Q1) In the first question we just have to print the

highlighted vales so first the list has been selected from the dictionary then only values having type int are selected.

Q2)

* First, the exploratory data analysis is done to get insights into the data.
* Data cleaning is done.

1. Only the age column had missing values

so missing values were filled by mean.

1. Categorical data is changed for numerical market segment, nationality and distribution channel.
2. Unnamed: 0 and Id contains many unique values and are not essential for predicting the output, so they were dropped.
3. Features having the value count of one class concerning the other is significantly less(less than 1%) were droppped.
4. Since the BookingsCheckedIn more than 2 were very less so they are not taken as the target classes.

* Test Train split is done.

StratifiedKFold is used since we were having imbalanced data set.

* The random forest method is used and fitted to the data.
* Classification\_report is printed, and roc\_curve is plotted.

precision recall f1-score support

0 1.00 1.00 1.00 3879

1 1.00 0.01 0.01 12347

2 0.02 1.00 0.04 228

accuracy 0.25 16454

macro avg 0.67 0.67 0.35 16454

weighted avg 0.99 0.25 0.25 16454

* The model is then tested on the test data and deployed using streamlit.

Bonus Section

1. Write about any difficult problem that you solved. (According to us difficult - is something which 90% of people would have only 10% probability in getting a similarly good solution).

I have solved the spaceship titanic problem which involved a lot of data cleaning and used linear regression and random forest for predicting the results. The data was taken from a Kaggle competition.

2. Explain backpropagation and tell us how you handle a dataset if 4 out of 30 parameters have null values of more than 40 per cent.

Backpropagation is an algorithm that back propagates the errors from output nodes to the input nodes. It is a learning algorithm to compute a gradient descent concerning weights.

Since the null values are more than 40 per cent so we can drop these columns, but if these columns are strongly correlated with the output, we can plug the mean depending upon the other columns.