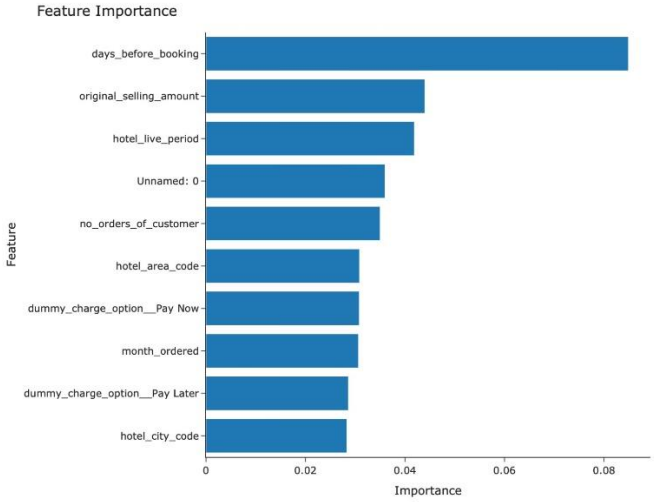
Question 3

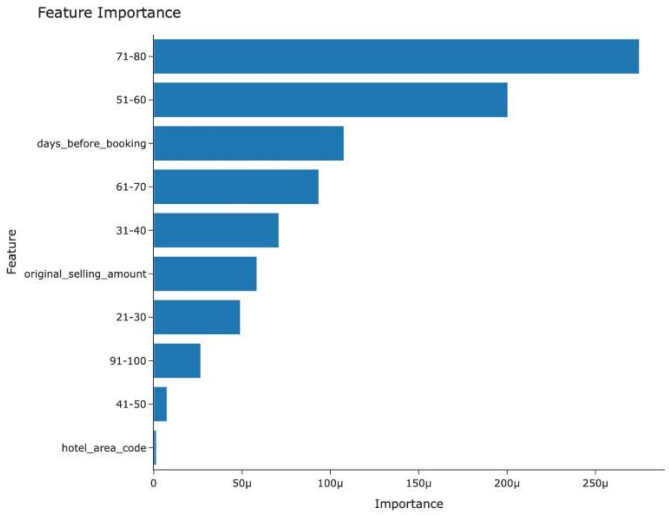
We've ploted the following graphs, each one describes features importance based on different models fitting.

Random Forest Classifier:



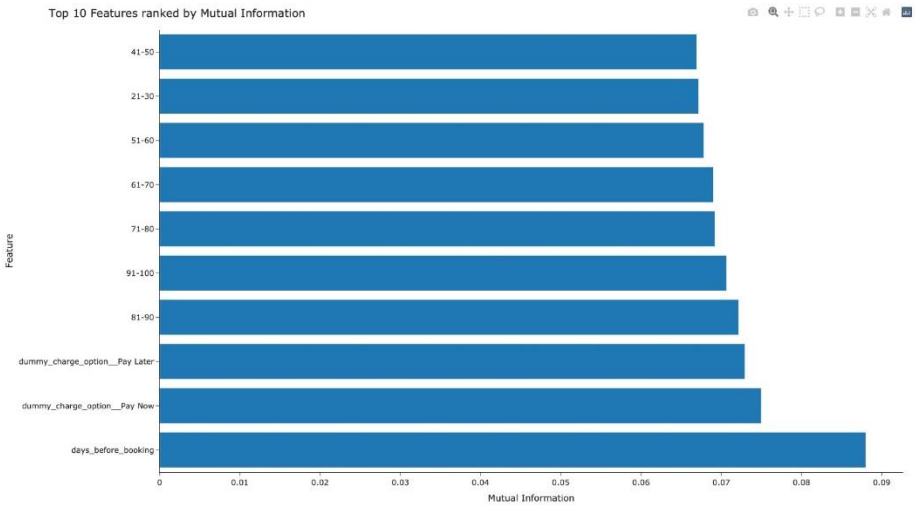
The Random Forest Classifier is an ensemble learning algorithm that combines multiple decision trees to make accurate predictions for classification tasks. As we can see, 'days before booking' is the feature considered to be the most important one based on the correlation to cancellation.

Lasso Classifier:



Lasso correlation refers to the use of the Lasso regularization technique in linear regression models to select relevant features and reduce the impact of irrelevant or correlated features on the model's predictions. Notice that '71-80' is the most important feature here.

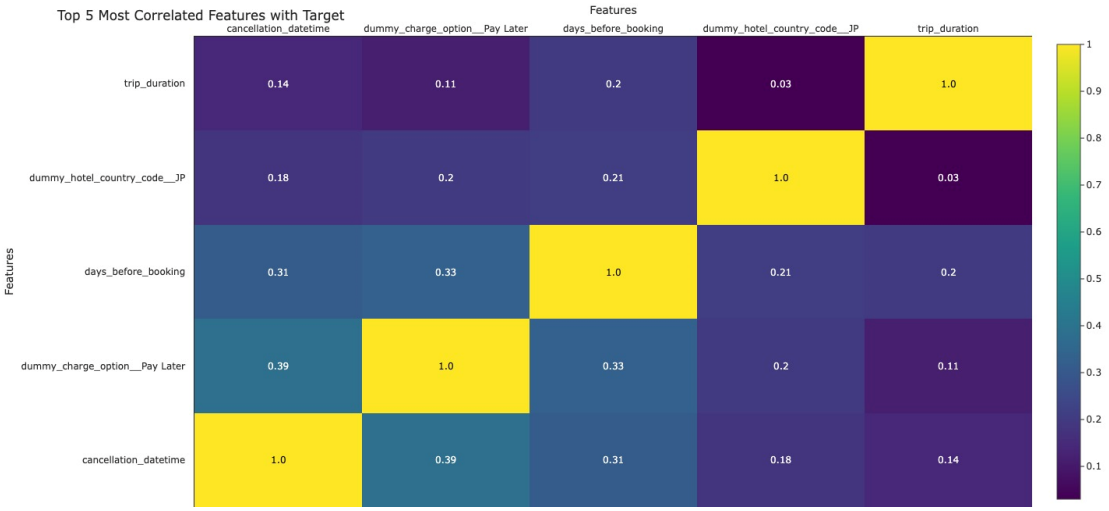
Mutual Information:



Mutual information is a statistical measure that quantifies the dependence between two variables by capturing the amount of information shared between them, often used in feature selection to identify the relevance of features for a specific target variable.

'days before booking' is the most important feature.

After evaluating these methods, we examined the correlation matrix:



The correlation matrix model calculates the pairwise correlation coefficients between variables, providing insights into the strength and direction of linear relationships, allowing us to identify patterns and dependencies in the data.

Thus, we infer that 'days before booking' is most relevant feature for cancellation prediction because it was the most significant one in 2 also was high in the third one.

Furthermore, we noticed that the features '71-80', '81-90' and '91-100' which describes in each entry the percentage of charge out of full price according to days before checking as in the provided cancellation policy of that entry also seem to have a great influence.