FLIGHT PRICE PREDICTION PROJECT

Submitted by:

ANISH ANTONY

ABSTRACT:

Traveling in an airplane is one of the most exciting and expensive modes of travel. Since it is expensive, we must plan well accordingly. In this project we will see how the ticket vary based on advance booking of the flight and also how the price varies as the dates come closer. In this project we will be predicting the flight price from machine learning models. To predict this model, we need to scrape the flight booking details for two months from online websites using selenium web driver, and the data have preprocessed, trained and tested using the regression algorithms. Then it is hypertuned and best algorithm with best parameters is obtained and finally the ticket price of the flight is predicted.

Keywords: Flight ticket ,Selenium ,Data cleaning ,Ticket Price ,Regression

The features used in the dataset are:

- Flight Company
- Flight No
- Flight Class
- Flight Model
- Travel
- Departure Time
- Departure Day
- Departure Date
- Departure Month
- Departure Duration
- Arrival Time
- Arrival Day
- Meal
- Flight Ticket Price
- Luggage
- Connecting Planes
- No of days

Data Cleaning:

- Flight Company – Grouped as str
- Flight No – Grouped as str
- Flight Class – Grouped as str
- Flight Model – Grouped as str
- Travel – Combined source and destination grouped as str
- Departure Time – Converted to hours
- Departure Day – Grouped as str
- Departure Date – Grouped as str
- Departure Month – Grouped as str
- Departure Duration – Converted to minutes as int
- Arrival Time – Grouped as str
- Arrival Day – Grouped as str
- Meal – Grouped as str
- Flight Ticket Price – Grouped as int
- Luggage – Grouped as str
- Connecting Planes – Grouped as str
- No of days – Grouped as int

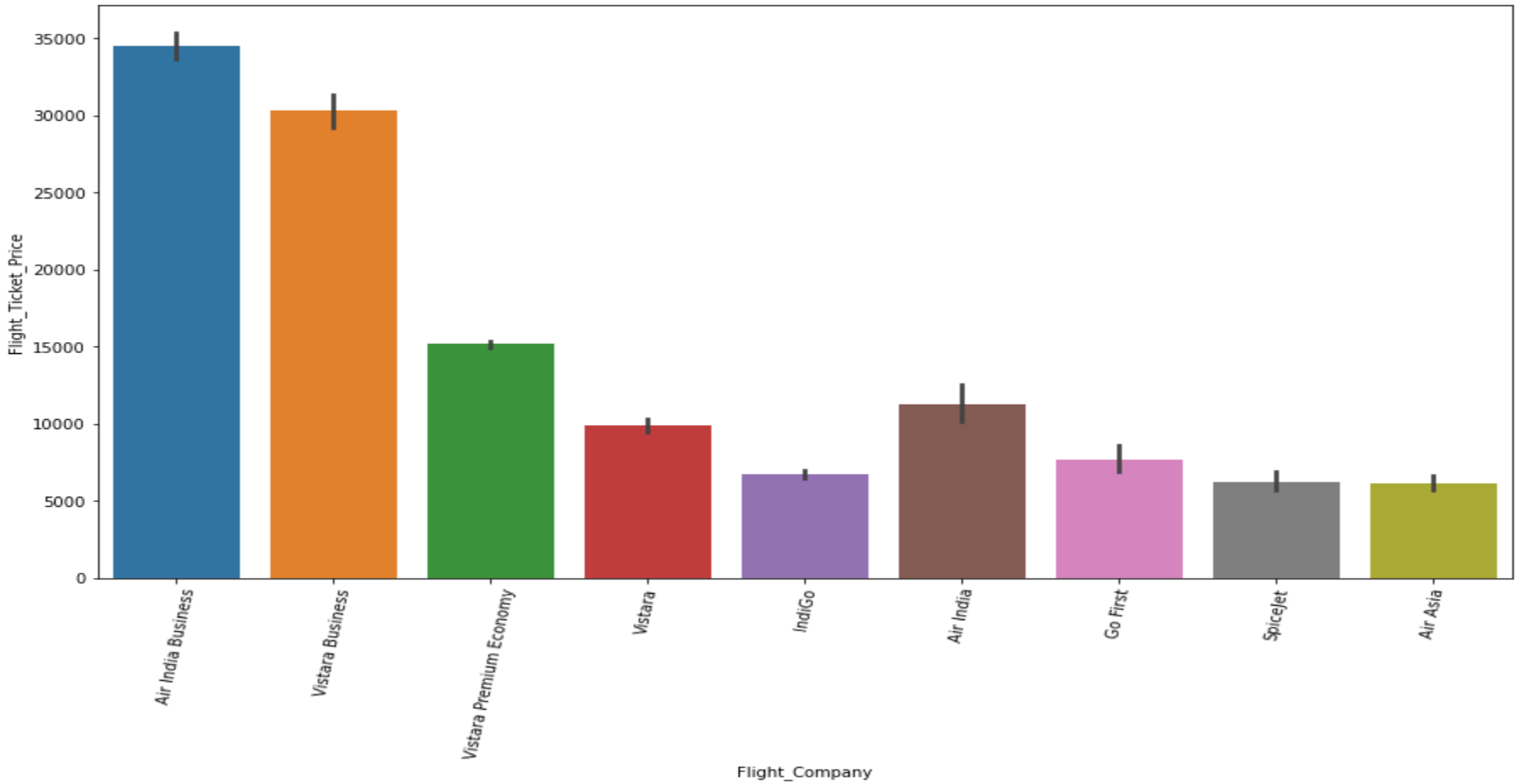
Dataset Description:

Columns	Datatype	Unique values	Mode Mean	Values
1	Flight Company	object	9	Vistara Premium Economy
2	Flight No	object	446	UK832
3	Flight Class	object	3	Economy
4	Flight Model	object	19	Vistara, Airbus A32-100
5	Travel	object	6	Bengaluru to Mumbai
6	Departure Time	object	22	7
7	Departure Day	object	7	Sat
8	Departure Date	object	28	29
9	Departure Month	object	2	Aug
10	Departure Duration	int32	68	126.74935
11	Arrival Time	int32	23	15.164451
12	Arrival Day	object	7	Sat
13	Meal	object	2	Free Meal
14	Flight Ticket Price	int64	652	19838.348439
15	Luggage	object	5	15
16	Connecting Planes	object	3	Has one connecting plane
17	No of days	int32	5	19.52242

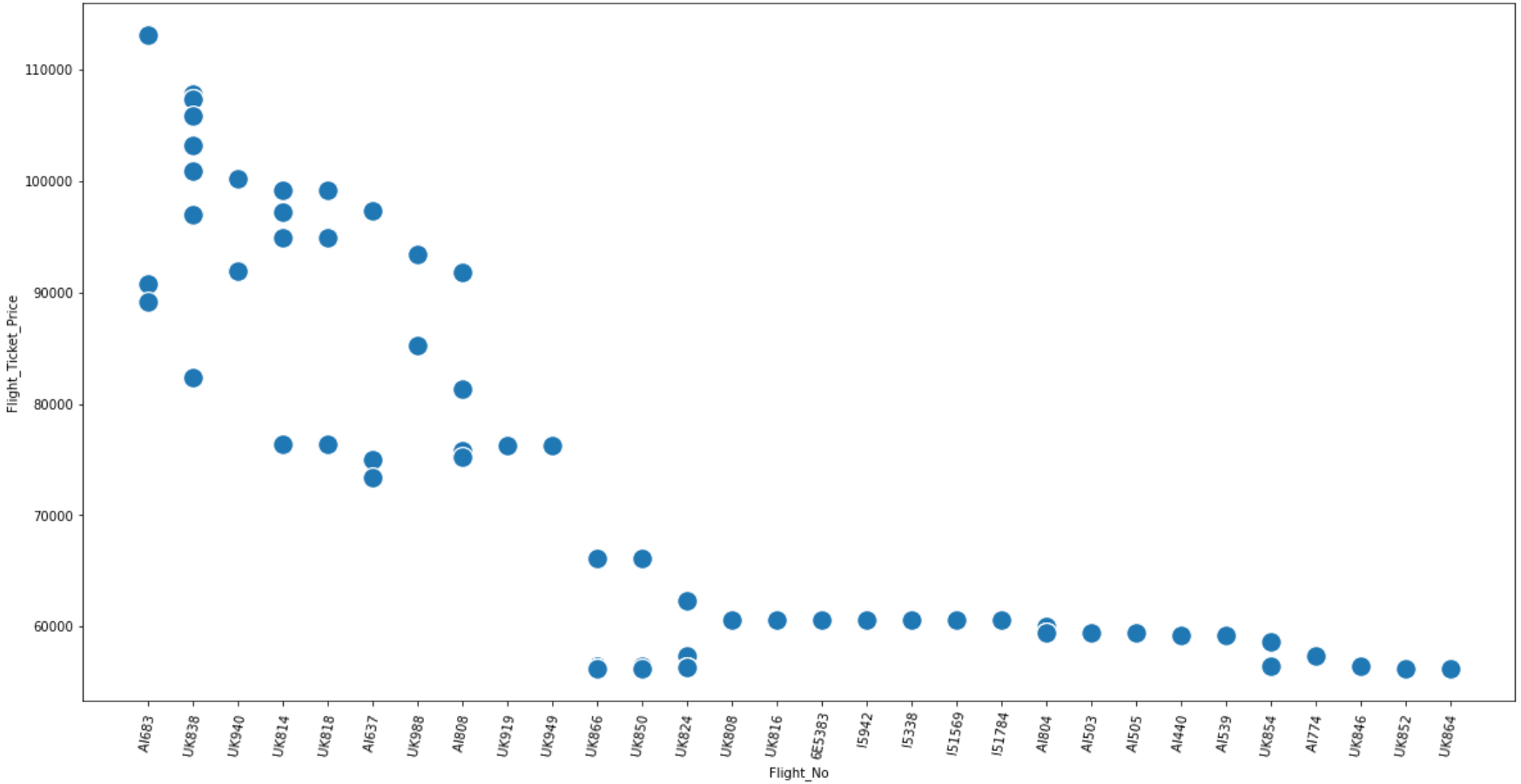
Hardware and Software Requirements and Tools Used:

- Hardware – PC Windows 10, 4 GB Ram
- Software – Google chrome, MS Excel, Python, Selenium webdriver
- Libraries – Pandas, NumPy, Matplotlib, Seaborn, sklearn, SciPy. Stats
    - ❑ Browsing – Google Chrome
    - ❑ Webscraping – Python, Selenium webdriver
    - ❑ Data cleaning – Python, Pandas, NumPy & SciPy. Stats
    - ❑ Data visualization – Matplotlib & Seaborn
    - ❑ Machine learning – Sklearn

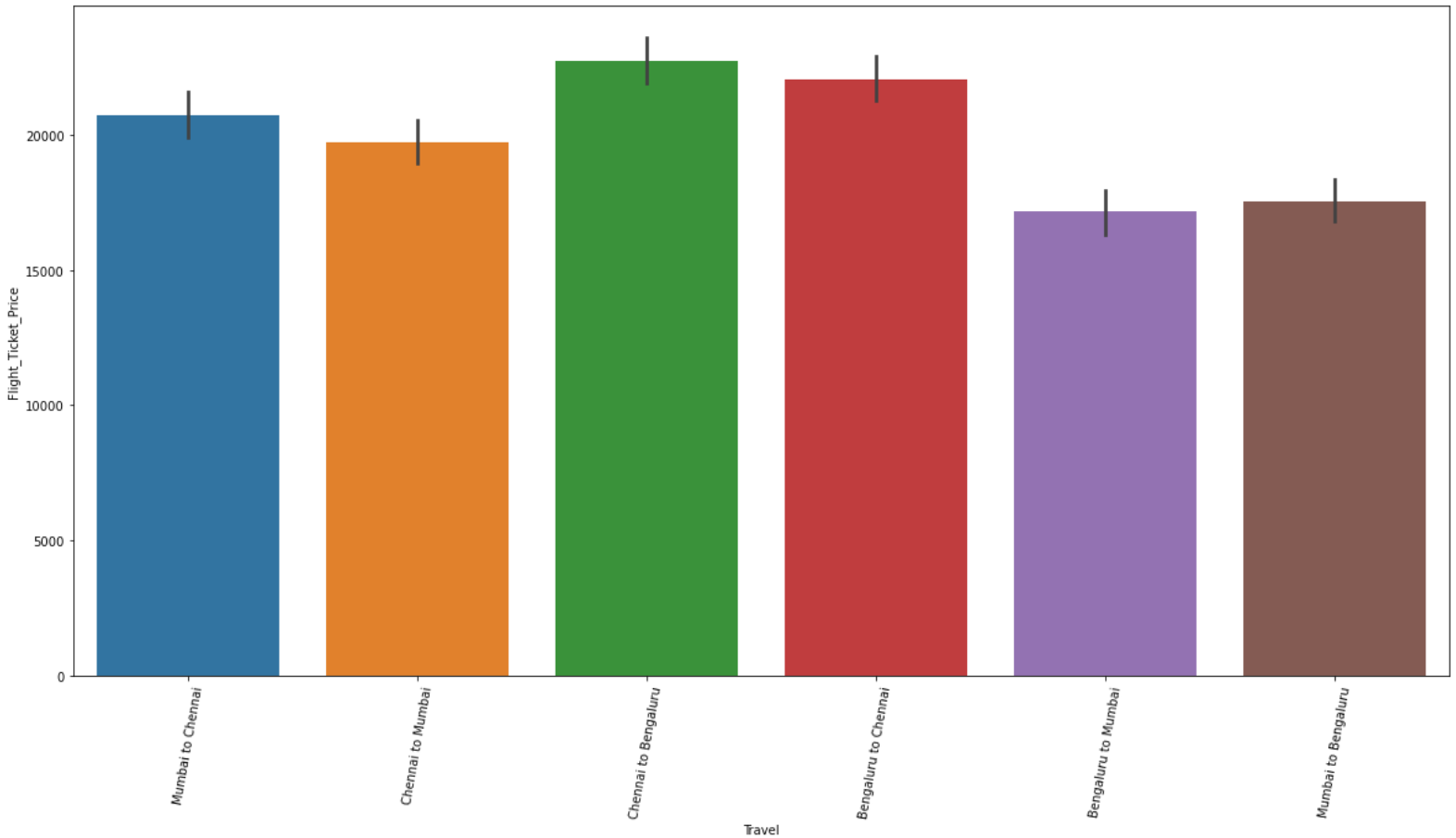
Flight Company :



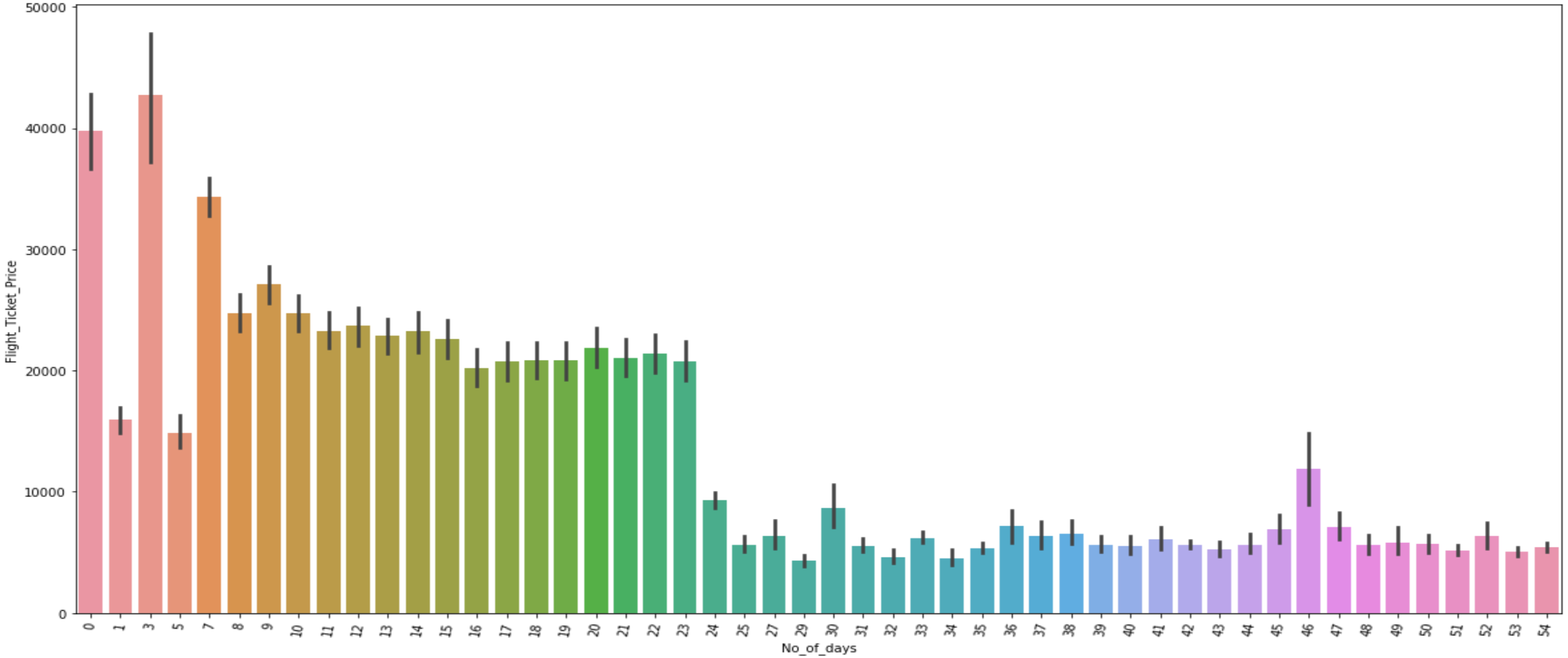
Flight No:


Flight_Ticket_Price
110000
100000
90000
80000
70000
60000
AI683
UK838
UK940
UK814
UK818
AI637
UK988
AI808
UK919
UK949
UK866
UK850
UK824
UK808
UK816
6E5383
I5942
I5338
I51569
I51784
AI804
AI503
AI505
AI440
AI539
UK854
AI774
UK846
UK852
UK864
Flight_No

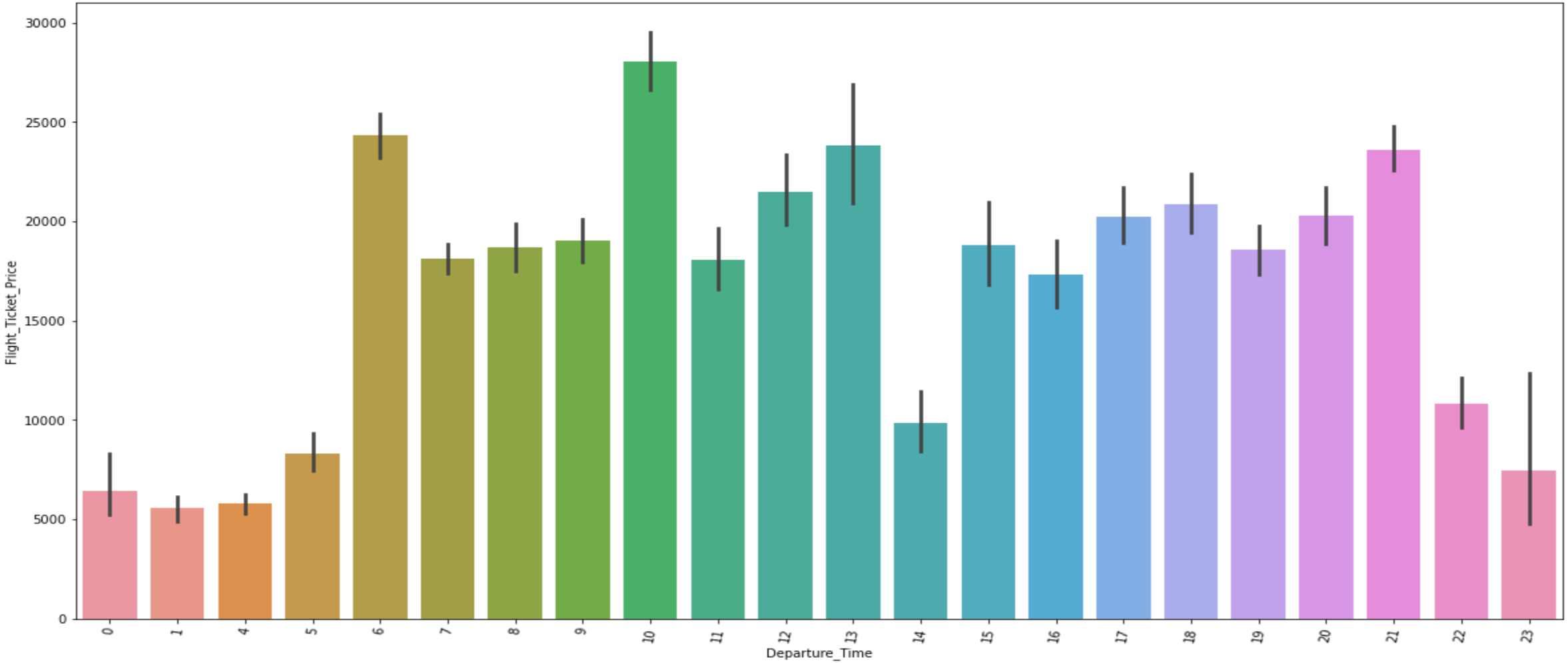
Travel:


Flight_Ticket_Price
0
5000
10000
15000
20000
Mumbai to Chennai
Chennai to Mumbai
Chennai to Bengaluru
Bengaluru to Chennai
Bengaluru to Mumbai
Mumbai to Bengaluru
Travel

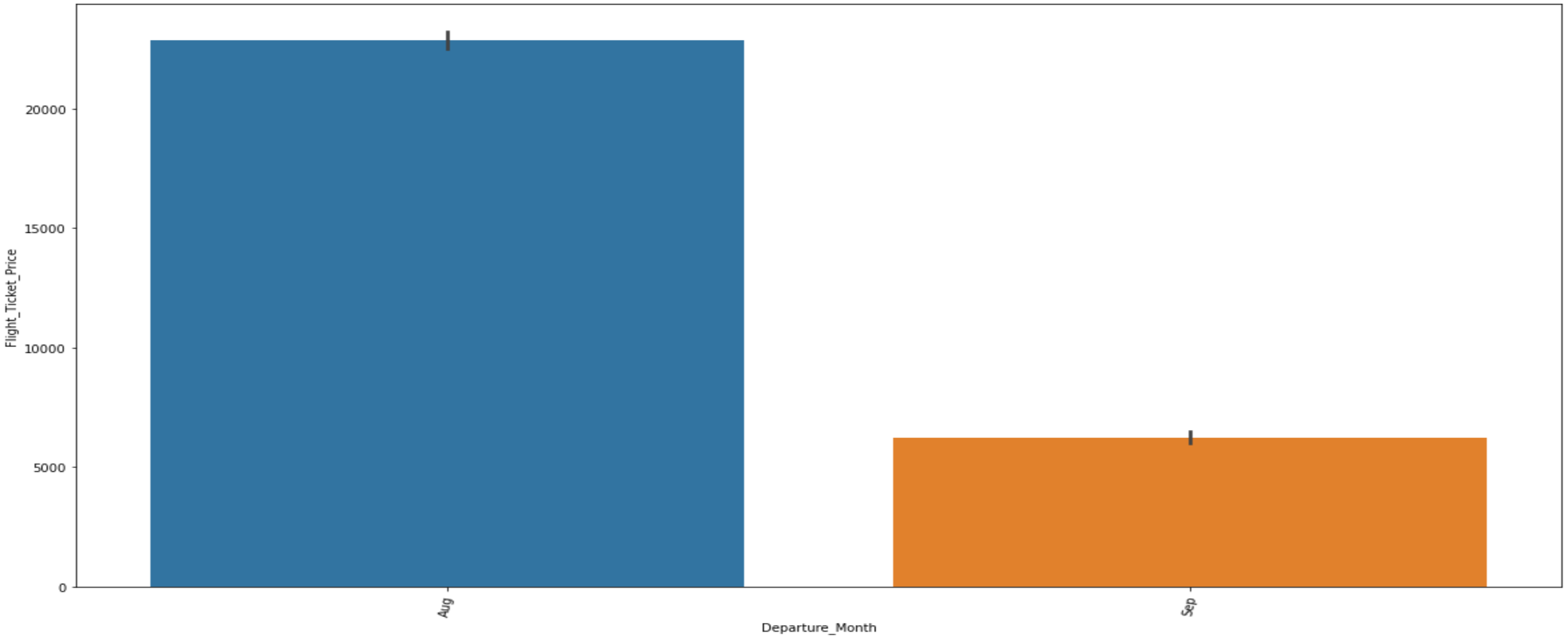
No of days:


Flight_Ticket_Price
50000
40000
30000
20000
10000
0
No_of_days
0 1 3 5 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 27 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54

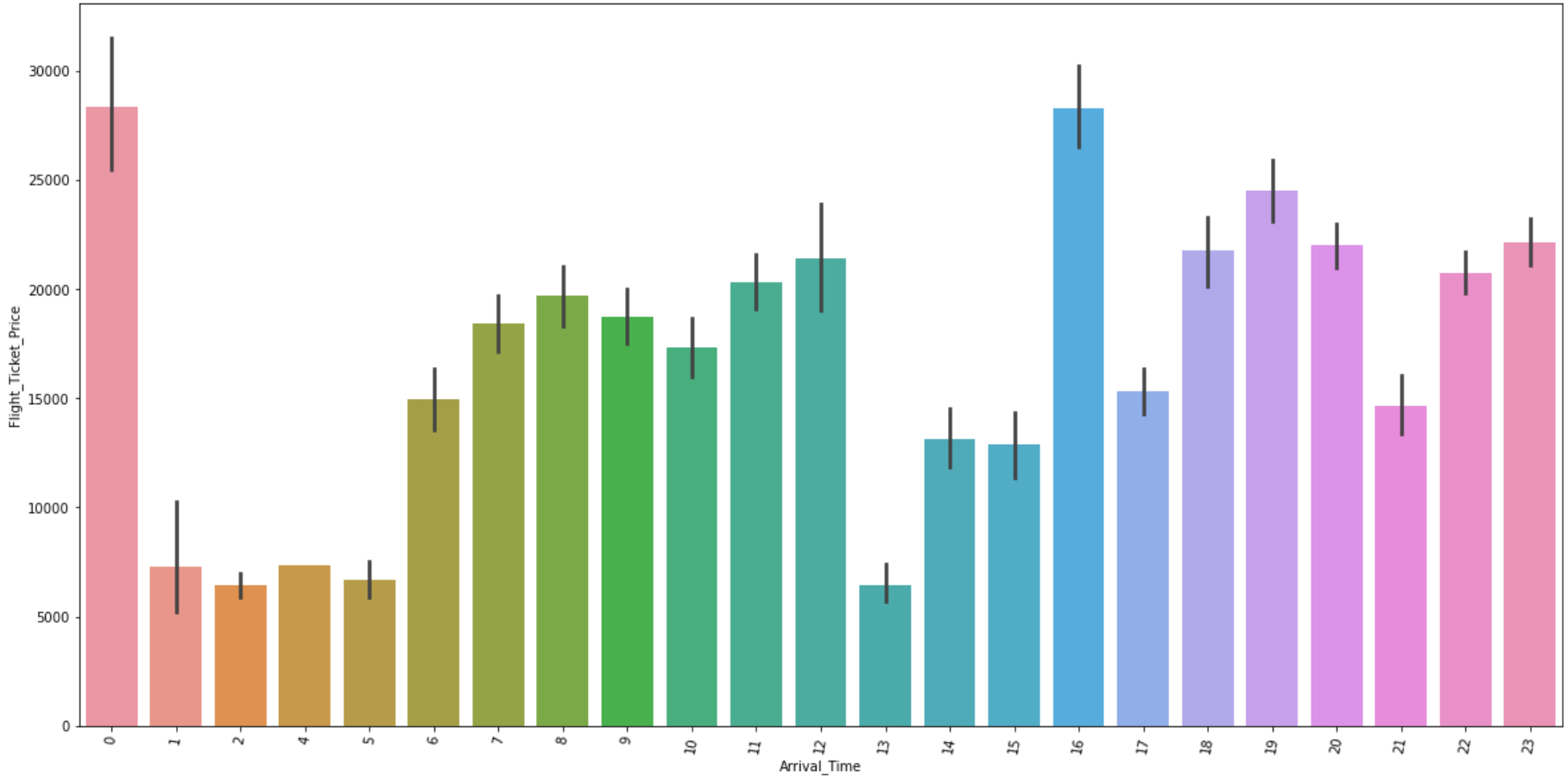
Departure Time:



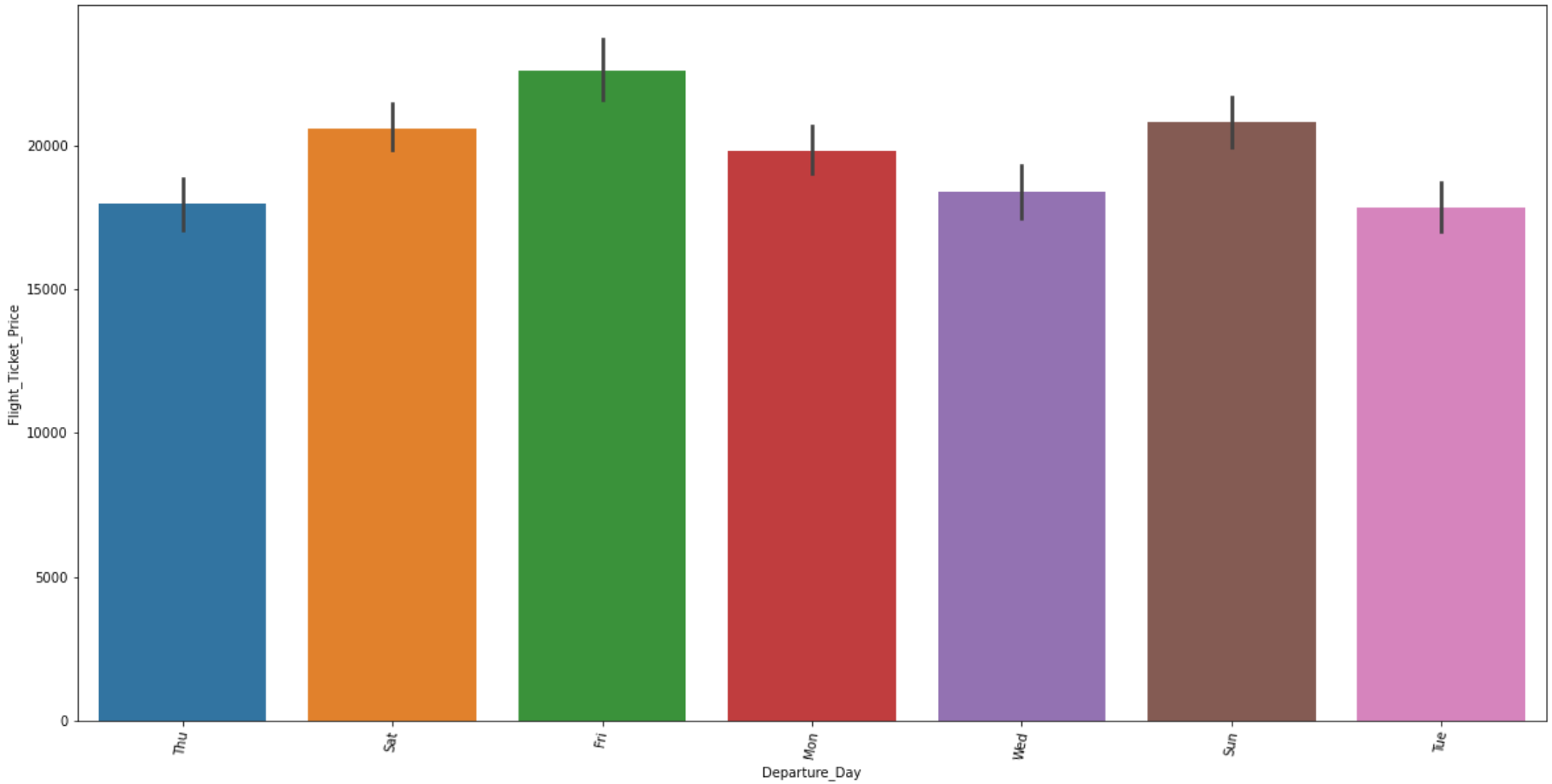
Departure Month:



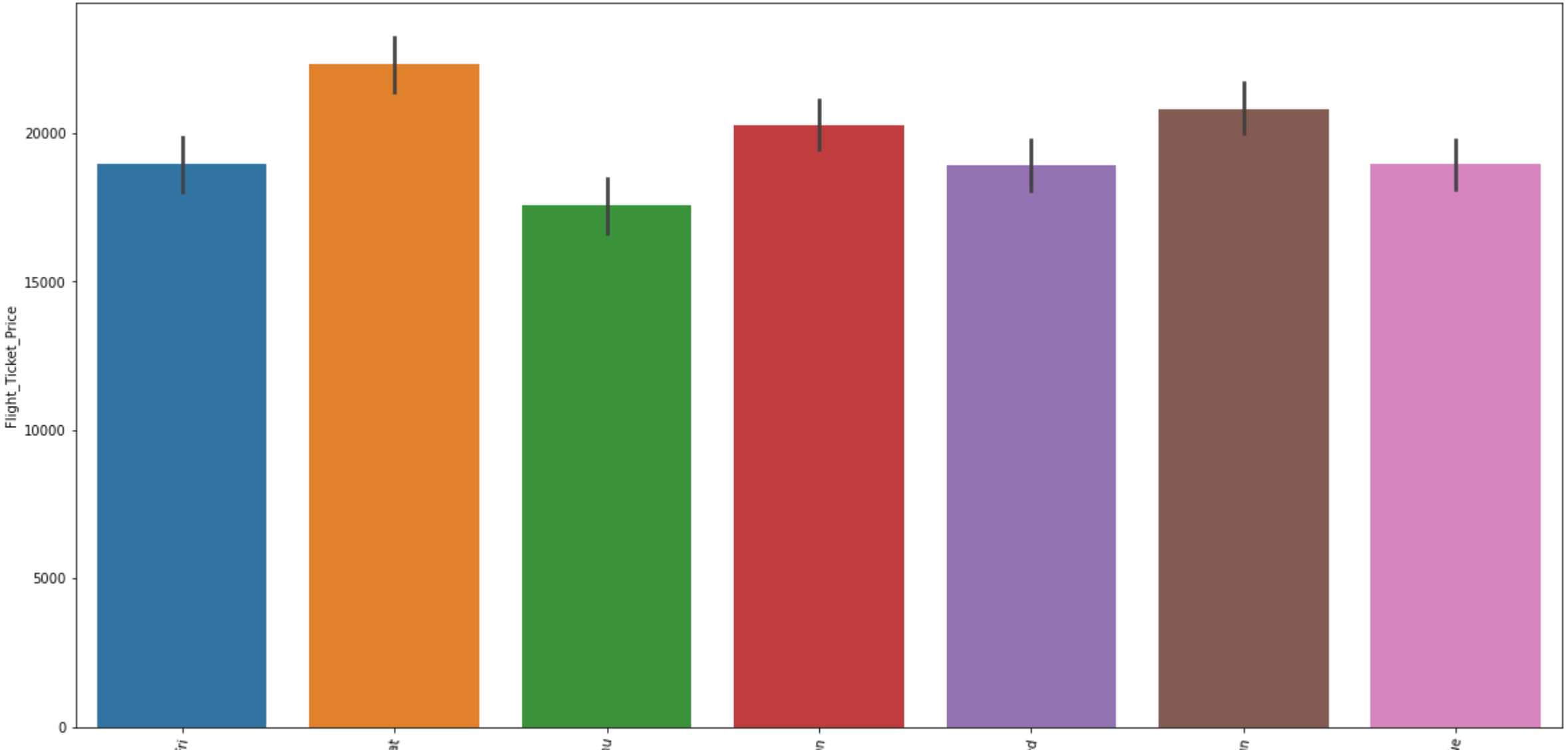
Arrival Time:



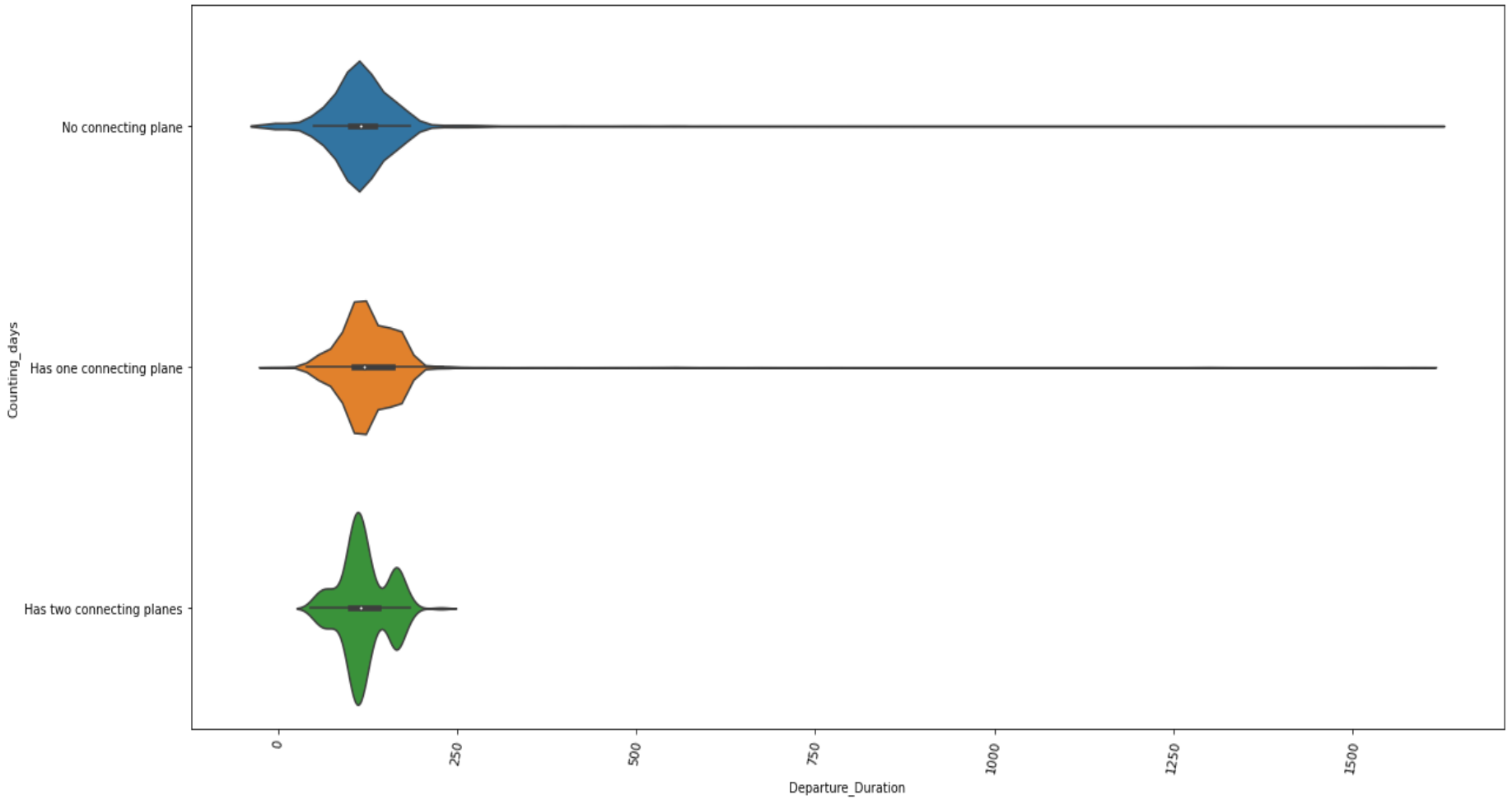
Departure Day:



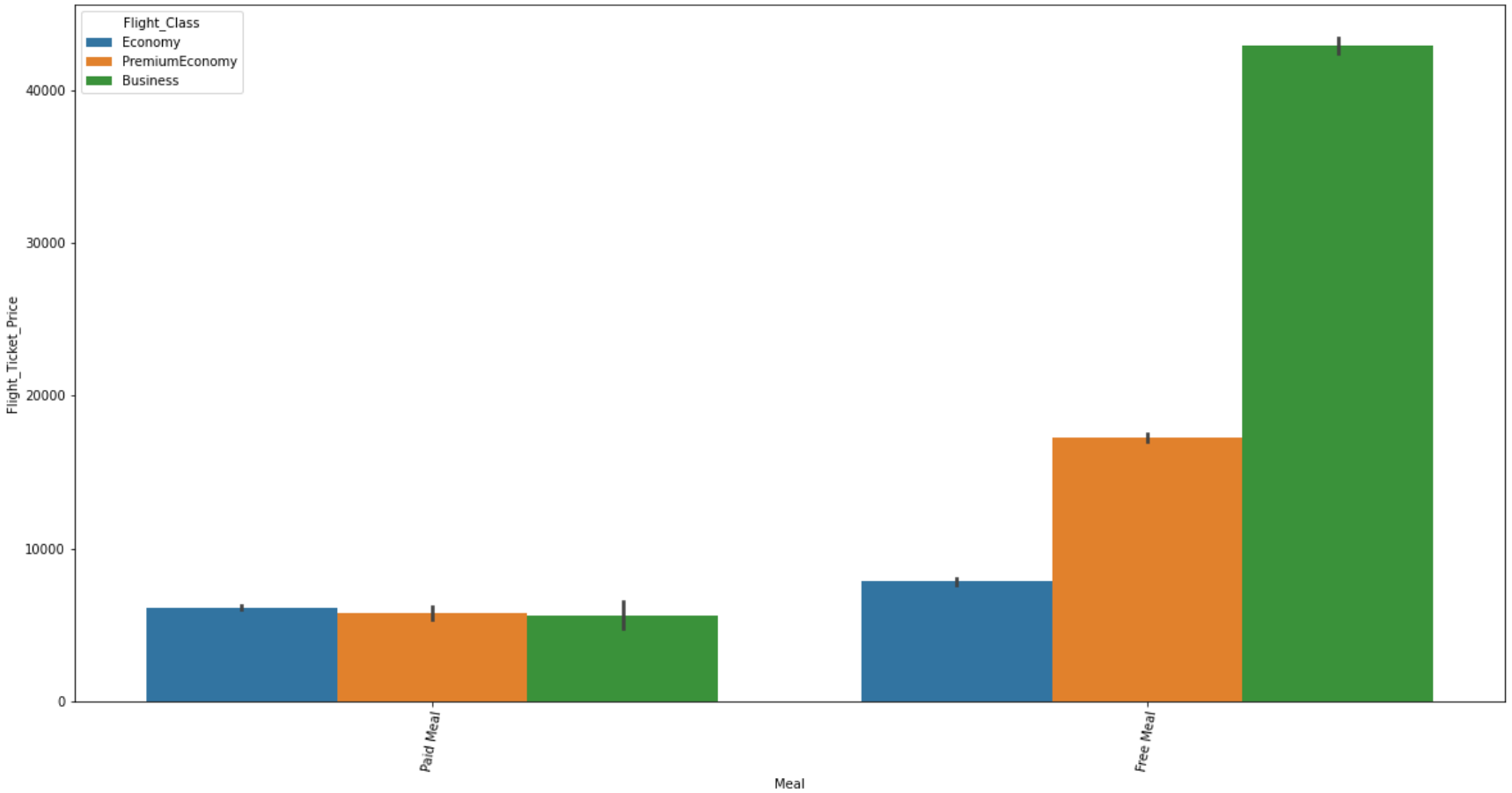
Arrival Day:



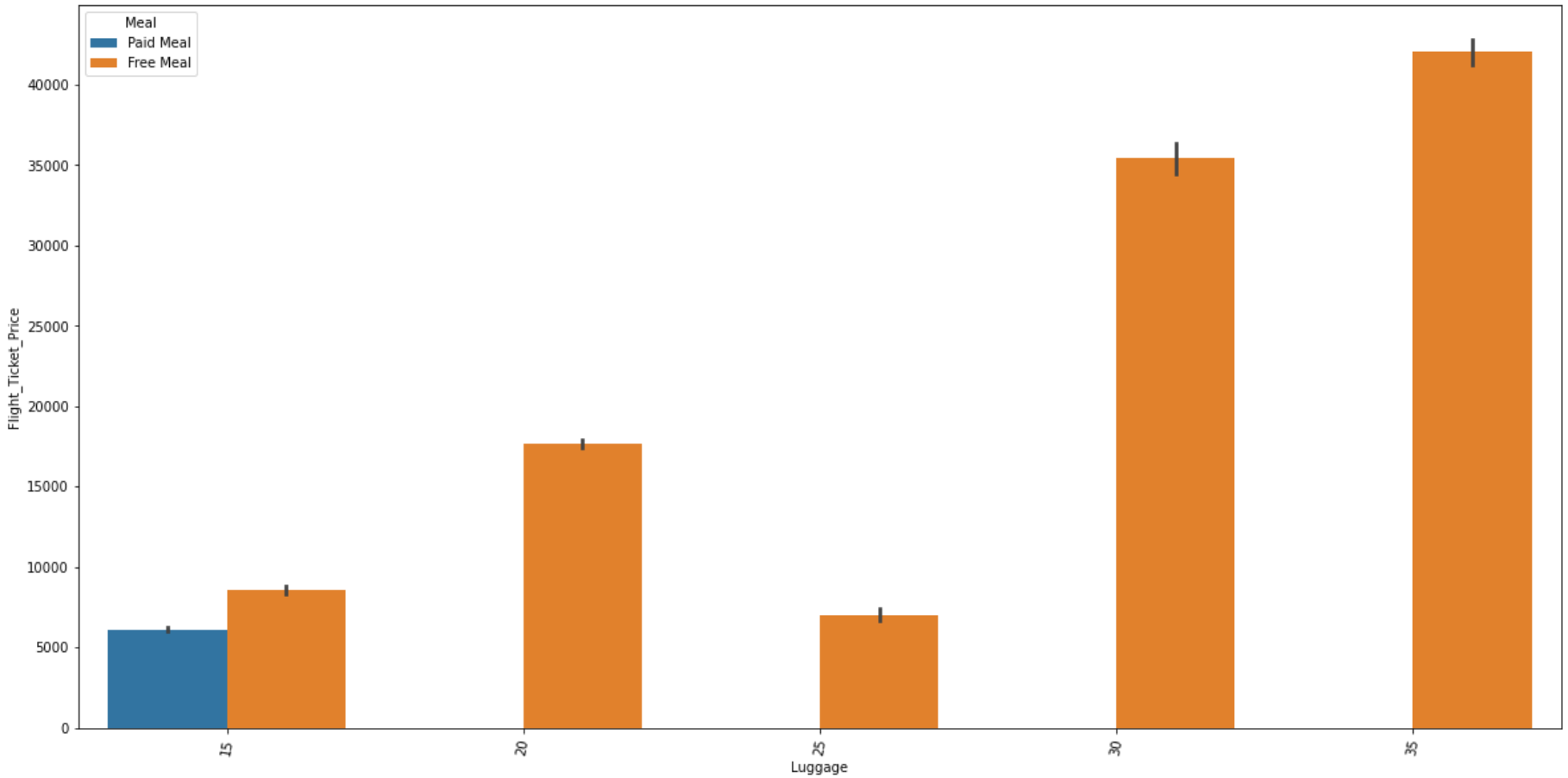
Duration:


No connecting plane
Has one connecting plane
Has two connecting planes
Counting_days
0
250
500
750
1000
1250
1500
Departure_Duration

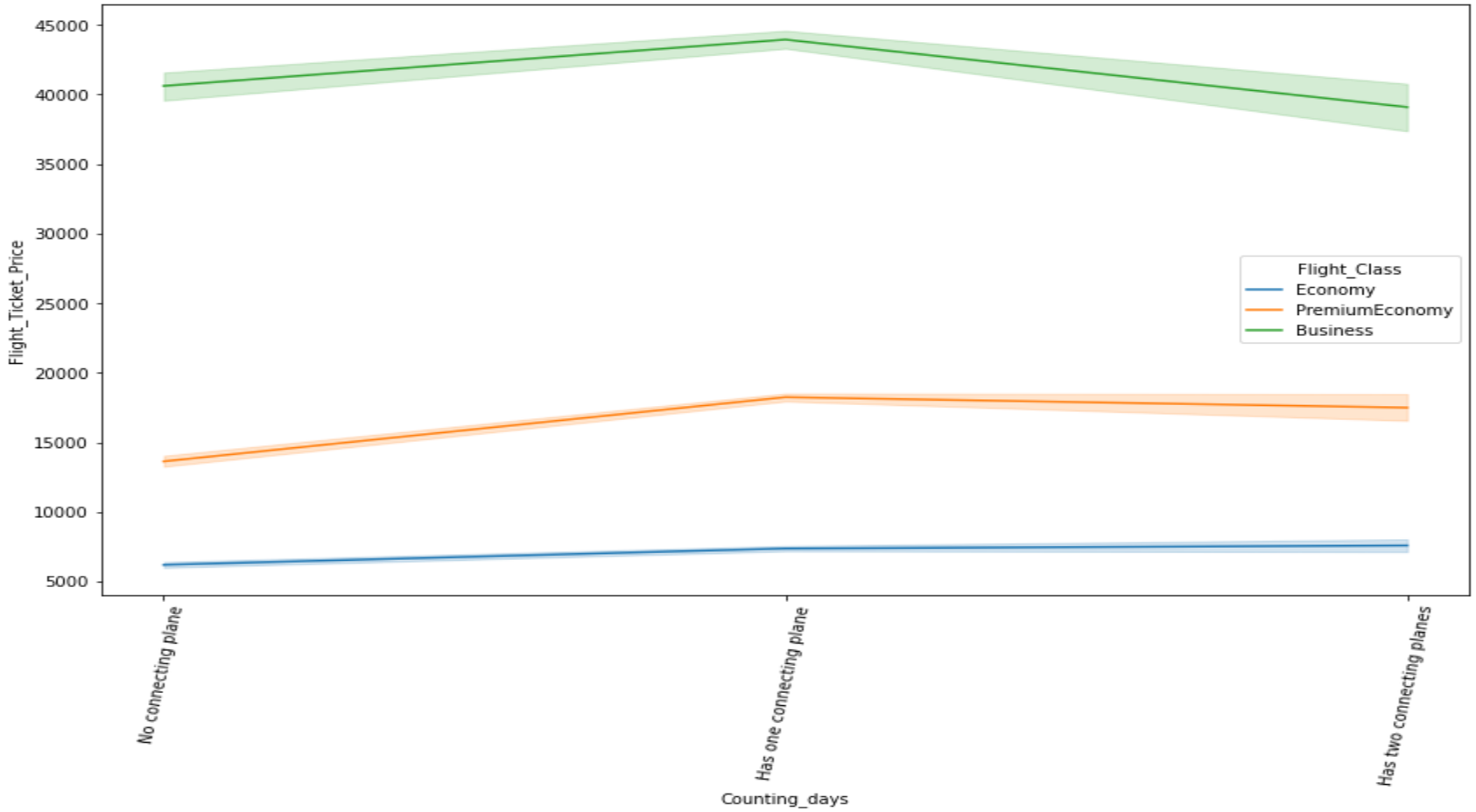
Meal:


Flight_Class
Economy
PremiumEconomy
Business
40000
30000
20000
10000
0
Flight_Ticket_Price
Paid Meal
Free Meal
Meal

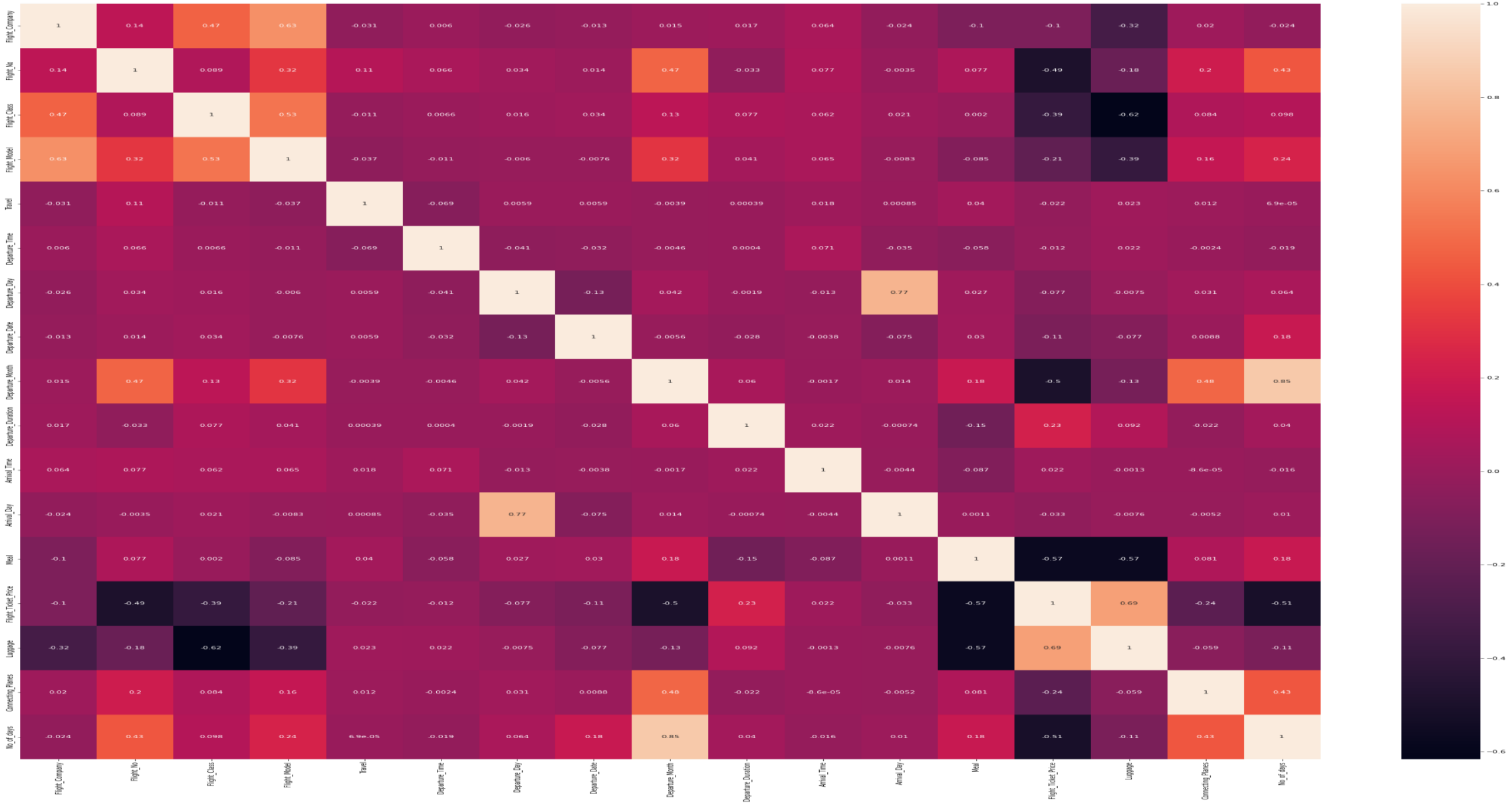
Luggage:


Meal
Paid Meal
Free Meal
40000
35000
30000
25000
20000
15000
10000
5000
0
Flight_Ticket_Price
15
20
25
30
35
Luggage

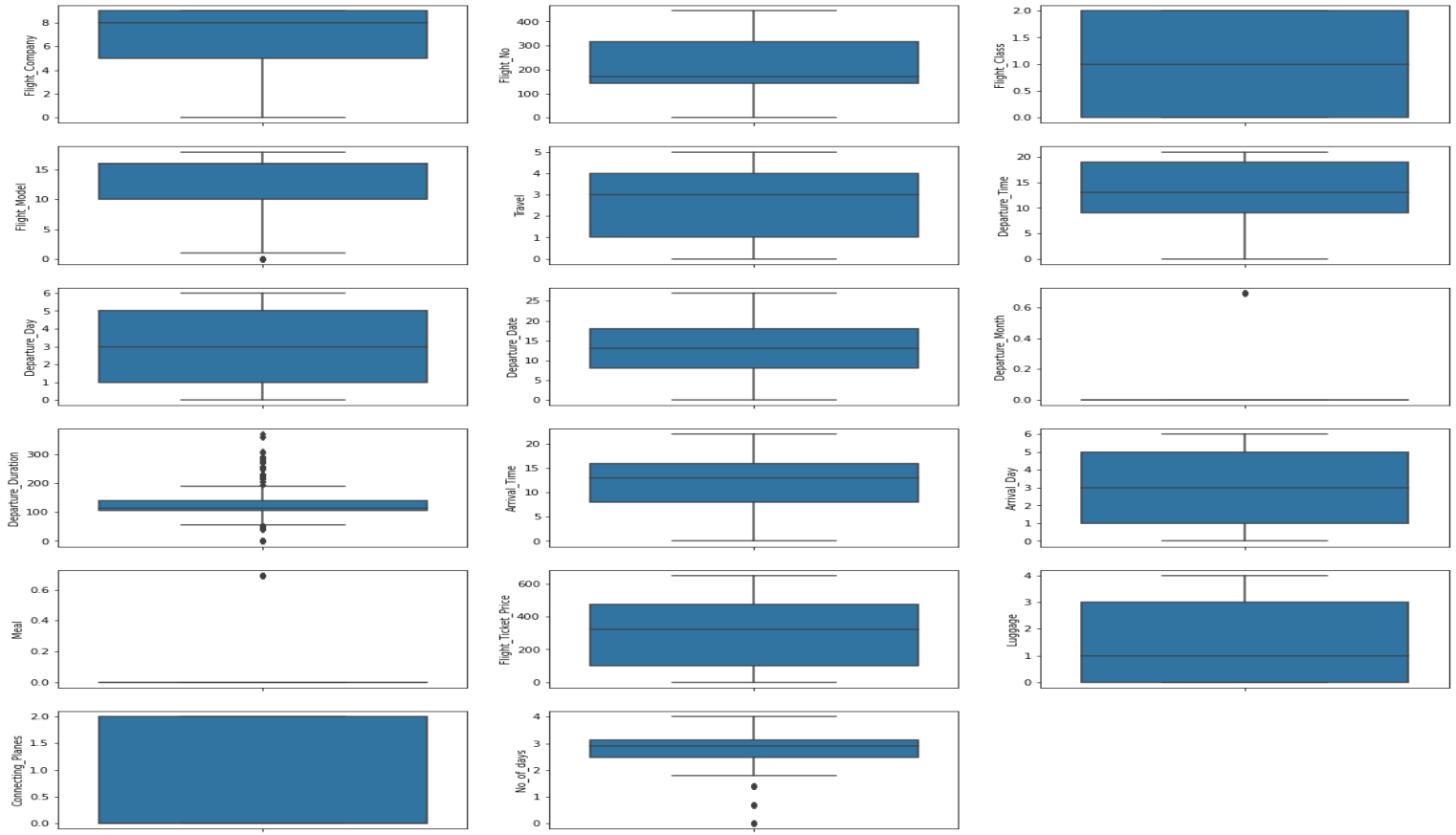
Connecting Flights:



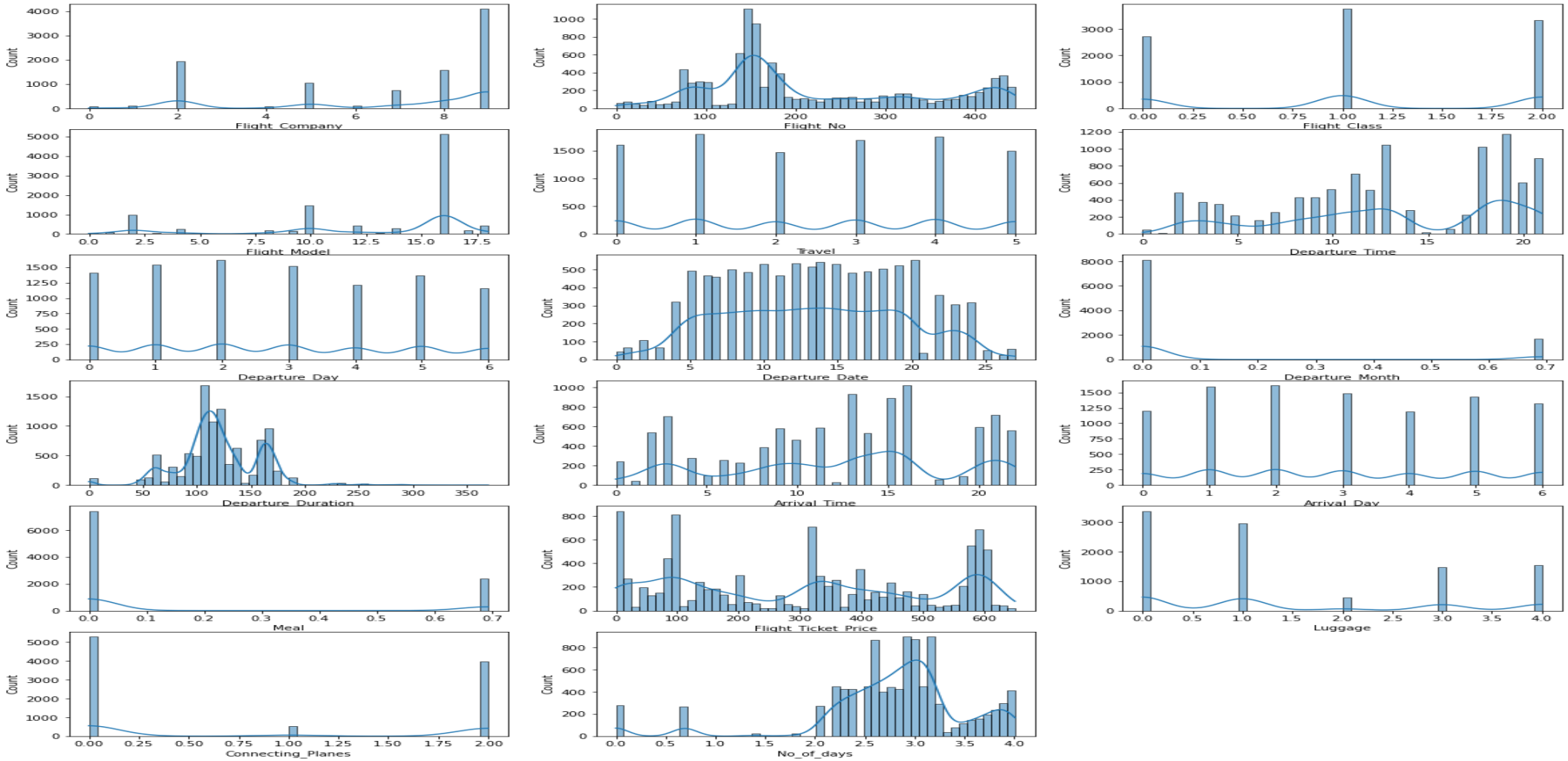
Heatmap:


Flight_Company
Flight_No
Flight_Class
Flight_Model
Travel
Departure_Time
Departure_Day
Departure_Date
Departure_Month
Departure_Duration
Arrival_Time
Arrival_Day
Meal
Flight_Ticket_Price
Luggage
Connecting_Planes
No_of_days
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Boxplot:


Flight_Company
Flight_No
Flight_Class
Flight_Model
Travel
Departure_Time
Departure_Day
Departure_Date
Departure_Month
Departure_Duration
Arrival_Time
Arrival_Day
Meal
Flight_Ticket_Price
Luggage
Connecting_Planes
No_of_days

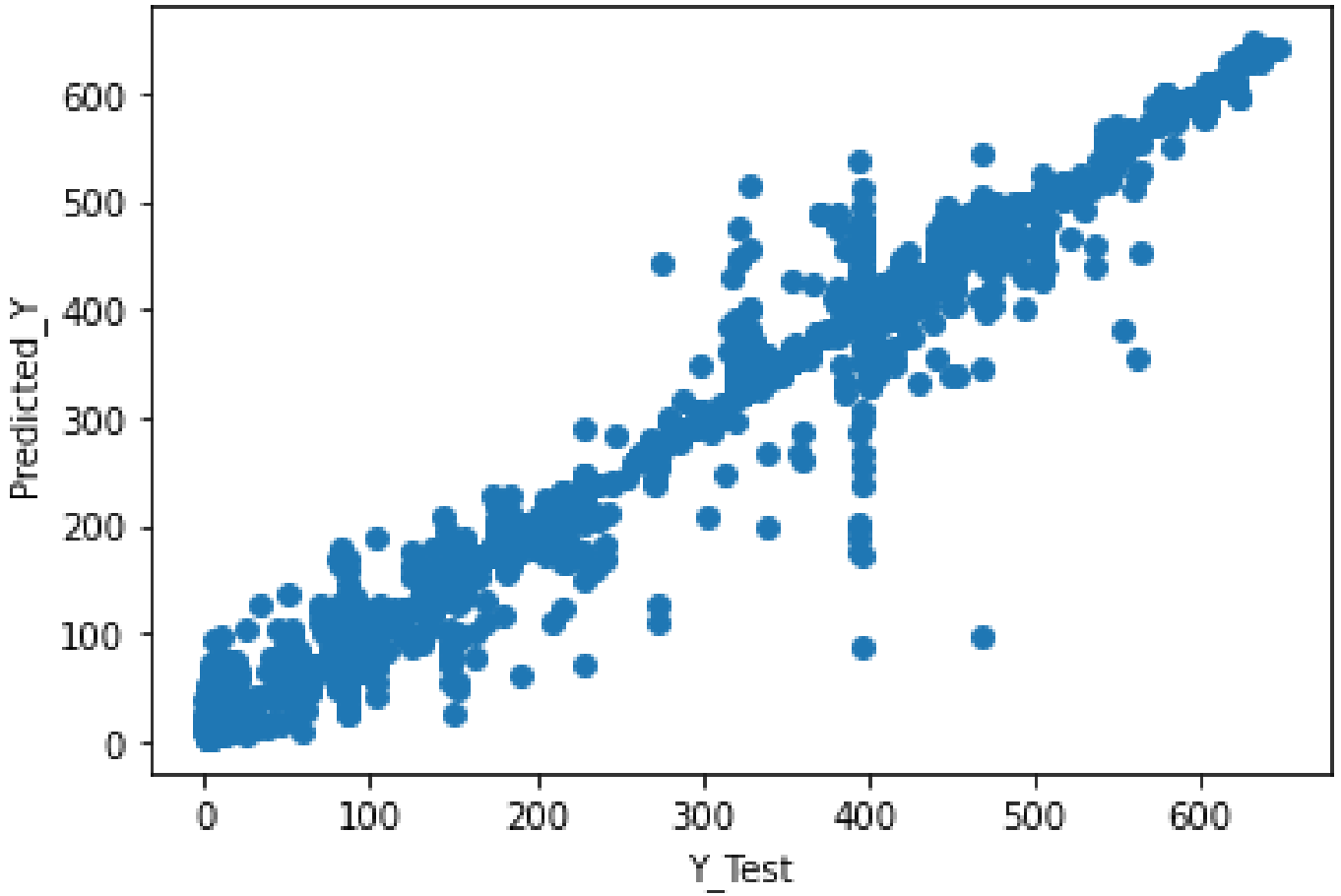
Histogram:


Flight_Company
Flight_No
Flight_Class
Flight_Model
Travel
Departure_Time
Departure_Day
Departure_Date
Departure_Month
Departure_Duration
Arrival_Time
Arrival_Day
Meal
Flight_Ticket_Price
Luggage
Connecting_Planes
No_of_days
Count

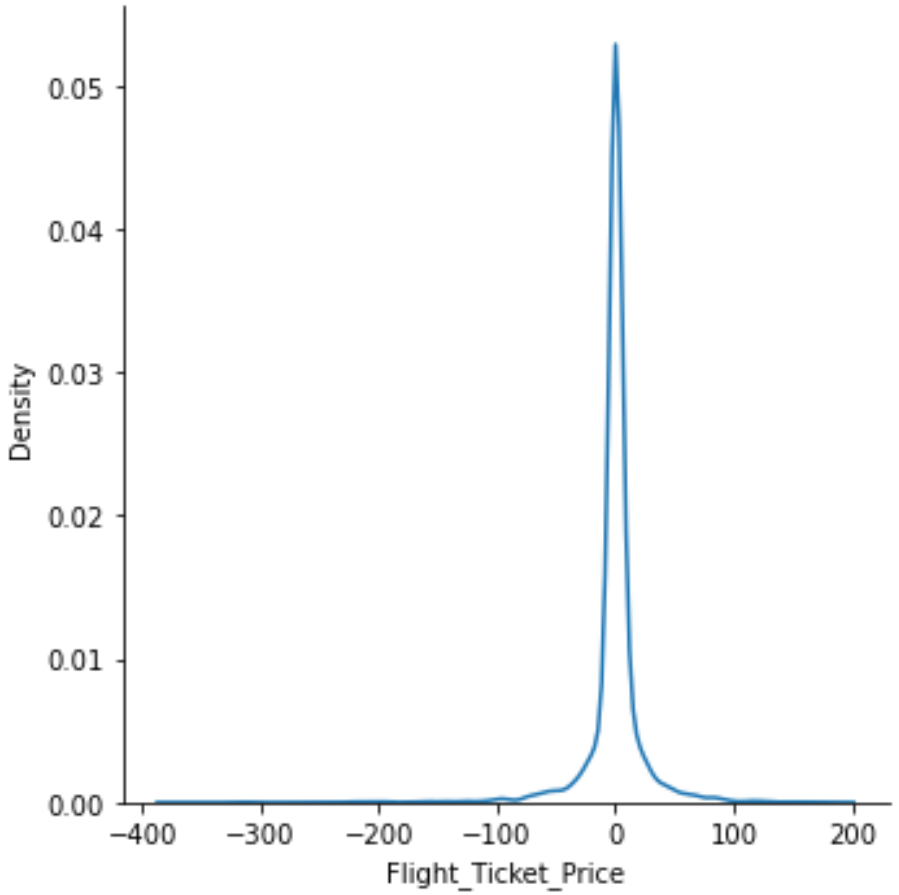
Testing of Identified Approaches (Algorithms):

- Linear Regression
- Gradient Boosting Regressor
- AdaBoost Regressor
- Decision Tree Regressor
- KNeighbors Regressor
- Extra Trees Regressor
- Random Forest Regressor

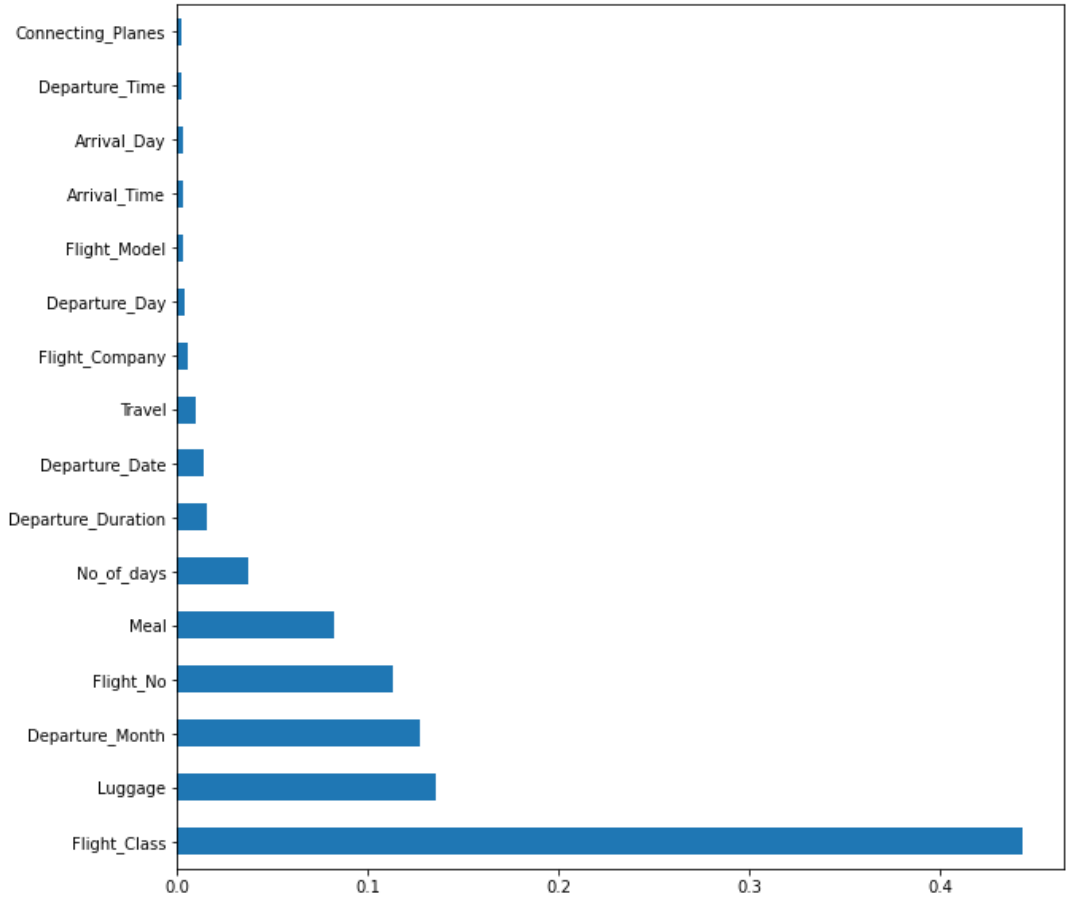
Y test plot:



Distribution of Predicted Values:



Feature importance’s



Accuracy Parameter:

- R2 Score: 98.48414998011373
- Mean Absolute Error: 10.476769408502774
- Extra Forest Regression gives an R2 score 0.98
- Flight class dominates the flight price more.

Key Findings and Conclusions of the Study:

- This dataset has been taken from 2 websites of this, Yatra and MakeMyTrip constitutes the majority of data
- Since the target feature is continuous data, this problem can be solved by regression algorithms
- Extra Forest Regression gives an R2 score 0.98
- Flight class dominates the flight price more.