**Project Summary**

### *Instructions*:

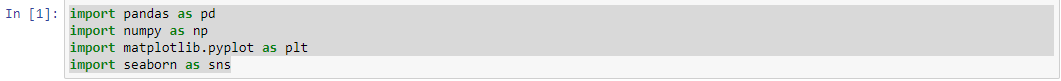
1. **Data Understanding and Pre-processing:**
   1. Load the data-set and provide a summary of its structure (number of rows, columns, data types, and missing values).
   2. Perform data cleaning:
      1. Handle missing values appropriately.
      2. Convert data types if necessary.
      3. Create 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)*

# **Data understanding and Preprocessing**

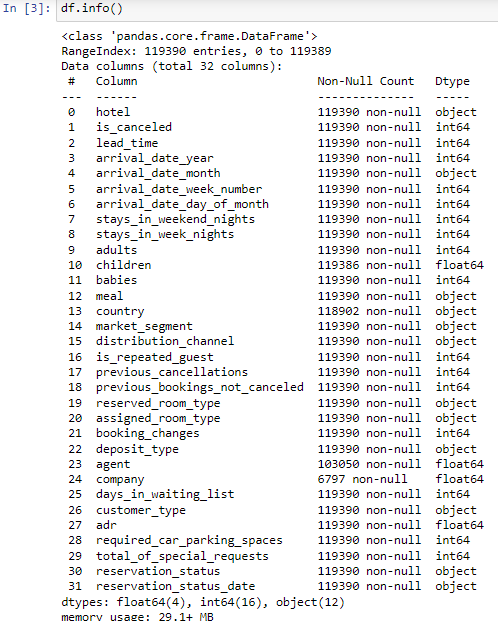
1. ***Load the data-set and provide a summary of its structure (number of rows, columns, data types, and missing values).***

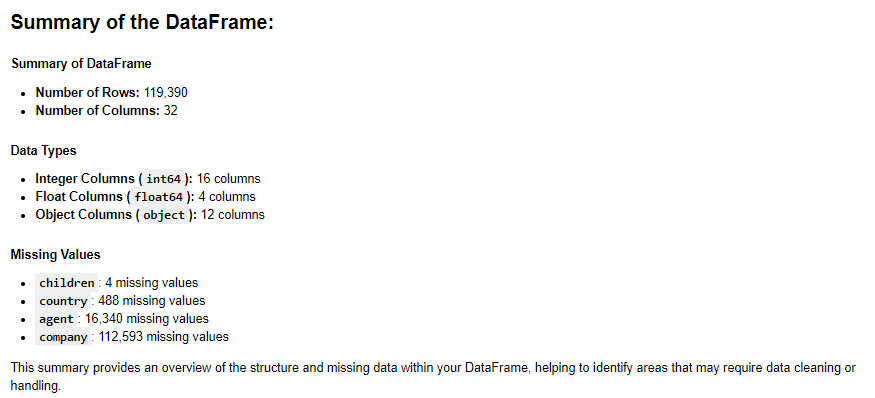
**Importing important libraries for Data Preprocessing:-**



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

**Getting Information of the DataFrame:-**





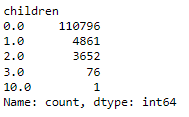
* This code shows the total number of **Rows** and **Columns** along with the data types of columns.
* Using this code we can also see if there are any **missing values** available in any of the columns or not.

1. ***Performing Data Cleaning***

**Handling missing values:-**

***“children”:-***

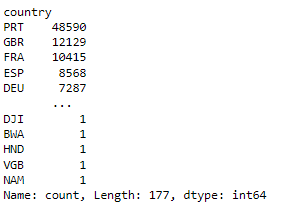
* Given below is the children count in **“children”** column.



* 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.



* 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**.

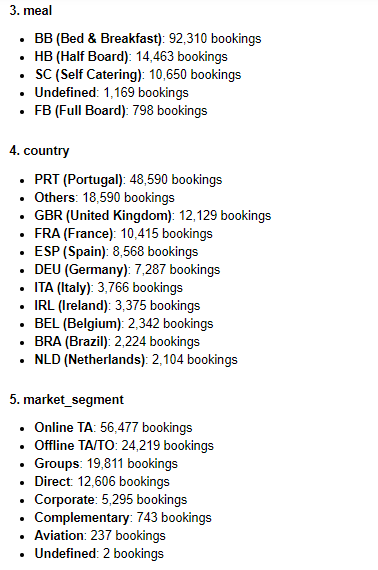
***“company”:-***

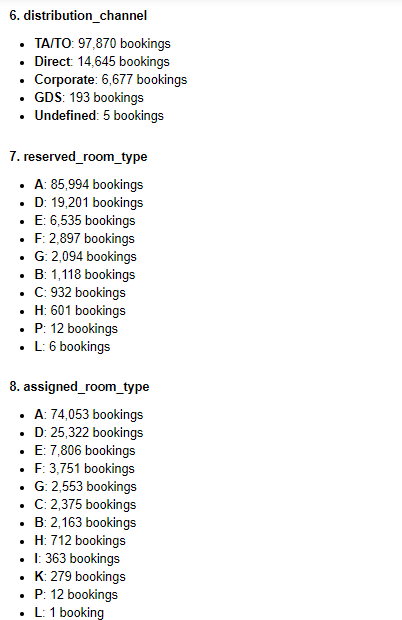
* There were **112593 missing values** in **“company”** column.
* Therefore, we will remove this column because replacing the missing values will not be useful.

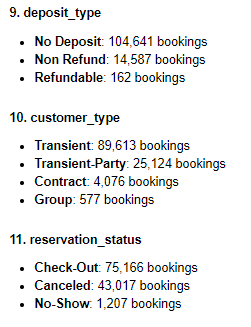
**Handling categorical Values**

* Given below are the categorical columns present in dataset.





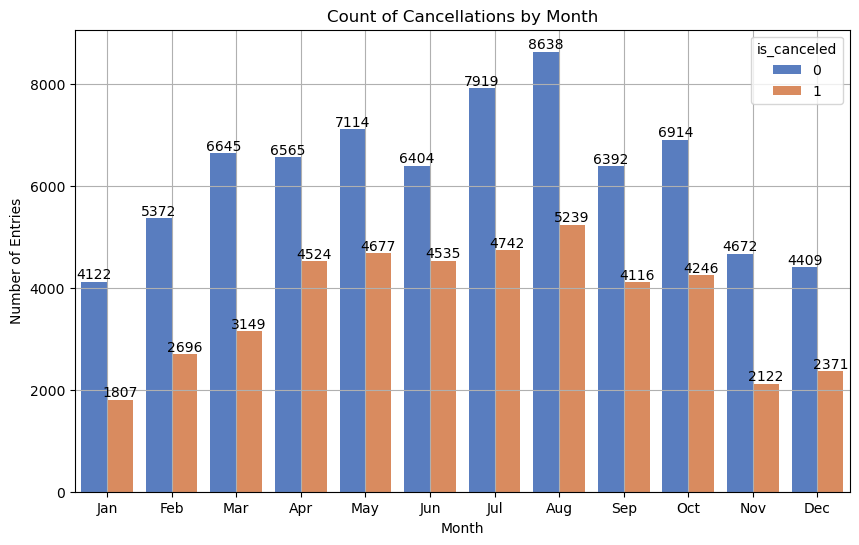




* We converted all the possible categorical columns into numeric columns.

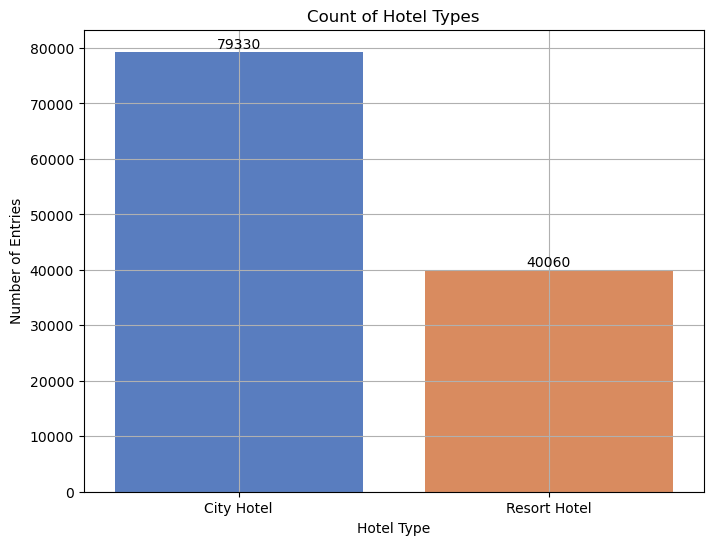
# **Exploratory Data Analysis (EDA)**

1. ***Seasonality in bookings.***



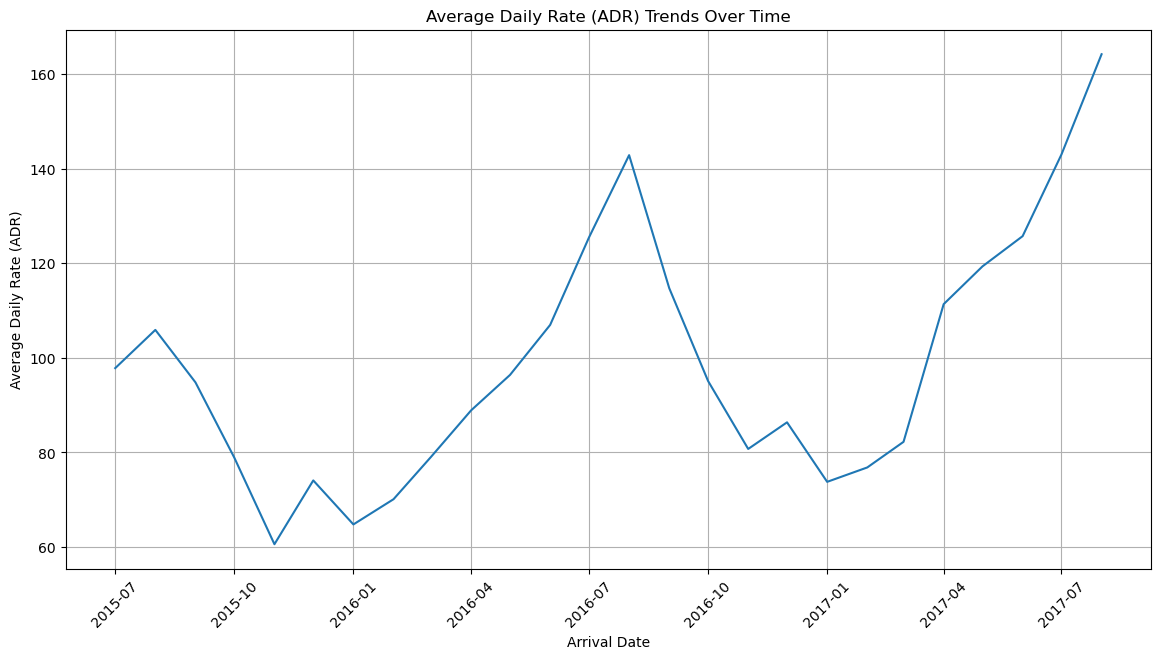
* This plot shows **top 4 months with the highest bookings**.
  + **August** - 8638 bookings
  + **July** - 7919 bookings
  + **May** - 7114 bookings
  + **October** - 6914 bookings

1. ***Distribution of stays across different hotel types.***



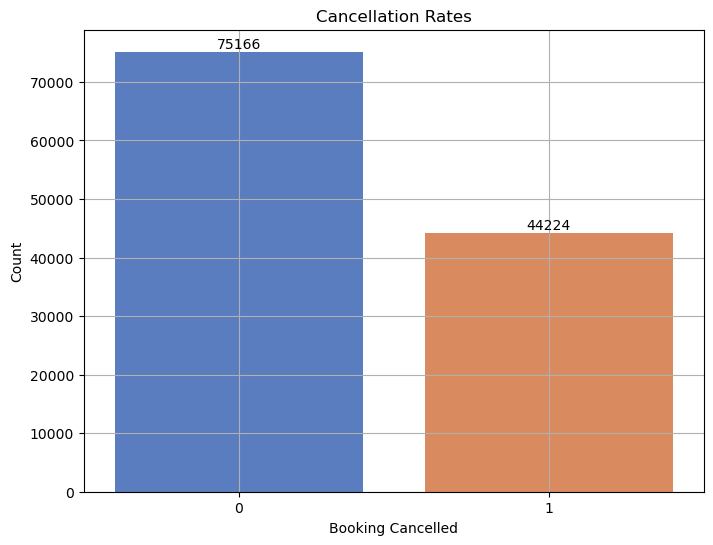
* This plot shows the bookings of Hotel.
  + Mostly hotel bookings were **“City Hotel”** type
  + **City Hotel** - 79330 bookings
  + **Resort Hotel** - 40060 bookings

1. ***Average Daily Trends (ADR)***

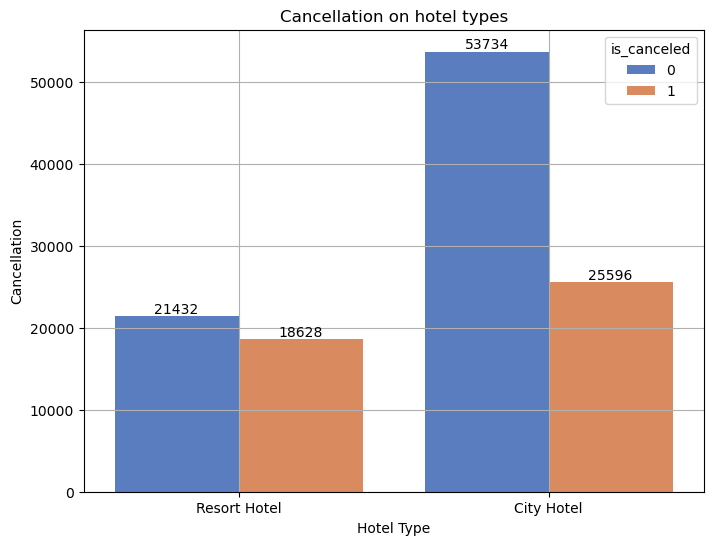


* This line-plot shows the ADR spikes from **"January" (07)** to **"August" (08)** of every year.
* The ADR dips is seen on every from **"August" (09)**to **"November" (11)** of each year.
* The Highest ADR recorded was on **"2017-07" (July 2017)**
* The Lowest ADR recorded was on **"2015-11" (November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

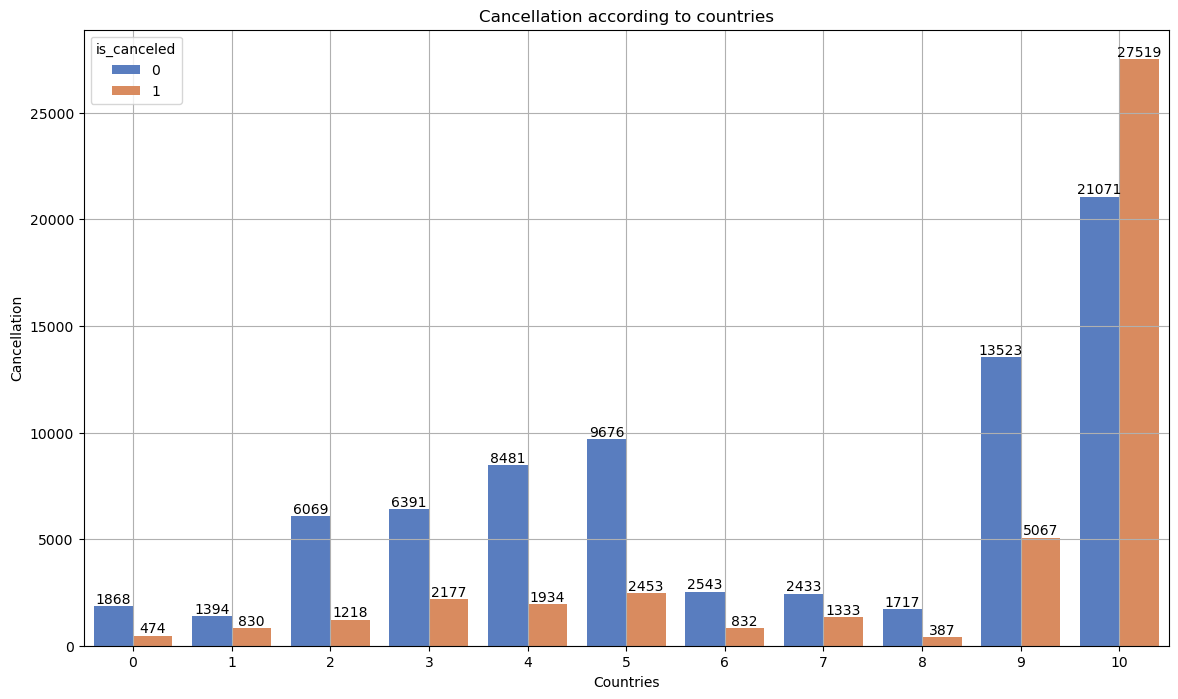
1. ***Cancellation rates and factors affecting cancellations***

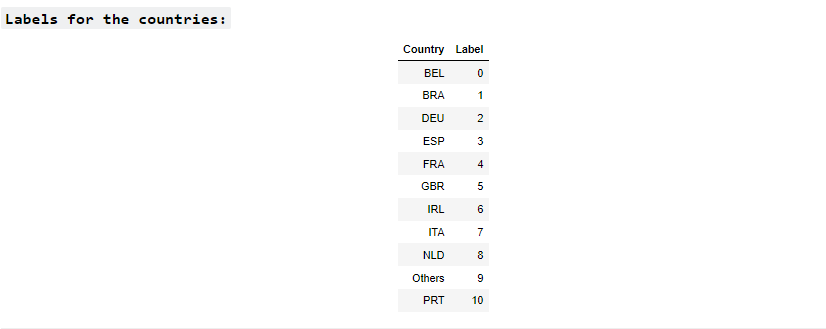


* This plot shows that the **44224 bookings** were canceled out of **119390 bookings**.

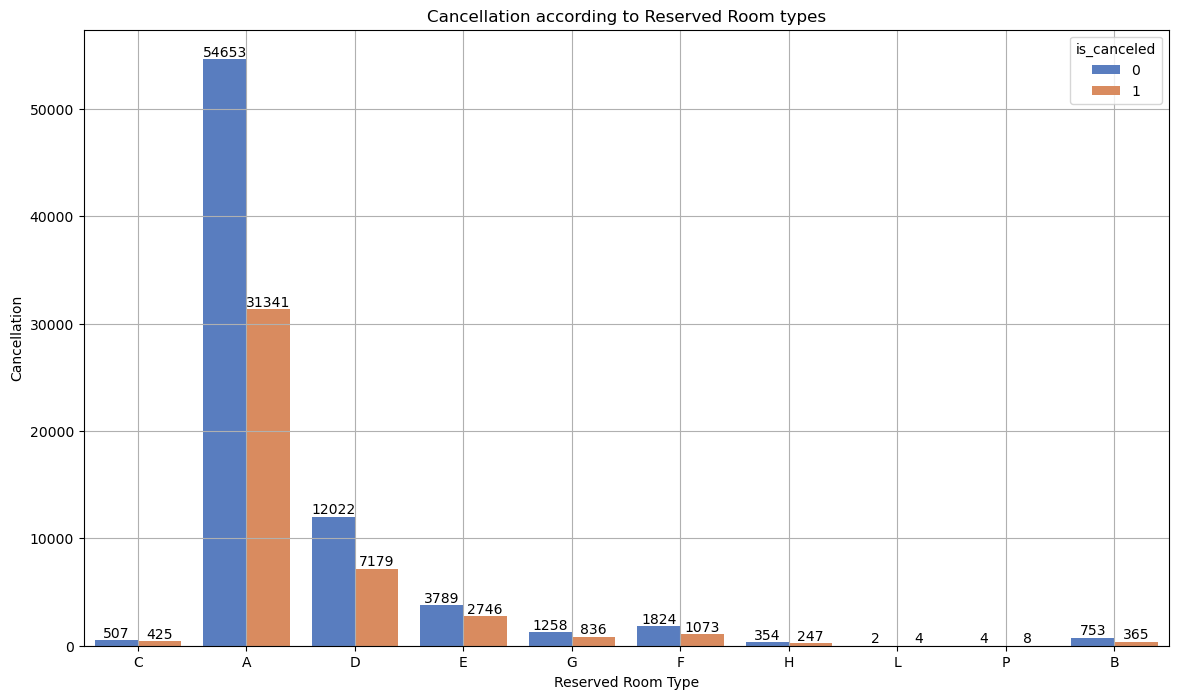


* This plot shows that **most of the bookings were of "City Hotel"**.
* Since **"Resort Hotel"** has **less bookings as compared to** **"City Hotel"**. But the **cancellation rate is higher as compared to "Resort Hotel”**.





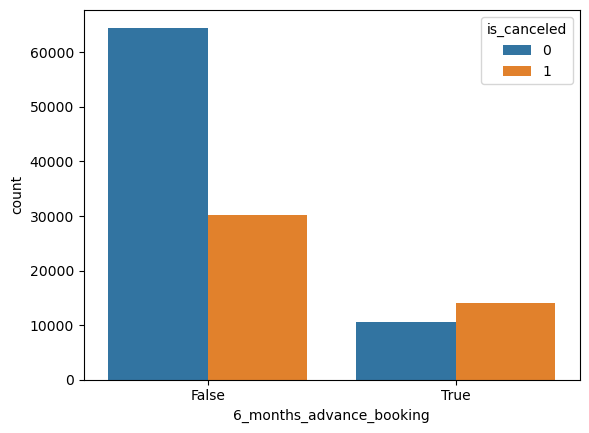
* The most of the bookings were from country code **"PRT" (Portugal)**.
* **"PRT" (Portugal)** is the only country which has **higher amount of cancellations (27519)** **even higher than non-cancellations (21071)**.
* Rest of the country has lower rate of cancellations.



* 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)**.

# **Hypothesis Testing**

1. ***Customers booking more than 6 months in advance are more likely to cancel.***

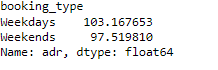


* As we can see in the above plot, the **"cancellations rate" in "6 months advance booking"** is **higher than the "non cancellations rate"** as **compared to the cancellations of "bookings before 6 months"**.

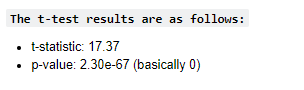
**Conclusion:-**

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

1. ***Weekday bookings have a higher average daily rate than weekend bookings***



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

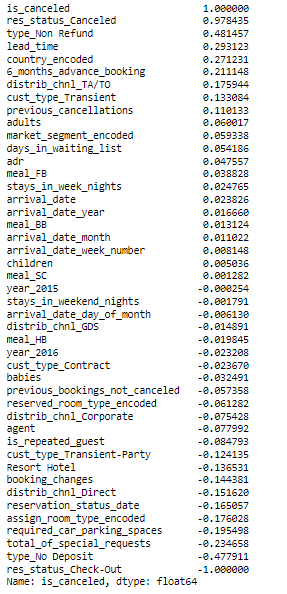


* Since the **P-value is less than 0.05**. Therefore, we can reject the null hypothesis.
* This supports the hypothesis that **"Weekday bookings have a higher average daily rate than weekend bookings."**

# **Predictive Modeling**

1. ***Select appropriate features.***

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



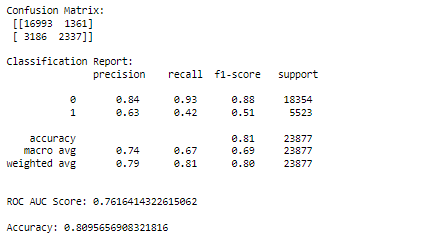
* Given below are the features that we will be taking while training the model because of the high correlations.



***Model Training***

***Random Forest Classifier***

**Model Evaluation**



**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****

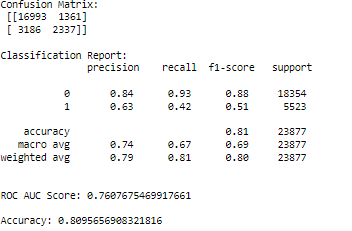
* ****Value:**** 0.76
* ****Explanation:**** A score of 0.76 indicates that the model has a moderate ability to differentiate between canceled and non-canceled bookings. A score closer to 1 would indicate excellent discrimination, while a score around 0.5 suggests no better performance than random guessing.

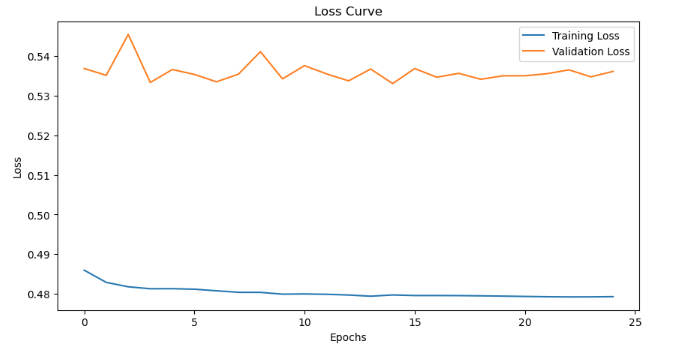
### ****Overall Summary****

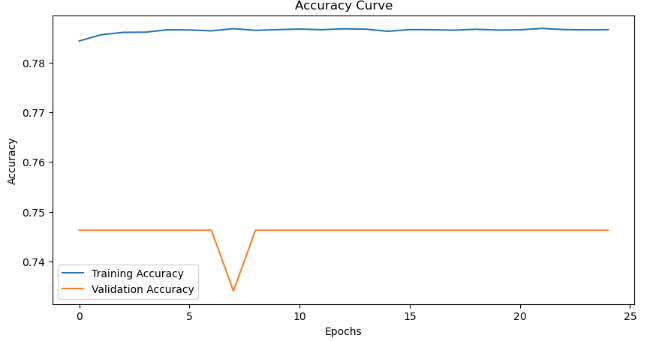
The Random Forest model has demonstrated a solid performance in predicting non-cancellations (Class 0) with high precision, recall, and F1-score. However, its performance is notably lower for predicting cancellations (Class 1), with lower precision, recall, and F1-score. The model tends to miss a significant 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**







### **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.

1. ***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.

1. ****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.

# **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.***

### **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.