Task 2

Team O

Overview

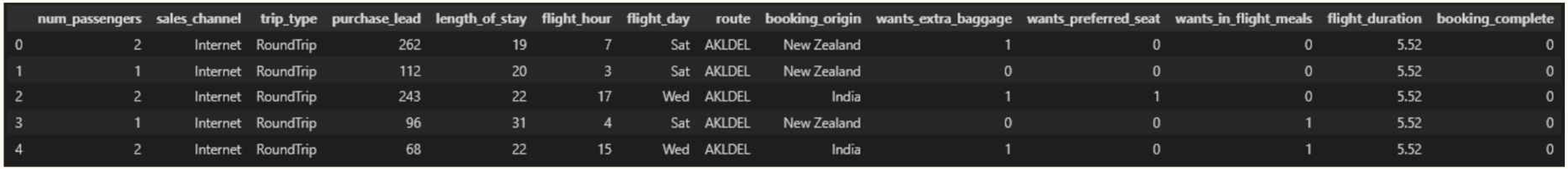
- Problem Statement
- Raw Data
- Data Pre-Processing
- Outlier Handling
- Correlation Matrix
- EDA
- Performance Metrics
- Feature Importance

Problem Statement

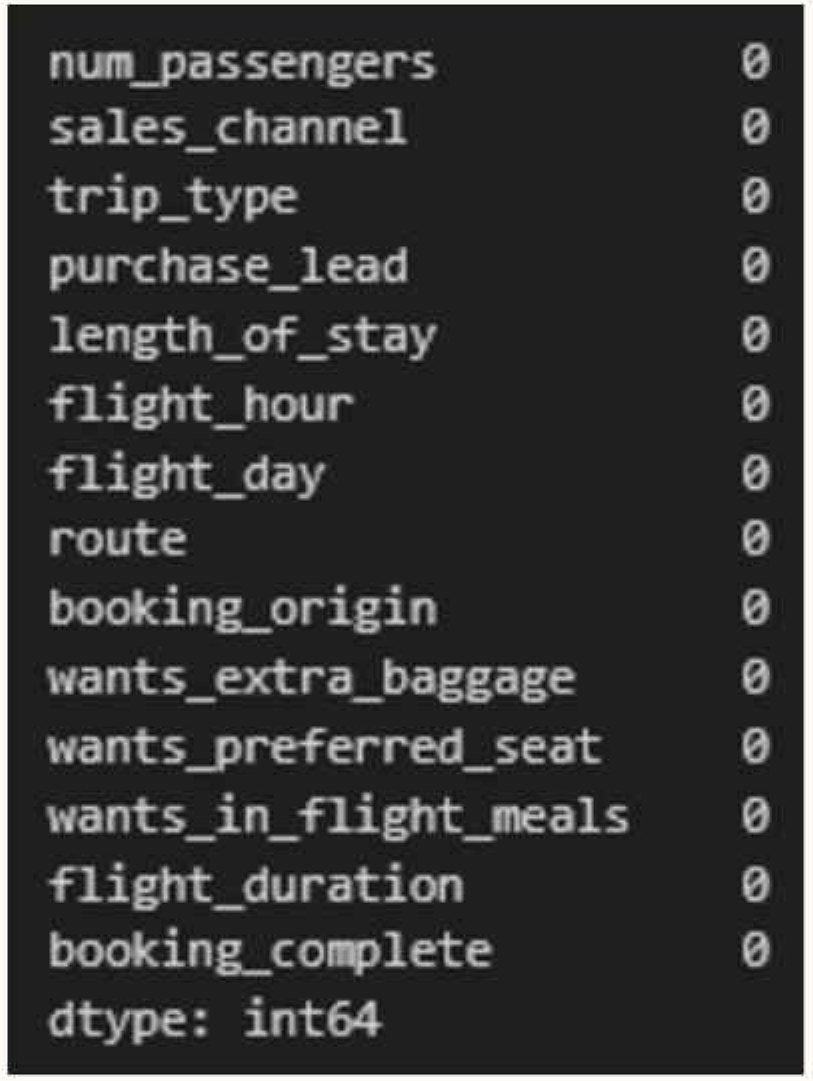
Objectives

- Explore and prepare the dataset
- Train a machine learning model
- Evaluate model and present findings

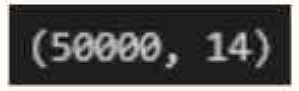
Raw Data



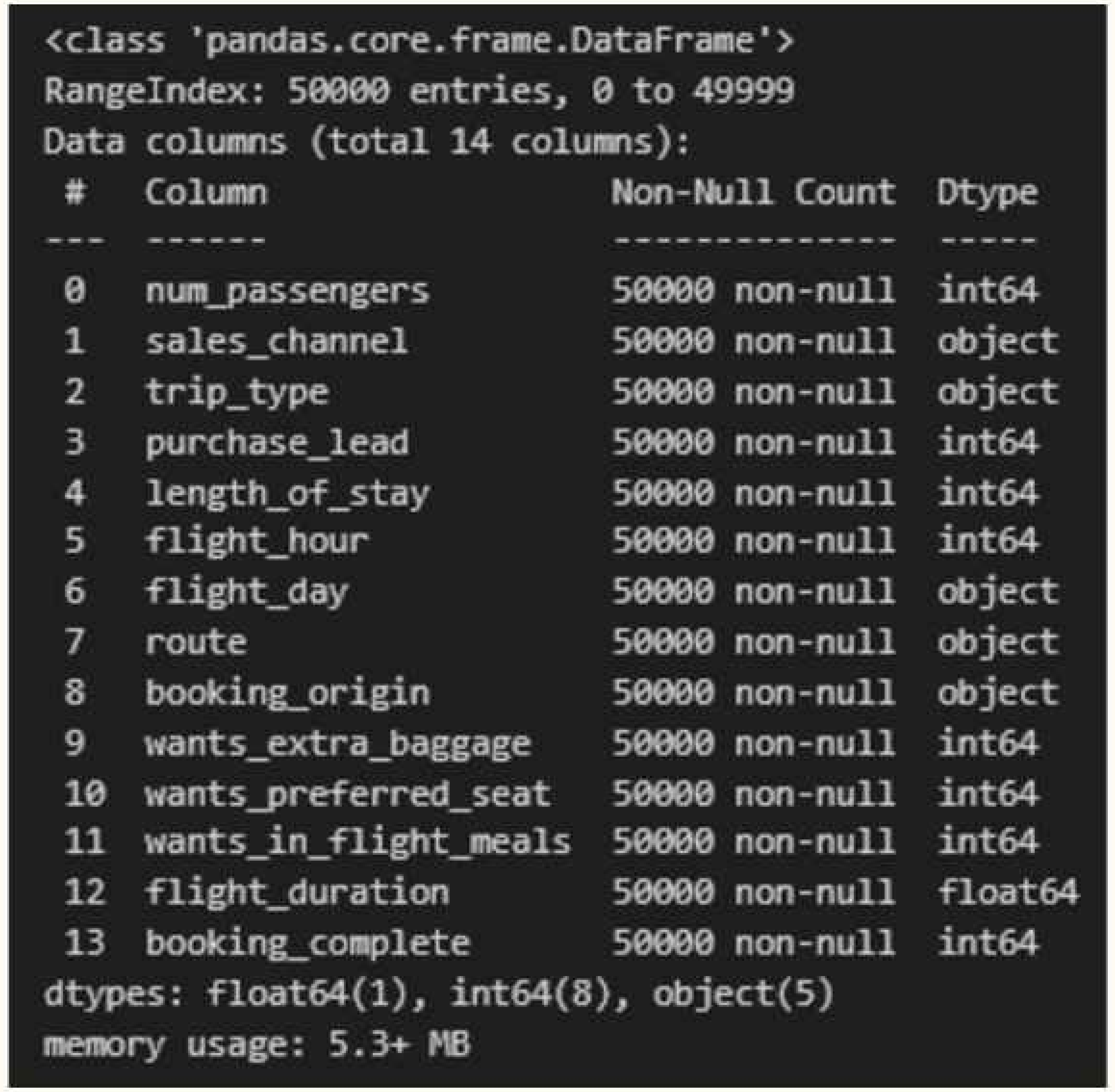
	num_passengers	sales_channel	trip_type	purchase_lead	length_of_stay	flight_hour	flight_day	route	booking_origin	wants_extra_baggage	wants_preferred_seat	wants_in_flight_meals	flight_duration	booking_complete
0	2	Internet	RoundTrip	262	19	7	Sat	AKLDEL	New Zealand	1	0	0	5.52	0
1	1	Internet	RoundTrip	112	20	3	Sat	AKLDEL	New Zealand	0	0	0	5.52	0
2	2	Internet	RoundTrip	243	22	17	Wed	AKLDEL	India	1	1	0	5.52	0
3	1	Internet	RoundTrip	96	31	4	Sat	AKLDEL	New Zealand	0	0	1	5.52	0
4	2	Internet	RoundTrip	68	22	15	Wed	AKLDEL	India	1	0	1	5.52	0



```
num_passengers           0
sales_channel            0
trip_type                0
purchase_lead            0
length_of_stay           0
flight_hour              0
flight_day               0
route                    0
booking_origin           0
wants_extra_baggage      0
wants_preferred_seat     0
wants_in_flight_meals    0
flight_duration          0
booking_complete         0
dtype: int64
```

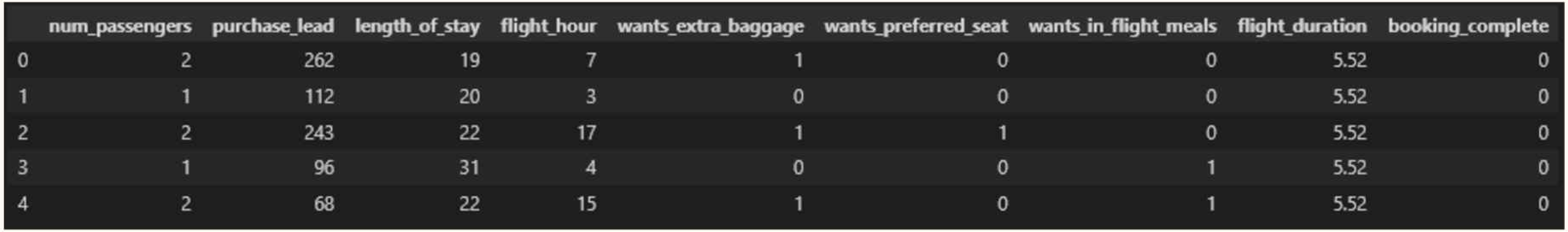


```
(50000, 14)
```

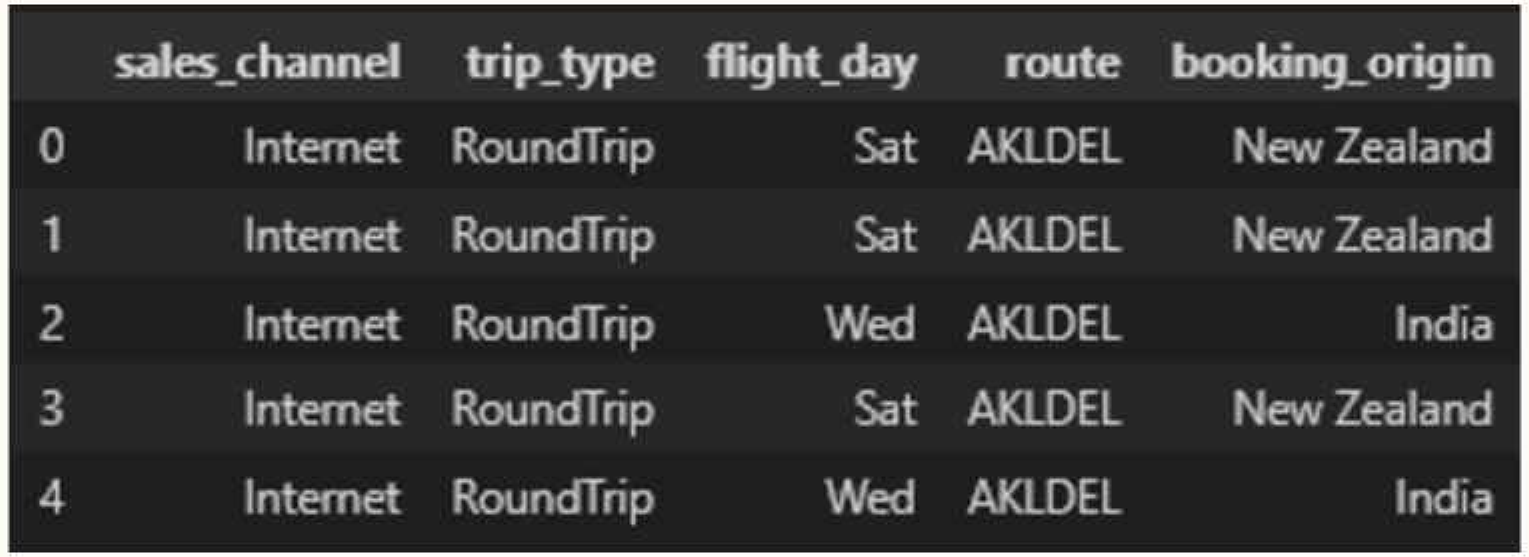


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50000 entries, 0 to 49999
Data columns (total 14 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   num_passengers         50000 non-null  int64  
 1   sales_channel          50000 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
```

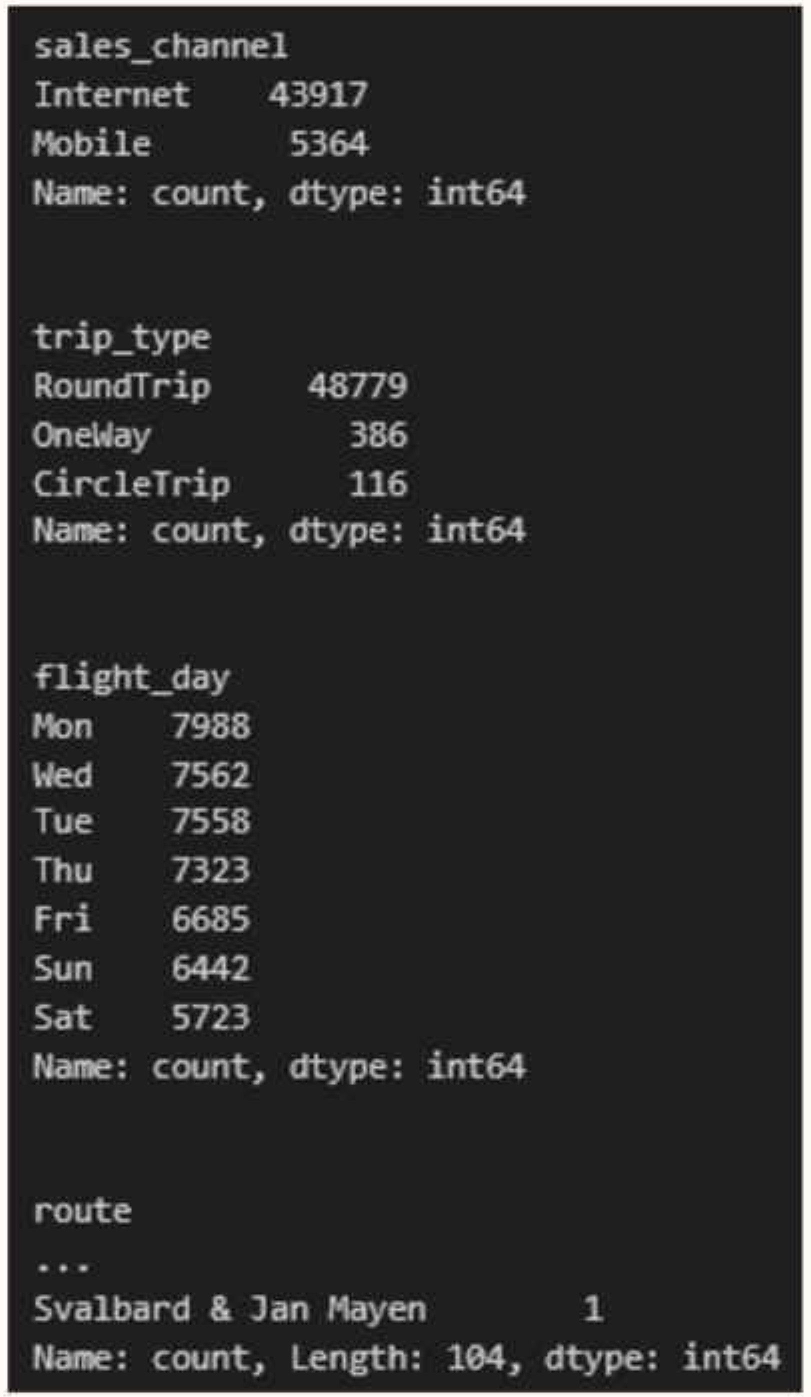
Data Pre-Processing



	num_passengers	purchase_lead	length_of_stay	flight_hour	wants_extra_baggage	wants_preferred_seat	wants_in_flight_meals	flight_duration	booking_complete
0	2	262	19	7	1	0	0	5.52	0
1	1	112	20	3	0	0	0	5.52	0
2	2	243	22	17	1	1	0	5.52	0
3	1	96	31	4	0	0	1	5.52	0
4	2	68	22	15	1	0	1	5.52	0

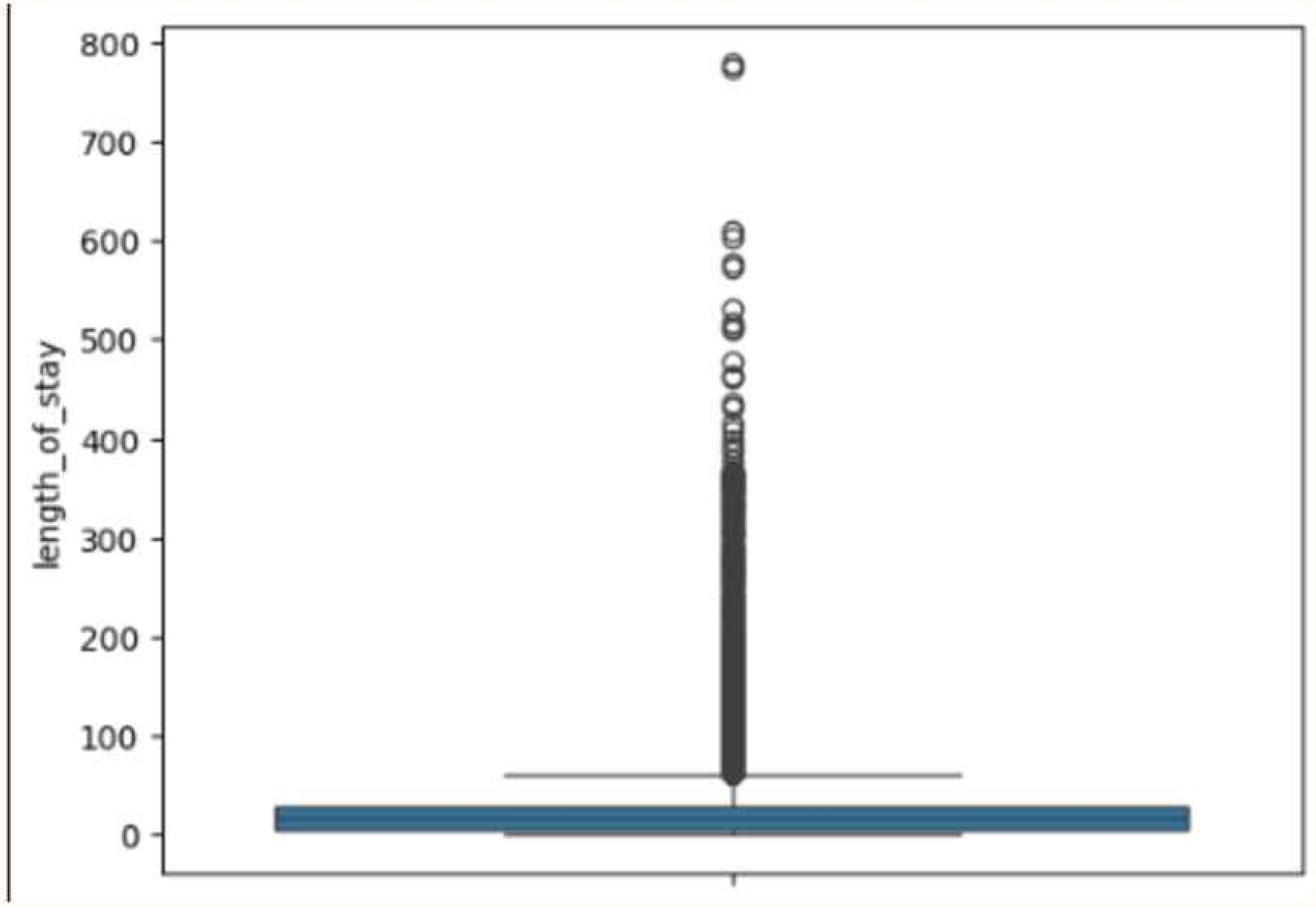


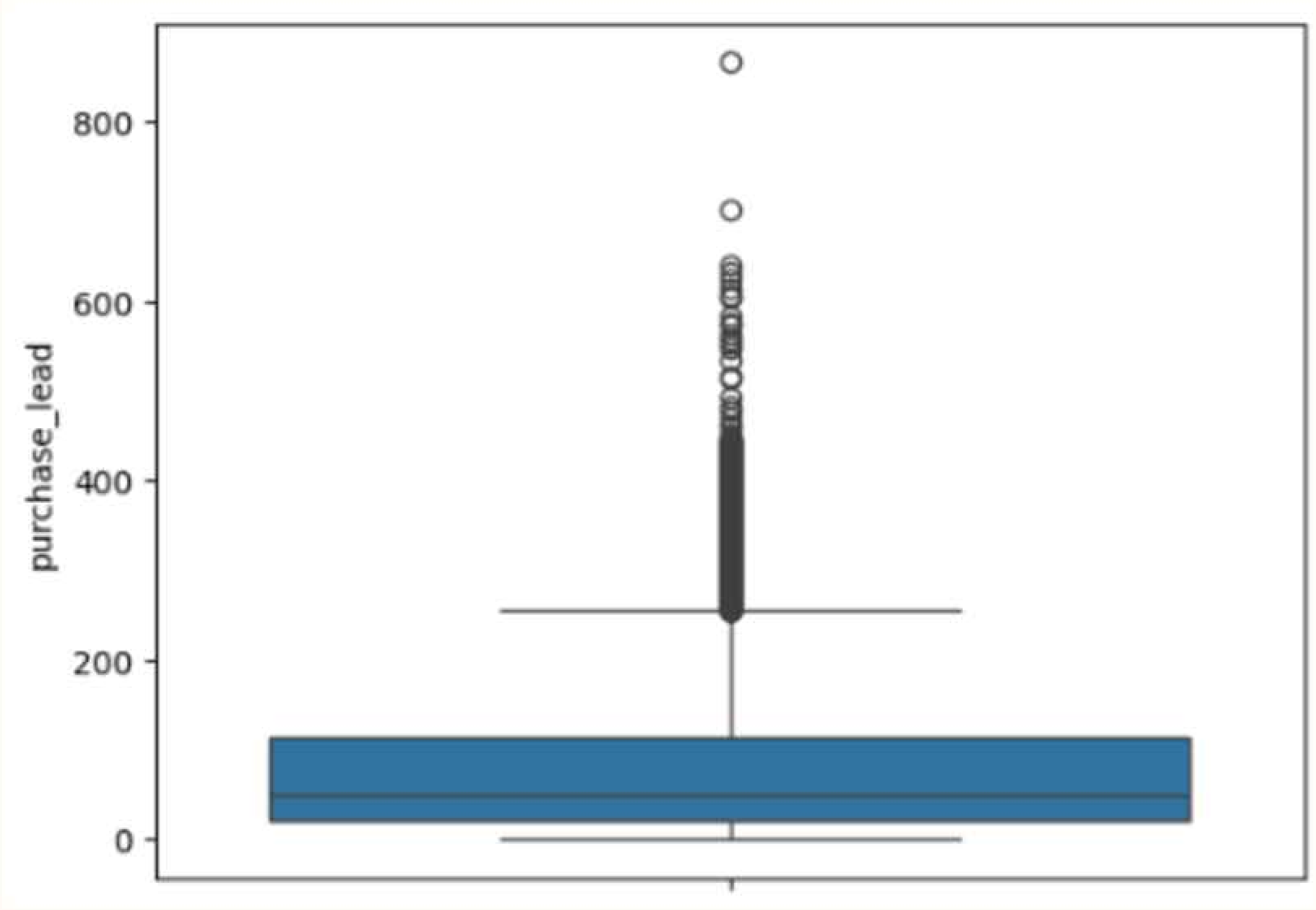
	sales_channel	trip_type	flight_day	route	booking_origin
0	Internet	RoundTrip	Sat	AKLDEL	New Zealand
1	Internet	RoundTrip	Sat	AKLDEL	New Zealand
2	Internet	RoundTrip	Wed	AKLDEL	India
3	Internet	RoundTrip	Sat	AKLDEL	New Zealand
4	Internet	RoundTrip	Wed	AKLDEL	India

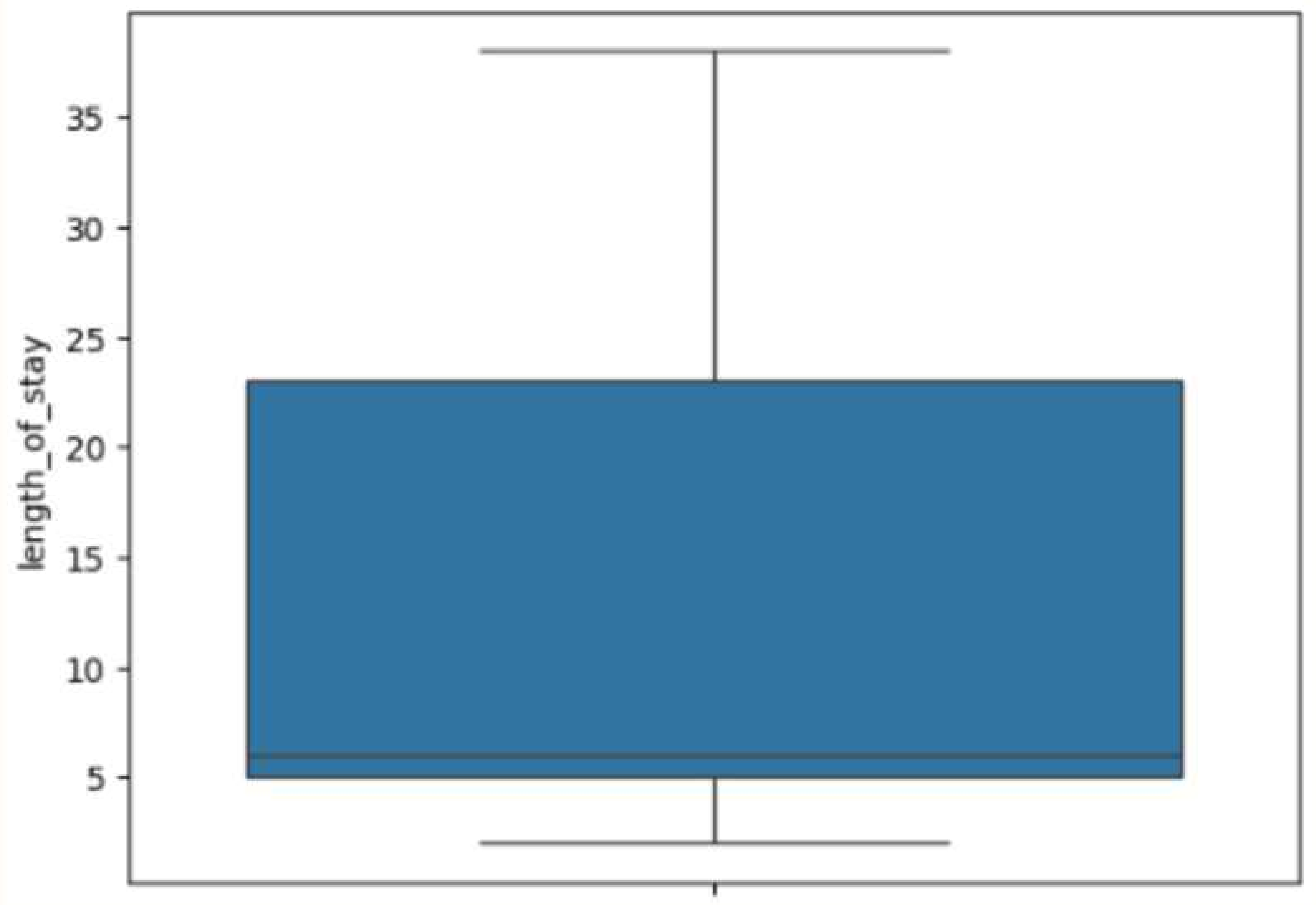


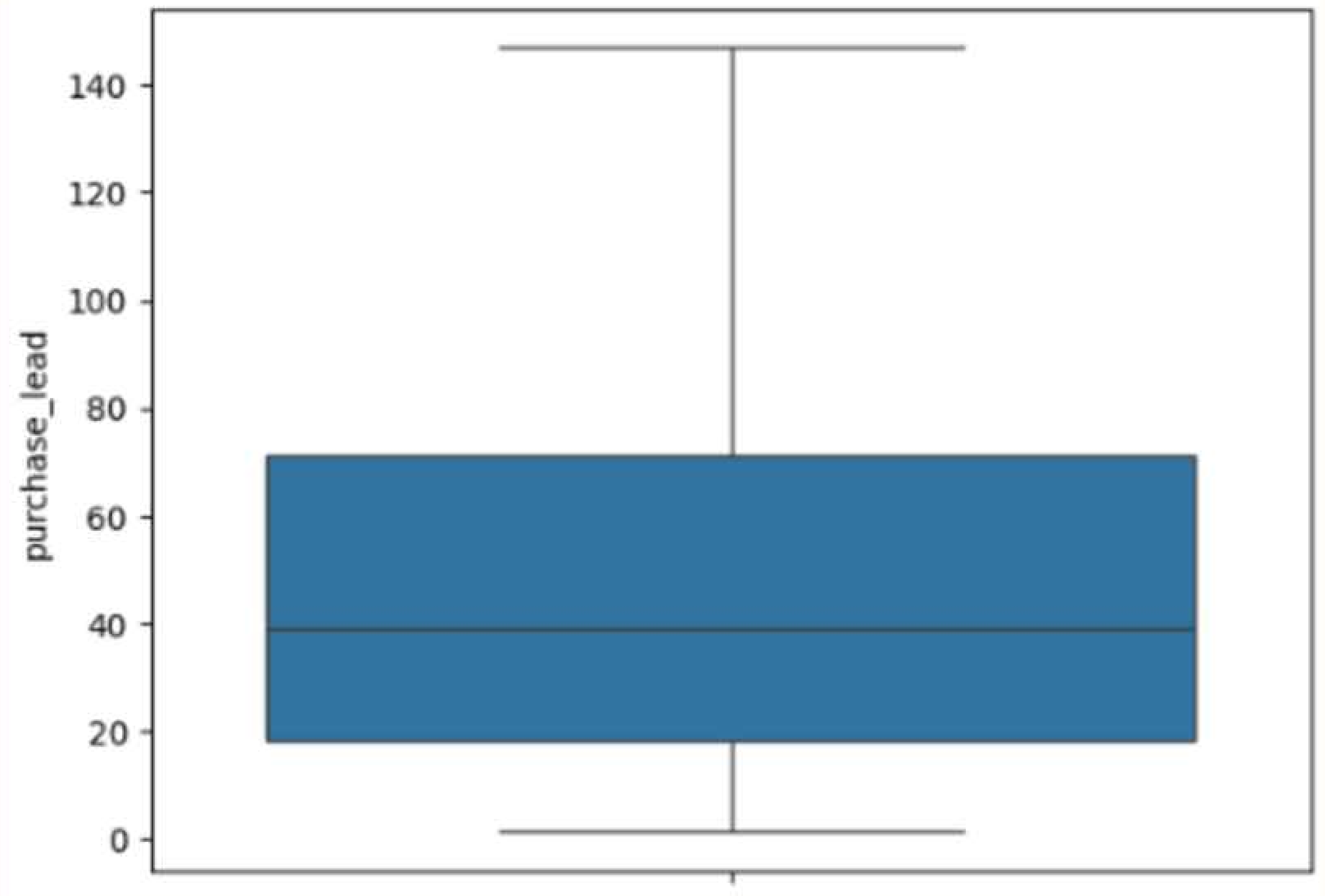
```
sales_channel
Internet    43917
Mobile       5364
Name: count, dtype: int64


trip_type
RoundTrip     48779
OneWay          386
CircleTrip      116
Name: count, dtype: int64


flight_day
Mon    7988
Wed    7562
Tue    7558
Thu    7323
Fri    6685
Sun    6442
Sat    5723
Name: count, dtype: int64


route
...
Svalbard & Jan Mayen        1
Name: count, Length: 104, dtype: int64
```

Outlier Handling

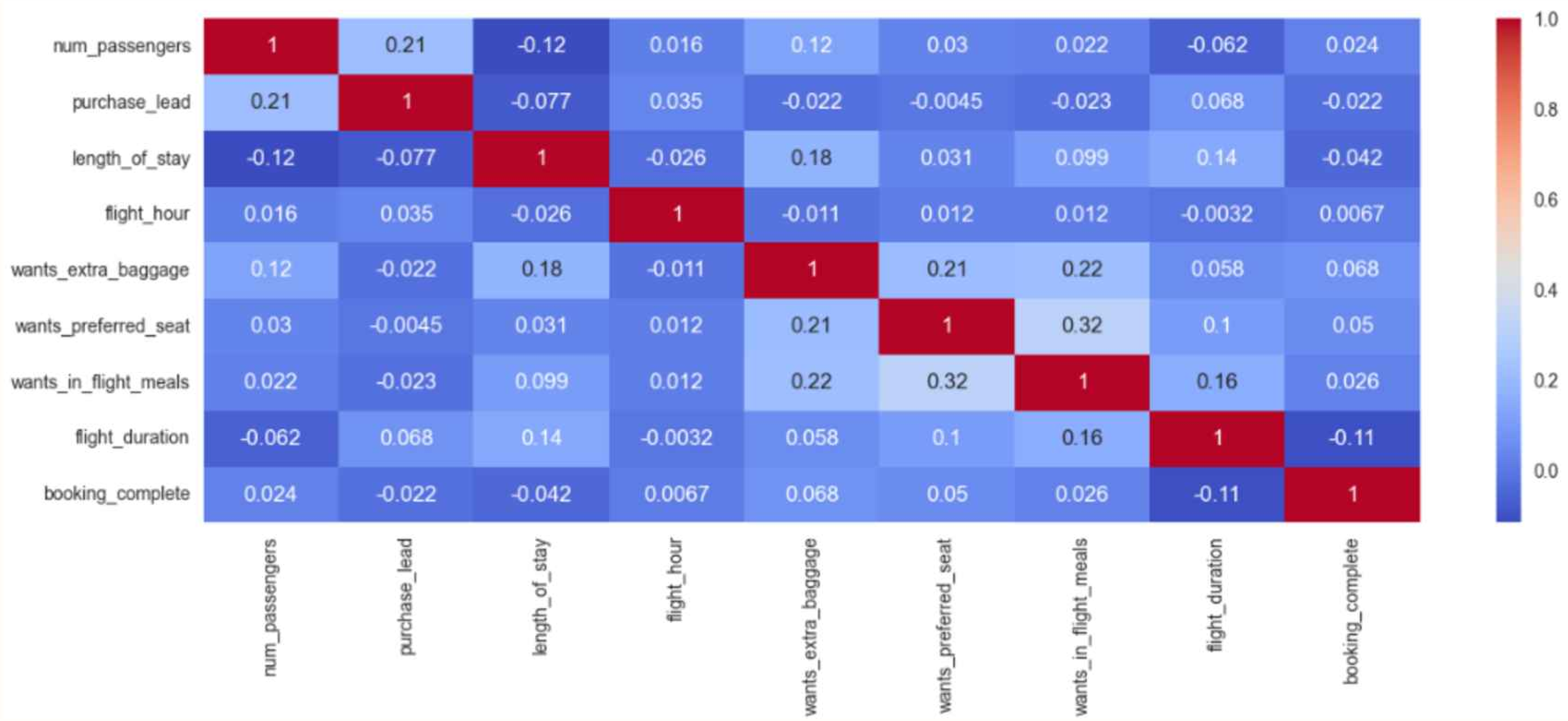

length_of_stay
800
700
600
500
400
300
200
100
0


purchase_lead
800
600
400
200
0


length_of_stay
35
30
25
20
15
10
5

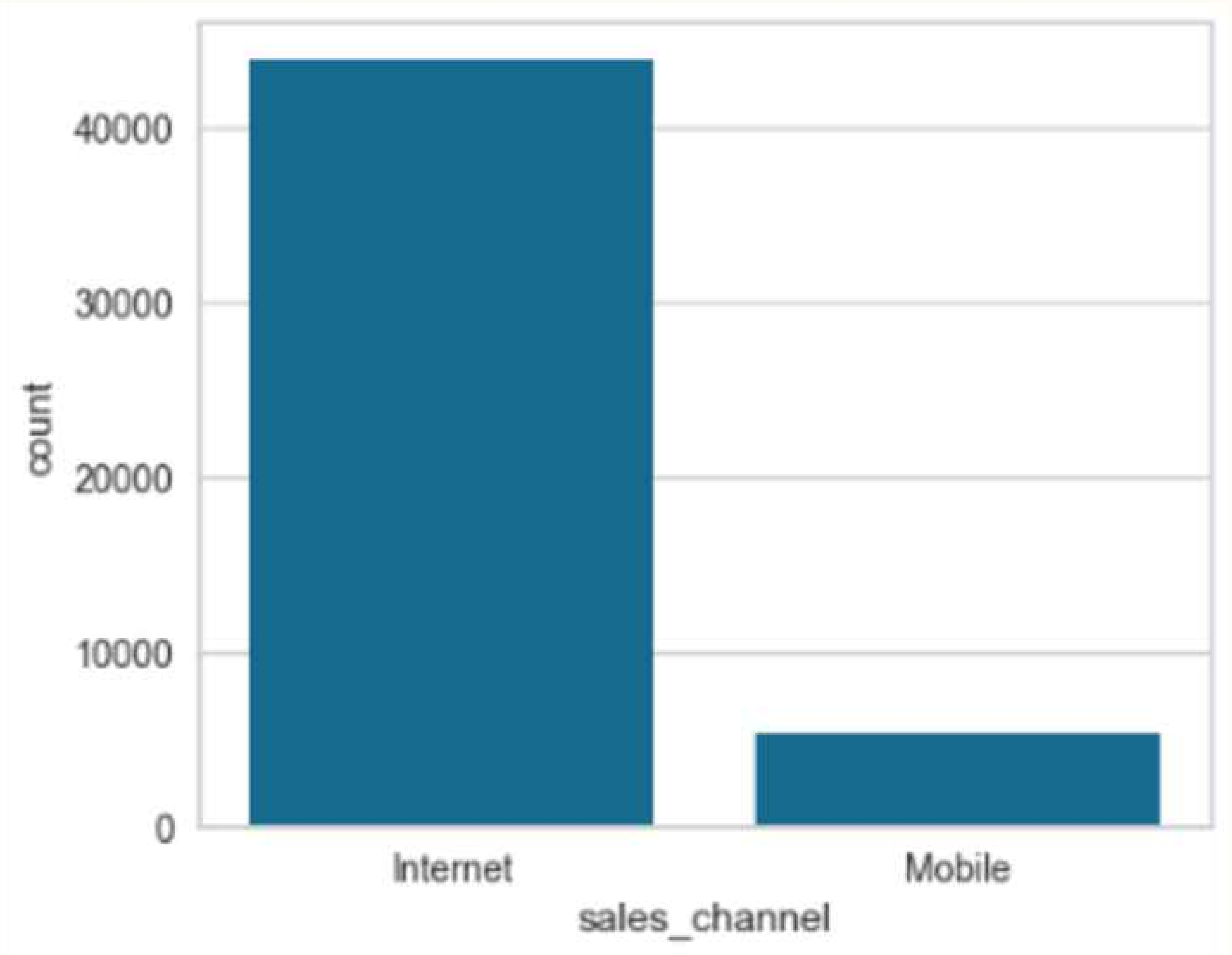

purchase_lead
140
120
100
80
60
40
20
0

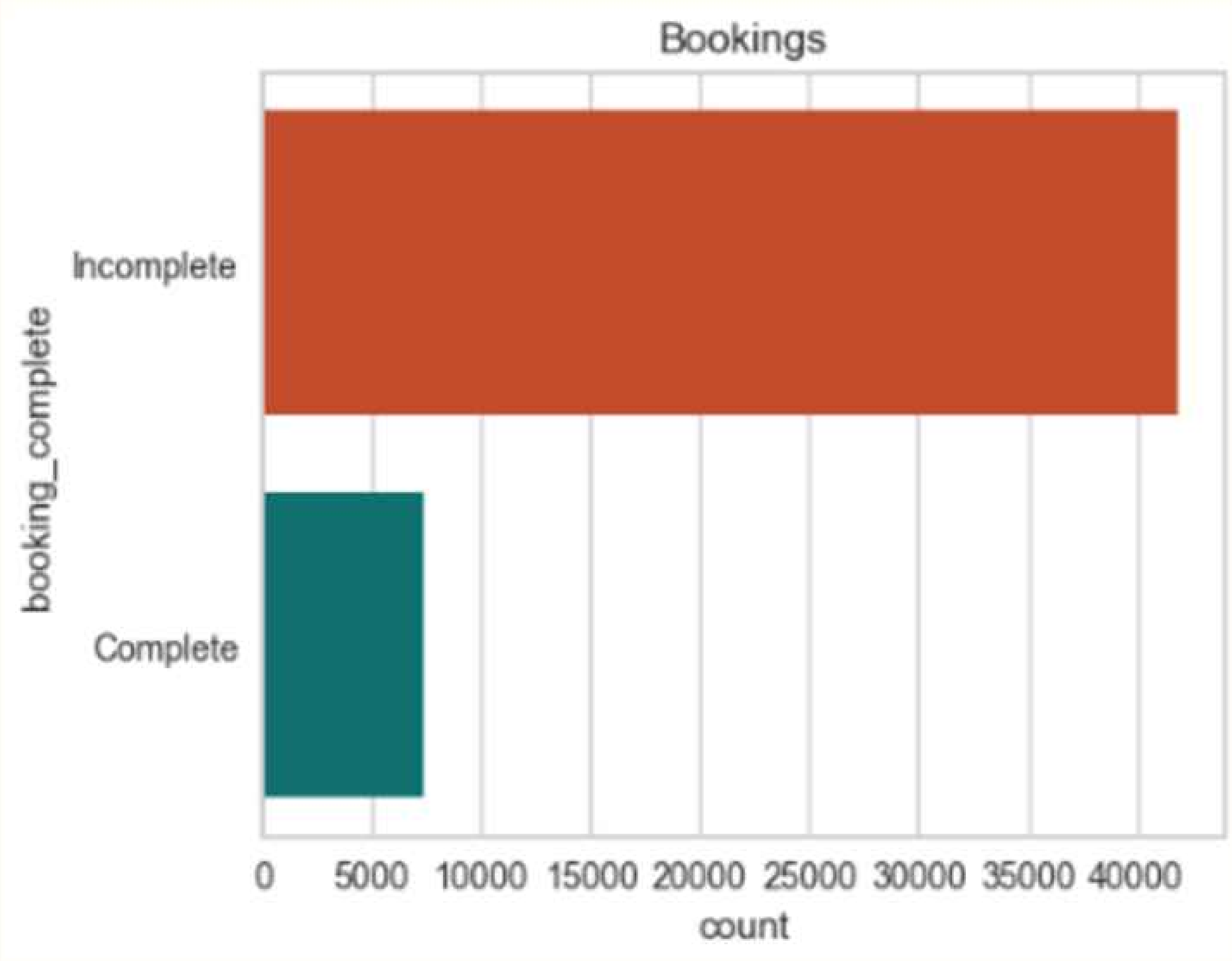
Correlation Matrix

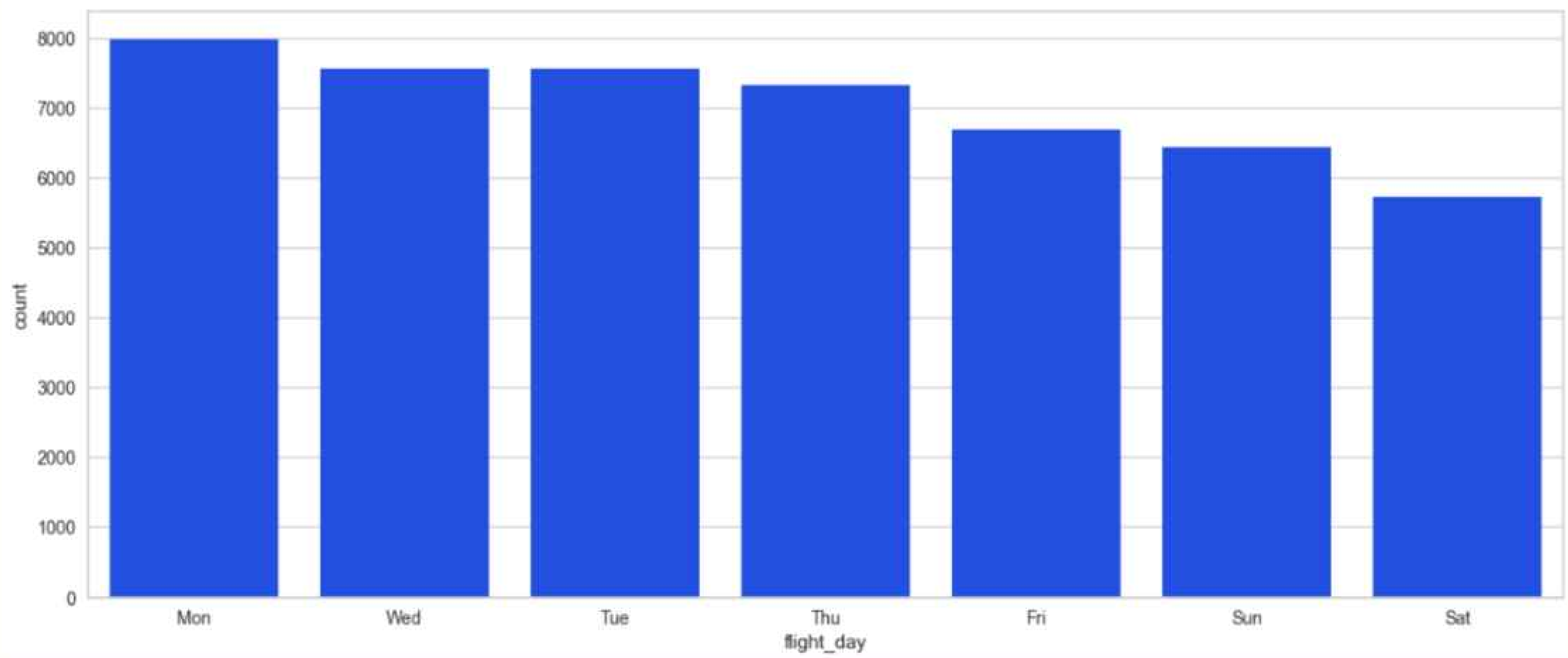


	num_passengers	purchase_lead	length_of_stay	flight_hour	wants_extra_baggage	wants_preferred_seat	wants_in_flight_meals	flight_duration	booking_complete
num_passengers	1	0.21	-0.12	0.016	0.12	0.03	0.022	-0.062	0.024
purchase_lead	0.21	1	-0.077	0.035	-0.022	-0.0045	-0.023	0.068	-0.022
length_of_stay	-0.12	-0.077	1	-0.026	0.18	0.031	0.099	0.14	-0.042
flight_hour	0.016	0.035	-0.026	1	-0.011	0.012	0.012	-0.0032	0.0067
wants_extra_baggage	0.12	-0.022	0.18	-0.011	1	0.21	0.22	0.058	0.068
wants_preferred_seat	0.03	-0.0045	0.031	0.012	0.21	1	0.32	0.1	0.05
wants_in_flight_meals	0.022	-0.023	0.099	0.012	0.22	0.32	1	0.16	0.026
flight_duration	-0.062	0.068	0.14	-0.0032	0.058	0.1	0.16	1	-0.11
booking_complete	0.024	-0.022	-0.042	0.0067	0.068	0.05	0.026	-0.11	1

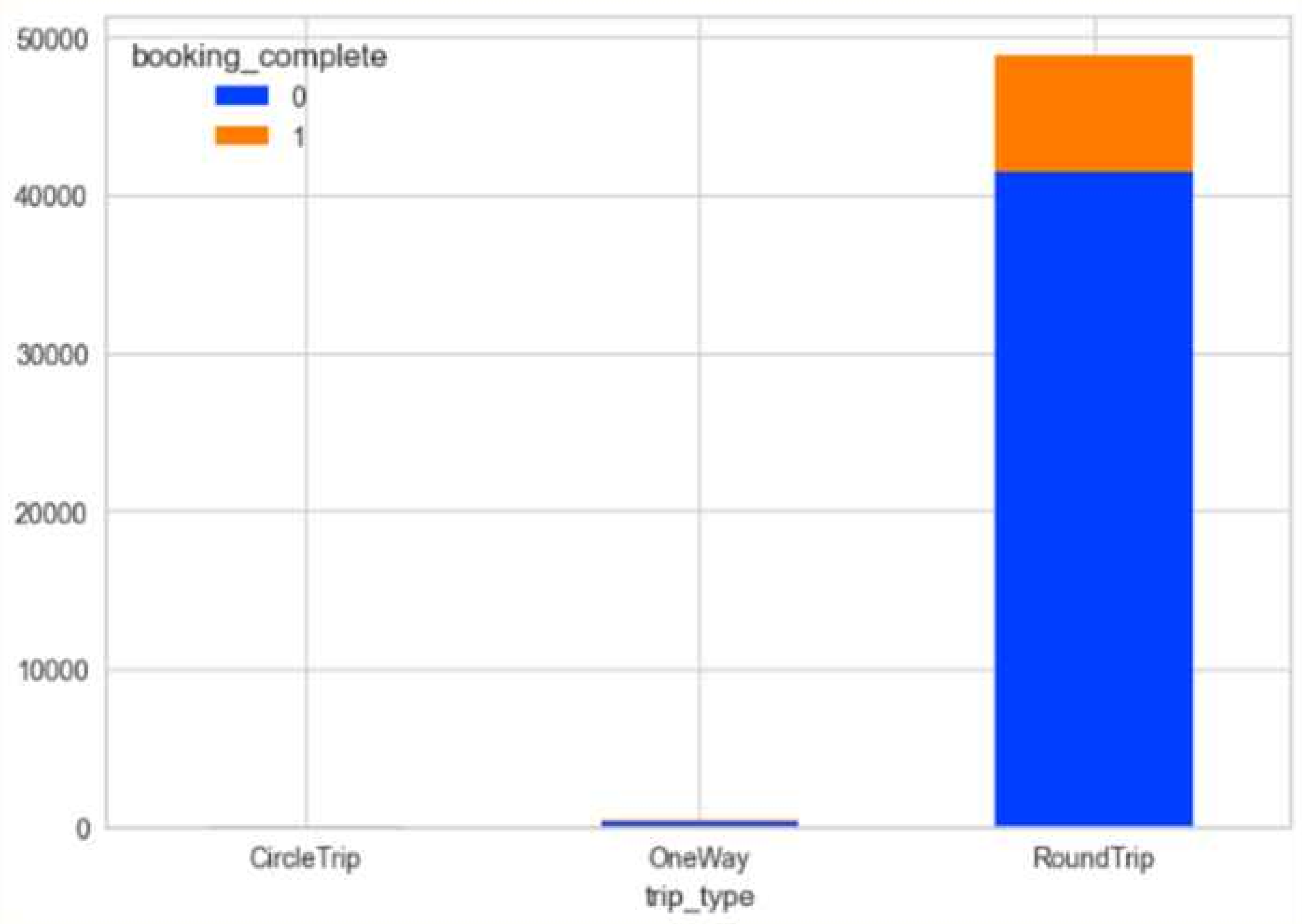
EDA


40000
30000
20000
10000
0
count
Internet
Mobile
sales_channel

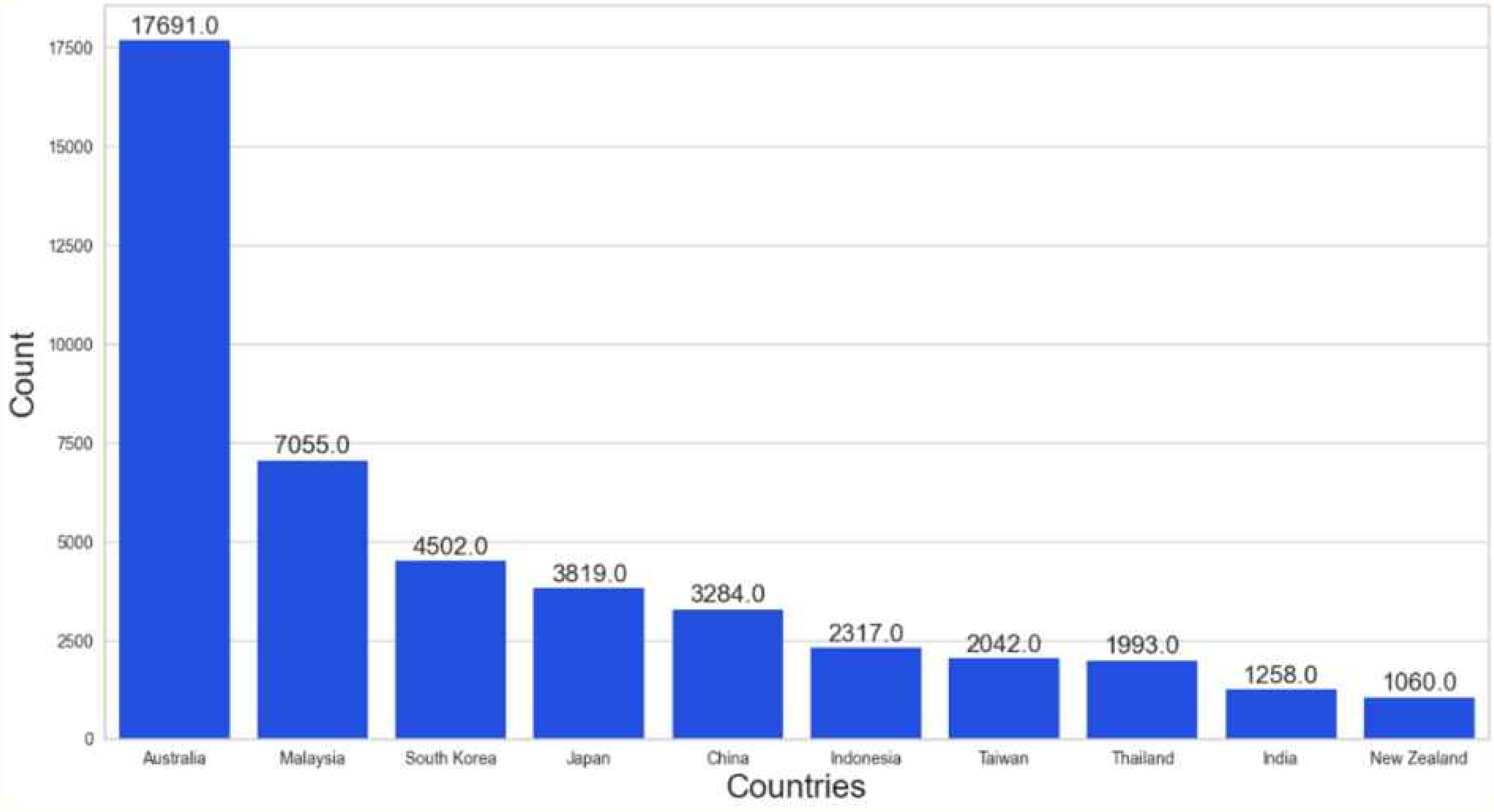

Bookings
booking_complete
Incomplete
Complete
0
5000
10000
15000
20000
25000
30000
35000
40000
count

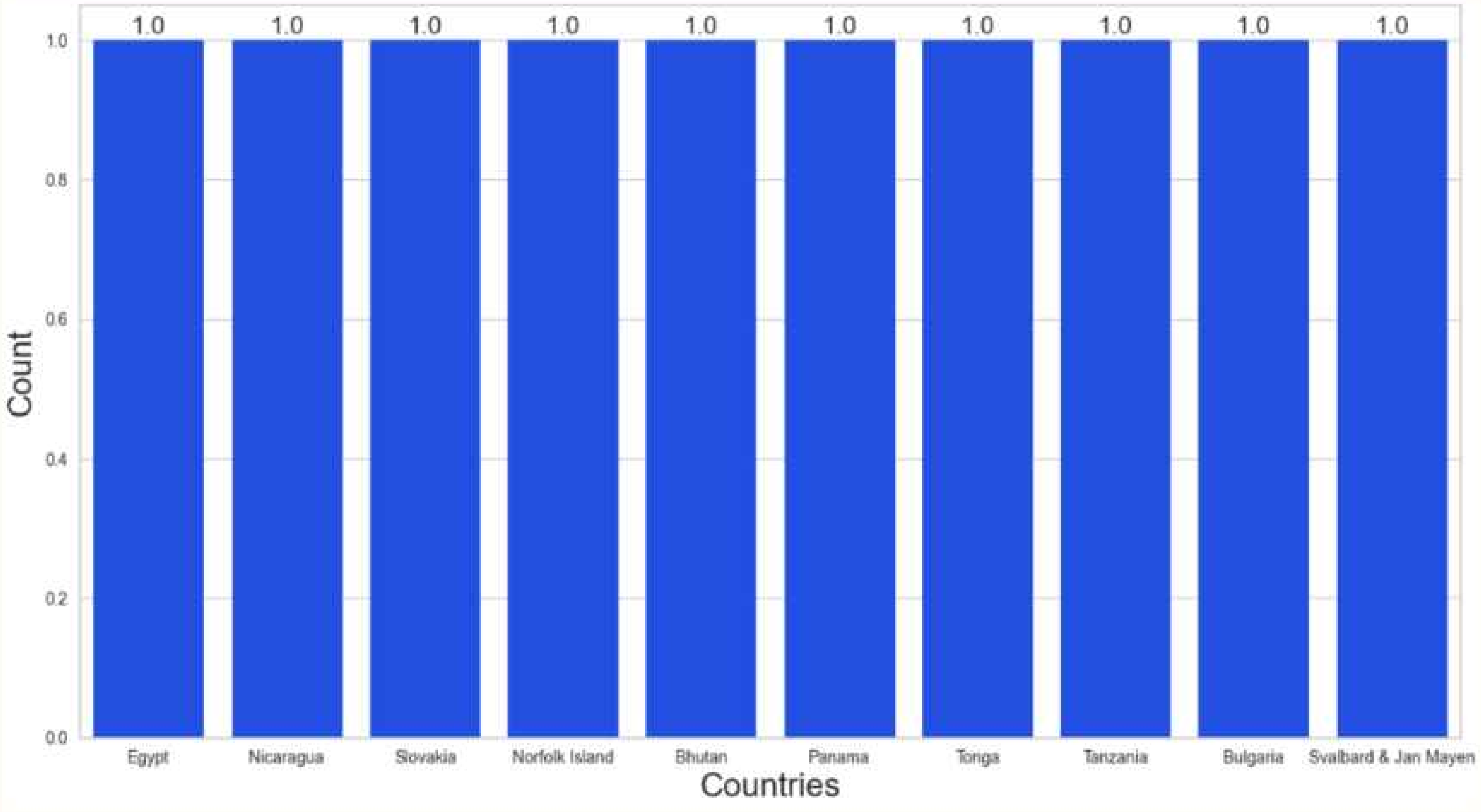


count
8000
7000
6000
5000
4000
3000
2000
1000
0
Mon
Wed
Tue
Thu
Fri
Sun
Sat
flight_day

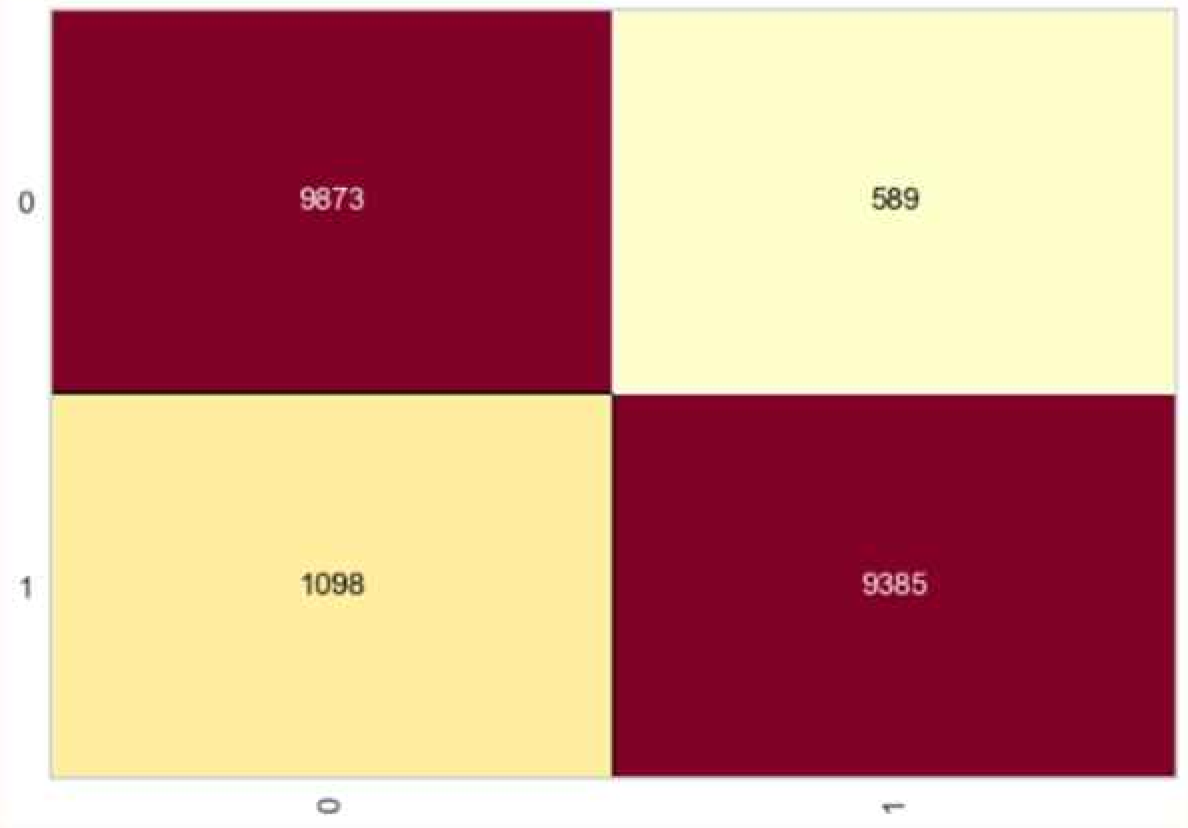


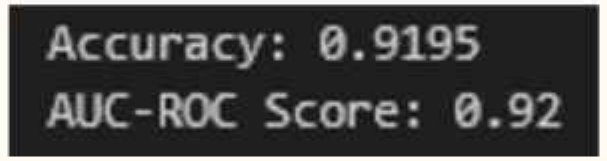
booking_complete
0
1
50000
40000
30000
20000
10000
0
CircleTrip
OneWay
RoundTrip
trip_type


Count
17500
15000
12500
10000
7500
5000
2500
0
17691.0
7055.0
4502.0
3819.0
3284.0
2317.0
2042.0
1993.0
1258.0
1060.0
Australia
Malaysia
South Korea
Japan
China
Indonesia
Taiwan
Thailand
India
New Zealand
Countries

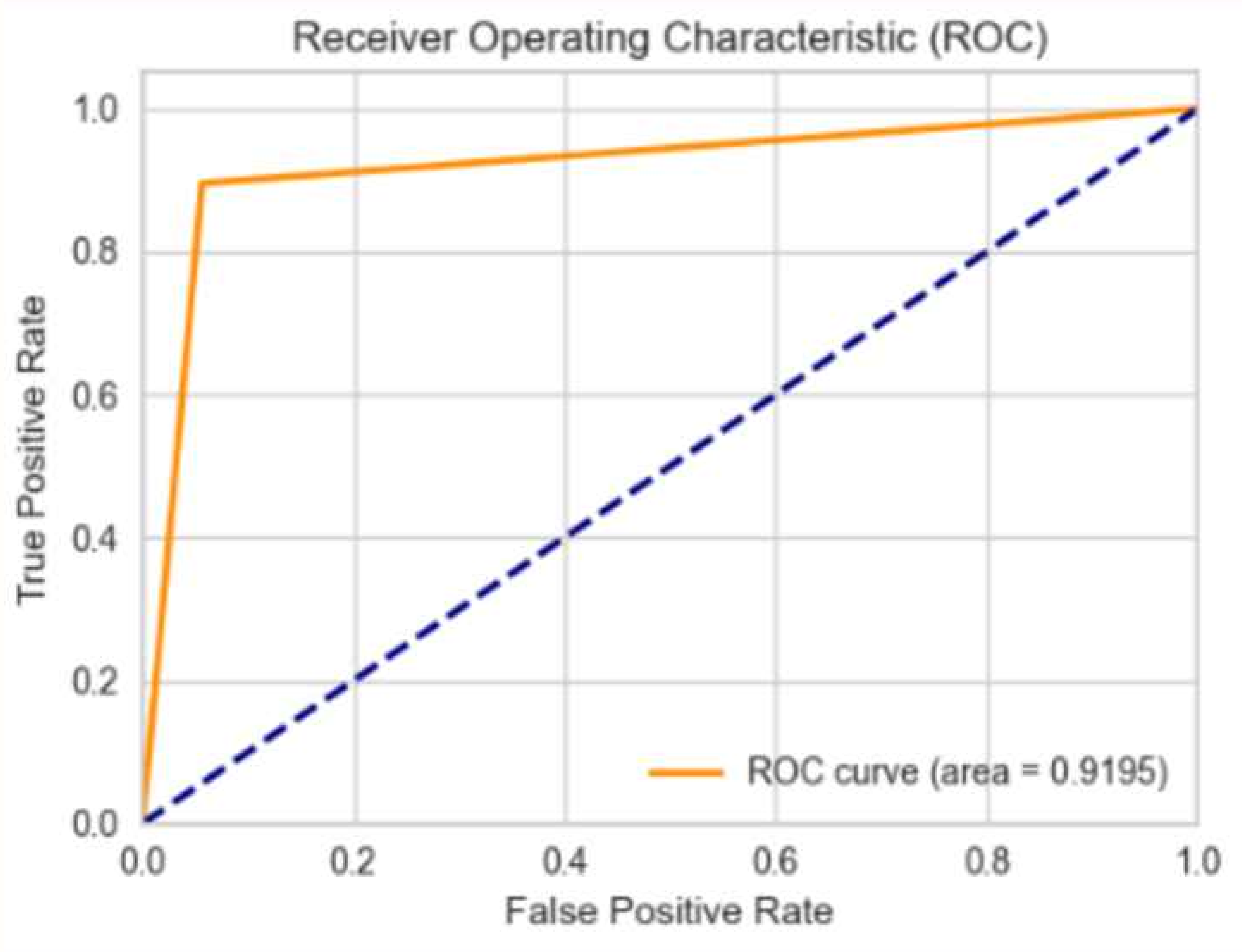

Count
1.0
0.8
0.6
0.4
0.2
0.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
Egypt
Nicaragua
Slovakia
Norfolk Island
Bhutan
Panama
Tonga
Tanzania
Bulgaria
Svalbard & Jan Mayen
Countries

Performance Metrics

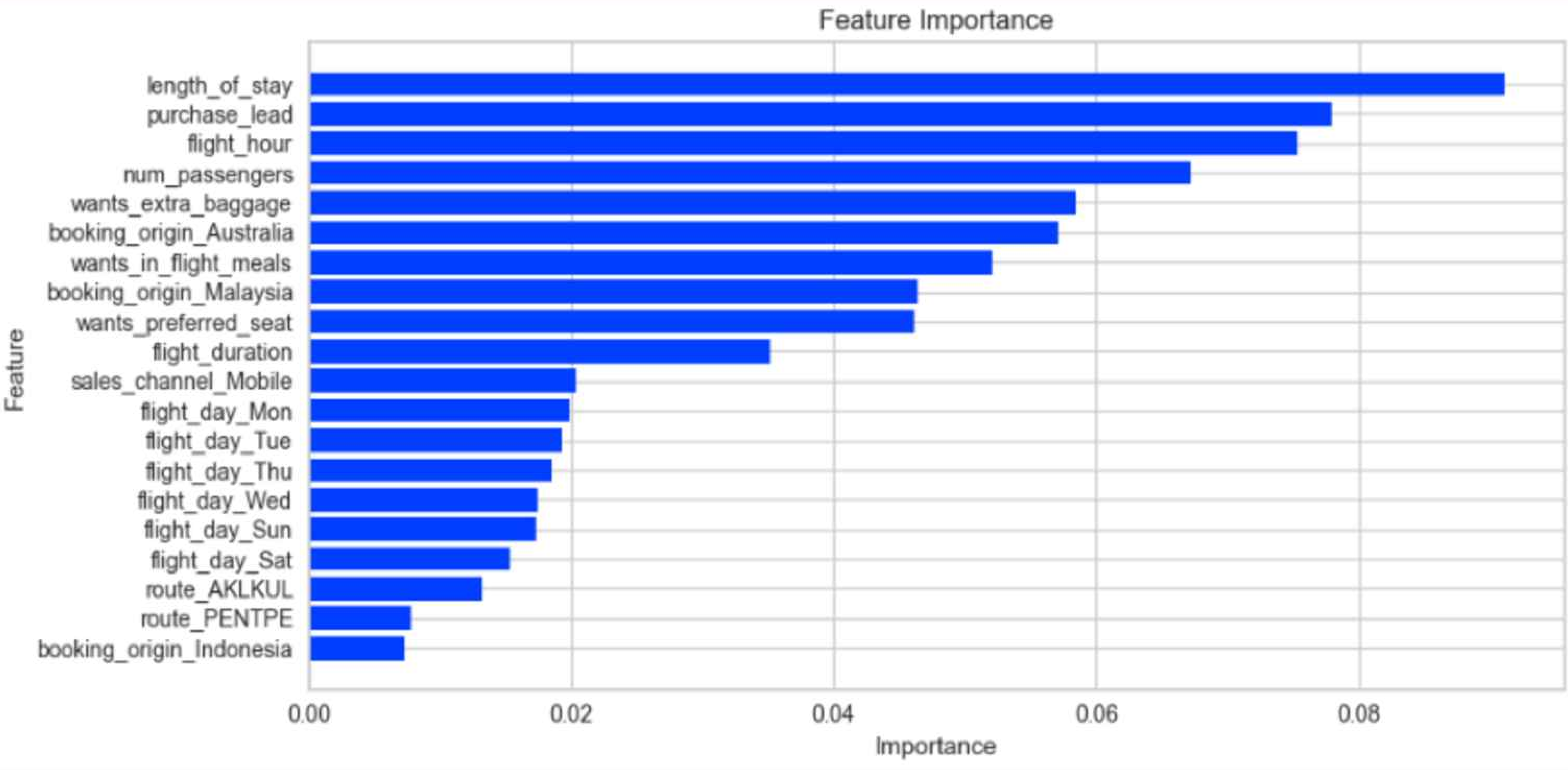




```
Accuracy: 0.9195
AUC-ROC Score: 0.92
```



Feature Importance


Feature Importance
length_of_stay
purchase_lead
flight_hour
num_passengers
wants_extra_baggage
booking_origin_Australia
wants_in_flight_meals
booking_origin_Malaysia
wants_preferred_seat
flight_duration
sales_channel_Mobile
flight_day_Mon
flight_day_Tue
flight_day_Thu
flight_day_Wed
flight_day_Sun
flight_day_Sat
route_AKLKUL
route_PENTPE
booking_origin_Indonesia
Feature
0.00
0.02
0.04
0.06
0.08
Importance

Thank You