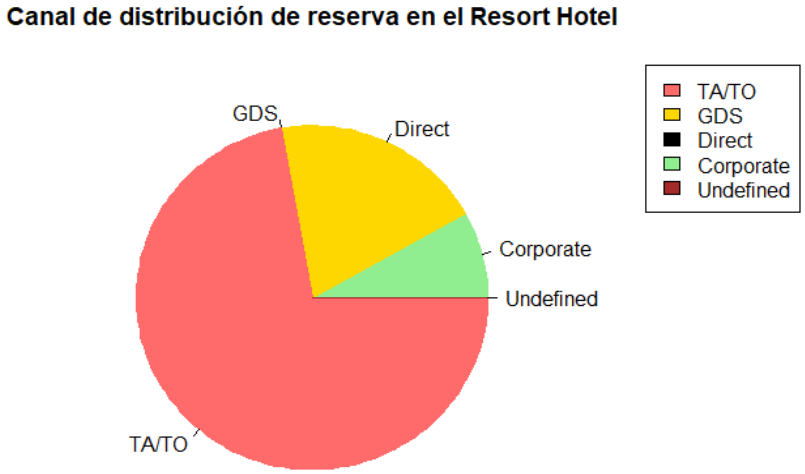
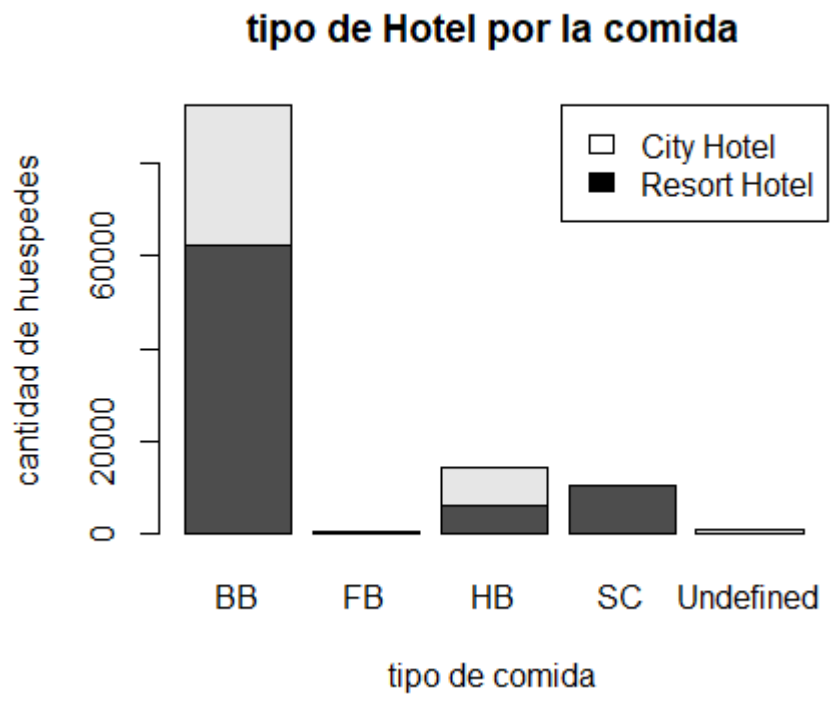


Gráfico que muestra el canal de reserva

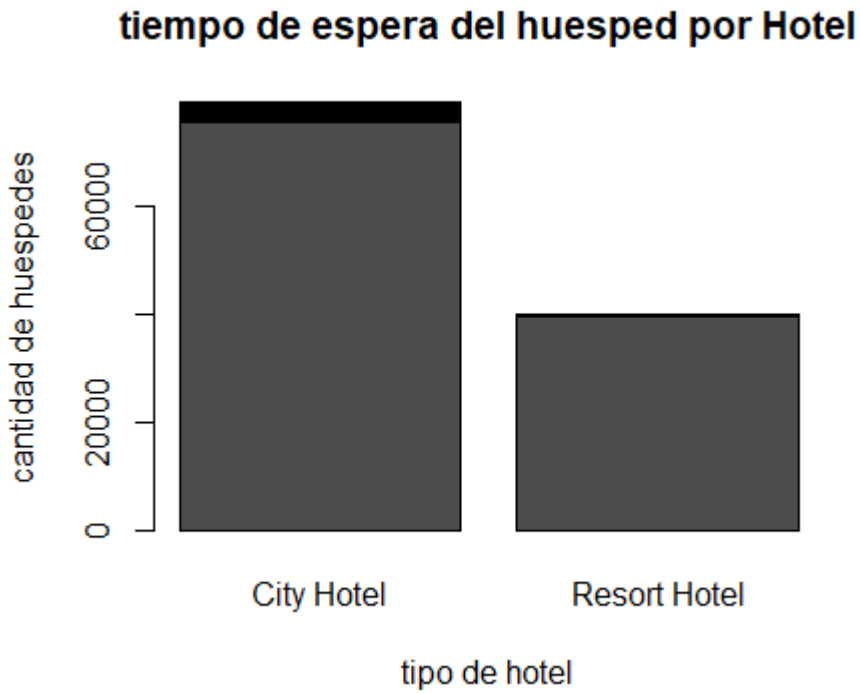




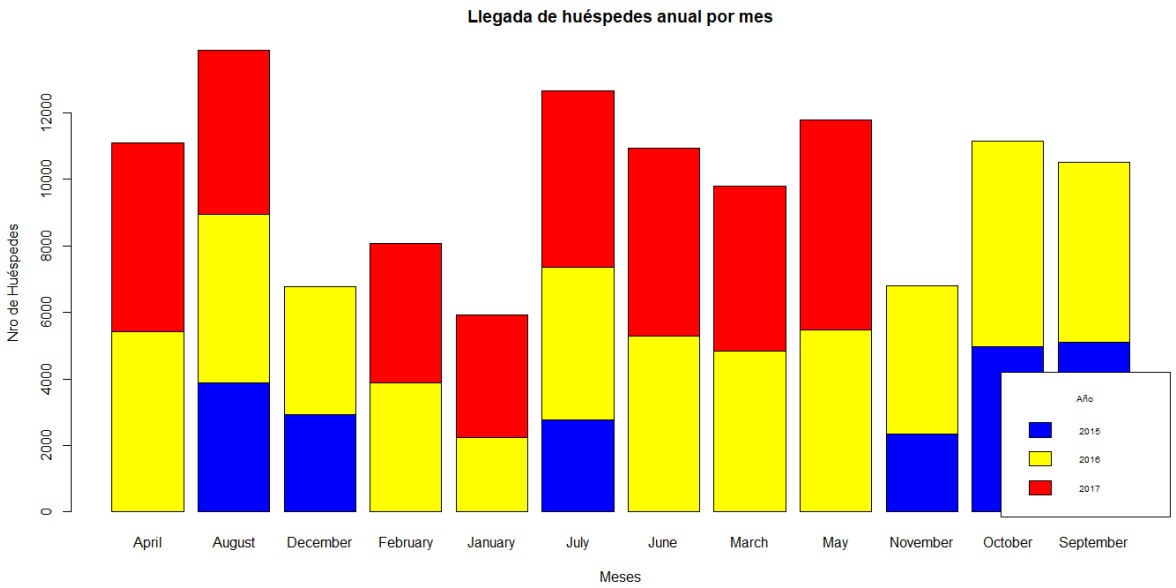
Comida pedida



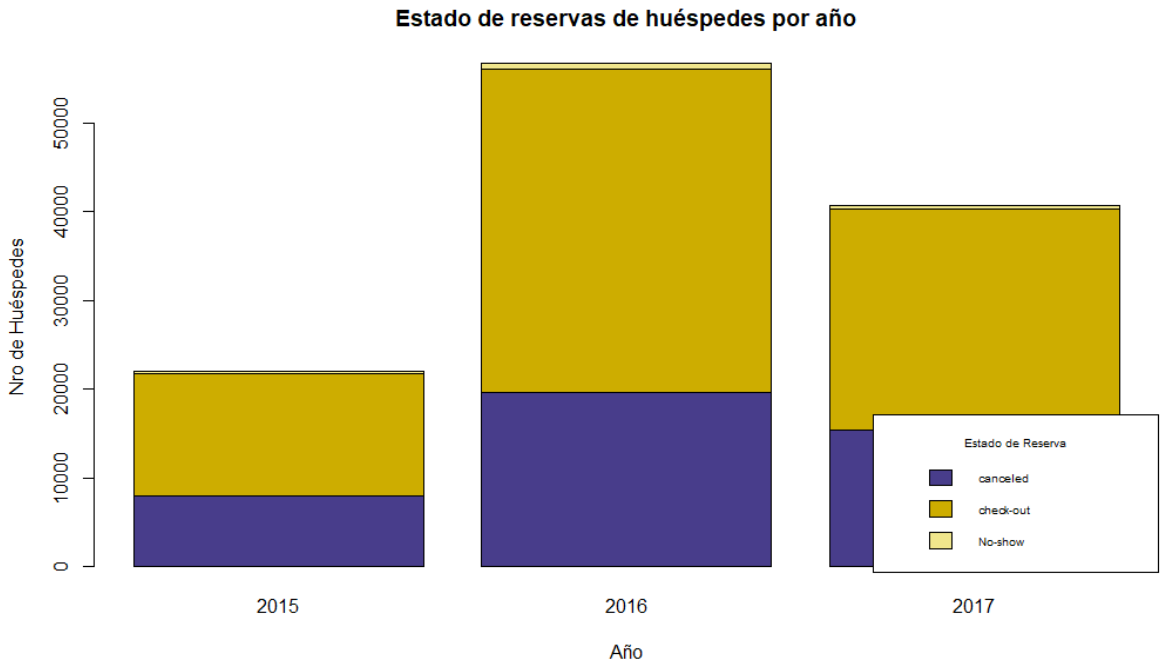
Tiempo de espera de un huésped



Llegada de huéspedes anual por mes



Estado de reservas de huéspedes por año



Conclusiones Preliminares

Luego del análisis de datos exploratorios, se pueden sacar conclusiones en base a las visualizaciones realizadas. Como se muestra en los gráficos de la sección de “Clientes repetidos”, se puede concluir que la gran mayoría de los clientes que van a ambos hoteles son nuevos y muy pocos ya han ido al hotel antes. En base a esto, nos dice que probablemente el marketing del hotel se enfoca en atraer a nuevos clientes en vez de crear clientes leales o recurrentes. Con esta información, podrían crear ofertas, membresías para que a las personas les sea más atractivo volver a ir al hotel.

De los gráficos en la sección “Canal de reserva”, se pueden realizar varias conclusiones. La primera es que la mayoría de clientes utiliza el canal de reserva TA/TO para hacer la reservación en el hotel. Con ello, se pueden realizar ofertas, paquetes, promociones especiales si el cliente hace la reserva con TA/TO. El hotel también puede realizar algún convenio o negocio en base a esta información con TA/TO. La segunda conclusión, es que muy pocos clientes utilizan GDS y muy pocos no definen el canal. Los canales de reserva podrían usar esta información para saber qué tan competitivos son o qué tan buenas sus estrategias son al ver cuántos usuarios tienen. Con ello pueden mejorar sus estrategias, realizar ofertas, mejorar el marketing, etc. La tercera conclusión es que GDS no tiene clientes en el Resort Hotel y en el City Hotel muy pocos, por lo que se puede suponer que el servicio no es muy bueno comparado con el de los otros canales de reserva.

Del gráfico en la sección “Comida pedida” podemos observar que el tipo de comida más pedida en ambos hoteles es el “BB” que es cama y desayuno por lo que se puede concluir que en los hoteles sus huéspedes tienen mayor actividad en el hotel por las mañanas que en otra hora del día, luego salen a realizar sus actividades diarias.

En la sección de “Tiempo de espera de un huésped” se visualiza que en el City hotel hay muchas personas mas esperando que en el Resort Hotel, por lo tanto se puede intuir que el City Hotel sus habitaciones se llenan más rápido y tienen mayor preferencia por este hotel por lo que también se puede decir que obtiene mayores ganancias.

Al revisar el gráfico de la sección “Llegada de huéspedes anual por mes”, se puede identificar a simple vista que el mes con mayor densidad de huéspedes es Agosto, mientras que el de menor número es Enero. El único año en el cual siempre hubieron huéspedes fue el 2016, que también es el año con mayor cantidad de huéspedes. De esta información se puede deducir que hay una mayor cantidad de huéspedes durante el segundo trimestre del año (Mayo, Junio, Julio, Agosto), contrastando mucho con los primeros y últimos meses del año (Noviembre, Diciembre, Enero, Febrero). Además, se nota que hubo un incremento de huéspedes entre el 2015 y 2016, pero estos disminuyeron entre 2016 y 2017, pese a no llegar a un punto tan bajo como el 2015.

Del gráfico de la sección “Estado de reservas de huéspedes por año” se observa que la mayoría de los huéspedes han realizado su check out, confirmando que sí estuvieron presentes en su hotel. Un menor grupo canceló sus reservas antes de ir al hotel, mientras que en muy pocos casos las personas no se han presentado pese a no cancelar su reserva. De la información se puede intuir que la mayoría de personas que deciden reservar una habitación en un hotel no se retractan y completan su estadía, mientras que por otro lado es muy raro que una persona no cancele su reserva y no se presente al hotel. Además, las proporciones de cada caso se han mantenido muy similares durante los tres años evaluados en el dataset.

Link del repositorio en github: https://github.com/OscarFloresP/Administracion

Bibliografía

Almeida. A, Antonio. N, Nunes. L (2019). *Hotel booking demand datasets*. Recuperado de https://www.sciencedirect.com/science/article/pii/S2352340918315191 [Consultado el 2 de mayo del 2021.]

Elsevier B.V, ScienceDirect (S/F). *Data in Brief*. Recuperado de https://www.sciencedirect.com/journal/data-in-brief [Consultado el 2 de mayo del 2021].

Mostipak. J (2020). *Hotel booking demand, From the paper: hotel booking demand datasets*. Recuperado de https://www.kaggle.com/jessemostipak/hotel-booking-demand?select=hotel_bookings.csv [Consultado el 2 de mayo del 2021].

ResearchGate (S/F). *Ana Maria De Almeida*. Recuperado de https://www.researchgate.net/profile/Ana-De-Almeida-6 [Consultado el 2 de mayo del 2021].

ResearchGate (S/F). *Nuno Antonio*. Recuperado de https://www.researchgate.net/profile/Nuno-Antonio [Consultado el 2 de mayo del 2021].

ResearchGate (S/F). *Luís Nunes*. Recuperado de https://www.researchgate.net/profile/Luis-Nunes-16 [Consultado el 2 de mayo del 2021].