R STUDIO

FLIGHT PRICING FORECAST

A dataset given in Excel and Python script are transformed into R to test correlations among variables to forecast new price predictions for competing airlines.

PRESENTED BY

Sara Galapo and Mary Fredericks

AGENDA

1. Background of the Study
2. Rearch Question
3. Methodology
5. Analysis
6. Insights

BACKGROUND OF STUDY

Our goal is to understand which variables impact the price of different flights and ways for customers to score the best deals for competing airlines

RESEARCH QUESTIONS

Price Impact on Flight Purchases

How does price impact the number of flights purchased?

Hypothesis: Customers are more likely to buy low priced flights with longer duration.

Duration Impact on Price

What impact does duration impact the price?

Hypothesis: Flights that are longer are more expensive.

METHODOLOGY

Syntax Used

- **Statistical Summary**
  - Mean, Range, Quartiles
  - Linear Regression Model
- **Charts**
  - Histogram, Scatter, Boxplot
  - Density Plots, ggplots
- **Comparison Modelings**
  - 10 Fold, AIC, BIC, MSPE, MAE, Cook Distance
  - Forward Variable Selection
- **Predictive Modeling**
  - Diagnostics

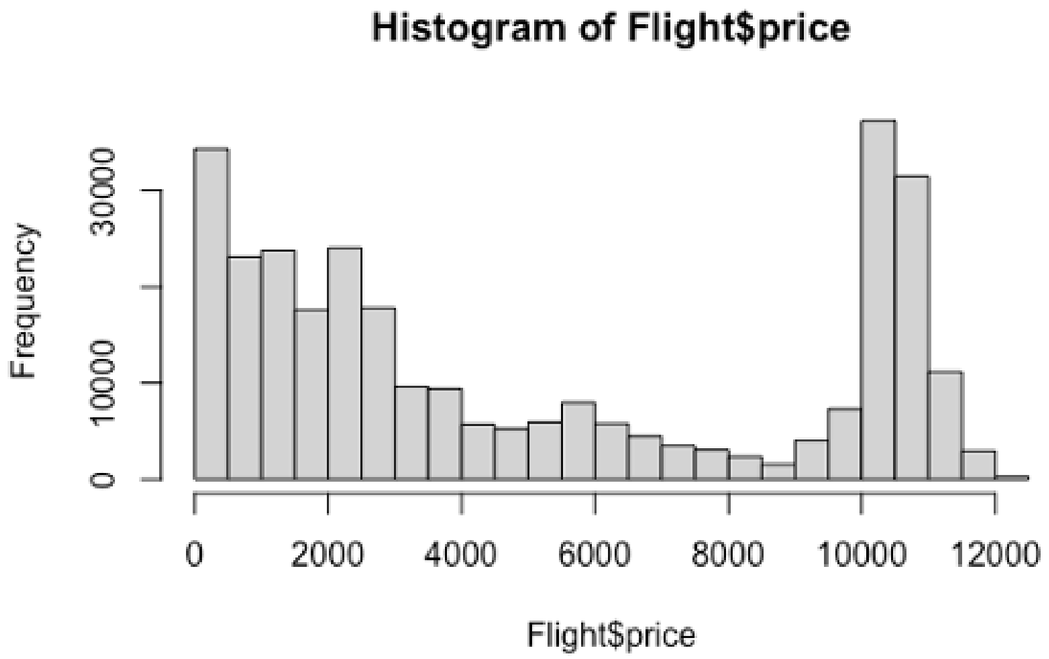
DEFINITIONS OF MAIN VARIABLES

Terms

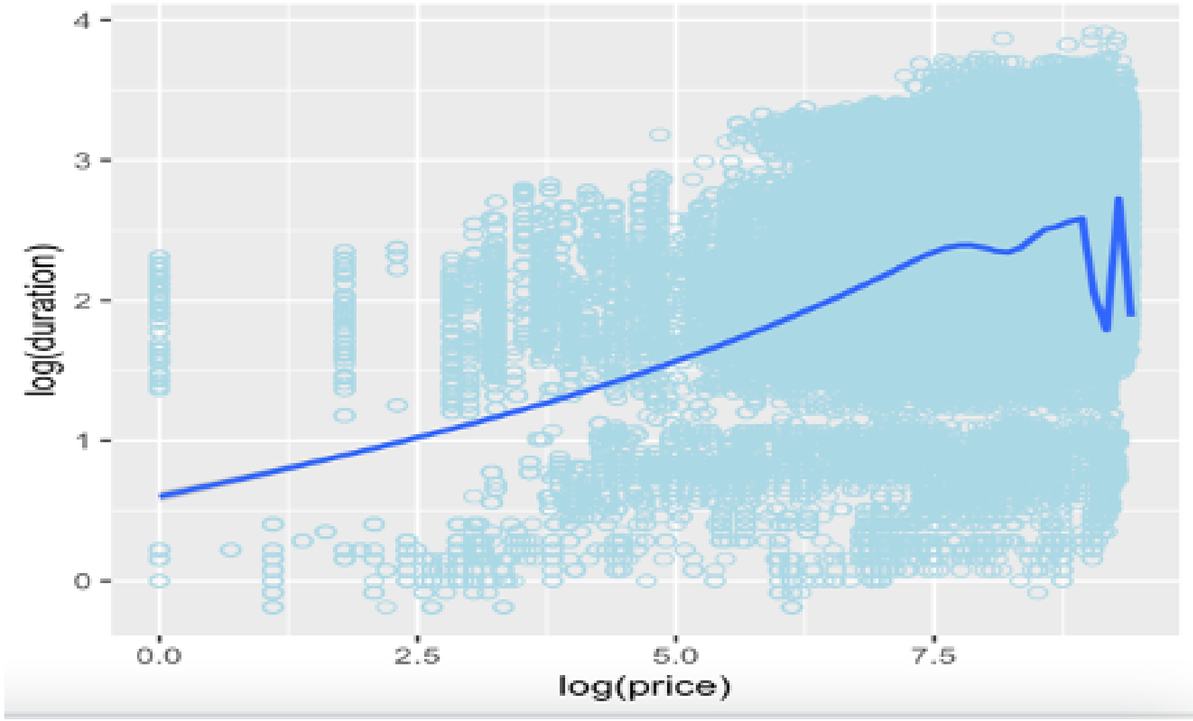
Variable Term	Meaning
airline	The name of the airline company is stored in the airline column. It is a categorical feature having 6 different airlines.
flight	Flight stores information regarding the plane's flight code. It is a categorical feature.
source_city	City from which the flight takes off. It is a categorical feature having 6 unique cities.
departure_time	This is a derived categorical feature obtained created by grouping time periods into bins. It stores information about the departure time and have 6 unique time labels.
stops	A categorical feature with 3 distinct values that stores the number of stops between the source and destination cities.
arrival_time	This is a derived categorical feature created by grouping time intervals into bins. It has six distinct time labels and keeps information about the arrival time.
destination_city	City where the flight will land. It is a categorical feature having 6 unique cities.
class	A categorical feature that contains information on seat class; it has two distinct values: Business and Economy.
duration	A continuous feature that displays the overall amount of time it takes to travel between cities in hours.
days_left	This is a derived characteristic that is calculated by subtracting the trip date by the booking date.
price	Target variable stores information of the ticket price.

EXPLORATORY ANALYSIS

Price is polarized at the tails (extremes).
Theory: Could have economic influence of wealth gap.

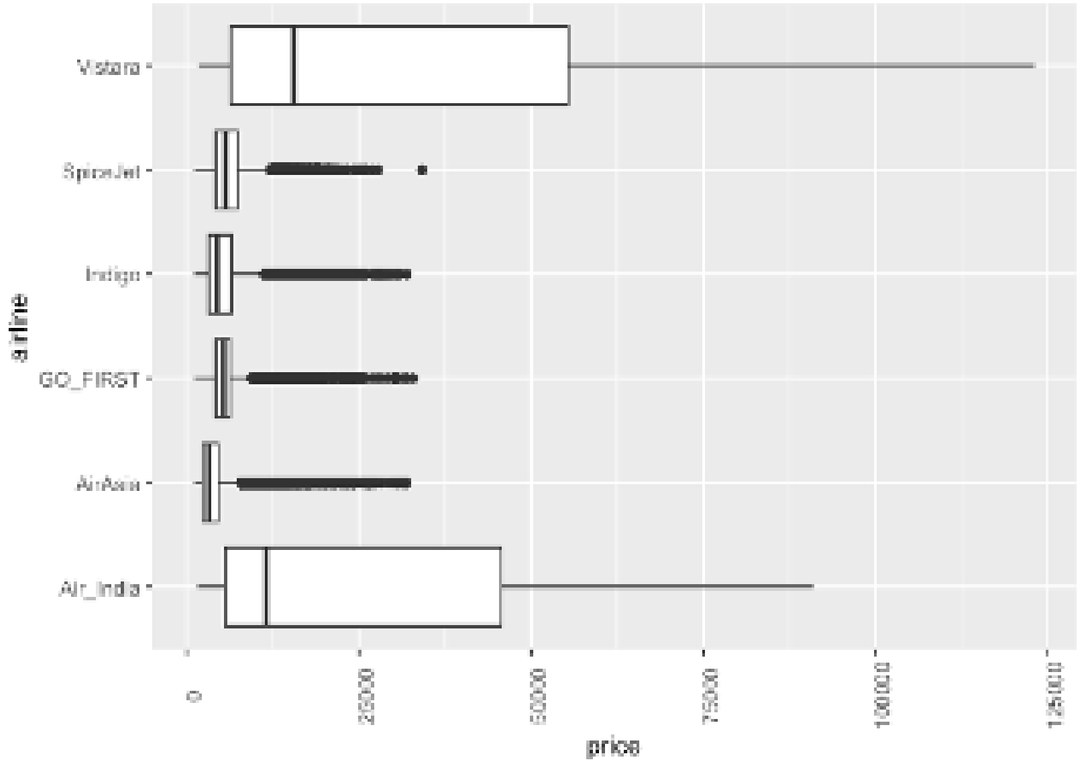


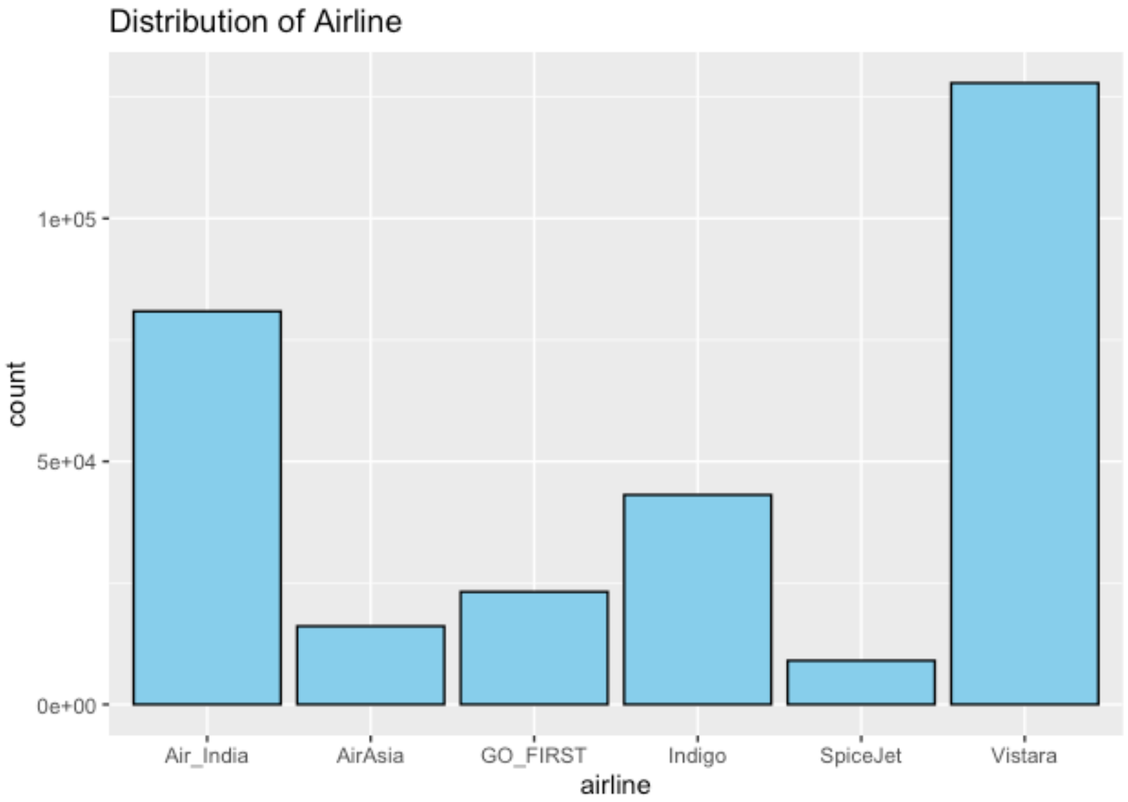
Log of Price and Duration show a linear relationship moving up.
As price increases, so does time, and vis versa.



EXPLORATORY ANALYSIS

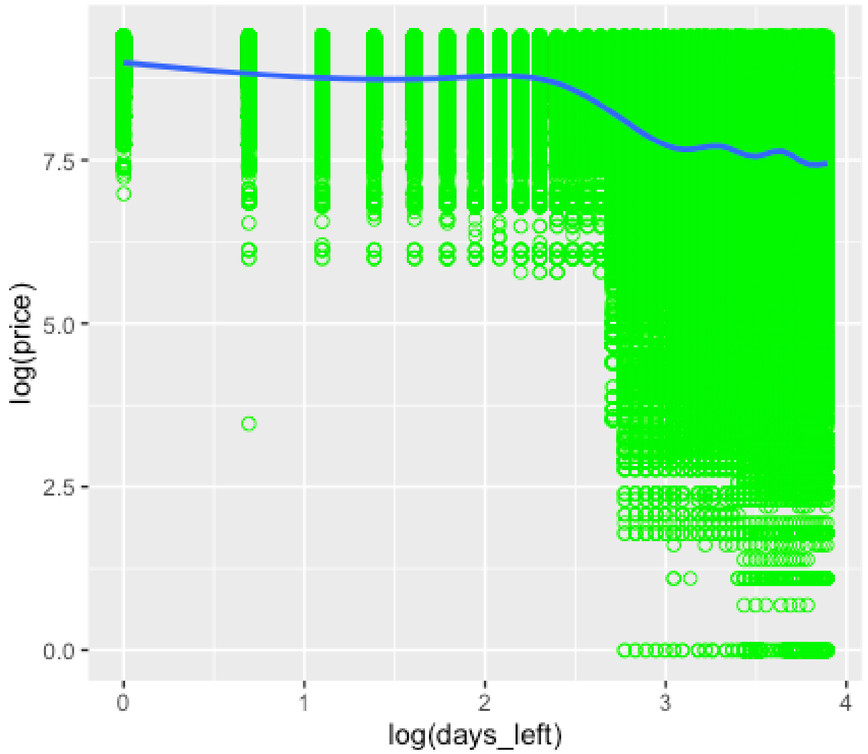
Air India and Vistara serve most passage within cities in India





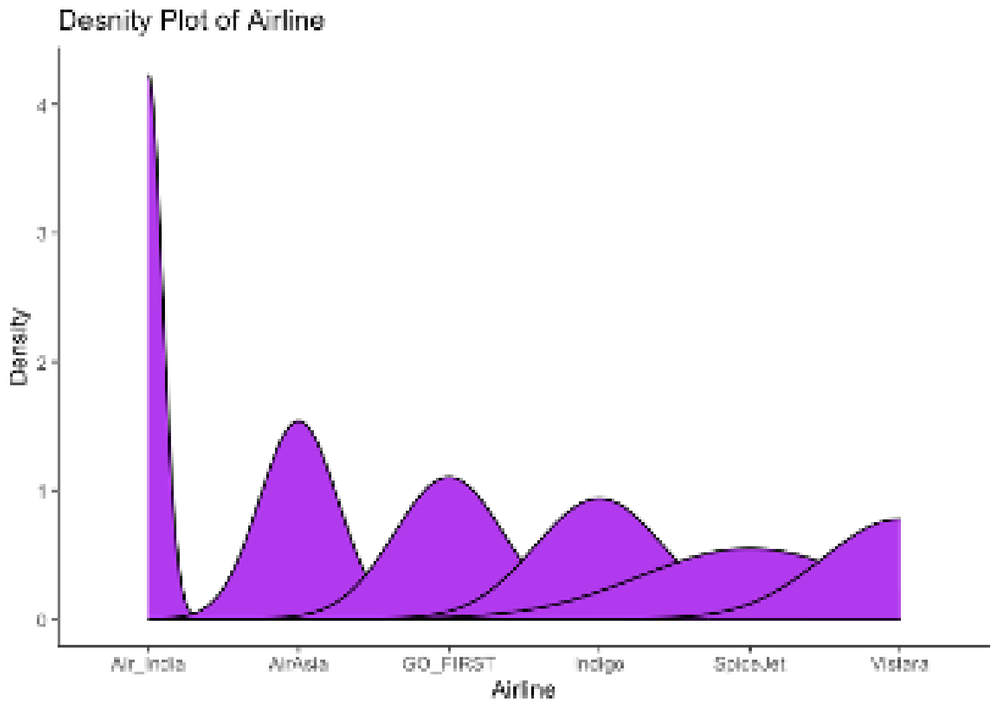
EXPLORATORY ANALYSIS

The more days left, the cheaper the flight.



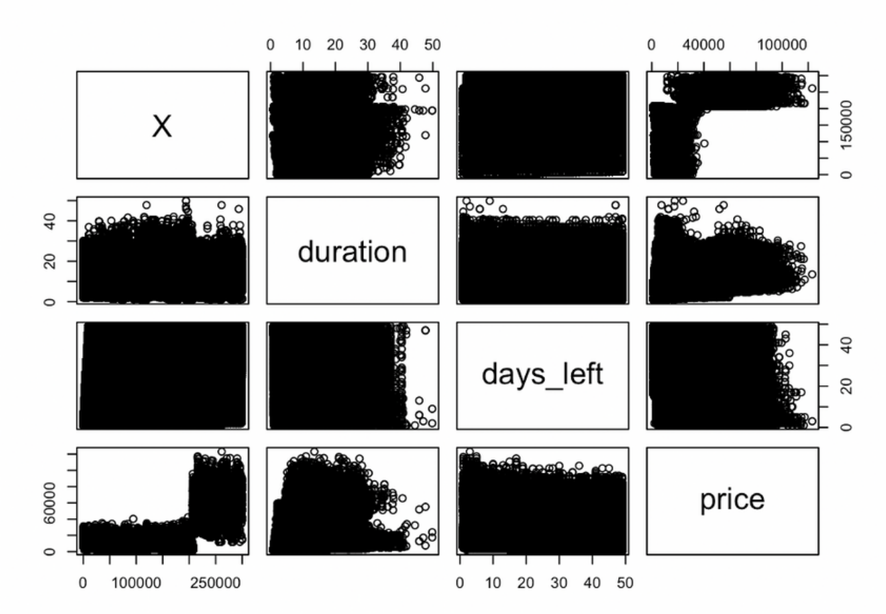
DENSITY OF PRICE AND AIRLINE

A price spike from 10,000 to 12,500 rupee occurs.
Price extremes still exist with drops at 5000 to 9000 rupee.
Possibly due to duration and flight arrival/ destination.

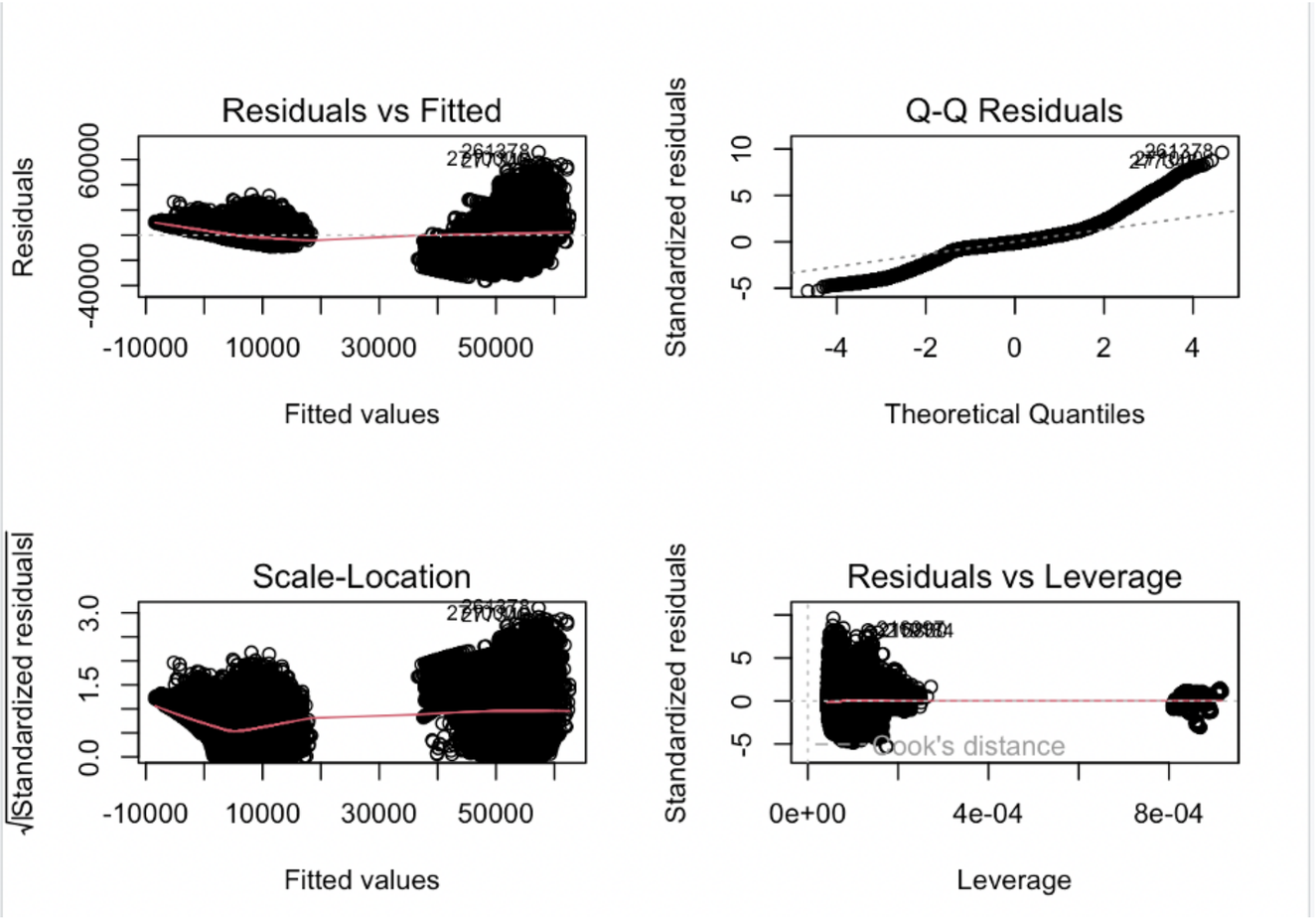


Air India is an outliner among the others. Flights at all ranges.

DIAGNOSTICS



Statistically Significant
P > 0.05
All variables significant



INSIGHTS

Questions Answered

Duration has an impact on Price.
Price has an impact on quantity of tickets purchased/sold.

- Booking a flight ahead of time will save you money.
- Visatara serves most passengers on flights across India
- Flight prices are on the extremes of low and high.
    - The flight class on your ticket greatly impacts the price of the flight.

RESOURCES

R Studio

Kraggle Dataset Samples

https://www.kaggle.com/code/afanys117/flight-price-prediction

APPENDIX

TABLES

	airline	n
6	Vistara	127859
2	Air_India	80892
4	Indigo	43120
3	GO_FIRST	23173
1	AirAsia	16098
5	SpiceJet	9011

	price	n
10555	10555	1445
82	82	1442
10560	10560	1390
10853	10853	1383
10831	10831	1230
10445	10445	1205
10352	10352	1205

	arrival_time	n
1	Afternoon	38
2	Early_Morning	15
3	Evening	78
4	Late_Night	14
5	Morning	62
6	Night	91

	departure_time	n
5	Morning	71146
2	Early_Morning	66790
3	Evening	65102
6	Night	48015
1	Afternoon	47794
4	Late_Night	1306

	class	n
1	Business	93487
2	Economy	206666

	stops	n
1	1	250863
2	2	13286
3	3	36004

	source_city	n
3	Delhi	61343
6	Mumbai	60896
1	Bangalore	52061
5	Kolkata	46347
4	Hyderabad	40806
2	Chennai	38700

	destination_city	n
6	Mumbai	59097
3	Delhi	57360
1	Bangalore	51068
5	Kolkata	49534
4	Hyderabad	42726
2	Chennai	40368

	duration	n
17	17	4242
18	18	4036
24	24	2879
16	16	2755
25	25	2323
137	137	2224
19	19	2053

	days_left	n
25	25	6633
18	18	6602
39	39	6593
32	32	6585
26	26	6573
24	24	6542
19	19	6537
31	31	6534
33	33	6532
40	40	6531
41	41	6525
28	28	6522

GETTING CONFIRMATION

	X	duration	days_left	price
197713	197712	0.83	49	1443
197714	197713	0.83	49	1443
197627	197626	0.83	48	1443
197628	197627	0.83	48	1443
197629	197628	0.83	48	1443

duration	days_left	price
2.00	49	2476
2.17	49	2476
2.17	49	2476
2.08	49	2700
2.17	49	2700
2.17	49	2700
2.17	49	2700
2.17	49	2700

	X	duration	days_left	price
206590	206589	1.00	49	1105
206591	206590	1.17	49	1105
206592	206591	1.17	49	1105
206593	206592	4.00	49	1105
206594	206593	4.33	49	1105
206595	206594	5.17	49	1105

Model: All Variables

Title

```
Call:
lm(formula = price ~ duration + stops + days_left + airline +
    departure_time + arrival_time + destination_city + class,
    data = Flight)

Residuals:
   Min     1Q Median     3Q    Max
-36248  -3058   -433   3119  65766

Coefficients:
                             Estimate Std. Error   t value Pr(>|t|)
(Intercept)                 5.198e+04  7.517e+01   691.446  < 2e-16 ***
duration                    5.599e+01  2.357e+00    23.750  < 2e-16 ***
stopstwo_or_more            2.299e+03  6.262e+01    36.714  < 2e-16 ***
stopszero                  -7.661e+03  4.636e+01  -165.273  < 2e-16 ***
days_left                  -1.311e+02  9.216e-01  -142.207  < 2e-16 ***
airlineAirAsia             -2.411e+01  6.345e+01    -0.380    0.704
airlineGO_FIRST             1.706e+03  5.507e+01    30.988  < 2e-16 ***
airlineIndigo               2.221e+03  4.752e+01    46.734  < 2e-16 ***
airlineSpiceJet             2.450e+03  7.768e+01    31.544  < 2e-16 ***
airlineVistara              3.986e+03  3.136e+01   127.096  < 2e-16 ***
departure_timeEarly_Morning 8.930e+02  4.167e+01    21.433  < 2e-16 ***
departure_timeEvening       7.839e+02  4.231e+01    18.527  < 2e-16 ***
departure_timeLate_Night    1.817e+03  1.936e+02     9.388  < 2e-16 ***
departure_timeMorning       9.487e+02  4.083e+01    23.235  < 2e-16 ***
departure_timeNight         7.769e+02  4.596e+01    16.905  < 2e-16 ***
arrival_timeEarly_Morning  -8.739e+02  6.691e+01   -13.060  < 2e-16 ***
arrival_timeEvening         9.863e+02  4.329e+01    22.785  < 2e-16 ***
arrival_timeLate_Night      9.874e+02  7.045e+01    14.015  < 2e-16 ***
arrival_timeMorning         5.099e+02  4.552e+01    11.200  < 2e-16 ***
arrival_timeNight           1.184e+03  4.240e+01    27.918  < 2e-16 ***
destination_cityChennai    -2.699e+02  4.586e+01    -5.885 3.99e-09 ***
destination_cityDelhi      -1.154e+03  4.202e+01   -27.469  < 2e-16 ***
destination_cityHyderabad  -1.398e+03  4.516e+01   -30.954  < 2e-16 ***
destination_cityKolkata     1.076e+03  4.344e+01    24.758  < 2e-16 ***
destination_cityMumbai      4.068e+01  4.157e+01     0.979    0.328
classEconomy               -4.495e+04  3.043e+01 -1476.993  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6829 on 300127 degrees of freedom
Multiple R-squared:  0.9095,    Adjusted R-squared:  0.9095
F-statistic: 1.206e+05 on 25 and 300127 DF,  p-value: < 2.2e-16
```