Project Summary

***Instructions*:**

1. **Data Understanding and Pre-processing:**
    1. Load the data-set and provide a summary of its structure (number of rows, columns, data types, and missing values).
    2. Perform data cleaning:
        1. Handle missing values appropriately.
        2. Convert data types if necessary.
        3. Create any new features you think might be useful.
2. **Exploratory Data Analysis (EDA):**

    1. Provide a detailed EDA including visualizations. Focus on understanding booking trends, customer demographics, and cancellation patterns.
    2. Use visualizations to highlight key insights, such as:
        1. Seasonality in bookings.
        2. Distribution of stays across different hotel types.
        3. Average daily rate (ADR) trends.
        4. Cancellation rates and factors affecting cancellations.
3. **Hypothesis Testing:**

    1. Formulate and test at least two hypotheses related to the data. For example:
        1. "Customers booking more than 6 months in advance are more likely to cancel."
        2. "Weekday bookings have a higher average daily rate than weekend bookings."
    2. Use appropriate statistical tests to validate these hypotheses.
4. **Predictive Modeling:**

    1. Build a predictive model to forecast hotel cancellations. Include the following steps:
        1. Select appropriate features.
        2. Split the data into training and test sets.
        3. Train at least two different models (e.g., Logistic Regression, Random Forest).
        4. Evaluate model performance using relevant metrics (accuracy, precision, recall, F1-score).
        5. Discuss any improvements you would recommend for the model.
5. **Operational Insights:**

    1. Provide actionable insights and recommendations for hotel management based on your analysis. Consider aspects like

pricing strategies, customer segmentation, and marketing focus.

6. **Report and Presentation:**

    1. Summarize your findings in a written report.
    2. Create a presentation highlighting key insights, methodologies, and recommendations. Ensure it is clear and concise, suitable for stakeholders with varying levels of technical expertise.

What is inside my Jupyter notebook?(Let's take a look)

1. Data understanding and Preprocessing

a) Load the data-set and provide a summary of its structure (number of rows, columns, data types, and missing values).

Importing important libraries for Data Preprocessing:-

```
In [1]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns
```

- This code is used to import all the necessary libraries which are used in data preprocessing and data analysis.

Getting Information of the DataFrame:-

In [3]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 32 columns):
 #   Column                          Non-Null Count   Dtype  
---  ------                          --------------   -----  
 0   hotel                           119390 non-null  object 
 1   is_canceled                     119390 non-null  int64  
 2   lead_time                       119390 non-null  int64  
 3   arrival_date_year               119390 non-null  int64  
 4   arrival_date_month              119390 non-null  object 
 5   arrival_date_week_number        119390 non-null  int64  
 6   arrival_date_day_of_month       119390 non-null  int64  
 7   stays_in_weekend_nights         119390 non-null  int64  
 8   stays_in_week_nights            119390 non-null  int64  
 9   adults                          119390 non-null  int64  
 10  children                        119386 non-null  float64
 11  babies                          119390 non-null  int64  
 12  meal                            119390 non-null  object 
 13  country                         118902 non-null  object 
 14  market_segment                  119390 non-null  object 
 15  distribution_channel            119390 non-null  object 
 16  is_repeated_guest               119390 non-null  int64  
 17  previous_cancellations          119390 non-null  int64  
 18  previous_bookings_not_canceled  119390 non-null  int64  
 19  reserved_room_type              119390 non-null  object 
 20  assigned_room_type              119390 non-null  object 
 21  booking_changes                 119390 non-null  int64  
 22  deposit_type                    119390 non-null  object 
 23  agent                           103050 non-null  float64
 24  company                         6797 non-null    float64
 25  days_in_waiting_list            119390 non-null  int64  
 26  customer_type                   119390 non-null  object 
 27  adr                             119390 non-null  float64
 28  required_car_parking_spaces     119390 non-null  int64  
 29  total_of_special_requests       119390 non-null  int64  
 30  reservation_status              119390 non-null  object 
 31  reservation_status_date         119390 non-null  object 
dtypes: float64(4), int64(16), object(12)
memory usage: 29.1+ MB
```

Summary of the DataFrame:

Summary of DataFrame

- **Number of Rows:** 119,390
- **Number of Columns:** 32

Data Types

- **Integer Columns (`int64`):** 16 columns
- **Float Columns (`float64`):** 4 columns
- **Object Columns (`object`):** 12 columns

Missing Values

- `children` : 4 missing values
- `country` : 488 missing values
- `agent` : 16,340 missing values
- `company` : 112,593 missing values

This summary provides an overview of the structure and missing data within your DataFrame, helping to identify areas that may require data cleaning or handling.

- This code shows the total number of **Rows** and **Columns** along with the data types of columns.
- Using this code we can also see if there are any **missing values** available in any of the columns or not.

b) Performing Data Cleaning

Handling missing values:-

"children":-

- Given below is the children count in **"children"** column.

```
children
0.0     110796
1.0       4861
2.0       3652
3.0         76
10.0         1
Name: count, dtype: int64
```

- There were only **4 missing values** in **"children"** column. So, we **replaced the missing values with "0"** as zero being the highest occurred number.

"country":-

- Given below is the countries count in **"country"** column and it's categories.

```
country
PRT    48590
GBR    12129
FRA    10415
ESP     8568
DEU     7287
       ...
DJI        1
BWA        1
HND        1
VGB        1
NAM        1
Name: count, Length: 177, dtype: int64
```

- There were **488 missing values** in **"country"** column. So, we **replaced the missing values with "unknown"**.

"agent":-

- There were **16340 missing values** in the column **"agent"**.
- Since, it's data type is **'int'**. We **replaced the missing values with the mean**.

"company":-

- There were **112593 missing values** in **"company"** column.
- Therefore, we will remove this column because replacing the missing values will not be useful.

Handling categorical Values

- Given below are the categorical columns present in dataset.

1. hotel

- **City Hotel**: 79,330 bookings
- **Resort Hotel**: 40,060 bookings

2. arrival_date_month

- **August**: 13,877 bookings
- **July**: 12,661 bookings
- **May**: 11,791 bookings
- **October**: 11,160 bookings
- **April**: 11,089 bookings
- **June**: 10,939 bookings
- **September**: 10,508 bookings
- **March**: 9,794 bookings
- **February**: 8,068 bookings
- **November**: 6,794 bookings
- **December**: 6,780 bookings
- **January**: 5,929 bookings

3. meal

- **BB (Bed & Breakfast)**: 92,310 bookings
- **HB (Half Board)**: 14,463 bookings
- **SC (Self Catering)**: 10,650 bookings
- **Undefined**: 1,169 bookings
- **FB (Full Board)**: 798 bookings

4. country

- **PRT (Portugal)**: 48,590 bookings
- **Others**: 18,590 bookings
- **GBR (United Kingdom)**: 12,129 bookings
- **FRA (France)**: 10,415 bookings
- **ESP (Spain)**: 8,568 bookings
- **DEU (Germany)**: 7,287 bookings
- **ITA (Italy)**: 3,766 bookings
- **IRL (Ireland)**: 3,375 bookings
- **BEL (Belgium)**: 2,342 bookings
- **BRA (Brazil)**: 2,224 bookings
- **NLD (Netherlands)**: 2,104 bookings

5. market_segment

- **Online TA**: 56,477 bookings
- **Offline TA/TO**: 24,219 bookings
- **Groups**: 19,811 bookings
- **Direct**: 12,606 bookings
- **Corporate**: 5,295 bookings
- **Complementary**: 743 bookings
- **Aviation**: 237 bookings
- **Undefined**: 2 bookings

6. distribution_channel

- **TA/TO**: 97,870 bookings
- **Direct**: 14,645 bookings
- **Corporate**: 6,677 bookings
- **GDS**: 193 bookings
- **Undefined**: 5 bookings

7. reserved_room_type

- **A**: 85,994 bookings
- **D**: 19,201 bookings
- **E**: 6,535 bookings
- **F**: 2,897 bookings
- **G**: 2,094 bookings
- **B**: 1,118 bookings
- **C**: 932 bookings
- **H**: 601 bookings
- **P**: 12 bookings
- **L**: 6 bookings

8. assigned_room_type

- **A**: 74,053 bookings
- **D**: 25,322 bookings
- **E**: 7,806 bookings
- **F**: 3,751 bookings
- **G**: 2,553 bookings
- **C**: 2,375 bookings
- **B**: 2,163 bookings
- **H**: 712 bookings
- **I**: 363 bookings
- **K**: 279 bookings
- **P**: 12 bookings
- **L**: 1 booking

9. deposit_type

- **No Deposit**: 104,641 bookings
- **Non Refund**: 14,587 bookings
- **Refundable**: 162 bookings

10. customer_type

- **Transient**: 89,613 bookings
- **Transient-Party**: 25,124 bookings
- **Contract**: 4,076 bookings
- **Group**: 577 bookings

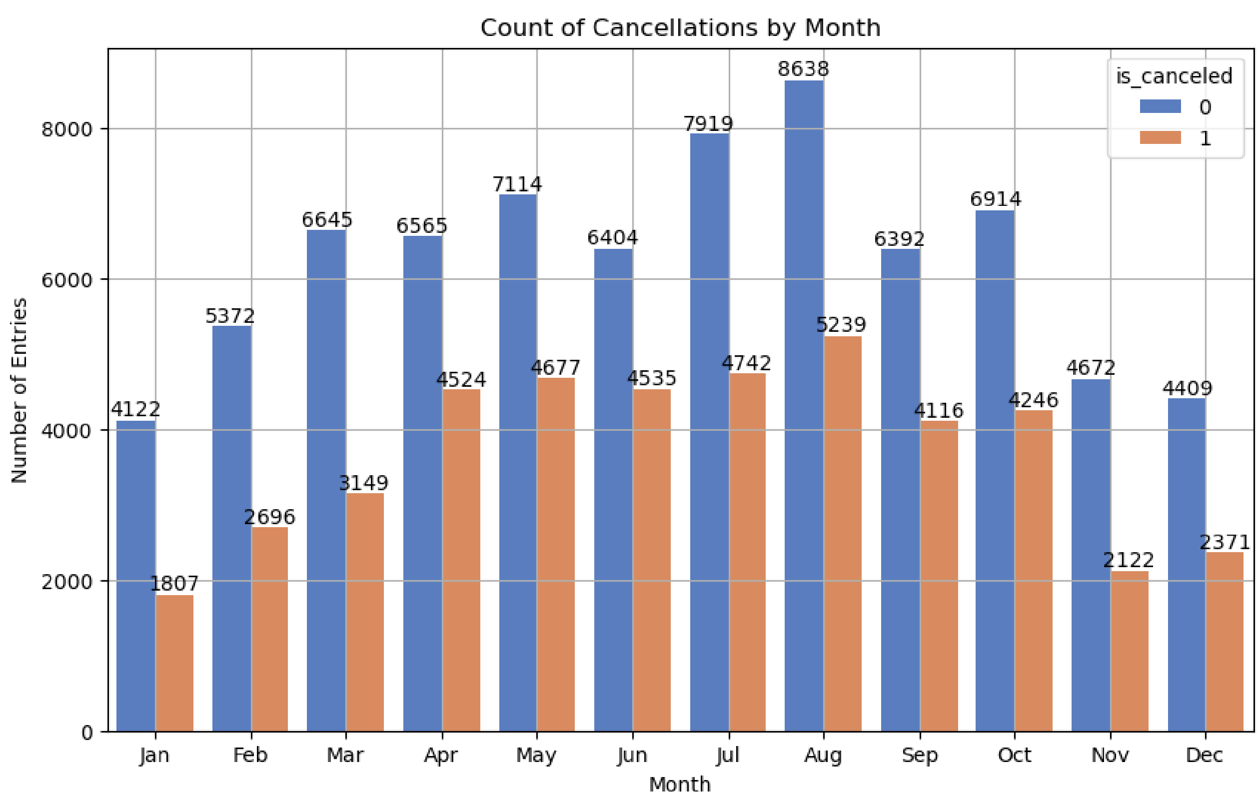
11. reservation_status

- **Check-Out**: 75,166 bookings
- **Canceled**: 43,017 bookings
- **No-Show**: 1,207 bookings

- We converted all the possible categorical columns into numeric columns.

2. Exploratory Data Analysis (EDA)

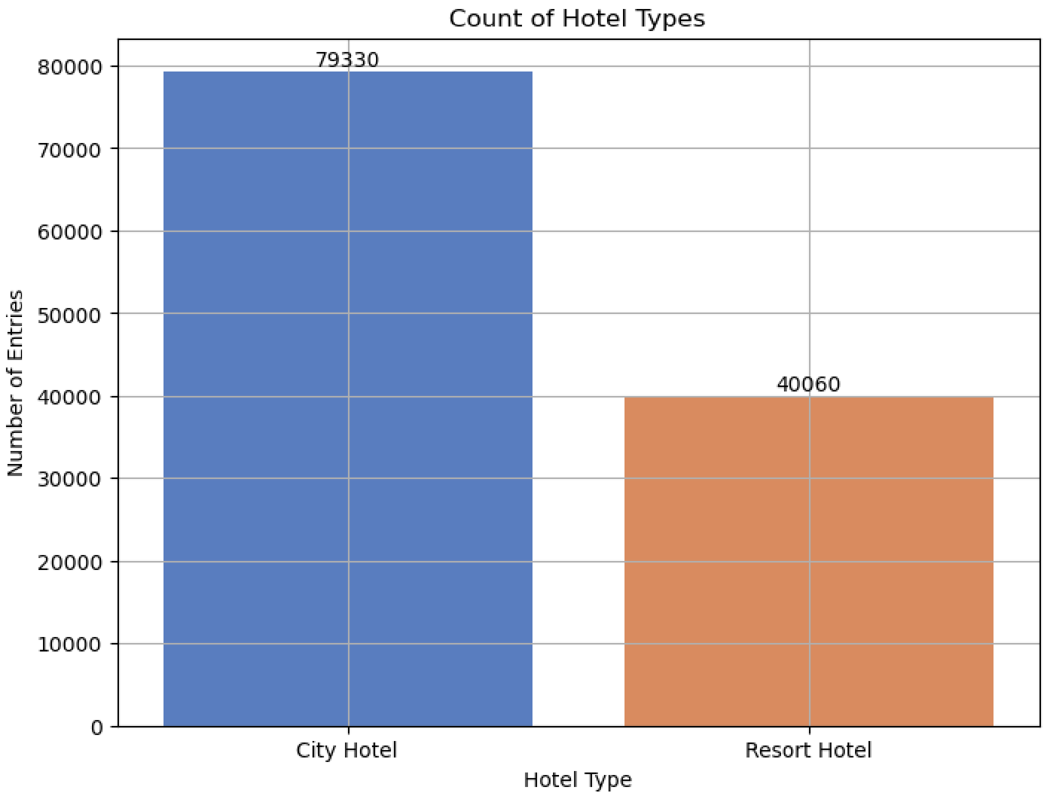
a) Seasonality in bookings.



- This plot shows **top 4 months with the highest bookings**.

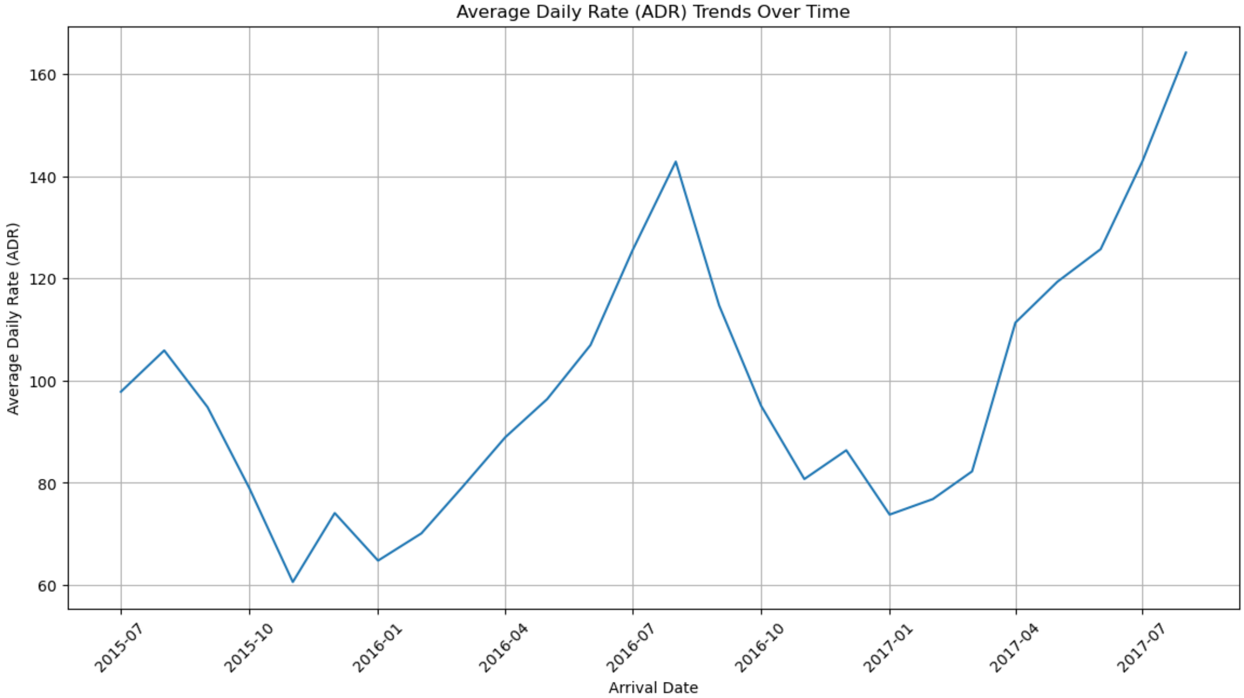
- **August** - 8638 bookings
- **July** - 7919 bookings
- **May** - 7114 bookings
- **October** - 6914 bookings

b) Distribution of stays across different hotel types.



- This plot shows the bookings of Hotel.
  - Mostly hotel bookings were **“City Hotel”** type
  - **City Hotel** - 79330 bookings
  - **Resort Hotel** - 40060 bookings

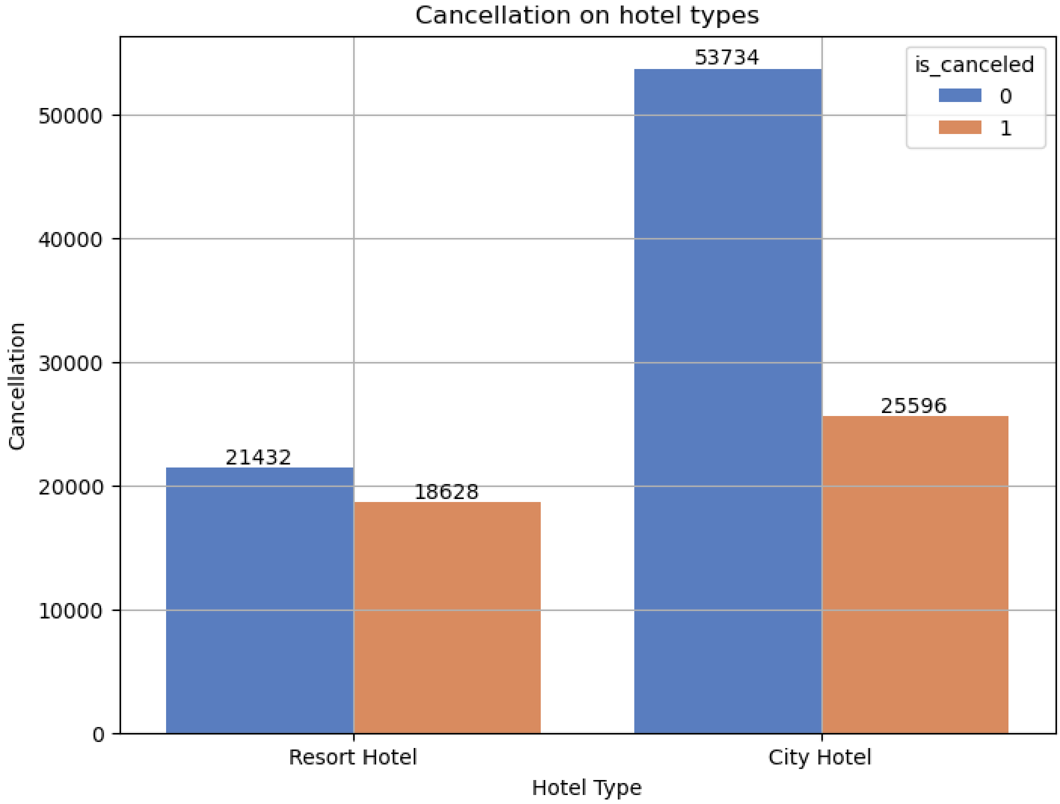
c) Average Daily Trends (ADR)



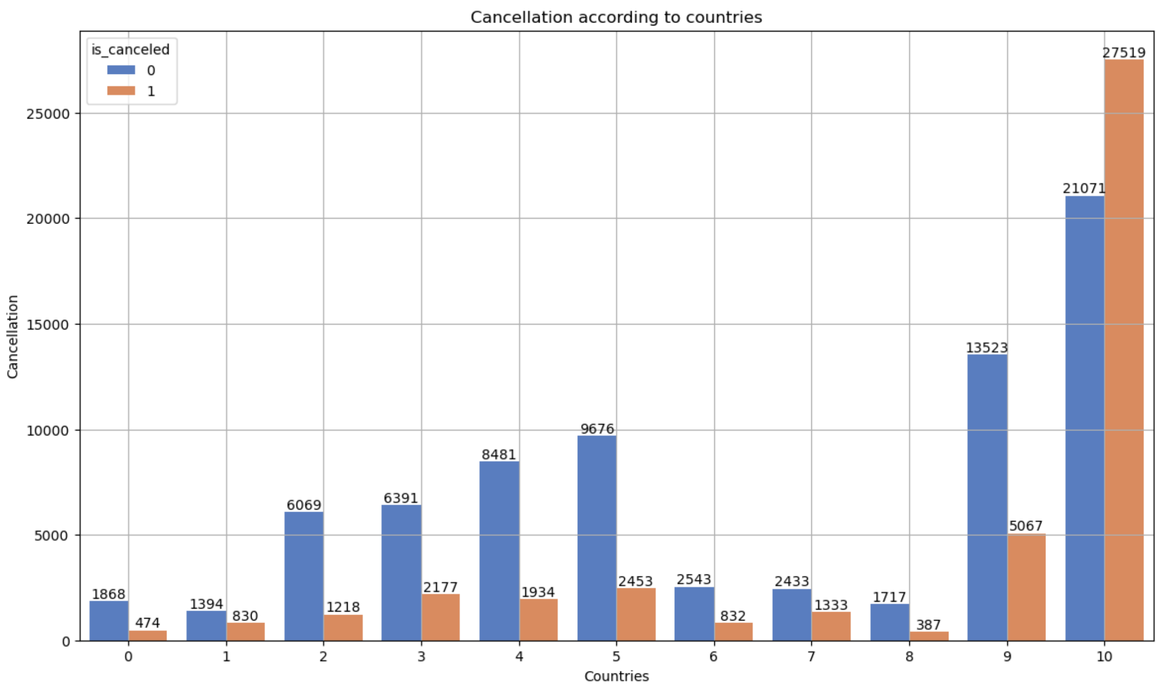
- This line-plot shows the ADR spikes from **"January" (07)** to **"August" (08)** of every year.
- The ADR dips is seen on every from **"August" (09)**to **"November" (11)** of each year.
- The Highest ADR recorded was on **"2017-07" (July 2017)**
- The Lowest ADR recorded was on **"2015-11" (November 2015)**

d) Cancellation rates and factors affecting cancellations

- This plot shows that the **44224 bookings** were canceled out of **119390 bookings**.



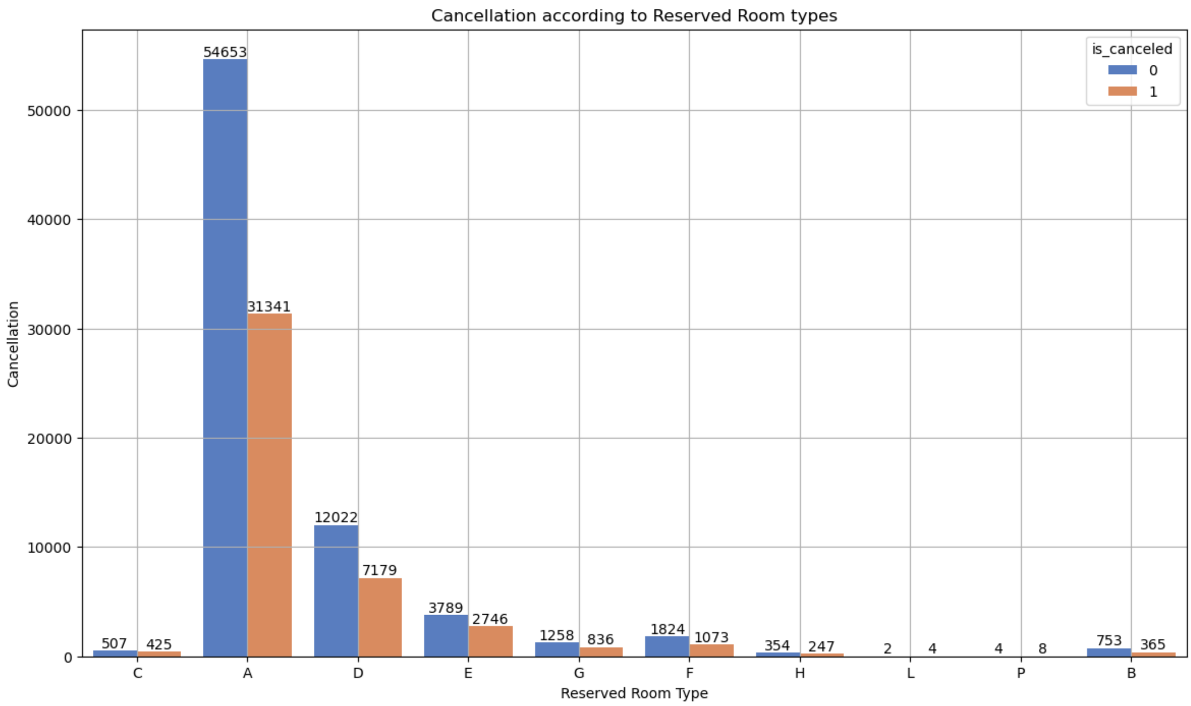
- This plot shows that **most of the bookings were of "City Hotel"**.
- Since **"Resort Hotel"** has **less bookings as compared to "City Hotel"**. But the **cancellation rate is higher as compared to "Resort Hotel"**.



```
Labels for the countries:
```

Country	Label
BEL	0
BRA	1
DEU	2
ESP	3
FRA	4
GBR	5
IRL	6
ITA	7
NLD	8
Others	9
PRT	10

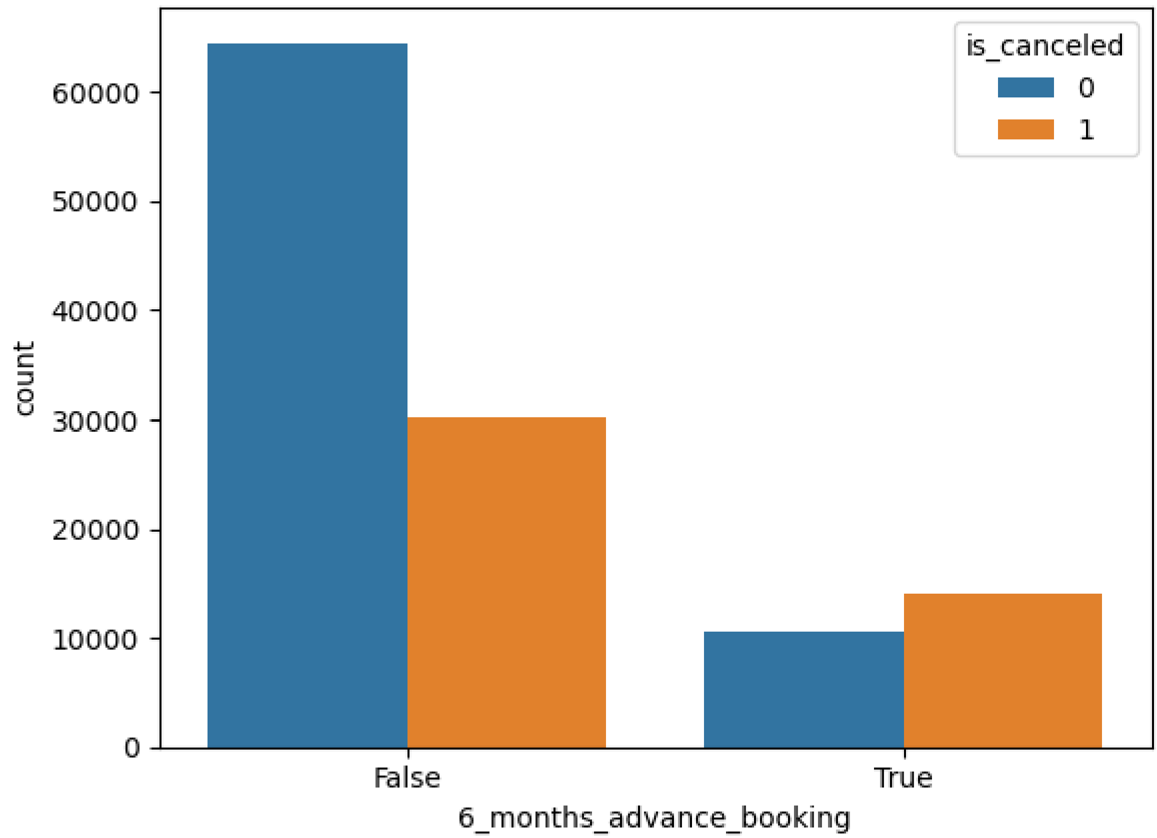
- The most of the bookings were from country code **"PRT" (Portugal)**.
- **"PRT" (Portugal)** is the only country which has **higher amount of cancellations (27519) even higher than non-cancellations (21071)**.
- Rest of the country has lower rate of cancellations.



- The **most of the bookings** were from **reserved room type (A)**.
- The **most cancellation** were also from **reserved room type (A) (31341 out of 85994)**.

3. Hypothesis Testing

a) Customers booking more than 6 months in advance are more likely to cancel.



- As we can see in the above plot, the **"cancellations rate" in "6 months advance booking"** is **higher than the "non cancellations rate"** as **compared to the cancellations of "bookings before 6 months"**.

Conclusion:-

- This proved that the **hypothesis is TRUE**.

b) Weekday bookings have a higher average daily rate than weekend bookings

```
booking_type
Weekdays    103.167653
Weekends     97.519810
Name: adr, dtype: float64
```

- This shows that **weekdays bookings have higher ADR** as compared to weekends.

The t-test results are as follows:

- t-statistic: 17.37
- p-value: 2.30e-67 (basically 0)

- Since the **P-value is less than 0.05**. Therefore, we can reject the null hypothesis.
- This supports the hypothesis that **"Weekday bookings have a higher average daily rate than weekend bookings."**

4. Predictive Modeling

a) Select appropriate features.

- Given below is the correlations of the dependent variables with respect to independent variables.

```
is_canceled                      1.000000
res_status_Canceled              0.978435
type_Non Refund                  0.481457
lead_time                        0.293123
country_encoded                  0.271231
6_months_advance_booking         0.211148
distrib_chnl_TA/TO               0.175944
cust_type_Transient              0.133084
previous_cancellations           0.110133
adults                           0.060017
market_segment_encoded           0.059338
days_in_waiting_list             0.054186
adr                              0.047557
meal_FB                          0.038828
stays_in_week_nights             0.024765
arrival_date                     0.023826
arrival_date_year                0.016660
meal_BB                          0.013124
arrival_date_month               0.011022
arrival_date_week_number         0.008148
children                         0.005036
meal_SC                          0.001282
year_2015                       -0.000254
stays_in_weekend_nights         -0.001791
arrival_date_day_of_month       -0.006130
distrib_chnl_GDS                -0.014891
meal_HB                         -0.019845
year_2016                       -0.023208
cust_type_Contract              -0.023670
babies                          -0.032491
previous_bookings_not_canceled  -0.057358
reserved_room_type_encoded      -0.061282
distrib_chnl_Corporate          -0.075428
agent                           -0.077992
is_repeated_guest               -0.084793
cust_type_Transient-Party       -0.124135
Resort Hotel                    -0.136531
booking_changes                 -0.144381
distrib_chnl_Direct             -0.151620
reservation_status_date         -0.165057
assign_room_type_encoded        -0.176028
required_car_parking_spaces     -0.195498
total_of_special_requests       -0.234658
type_No Deposit                 -0.477911
res_status_Check-Out            -1.000000
Name: is_canceled, dtype: float64
```

- Given below are the features that we will be taking while training the model because of the high correlations.

- res_status_Canceled - `0.978435`
- type_Non Refund - `0.481457`
- lead_time - `0.293123`
- country_encoded - `0.271231`
- 6_months_advance_booking - `0.211148`
- distrib_chnl_TA/TO - `0.175944`
- cust_type_Transient - `0.133084`
- booking_changes - `-0.144381`
- distrib_chnl_Direct - `-0.151620`
- reservation_status_date - `-0.165057`
- assign_room_type_encoded - `-0.176028`
- required_car_parking_spaces - `-0.195498`
- total_of_special_requests - `-0.234658`
- type_No Deposit - `-0.477911`
- res_status_Check-Out - `-1.000000`

Model Training

Random Forest Classifier

Model Evaluation

```
Confusion Matrix:
 [[16993  1361]
 [ 3186  2337]]

Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.93      0.88     18354
           1       0.63      0.42      0.51      5523

    accuracy                           0.81     23877
   macro avg       0.74      0.67      0.69     23877
weighted avg       0.79      0.81      0.80     23877


ROC AUC Score: 0.7616414322615062

Accuracy: 0.8095656908321816
```

Model Summary

Based on the results from your Random Forest Classifier model, here's a summary of the key evaluation metrics:

1. Accuracy

- **Value:** 0.81 (or 81%)

- **Explanation:** In this case, the model correctly predicted whether a booking would be canceled or not 81% of the time.

2. Precision

- **Class 0 (Not Canceled):** 0.84
- **Class 1 (Canceled):** 0.63
- **Explanation:** A high precision for Class 0 means that when the model predicts a booking will not be canceled, it is correct 84% of the time. For Class 1, the model is correct 63% of the time when it predicts a cancellation. This indicates that the model has more confidence in predicting non-cancellations than cancellations.

3. Recall (Sensitivity)

- **Class 0 (Not Canceled):** 0.93
- **Class 1 (Canceled):** 0.42
- **Explanation:** For Class 0, the recall is 0.93, meaning the model successfully identified 93% of the actual non-cancelled bookings. For Class 1, the recall is much lower at 0.42, indicating the model only correctly identified 42% of the actual canceled bookings.

4. F1-Score

- **Class 0 (Not Canceled):** 0.88
- **Class 1 (Canceled):** 0.51
- **Explanation:** For Class 0, the F1-score is 0.88, showing a good balance between high precision and high recall. For Class 1, the F1-score is 0.51, indicating that while the model has reasonable precision, its lower recall for cancellations affects the overall balance.

5. ROC AUC Score

- **Value:** 0.76
- **Explanation:** A score of 0.76 indicates that the model has a moderate ability to differentiate between canceled and non-canceled bookings. A score closer to 1 would indicate excellent discrimination, while a score around 0.5 suggests no better performance than random guessing.

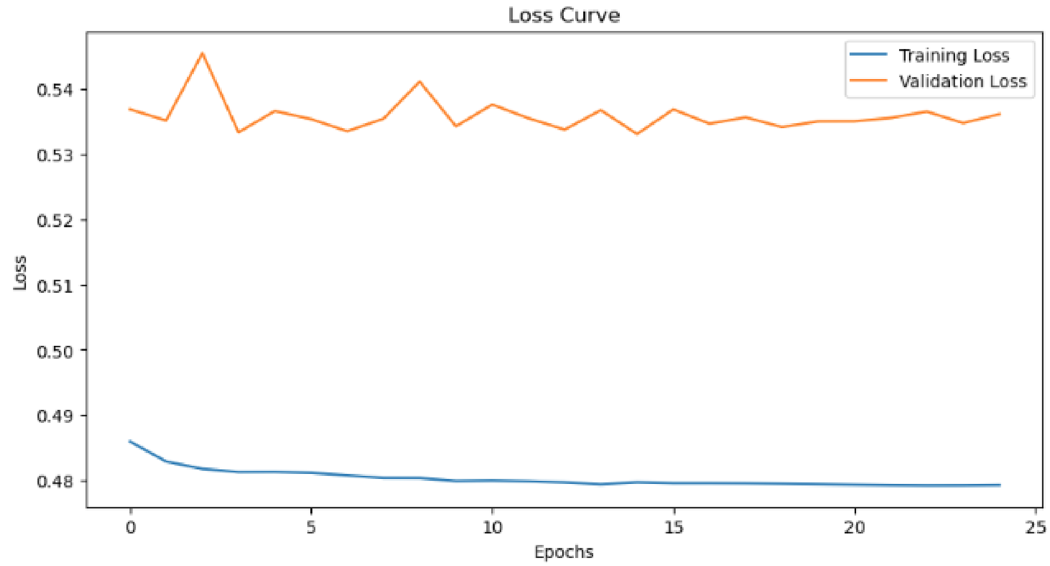
Overall Summary

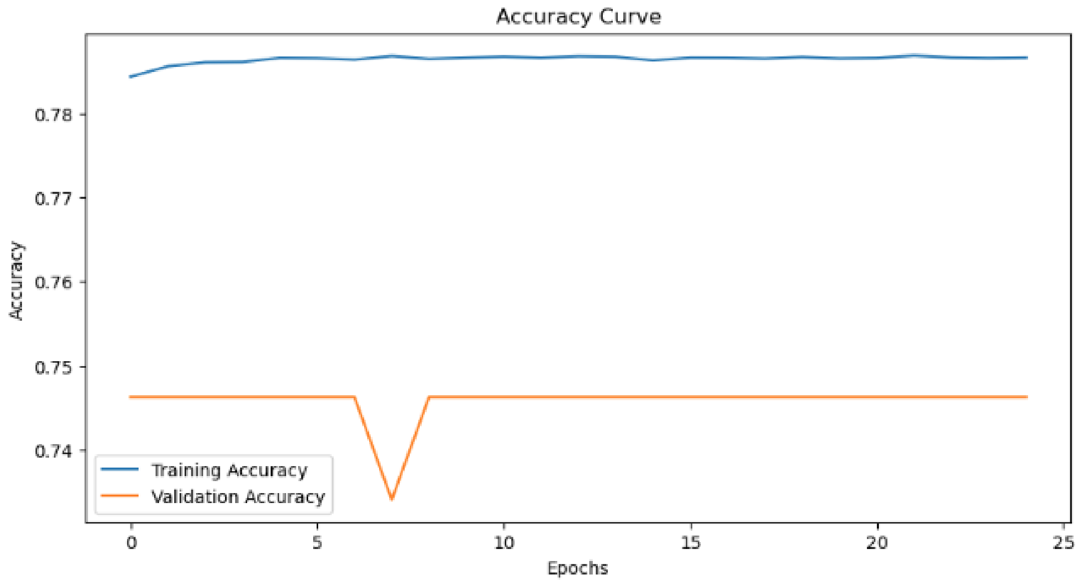
The Random Forest model has demonstrated a solid performance in predicting non-cancellations (Class 0) with high precision, recall, and F1-score. However, its performance is notably lower for predicting cancellations (Class 1), with lower precision, recall, and F1-score. The model tends to miss a significant number of actual cancellations, as indicated by the recall for Class 1 (0.42). Improving the recall for canceled bookings could be an area of focus for enhancing the model's overall effectiveness.

ANN (Artificial Neural Network) Classifier

Model Evaluation

```
Confusion Matrix:
 [[16993  1361]
 [ 3186  2337]]

Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.93      0.88     18354
           1       0.63      0.42      0.51      5523

    accuracy                           0.81     23877
   macro avg       0.74      0.67      0.69     23877
weighted avg       0.79      0.81      0.80     23877


ROC AUC Score: 0.7607675469917661

Accuracy: 0.8095656908321816
```





Summary of Model Performance

1. Confusion Matrix:

- **True Negatives (TN):** 16,993 - These are correctly predicted non-cancellations.
- **False Positives (FP):** 1,361 - These are bookings that were predicted as canceled but were not actually canceled.
- **False Negatives (FN):** 3,186 - These are bookings that were not predicted as canceled but were actually canceled.
- **True Positives (TP):** 2,337 - These are correctly predicted cancellations.

2. Accuracy:

- **Value:** 0.81 (or 81%)
- **Explanation:** This represents the proportion of correctly predicted instances (both cancellations and non-cancellations) out of the total predictions. The model is correct 81% of the time, which indicates good overall performance.

3. Precision:

- **Class 0 (Not Canceled):** 0.84
- **Class 1 (Canceled):** 0.63
- **Explanation:** Precision measures the accuracy of positive predictions. For Class 0, when the model predicts a booking will not be canceled, it is correct 84% of the time. For Class 1, when the model predicts a booking will be canceled, it is correct 63% of the time. This shows that the model is more confident in predicting non-cancellations.

4. Recall (Sensitivity):

- **Class 0 (Not Canceled):** 0.93
- **Class 1 (Canceled):** 0.42
- **Explanation:** Recall measures how well the model identifies actual positive cases. For Class 0, the recall is 0.93, meaning the model correctly identifies 93% of actual non-cancellations. For Class 1, the recall is 0.42, indicating the model correctly identifies 42% of actual cancellations. The lower recall for cancellations suggests the model misses a significant number of actual cancellations.

5. F1-Score:

- **Class 0 (Not Canceled):** 0.88
- **Class 1 (Canceled):** 0.51
- **Explanation:** The F1-score is the harmonic mean of precision and recall. It provides a balance between the two. For Class 0, the F1-score is 0.88, indicating a strong performance in predicting non-cancellations. For Class 1, the F1-score is 0.51, reflecting a lower ability to predict cancellations accurately.

6. ROC AUC Score:

- **Value:** 0.76
- **Explanation:** The ROC AUC score measures the model's ability to distinguish between the two classes (canceled and not canceled). A score of 0.76 indicates a moderate ability to differentiate between the classes. A perfect model would have a score of 1.0, while a score of 0.5 indicates no discriminatory power (equivalent to random guessing).

Overall Summary:

- The ANN model demonstrates good performance in predicting non-cancelled bookings, with high accuracy, precision, recall, and F1-score for Class 0 (Not Canceled).
- However, the model has a relatively lower performance in predicting cancellations (Class 1). It has a moderate precision for cancellations but a low recall, indicating it misses many actual cancellations.

b) Discuss any improvements you would recommend for the model.

Summary of Recommendations for Model Improvement

1. **Address Class Imbalance:**
   - **Resampling Techniques:** Use oversampling (e.g., SMOTE) or undersampling to balance the dataset.
   - **Class Weight Adjustment:** Increase the importance of the minority class in the loss function.
2. **Improve Feature-Dependent Variable Correlation:**
   - **Feature Engineering:** Enhance or create new features to better capture relationships with the dependent variable.
   - **Feature Selection:** Identify and use the most relevant features to improve model performance.

5. Operational Insights

a) Provide actionable insights and recommendations for hotel management based on your analysis. Consider aspects like pricing strategies, customer segmentation, and marketing focus.

Recommendations for Hotel Management

- The hotel in country PRT (Portugal) should focus more on their management because of the highest number of hotel cancellations compared to its own non-cancellations.
- The management of Resort type hotel should focus more on their bookings as it has comparatively less bookings than the City type hotels.
- Reserved room and Assigned room types has highest bookings only on Room Type A meaning there is a room for improvement on other room types.

Note:-

- The customer is more likely to cancel the bookings if the booking is done in 6 months advance or more than that. That can be prevented by giving special offer on advance booking that attracts customers.