project hotell

presented by

karim Atef

PROBLEM **STATEMENT**

OBJECTIVE

It is the design and implementation of a machine learning model for a hotel to predict whether the booking will be canceled or not.

- **Goal:** He did the classification correctly.

FIRST
STEP

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import classification_report, confusion_matrix
```

```python
data = pd.read_csv(r"C:\Users\karem\Downloads\first inten project.csv")
data.head(10)
```

IMPORT DATA AND LIBRARYS

- **pandas** : To maintain and retrieve the data
- **seaborn , matplotlib :** To create a visualization for the data
- **train_test_splite :** To split the data into test data and train data
- **KNeighborsClassifier** : To create a model of type KNN
- **Confusion_matrix** : He is doing a test for the model.

DATA **EXAMINATION AND PROCESSING**

```python
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36285 entries, 0 to 36284
Data columns (total 17 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Booking_ID                36285 non-null  object 
 1   number of adults          36285 non-null  int64  
 2   number of children        36285 non-null  int64  
 3   number of weekend nights  36285 non-null  int64  
 4   number of week nights     36285 non-null  int64  
 5   type of meal              36285 non-null  object 
 6   car parking space         36285 non-null  int64  
 7   room type                 36285 non-null  object 
 8   lead time                 36285 non-null  int64  
 9   market segment type       36285 non-null  object 
 10  repeated                  36285 non-null  int64  
 11  P-C                       36285 non-null  int64  
 12  P-not-C                   36285 non-null  int64  
 13  average price             36285 non-null  float64
 14  special requests          36285 non-null  int64  
 15  date of reservation       36285 non-null  object 
 16  booking status            36285 non-null  object 
dtypes: float64(1), int64(10), object(6)
memory usage: 4.7+ MB
```

```python
data[data.duplicated() == True]
```

NULL CHECK

- Null does not exist in the data.

DUPLICATED CHECK

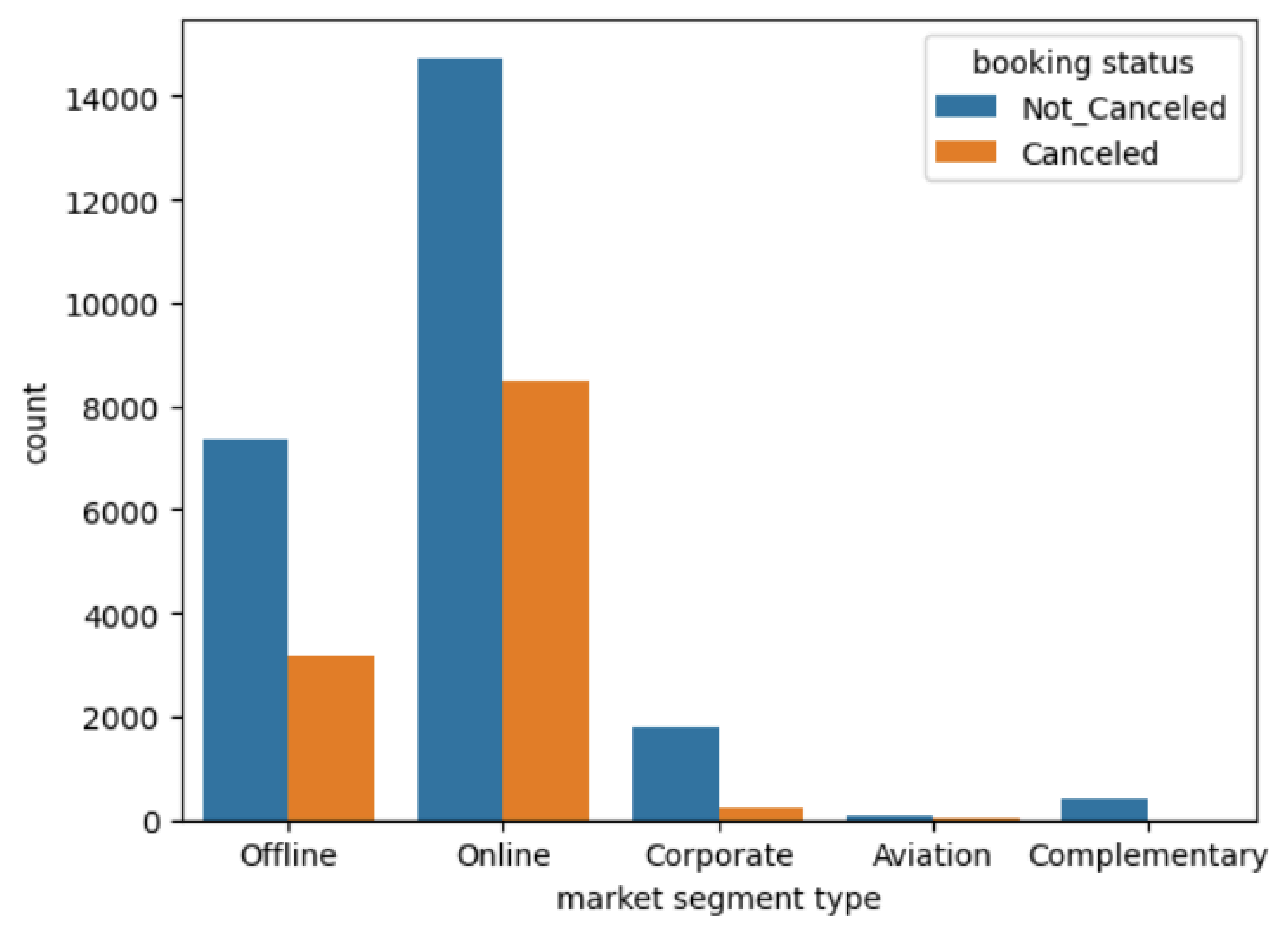
- There are no duplicates.

CHECK DATA TYPE

- The type of data for the date of reservation is incorrect and we will change it to date, time, but we faced some problems that there is a date of 2018-2-29 which is wrong, so I replaced it with 1/3/2018.

DATA
ANALYSIS

```
sns.countplot(data=data, x='market segment type', hue='booking status')
plt.show()
```

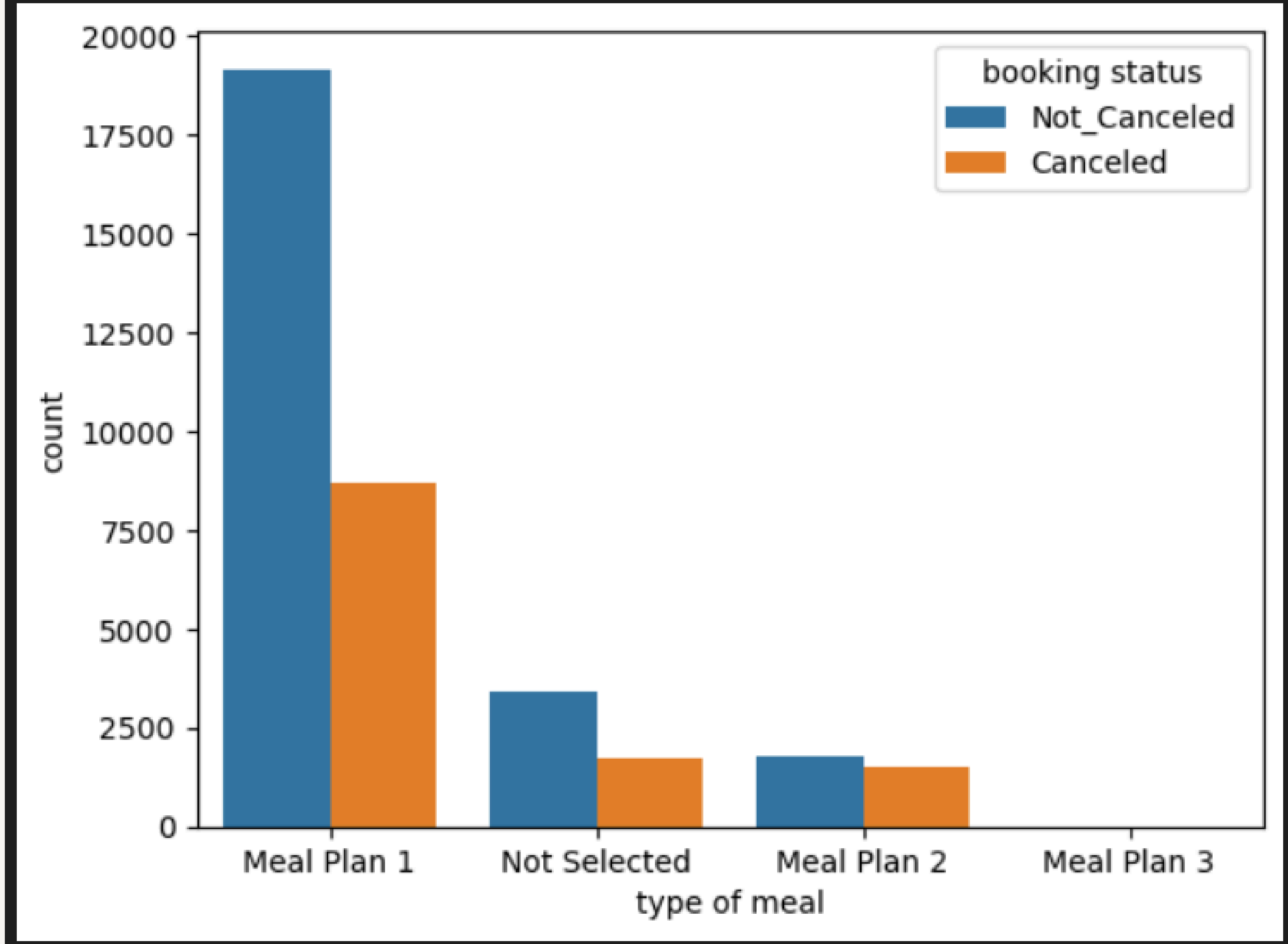


 # Here I wanted to study if there is a relationship between 'booking status' and 'market segment type'.

- In that case, it turned out that the highest number of bookings was online, and there was a higher number of cancellations.
- Also, the difference between 'canceled' and 'not canceled' is in 'complementary.'

 # It is clear that the 'market segment type' affects the 'booking status,' so we will take the 'market segment type.'

```python
sns.countplot(data=data, x='type of meal', hue='booking status')
plt.show()
```



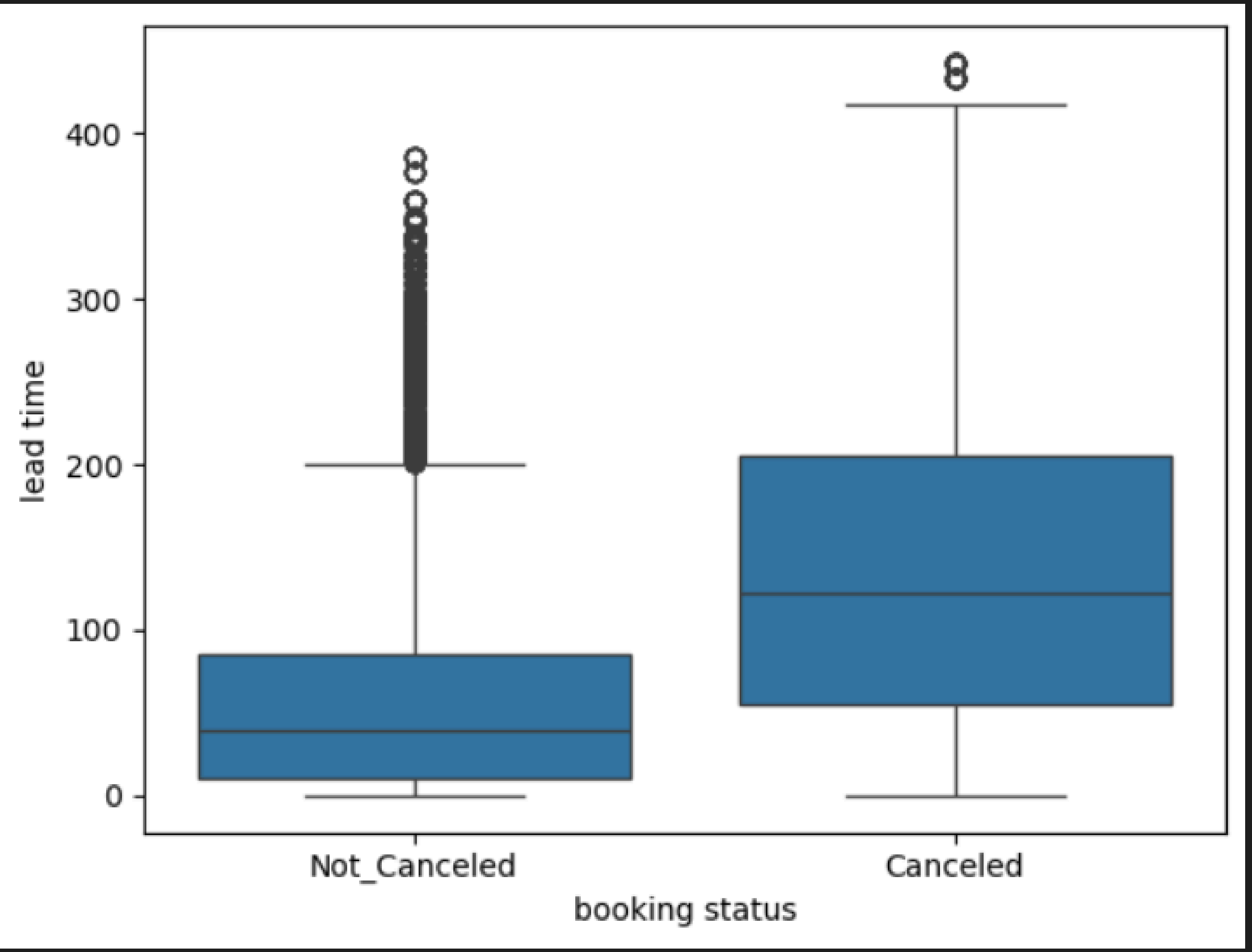
HERE I WANTED TO STUDY IF THERE IS A RELATIONSHIP BETWEEN 'BOOKING STATUS' AND 'TYPE OF MEAL' BY BAR PLOT.

- It is clear that plan 1 is the best in terms of performance and that plan 2 is less efficient regarding the difference between canceled and not canceled.

- Plan 3 is to be the least plan in terms of quantity and may reach zero.

 # It is clear that the 'type of meal' affects the 'booking status'.

```
sns.boxplot(data=data, x='booking status', y='lead time')
plt.show()
```



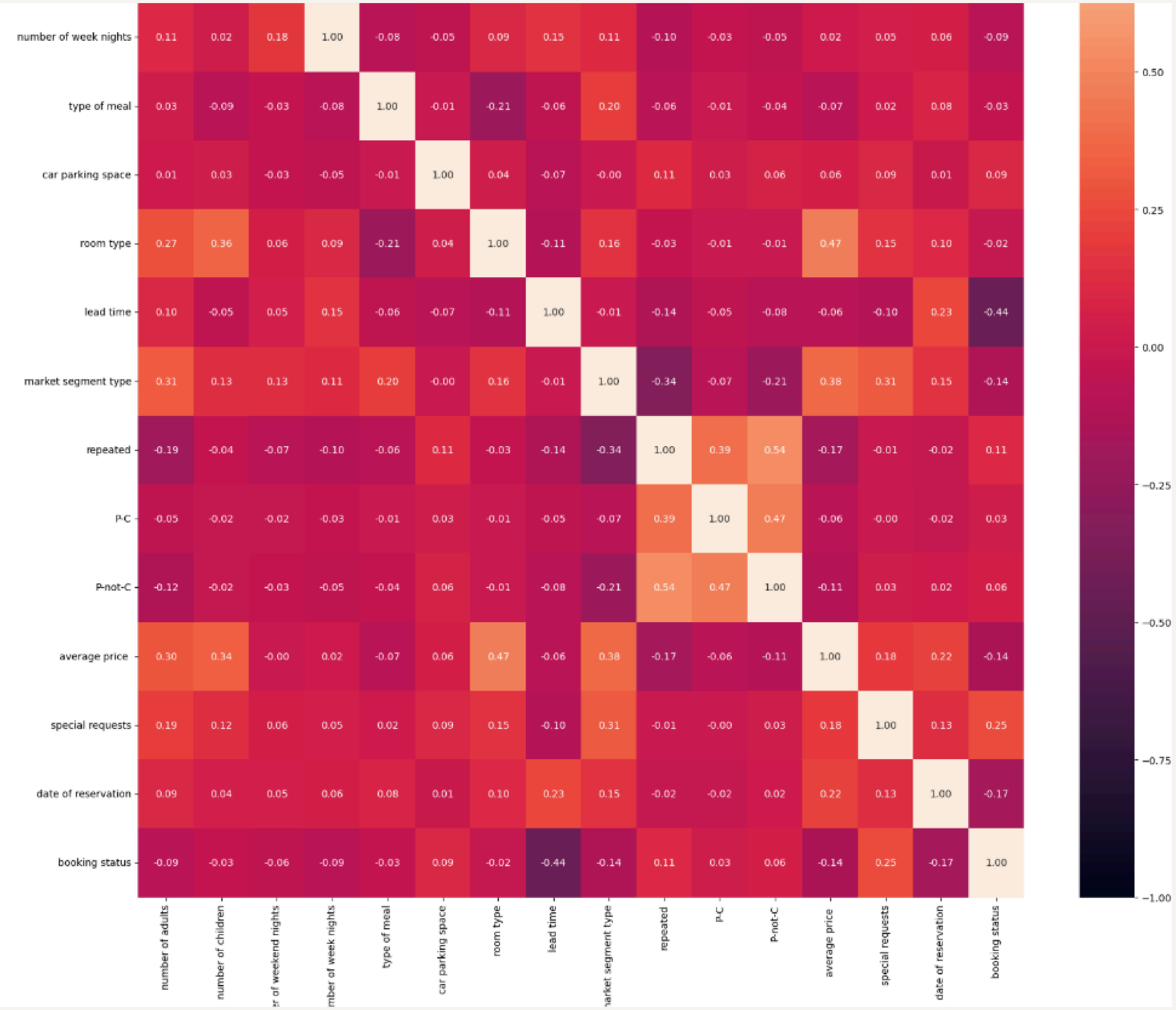
HERE I WANTED TO STUDY IF THERE IS A RELATIONSHIP BETWEEN 'BOOKING STATUS' AND 'LEAD TIME' BY BOX PLOT.

- It is clear that the longer the lead time, the higher the number of cancellations.

- It is clear that the shorter the lead time, the higher the number of Not_canceled.

```python
le = LabelEncoder()
data['type of meal'] = le.fit_transform(data['type of meal'])
data['market segment type'] = le.fit_transform(data['market segment type'])
data['booking status'] = le.fit_transform(data['booking status'])
data['room type'] = le.fit_transform(data['room type'])
data.drop(['Booking_ID','month'] , inplace= True ,axis= 1)
```

- This is to do encoding for the column categories because the KNN model cannot accept non-numeric data.



- This heatmap represents the correlation between the column and then.

colum	**correlation**
date of reservation	-0.17
special requests	0.25
average price	-0.14
p-not-c	0.06
p-c	0.03
repeated	0.11
market segment type	-0.14
lead time	-0.44

SELECT BEST **COLUMES**

```
data_traning =  data.loc[ :,[ 'type of meal','repeated', 'car parking space','lead time', 'market segment type', 'average price ', 'special requests']]
targetcolum = data['booking status']
data_traning
```

```
X_train, X_test, y_train, y_test = train_test_split(data_traning, targetcolum, test_size=0.3, random_state=42)
```

```
model = KNeighborsClassifier(n_neighbors=5)
model.fit(X_train, y_train)
```

BEST COLUMS

- **'type of meal' , 'repeated' , 'car parking space' , 'lead time' , 'market segment type' , 'average price ' , 'special requests'**

TRAIN MODEL

- I chose here that it takes 5 from neighbors because it was giving the best result.

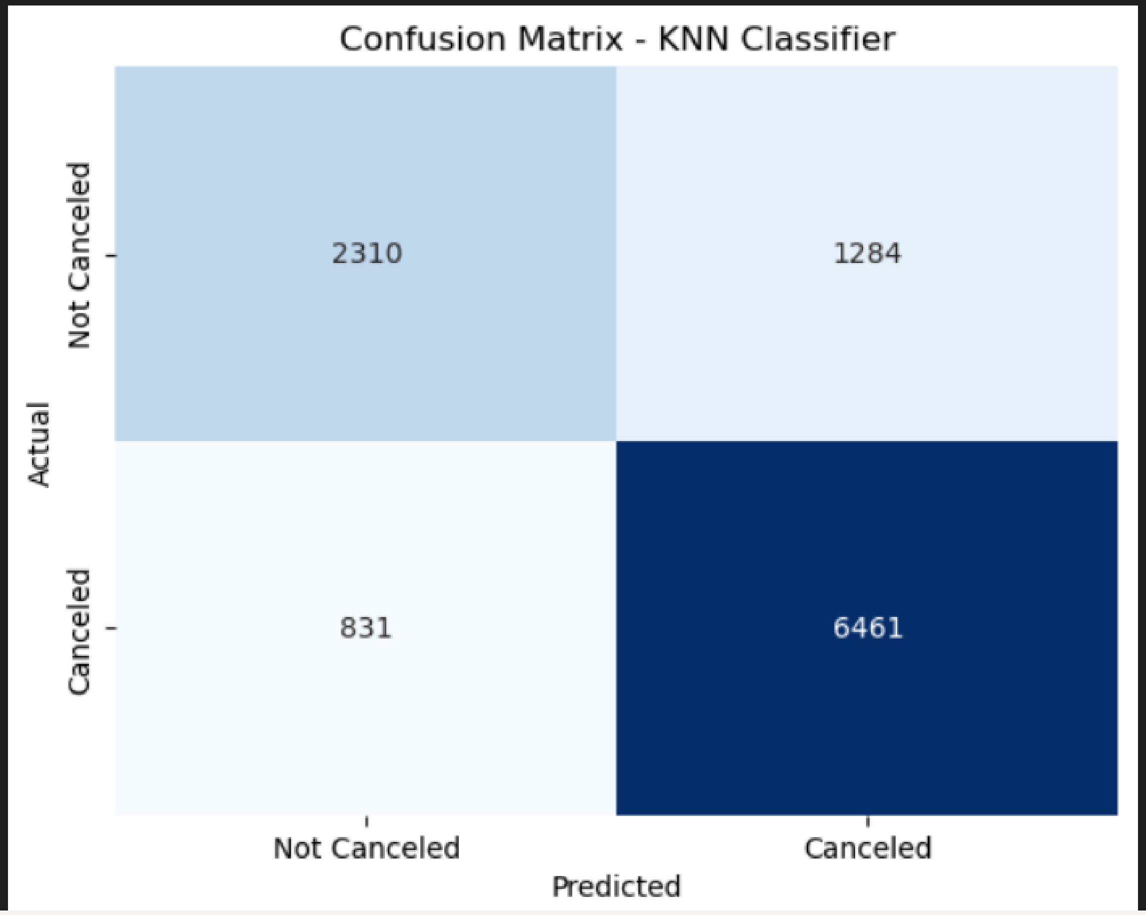
SPLIT DATA

- **Here I did a split of the data with a 70 to 30 ratio, and this was the best ratio used.**

EVALUATING THE **MODEL**

```python
y_pred = model.predict(X_test)
report = classification_report(y_test, y_pred )
sns.heatmap(confusion_matrix(y_test, y_pred), annot=True, fmt='d', cmap='Blues', cbar=False,xticklabels=['Not Canceled', 'Canceled'],yticklabels=['Not Canceled', 'Canceled'])
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title('Confusion Matrix - KNN Classifier')
plt.show()
```

✓ 0.4s



HIS RESULTS

- It is clear that the model predicts canceled cases very accurately, as it predicted 6461 correctly and 2310 correctly for not canceled cases.

PRESENTED BY :

karem atef

THANKS

karem atef