Logistic Regression

1. Using the same dataset as you did for the Linear Regression portion of the assignment, please estimate a logistic regression to predict whether a customer booked a suite (Room Type = S) as opposed to any other type of room. (HINT: You need to create a new dependent variable which =1 if Room Type = S and 0 otherwise) Select 2 independent variables which you believe to be most important and explain why you chose those variables in your model.

i. Import the necessary packages such as:

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import confusion_matrix, classification_report
```

ii. Convert the dependent variable from categorical to integer (0 and 1). You may use the 'astype(int)' function.

iii. It is a good practice to use the head() or tail() function to view that your data frame contains the new dependent variable.

iv. Develop a baseline model first, with intercept as the only predictor, e.g.

```
X_intercept = pd.DataFrame({'Intercept': 1}, index=df.index)
```

Split the data into training and testing sets

```
X_train, X_test, y_train, y_test = train_test_split(X_intercept, y, test_size=0.3, random_state=42)
```

Fit the model (logistic regression with only intercept)

```
logit_model = sm.Logit(y_train, X_train)
result = logit_model.fit()
```

Print the summary

```
print(result.summary())
```

Predict on test set

```
y_pred_prob = result.predict(X_test)
y_pred_binary = (y_pred_prob >= 0.5).astype(int)
```

Calculate misclassification rate

```
misclassification_rate = (y_pred_binary != y_test).mean()
print("\nMisclassification Rate:", misclassification_rate)
```

v. Now, add one or two variables of your choice to the model, either individually or together, and follow the same process as before to calculate the misclassification rate.

2. Add 2 additional variables. How did your misclassification rate change? Which model do you believe is better and why? Please be sure your results screenshot includes anything you reference in your answer.

i. Add two additional variables of your choice, and calculate the misclassification rate by following the steps described in question 1 instructions. Compare the models. Remember that high Variance Inflation Factor (VIF) values indicate multicollinearity, which can negatively impact the model by causing instability in the coefficient estimates.