Task 2 (70 marks):

You are given a CSV file (“*hotel_bookings.csv*”). This data set contains booking information for a city hotel and a resort hotel and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things.

You will have to write a **PySpark program** in a **Databricks Notebook** to meet all the requirements as below**:**

1. Read the given CSV file. (2 marks)

2. Perform data validation that includes:

   a. Preview the first five rows of records. (2 marks)
   b. Checking any inappropriate column type from data schema (e.g. is there any column should be a number instead of a string) (2 marks)
   c. Checking any missing values or null values (2 marks)
   d. Checking any outliers (2 marks)

   (Write your observation on the result in Python comments)

3. Perform necessary data cleaning on those missing values, null values and outliers. (5 marks)

4. By applying appropriate data filtering, data transformation & data visualization (e.g. bar chart, pie chart, line chart, etc) approach, perform the exploratory data analysis to address the following question:
   a. Where do the hotel guests come from? (5 Marks)
   b. What is the cost of stay for a guest for one night? (5 Marks)
   c. Which room type is most popular? (5 Marks)
   d. What are the peak season months? Is there any seasonality pattern? (5 Marks)

e. What is the common length of stay at the hotels? (5 Marks)
f. What are the compositions of bookings by market segment? (5 Marks)
g. Which month have the highest number of cancelations? (5 Marks)
h. Which type of room suffer more from cancellation? (5 Marks)
i. Do people from a specific country tend to cancel their booking more than the others? (5 Marks)
j. Apart from the booking month, room type and country of origin, what are the other features that are possibly correlated with booking cancellation? Give at least two more examples. (10 Marks)

For all the questions above, please write your answers/observations from the graphs in Python comments.

Important Notes:

Please study the given dataset carefully and determine the target columns required for your study. You will need to remove unwanted columns, perform necessary data filtering & aggregation (e.g. group by category, average by category, etc) wherever necessary for your analysis to address the questions above.