REPORT HOTEL BOOKING DEMAND

Did you think anytime about the best time of the year to book a room in a hotel? Or what are the best countries to travel in each season? Have you ever wondered about how many people have cancelled the hotel booking in a year? In this report, I want to cover these questions and try to explain a basic EDA of the dataset used.

1. Reference

The dataset used contains information about several bookings for 2 types of hotel: city and resort and includes information about the date of the booking, number of days, number of people or even "special guests" (in the Appendix, there is an explanation of each variable). I need to say that personal information is not included due to privacy legislation.
The data was obtained in Kaggle.com and originally from the article *Hotel Booking Demand Datasets*, written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22 in February 2019. The data was downloaded and cleaned by Thomas Mock and Antoine Bichat for *#TidyTuesday* during the week of February 11th, 2020.

2. Business Needs

In terms of the business needs, we can think about different questions that are relevant to a hotel booking company (for example, Booking.com) and perform an analysis based on that. In my case, the 3 questions are:

- What is **the ratio of bookings** per season?
- What are **the main statistics about Season cancellation** by different characteristics (days, people, changes and lead time)?
- What are **the top 20 countries** with more bookings in **Warm Seasons** (Spring and Summer)?

3. Analysis

In order to take the analysis of the bookings, I did the following steps:

1. Start the PySpark environment

   Firstly, I start the VM and open a new Python notebook. As we studied, Spark is based on Scala and by using the PySpark API, we can access all the basic functions available for Spark. Also, I import a library called *pyspark.sql.session* to create later other SQL queries.

2. Import the dataset

Then, I need to upload my dataset, with the following options:

- InferSchema = True: to analyze the type of variable that each column has
- Header = True: to say that my dataset contains in the first row the name of the columns
- CSV: to specify the format of my dataset

Also, I have named my dataset *bookingsDF* because it is a Data Frame and contains information about bookings.

3. First "taste" of the data

a. Schema and Size

It shows the type of variable of each column (string or integer in my case) and also if nulls are allowed. Additionally, it appears the number of rows (119390).

b. Data Cleaning and Formatting

One of the problems that my dataset has is that the column month is written as a string ("January", "February"…), so I converted to an integer with the function *.withColumn* and *.when*. The objective is to create later an analysis and use the new column called *month*.

Also, I dropped some columns that are irrelevant or contain lot of null values. I did with a creation of a new variable with the columns to drop and then, using the *.drop* function and renaming my dataset.

c. Obtaining a random sample

Then, I used the function *.cache* to optimize the dataset and make the processing faster. Additionally, I get 2 random samples of the dataset to confirm that the changes and the data are in the correct format.

d. Entities, Metrics and Dimensions

This section is more informative to understand better the main parts of the analysis (entities), their measures (metrics) and related attributes (dimensions).

e. Column categorization

In order to do a further analysis, I divide (categorize) my variables into 2 main parts: timing (dates) and booking (hotel related).

4. Columns basic profiling

In this section, I divided the analysis based on the column categorization that I did before.

a. Timing

Firstly, I import again the libraries that I will use later for the analysis (it is not necessary because previously I did the same). Then, I obtain the summary statistics for the columns with the *.summary* function and also, I check for null and distinct values with the functions *count* & *.isNull* and *countDistinct* and grouping the columns in an array. At the end, I create a "matrix" to know the most and least frequency of 2 important columns: month and day of the month.

Some of the most interesting conclusions are:

- The majority of the records are in 2016, which makes sense because it's the only year that has all the months (in 2015, only 6 months and in 2017, only 8).
- The average month is June with 3 moths of S.D. This is the reason of why I will choose later the warm seasons (April to September) to obtain more insights. Also, it makes sense that the majority of bookings are with good weather and holidays.
- There are not any nulls in any of the columns, which it is always a good signal of the quality of the data and also, to make a more valid analysis.
- The number of distinct values makes sense and it is realistic: only 3 in years, 12 in months, 31 in days and 53 in weeks.
- In terms of the frequency, we can verify that the most frequent month is August, while the least is January. Also, we can observe that there are more people in a hotel in the middle of the month that at the end.

b. Booking

For the booking columns I did the same structure (summary, null values, distinct and frequency) but in this case, I split in 2 parts because there are more variables. The main conclusions are:

- The average night in the weekend is 1 but with an S.D. of 1 night, which makes sense. There is an outlier (19), which is non-significant because it represents only 0.000001% of the dataset.
- In terms of week nights, the average is 2.5 and most of the people stays a maximum of 3 nights.
- The "persona" of the dataset is 2 adults without children nor babies.
- No nulls in any of the columns.
- An interesting fact is the average number of days in the waiting list, which is 2. This means that most of the people wait 2 days to confirm the reservation and the companies take advantage of that situation (advertising, cookies).

5. Business questions

In order to answer the business questions, first I need to divide the dataset in each season, using the following criteria: Summer (July-September), Autumn (October-December), Winter (January-March) and Spring (April-June). With the creation of a new name of the dataset and using *.withColumn* & *.when* functions, I categorize each month into a season.
Now, we can start to analyze each question:

a. Ratio of bookings per season

For this one I did a very similar query than in SQL (*SELECT SEASON FROM SEASONCATEGORIZATIONDF GROUP BY SEASON*). Select and group by season and then, counting how many bookings were in each season. Finally, I create the ratio of the number of bookings per season over the total bookings (I did previously with the *.count* function). Finally, I select the season, the number of rooms per season and the ratio (I created an alias) rounded to 2 decimals.

b. Season cancellation statistics

First able, I imported some functions from SQL and also to change to integer 1 column. Then, I rename my dataset (bookingCancellationDF) taking into account only the cancelled bookings, converting children (it was a string) into an integer and selecting the categories that I will use for the statistics. Also, I include the *.cache* function to optimize the dataset.
Finally, I did the individual analysis of each column (average, min, max and S.D.), groping and ordering by season and including an alias for each parameter.

c. Top 20 countries with more bookings in warm seasons

To address the question, the first thing is to create 2 data frames: one with the total bookings per country (grouping by country) and the other with the bookings in the warm seasons (spring and summer, grouping by country and season). Then, I did a join of both tables with the column country (it is in both DF), with at least 2 bookings in each season (if I allow all the bookings, the majority of countries with only 1 booking will be in the top 20 and it is not realistic) and with the creation of the ratio between the bookings in the warm seasons and the total bookings to know the preferences of the customers per season and country.
With the join table, I can display the top 20 countries with the highest booking ratio in warm seasons. Finally, I show the top 20 countries with the highest booking ratio order by the season summer and using the *pivot* function.

4. Conclusions

The main insights of the hotel booking analysis are:

- **6 out of 10** bookings were in **Summer and Spring**.
- The **highest cancellation in summer** was a period of **20 days**.
- Do not go to **Georgia, Gabon, Andorra and Vietnam** in **spring** (any booking in that period).
- **Barbados, Faroe Islands and Bolivia** are the ideal destinations to go in **warm seasons**.

5. Appendix

VARIABLE	DESCRIPTION
hotel	Type of 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 (Property Management System) 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 weeknights (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	Type of meal booked. 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 to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons
booking_changes	Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation
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	Type of booking, assuming one of four categories: 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 the booking was canceled or when did the customer checked-out of the hotel