I – Data selection

1. The dataset

The dataset selected for analysis is the 'Hotel Booking Demand' dataset published by Antonio et al. (2018). Source: https://www.sciencedirect.com/science/article/pii/S2352340918315191

This dataset comprehends real-life booking data at two hotels in Portugal, namely 'City Hotel' (H1) and 'Resort Hotel' (H2), from the 1st of July 2015 to the 31st of August 2017. This includes both reservations that were effectively arrived and those being canceled.

2. Dataset description

The dataset contains 40,060 observations of H1 and 79,330 observations of H2 (119390 observations in total). Each row represents a hotel reservation.
The dataset has 33 columns with description as following:

#	Column name	Data type	Description
1	Hotel	Str	Hotel (H1 = Resort Hotel or H2 = City Hotel)
2	is_canceled	Int	Value indicating if the booking was canceled (1) or not (0)
3	lead_time	Int	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
4	arrival_date_year	Int	Year of arrival date
5	arrival_date_month	Str	Month of arrival date
6	arrival_date_week_number	Int	Week number of year for arrival date
7	arrival_date_day_of_month	Int	Day of arrival date
8	stays_in_weekend_nights	Int	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
9	stays_in_week_nights	Int	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
10	sdults	Int	Number of adults
11	children	Int	Number of children
12	babies	Int	Number of babies
13	meal	Str	Type of meal booked. Categories are presented in standard hospitality meal packages: Undefined/SC – no meal
14	country	Str	Country of origin. Categories are represented in the ISO 3155–3:2013 format
15	market_segment	Str	Market segment designation. In categories, the term "TA" means "Travel Agents" and "TO" means "Tour Operators"

16	distribution_channel	Str	Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”
17	is_repeated_guest	Int	Value indicating if the booking name was from a repeated guest (1) or not (0)
18	previous_cancellations	Int	Number of previous bookings that were cancelled by the customer prior to the current booking
19	previous_booking_	Int	Number of previous bookings not cancelled by the customer prior to the current booking
20	reserved_room_type	Str	Code of room type reserved. Code is presented instead of designation for anonymity reasons.
21	assigned_room_type	Str	Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due
22	booking_changes	Str	Number of changes/amendments made to the booking from the moment the booking was entered on the PMS
23	deposit_type	Str	Indication on if the customer made a deposit to guarantee the booking.
24	agent	Str	ID of the travel agency that made the booking
25	company	Str	ID of the company/entity that made the booking or responsible for paying the booking.
26	days_in_waiting_list	Int	Number of days the booking was in the waiting list before it was confirmed to the customer
27	customer_type	Str	Type of booking
28	adr	Int	Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights
29	required_car_parking_spaces	Int	Number of car parking spaces required by the customer
30	total_of_special_requests	Int	Number of special requests made by the customer (e.g. twin bed or high floor)
31	reservation_status	Str	Reservation last status, assuming one of three categories: Canceled – booking was canceled by the customer; Check-Out
32	reservation_status_update	datetime	Date at which the last status was set.

II – Approach and methodology

1. Why this dataset?

The main reason for choosing this dataset is because this dataset is in the hotel and tourism industry.

The analysis of the dataset can produce relevant insights and recommendations for Vinpearl, given the same nature of business.

2. Observation process

Step 1: Data preprocessing

- Examine the shape and size of the dataset, data types of the columns.
- Transform data where necessary.
- Detect and handle missing values.

Step 2: Explanatory data analysis

- Explore every single data column
- Observe the statistical distribution of numerical columns or count of values in categorical columns.
- Perform summary analysis for the whole dataset and by groups.
- Apply statistical methods to further examine the variable distributions and correlation between variables.
- Build visualizations to facilitate analysis.
- Reconsider outputs of analysis if they align with common business sense.

Step 3: Recommendations

- Provide suggestions based on insights from EDA, experience and knowledge from market research.