Information of Data

num_passengers = number of passengers travelling

sales_channel = sales channel booking was made on

trip_type = trip Type (Round Trip, One Way, Circle Trip)

purchase_lead = number of days between travel date and booking date

length_of_stay = number of days spent at destination

flight_hour = hour of flight departure

flight_day = day of week of flight departure

route = origin -> destination flight route

booking_origin = country from where booking was made

wants_extra_baggage = if the customer wanted extra baggage in the booking

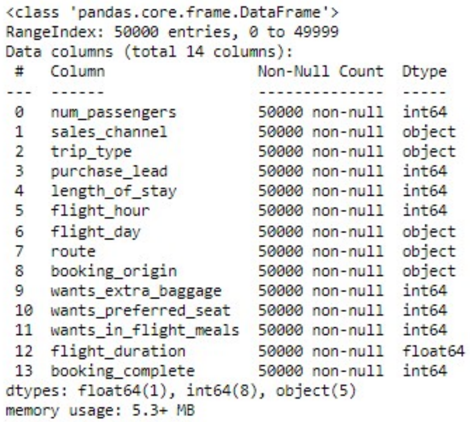
wants_preferred_seat = if the customer wanted a preferred seat in the booking

wants_in_flight_meals = if the customer wanted in-flight meals in the booking

flight_duration = total duration of flight (in hours)

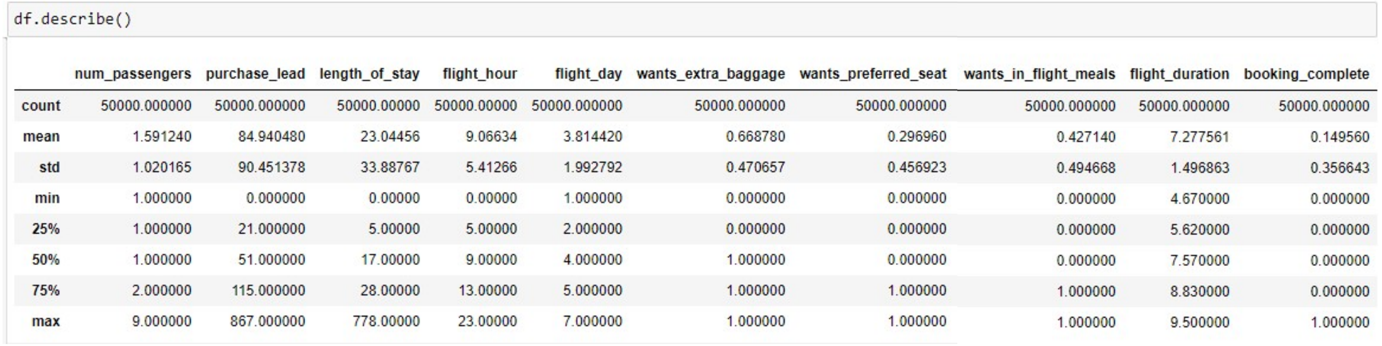
booking_complete = flag indicating if the customer completed the booking

The data was loaded and information of the data.
The information of the data shows datatypes, null values in data with count.
Number of columns, Number of records and Names of the columns.
There is no null values.there is a clear data without any missing values.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50000 entries, 0 to 49999
Data columns (total 14 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   num_passengers         50000 non-null  int64  
 1   sales_channel          50000 non-null  object 
 2   trip_type              50000 non-null  object 
 3   purchase_lead          50000 non-null  int64  
 4   length_of_stay         50000 non-null  int64  
 5   flight_hour            50000 non-null  int64  
 6   flight_day             50000 non-null  object 
 7   route                  50000 non-null  object 
 8   booking_origin         50000 non-null  object 
 9   wants_extra_baggage    50000 non-null  int64  
 10  wants_preferred_seat   50000 non-null  int64  
 11  wants_in_flight_meals  50000 non-null  int64  
 12  flight_duration        50000 non-null  float64
 13  booking_complete       50000 non-null  int64  
dtypes: float64(1), int64(8), object(5)
memory usage: 5.3+ MB
```

Here we can see the description of the data mean,standarddeviation,minimum,maximum



```
df.describe()
```

	num_passengers	purchase_lead	length_of_stay	flight_hour	flight_day	wants_extra_baggage	wants_preferred_seat	wants_in_flight_meals	flight_duration	booking_complete
count	50000.000000	50000.000000	50000.00000	50000.00000	50000.000000	50000.000000	50000.000000	50000.000000	50000.000000	50000.000000
mean	1.591240	84.940480	23.04456	9.06634	3.814420	0.668780	0.296960	0.427140	7.277561	0.149560
std	1.020165	90.451378	33.88767	5.41266	1.992792	0.470657	0.456923	0.494668	1.496863	0.356643
min	1.000000	0.000000	0.00000	0.00000	1.000000	0.000000	0.000000	0.000000	4.670000	0.000000
25%	1.000000	21.000000	5.00000	5.00000	2.000000	0.000000	0.000000	0.000000	5.620000	0.000000
50%	1.000000	51.000000	17.00000	9.00000	4.000000	1.000000	0.000000	0.000000	7.570000	0.000000
75%	2.000000	115.000000	28.00000	13.00000	5.000000	1.000000	1.000000	1.000000	8.830000	0.000000
max	9.000000	867.000000	778.00000	23.00000	7.000000	1.000000	1.000000	1.000000	9.500000	1.000000

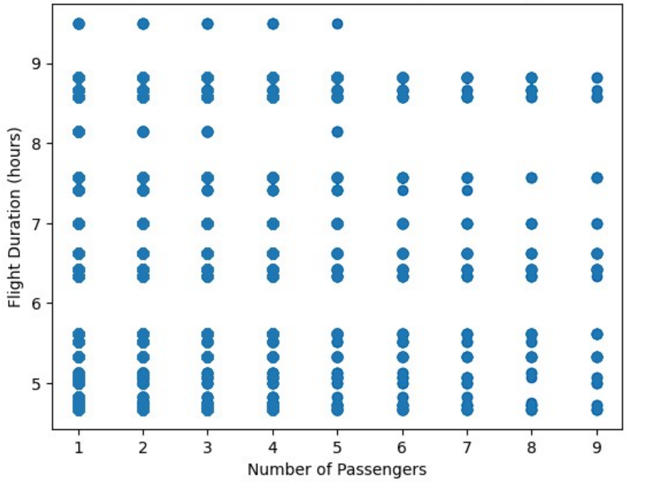
Data analysis:

1st plot : here is a distribution of no.of passengers its a distrubution plot the most no.of passengers are 1 to 4 and 5, to 9 are very rare.

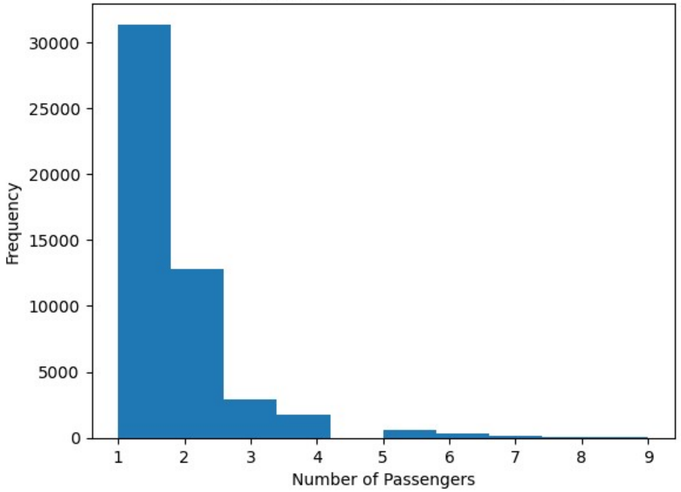
2nd plot: the duration of filght in hours the average duration is 7 to 8 hours.

3rd plot : the density of purchase lead time is mostly at 0 to 250 and 500 is less.

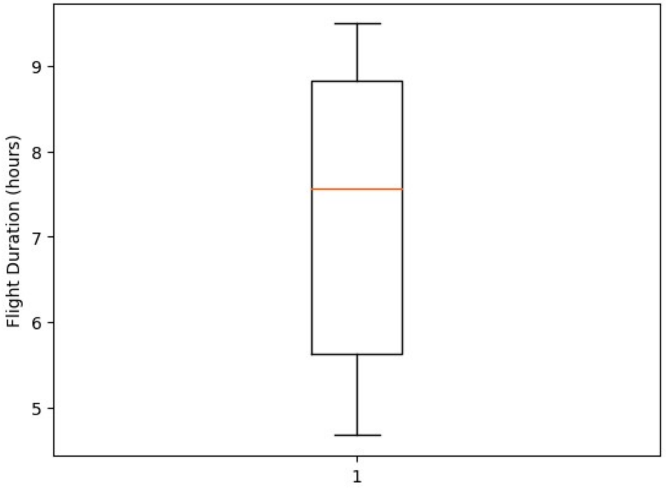
Relationship between Number of Passengers and Flight Duration



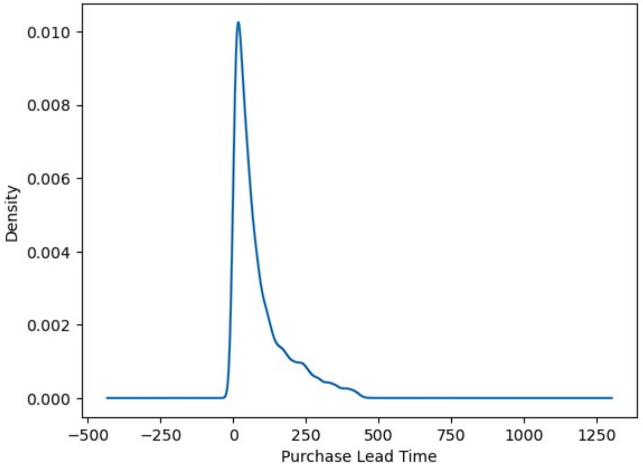
Distribution of Number of Passengers



Box Plot of Flight Duration



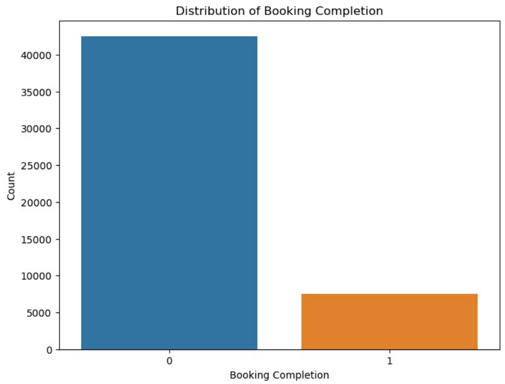
Density Plot of Purchase Lead Time

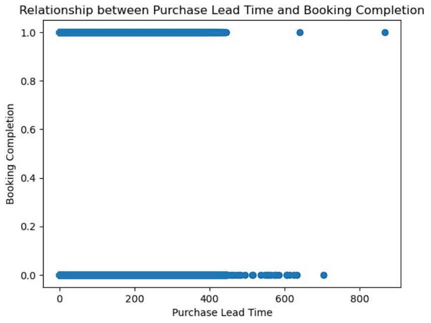


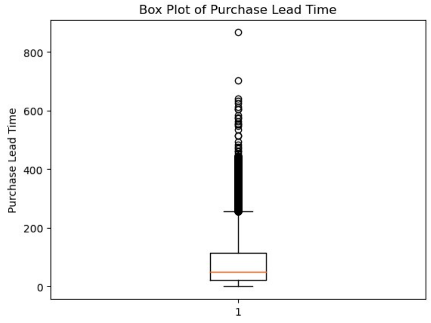
- Here is the distributions andrelationships of the coustomer data.
- The distributions of sales channels internet is high as compared to mobile.
- The pie plot of trip type Here is a count of roundtrip,oneway,circletrip

RoundTrip	49497
OneWay	387
CircleTrip	116

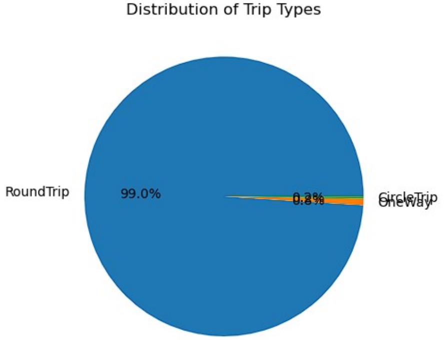
- where as we can see the relationship between purchase lead time and booking completion.They are related to each other the purchase lead time from500 and above is less booked when compared to the purchase lead time 1 to 500.

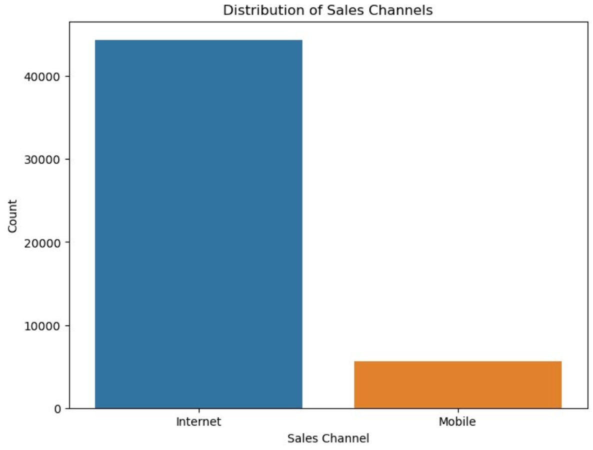

Distribution of Booking Completion

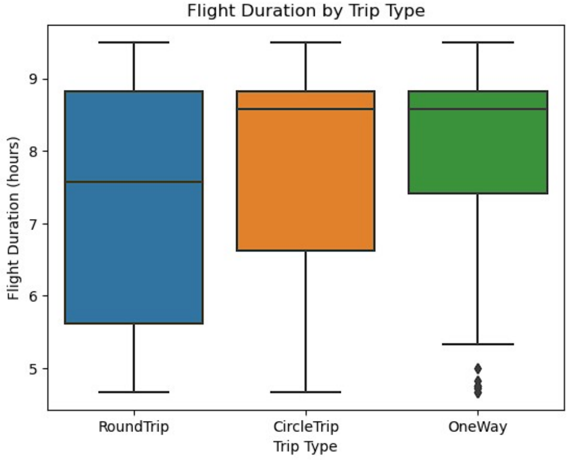

Relationship between Purchase Lead Time and Booking Completion


Box Plot of Purchase Lead Time

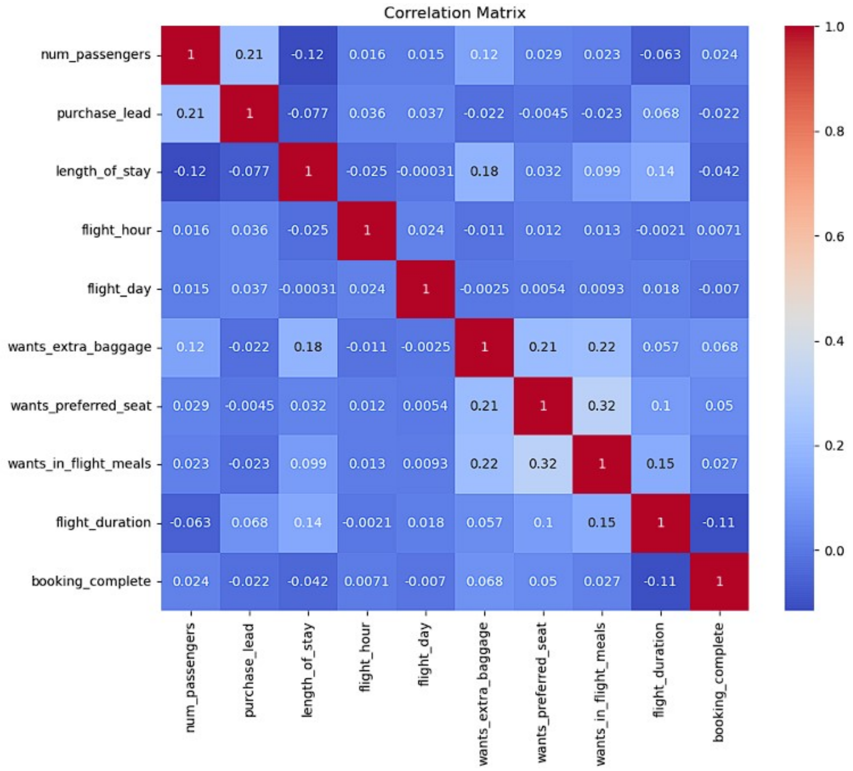
```
Proportion of Completed Bookings:
 0    85.044
1    14.956
Name: booking_complete, dtype: float64
```

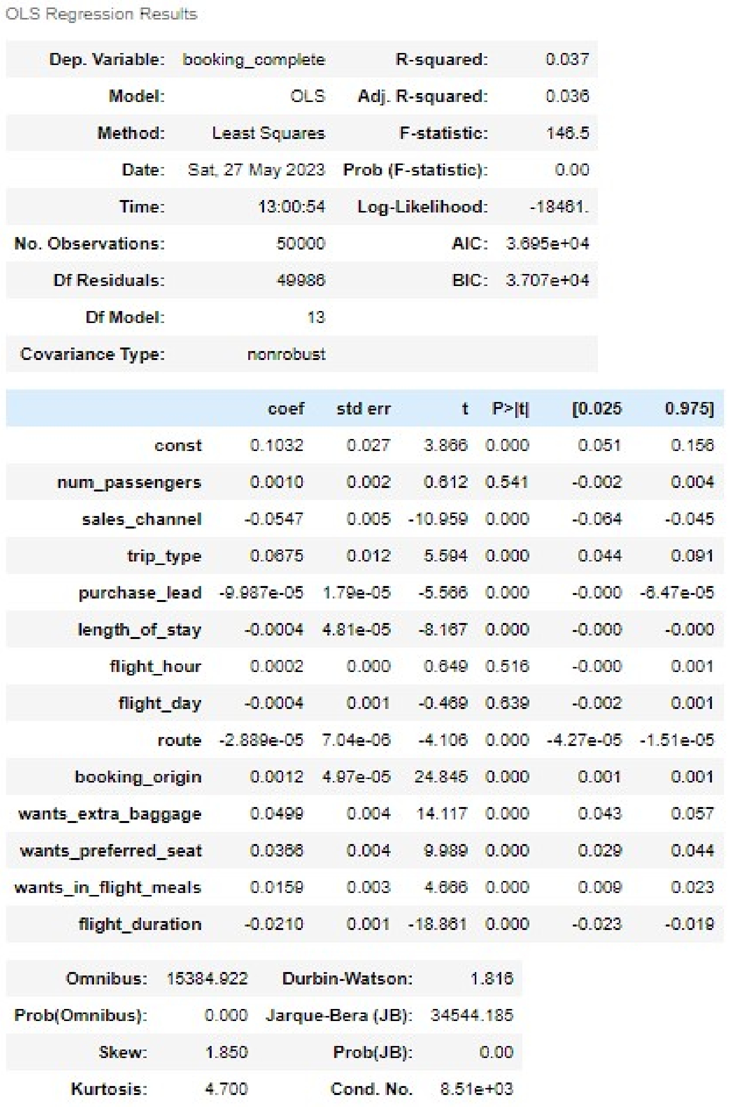

Distribution of Trip Types


Distribution of Sales Channels


Flight Duration by Trip Type

Here is a ols model which is basic statistical model and the correlation plot with heat map.
The R squared score is 0.037 and adj r_squared score is 0.038.
There is no such big relationships between the columns.





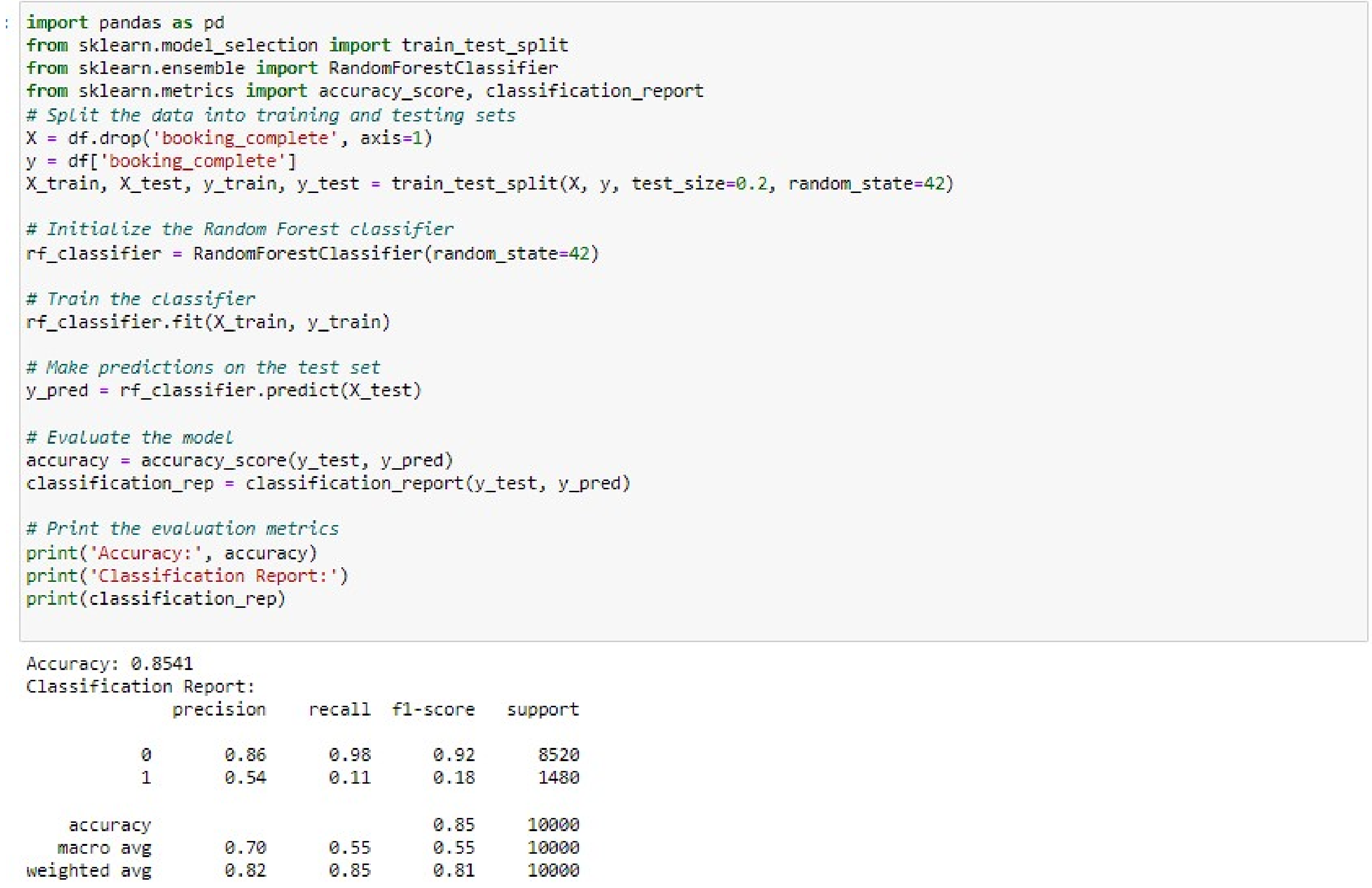
OLS Regression Results

Dep. Variable:	booking_complete	R-squared:	0.037
Model:	OLS	Adj. R-squared:	0.036
Method:	Least Squares	F-statistic:	146.5
Date:	Sat, 27 May 2023	Prob (F-statistic):	0.00
Time:	13:00:54	Log-Likelihood:	-18461.
No. Observations:	50000	AIC:	3.695e+04
Df Residuals:	49986	BIC:	3.707e+04
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1032	0.027	3.866	0.000	0.051	0.156
num_passengers	0.0010	0.002	0.612	0.541	-0.002	0.004
sales_channel	-0.0547	0.005	-10.959	0.000	-0.064	-0.045
trip_type	0.0675	0.012	5.594	0.000	0.044	0.091
purchase_lead	-9.987e-05	1.79e-05	-5.566	0.000	-0.000	-6.47e-05
length_of_stay	-0.0004	4.81e-05	-8.167	0.000	-0.000	-0.000
flight_hour	0.0002	0.000	0.649	0.516	-0.000	0.001
flight_day	-0.0004	0.001	-0.469	0.639	-0.002	0.001
route	-2.889e-05	7.04e-06	-4.106	0.000	-4.27e-05	-1.51e-05
booking_origin	0.0012	4.97e-05	24.845	0.000	0.001	0.001
wants_extra_baggage	0.0499	0.004	14.117	0.000	0.043	0.057
wants_preferred_seat	0.0366	0.004	9.989	0.000	0.029	0.044
wants_in_flight_meals	0.0159	0.003	4.666	0.000	0.009	0.023
flight_duration	-0.0210	0.001	-18.861	0.000	-0.023	-0.019

Omnibus:	15384.922	Durbin-Watson:	1.816
Prob(Omnibus):	0.000	Jarque-Bera (JB):	34544.185
Skew:	1.850	Prob(JB):	0.00
Kurtosis:	4.700	Cond. No.	8.51e+03

The target variable is booking complete.
Here is a machine learning model by the random forest classifier algorithm.
Done train with 80% of data and testing with 20% of data and the random state as 42.
There we can see the classification report with matrics like precision,recall,f1score,accuracy.
The best metric is accuracy it was with 85% accuracy and f1 score of 0’s as 0.92 and 1’s as 0.18.



```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report
# Split the data into training and testing sets
X = df.drop('booking_complete', axis=1)
y = df['booking_complete']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Initialize the Random Forest classifier
rf_classifier = RandomForestClassifier(random_state=42)

# Train the classifier
rf_classifier.fit(X_train, y_train)

# Make predictions on the test set
y_pred = rf_classifier.predict(X_test)

# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
classification_rep = classification_report(y_test, y_pred)

# Print the evaluation metrics
print('Accuracy:', accuracy)
print('Classification Report:')
print(classification_rep)
```

```
Accuracy: 0.8541
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.98      0.92      8520
           1       0.54      0.11      0.18      1480

    accuracy                           0.85     10000
   macro avg       0.70      0.55      0.55     10000
weighted avg       0.82      0.85      0.81     10000
```