A business report on Indian Domestic Flight ticket prices

This report is designed for the chief R&D officer of Reliance Industries, Phil Beaver, by Sr. Research analyst, Jane Joseph

Executive Summary

Business Scenario

Next year, Reliance Industries plans to enter the Airline business and launch its airline as Reliance Airlines. Keeping in mind that most people belong to the middle-class income category and the average monthly salary of an Indian is INR 31000 ($400), Reliance airlines has decided to operate as a reliable and low-cost airline. The strategy is to focus on the volume of the target audience rather than targeting rich people. The company is currently in its market research phase.

Problem Summary

The R&D team of Reliance airlines is expected to answer the following questions based on the given dataset.

Which factors influence flight ticket prices?

How can we meet the target of Reliance Airlines to make domestic flights affordable for most of the population of India?

How can we ensure that our flights go full?

Decision Summary

Our model proposes certain measures that will help Reliance Airlines achieve its goal. Reliance Airlines can be a successful low-cost airline if it launches short-haul flights that directly connect the source and destination with no halts. The flight should have only economy class seats. Most of the flights should arrive at their destination early in the morning. Potential customers should be targeted smartly so that they are encouraged to book flights in advance.

Report

To emerge as a successful low-cost airline, it is **recommended** that Reliance Airlines launches short-haul flights with only one class, i.e., the economy class. The flights should be direct flights with no halts between the source and destination locations. Potential customers should be encouraged to buy tickets in advance with the help of marketing strategies. The model also says that arrival times of flights impact the flight ticket prices. Reliance Airlines can sell tickets for a cheaper rate if it can plan all its domestic flights to arrive early in the morning. Suppose all the major metro cities are connected as per the suggestions given above. In that case, there are high chances Reliance Airlines will emerge as a popular and most preferred low-cost domestic airline in India.

We have a pretty good model with an R square greater than 0.85. This means that our input variables explain more than 85% of the variability in flight ticket prices. Some of the significant findings of this model are as follows.

- More the flight duration, expensive the flight ticket. Hence the suggestion of short-haul flights. The regression model shows that when all factors remain constant, the ticket price goes up by INR 134 every hour the flight travels.
- Tickets that are booked in advance are generally given for a lower price. The regression model shows that when all factors remain constant, the ticket price goes up by INR 170 each day as you get closer to the flight take-off date.
- The model says the class type significantly impacts flight ticket prices. Therefore, it is suggested to have a single class configuration airline that caters only to economy class passengers.
- The model highlights that direct flights and flights that have two or more stops between the source and the destination are cheaper than one-stop flights. Keeping public interest in mind it is suggested to have direct flights.

MODEL EQUATION

Flight_ticket_price = 10193.646 + (-4217.052*Air_India) + (-5288.878*AirAsia) + (-4525.402*GO_FIRST) + (-3054.036*Indigo) + (-3230.177*SpiceJet) + (4761.446*Stops_one) + (2400.2907*Stops_two_or_more) + (Class*34169.729) + (duration*134.22095) + (-170.286* Booking_day) + (Departure_Morning*961.54253) + (-1255.341* Arrival_Early_Morning)

The model was built through a process of both – backward elimination and forward selection. More details are in Appendix C.

Please refer:

Appendix A for more information on the Model and its Interpretation.

Appendix B for more information on the Model Statistical Analysis.

Appendix C for more information on the Modelling Process.

Appendix D for more information on the Initial Data Analysis

Appendix E for more information on the Data

Appendix A - Model and Interpretation

13 Input Variables are a part of the final model.

Variables	***Coefficients***
Intercept	10193.64581
Air_India	-4217.052717
AirAsia	-5288.87766
GO_FIRST	-4525.40152
Indigo	-3054.035942
SpiceJet	-3230.176709
Stops_one	4761.44617
Stops_two_or_more	2400.290689
Class	34169.72905
duration	134.2209495
Booking_day	-170.2860447
Departure_Morning	961.5425265
Departure_Evening	663.4942835
Arrival_Early_Morning	-1255.341487

Impact analysis

	MIN	MAX	RANGE	CORFF	IMPACT
duration	0.83	30.08	29.25	134.2209	3925.963
Booking_day	1	49	48	-170.286	-8173.73

Note that all other variables are categorical variables with a Min value of 0 and a Max value of 1 therefore their Impact will be the same as their coefficients.

Coefficient Analysis

The positive coefficient shows a positive relationship with Flight ticket prices.

The negative coefficient shows a negative relationship with the Flight ticket prices.

Note that all the airlines show a negative coefficient value. This means that the base airline which is VISTARA is the most expensive airline in comparison to these airlines.

Note that duration has a positive coefficient this means that the higher the duration of the flight, the more expensive the flight ticket.

Note that booking day has a negative value. This means the booking day is inversely proportional to the flight ticket price.

Note the Class has the highest impact. Therefore, having business class flights will increase the cost of the ticket in general for everyone.

Appendix B – Model Statistical Analysis

Regression Statistics

The regression Statistics of our final model are as follows

Regression Statistics	
Multiple R	0.923626
R Square	0.853084
Adjusted R Square	0.852046
Standard Error	5814.802
Observations	1854

Note the difference between R square and adjusted R square is 0.0010.

This indicates that the model is good.

More than 85% of the variability in flight ticket prices (target variable) is explained by this model.

Anova Table

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	13	3.61252E+11	2.78E+10	821.8595449	0
Residual	1840	62213927496	33811917		
Total	1853	4.23466E+11			

Null Hypothesis: Coefficients of any one input variable = 0

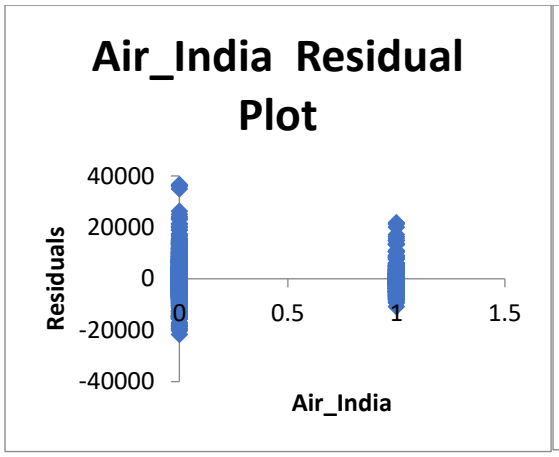
Alternative Hypothesis: At least one of the coefficients of the Input Variables is not equal to 0

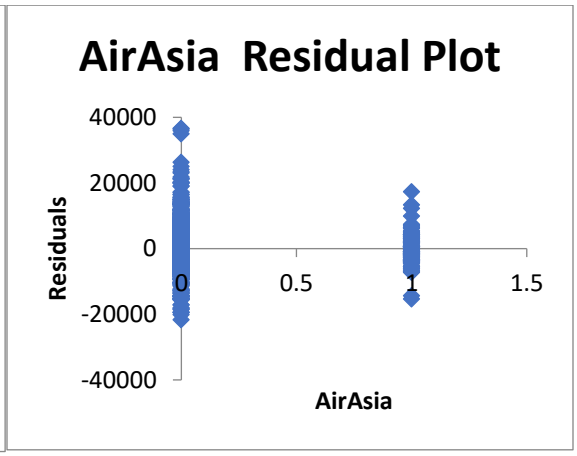
Note Significance F here is the P-value and $0 < 0.05$ so we reject the Null Hypothesis.

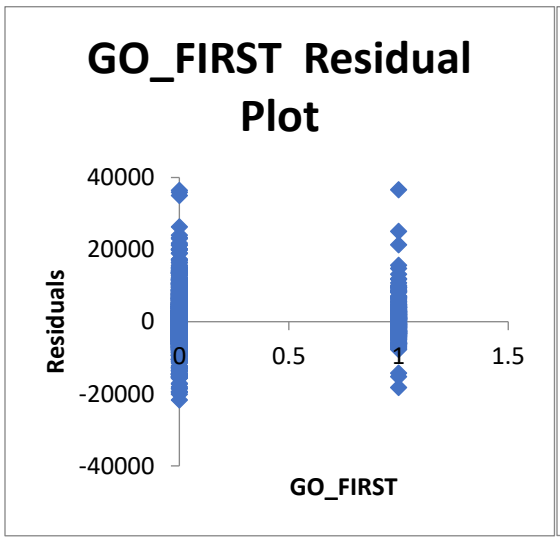
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	10193.6458	500.3595223	20.37264	2.32E-83	9212.31365	11174.978	9212.313646	11174.97797
Air_India	-4217.0527	476.0057205	-8.85925	1.86E-18	-5150.62089	-3283.4845	-5150.62089	-3283.484548
AirAsia	-5288.8777	716.0163352	-7.38653	2.27E-13	-6693.16763	-3884.5877	-6693.16763	-3884.587689
GO_FIRST	-4525.4015	608.6169273	-7.43555	1.59E-13	-5719.05396	-3331.7491	-5719.05396	-3331.749079
Indigo	-3054.0359	507.7300701	-6.01508	2.16E-09	-4049.82362	-2058.2483	-4049.82362	-2058.248262
SpiceJet	-3230.1767	610.8504051	-5.288	1.38E-07	-4428.20957	-2032.1438	-4428.20957	-2032.14385
Stops_one	4761.44617	395.9554363	12.02521	4.05E-32	3984.87695	5538.01539	3984.876948	5538.015391
Stops_two_or_more	2400.29069	994.2911711	2.414072	0.015872	450.233057	4350.34832	450.233057	4350.348322
Class	34169.7291	460.5249766	74.19734	0	33266.5226	35072.9355	33266.52255	35072.93555
duration	134.220949	29.03499772	4.62273	4.05E-06	77.2759412	191.165958	77.27594124	191.1659577
Booking_day	-170.28604	8.65004539	-19.6861	1.91E-78	-187.250982	-153.32111	-187.250982	-153.3211077
Departure_Morning	961.542527	352.6774159	2.726408	0.006464	269.8525	1653.23255	269.8524999	1653.232553
Departure_Evening	663.494284	339.1849447	1.956143	0.050599	-1.7335788	1328.72215	-1.7335788	1328.722146
Arrival_Early_Morning	-1255.3415	683.5986446	-1.83637	0.066464	-2596.05213	85.3691555	-2596.05213	85.3691555

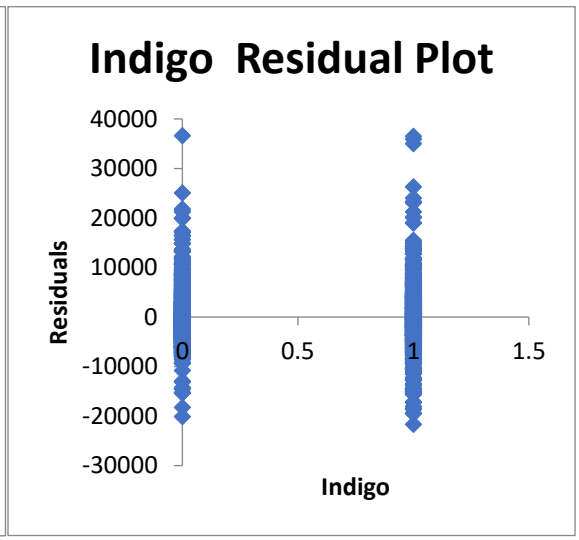
Note that most of our P-values are below 0.05. This means that most of our Input variables have a significant relation with the flight ticket price (output variable).

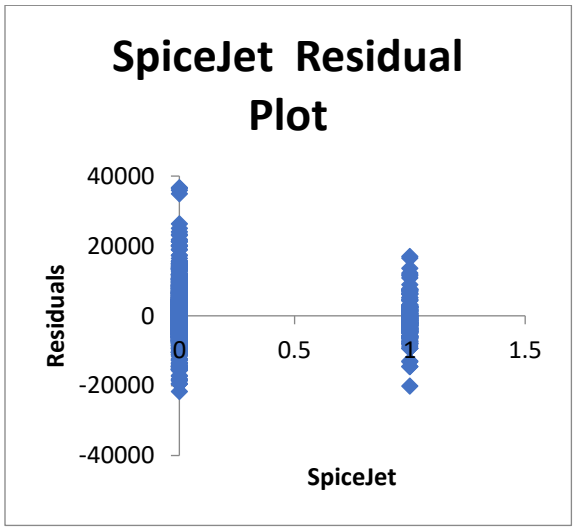
Residual Plots for different Airlines



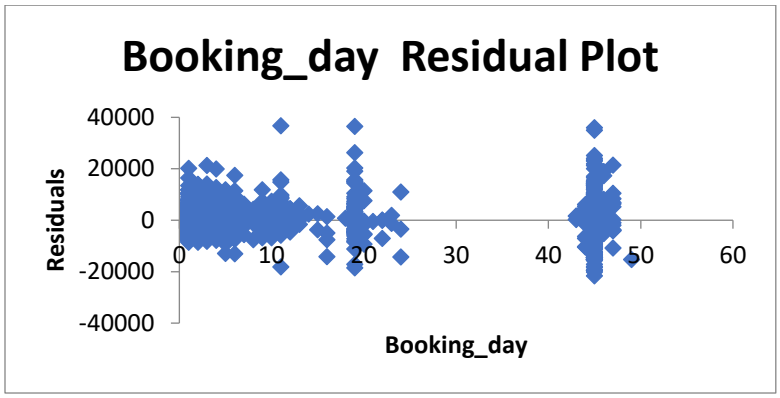




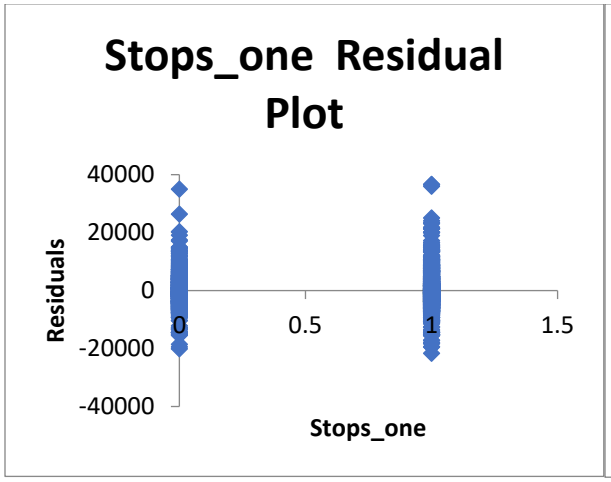


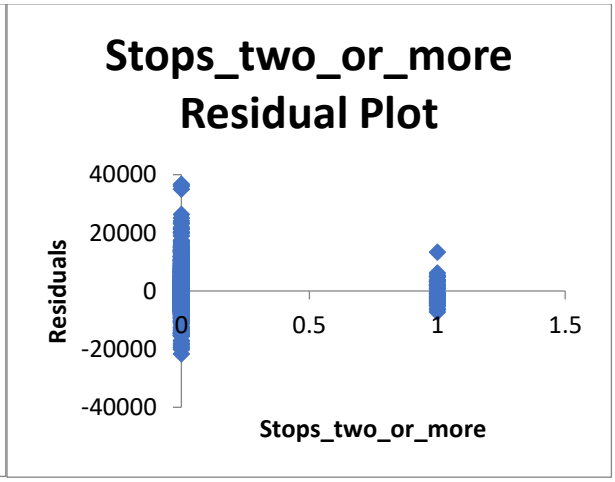


Residual Plot for Booking

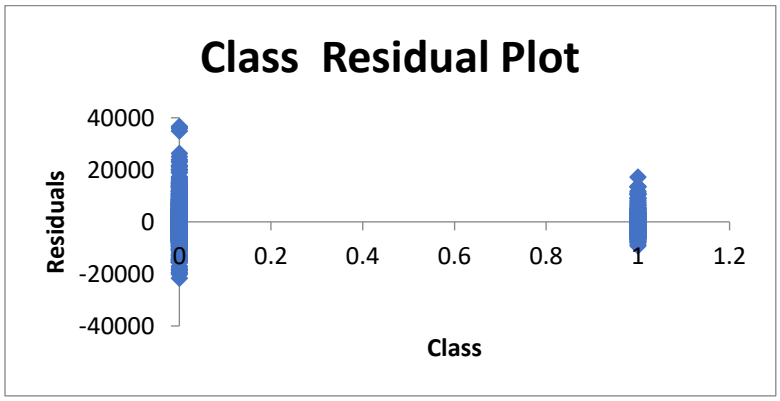


Residual Plots for Stops

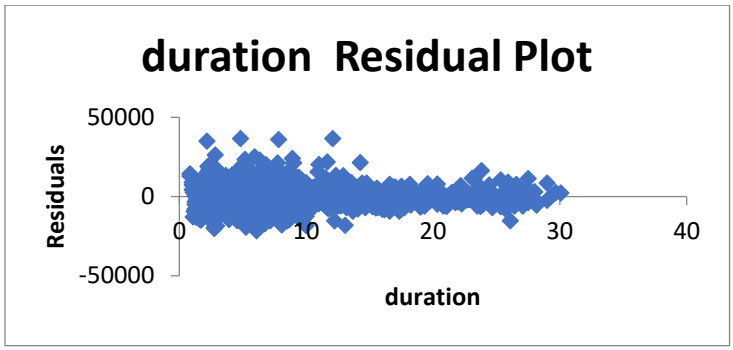




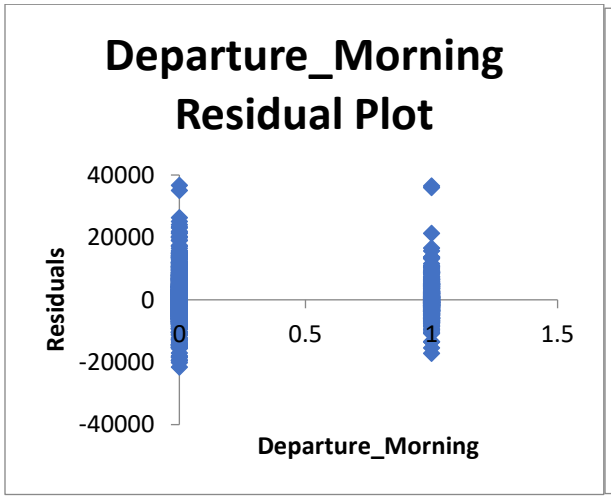
Residual Plot for Class

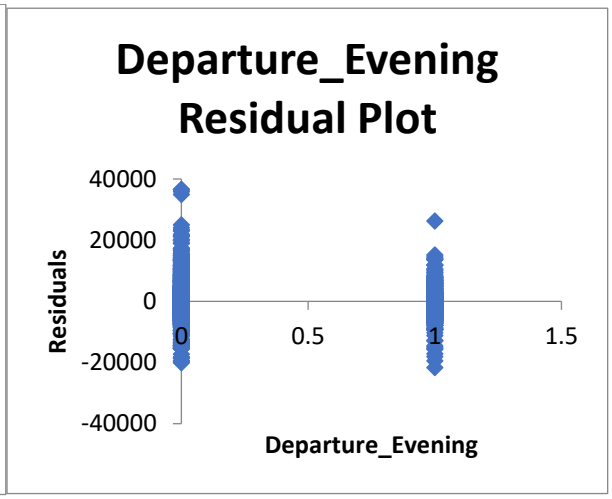


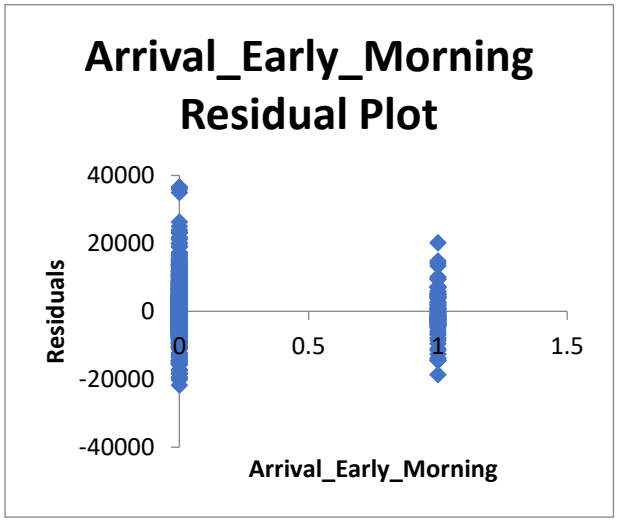
Residual Plot for flight duration



Residual Plots for Departure and Arrival times







From all these graphs we can say that we do have some critical outliers.

Appendix C - Model Process

We Started Regression Analysis with one Dependent Variable and thirty Independent Variables.

After running five regression models with different sets of input variables we could determine 10 Input variables that were constantly showing good P-values.

3 more regression analyses were done to finalize our top 16 input variables. The regression statistics and P-values of our model at this stage are as follows

Regression Statistics	
Multiple R	0.923618276
R Square	0.85307072
Adjusted R Square	0.851790987
Standard Error	5819.813177
Observations	1854

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	9162.966531	835.5642256	10.96620254	3.78131E-27
Air_India	-4220.138605	476.5139103	-8.856275783	1.90663E-18
AirAsia	-5392.828153	719.2348485	-7.498007312	1.00258E-13
GO_FIRST	-4568.178362	610.0982605	-7.487610861	1.08271E-13
Indigo	-3151.838675	529.4688245	-5.95283146	3.14941E-09
SpiceJet	-3233.156288	631.1005472	-5.12304466	3.3221E-07
Stops_one	4842.116933	396.12388	12.22374408	4.31312E-33
Stops_two_or_more	2544.221325	996.5462158	2.553038971	0.010758962
Class	34157.42667	460.9518594	74.10193922	0
duration	130.1611927	29.18714175	4.459538854	8.71043E-06
Booking_day	-171.0912229	8.641440168	-19.79892466	3.10287E-79
Destination_Bangalore	1049.188202	743.7921473	1.410593276	0.158533885
Destination_Delhi	1001.028223	682.5719212	1.466553475	0.142668723
Destination_Kolkata	1336.617953	754.8034193	1.770815975	0.076757109
Destination_Mumbai	965.7276316	707.7808711	1.36444438	0.172594868
Departure_Morning	1023.87785	351.1609682	2.915693777	0.003592
Departure_Evening	717.0842441	338.1330958	2.120715934	0.034079373

After this step, backward elimination was used to eliminate one variable at a time and we ended up deleting all the Destination Variables. At this point, the regression statistics and P-values of our model were as follows

Regression Statistics	
Multiple R	0.92348
R Square	0.852815
Adjusted R Square	0.851855
Standard Error	5818.547
Observations	1854

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	10147.653	500.0541701	20.29311	8.64E-83
Air_India	-4218.23858	476.3118675	-8.85604	1.91E-18
AirAsia	-5358.56284	715.4705943	-7.48956	1.07E-13
GO_FIRST	-4564.71555	608.6320354	-7.49996	9.87E-14
Indigo	-3136.04484	506.0880248	-6.19664	7.1E-10
SpiceJet	-3266.77456	610.9184391	-5.34732	1E-07
Stops_one	4828.80684	394.5066198	12.24012	3.56E-33
Stops_two_or_more	2532.93683	992.3026491	2.552585	0.010773
Class	34160.2637	460.7927266	74.13369	0
duration	131.598749	29.01854243	4.534988	6.13E-06
Booking_day	-171.831683	8.614543625	-19.9467	2.65E-80
Departure_Morning	1032.00946	350.809431	2.941795	0.003304
Departure_Evening	720.888008	337.9594794	2.13306	0.033052

Note all P-values were below 0.05 and we could use this as our final model but to experiment with the model and have more input variables to explain the flight ticket prices, the forward selection method was used at this stage.

The final model

Regression Statistics	
Multiple R	0.92362553
R Square	0.85308412
Adjusted R Square	0.85204613
Standard Error	5814.80155
Observations	1854

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	10193.64581	500.3595	20.37264	2.32081E-83
Air_India	-4217.052717	476.0057	-8.85925	1.85605E-18
AirAsia	-5288.87766	716.0163	-7.38653	2.27348E-13
GO_FIRST	-4525.40152	608.6169	-7.43555	1.58783E-13
Indigo	-3054.035942	507.7301	-6.01508	2.1638E-09
SpiceJet	-3230.176709	610.8504	-5.288	1.38401E-07
Stops_one	4761.44617	395.9554	12.02521	4.04936E-32
Stops_two_or_more	2400.290689	994.2912	2.414072	0.015872404
Class	34169.72905	460.525	74.19734	0
duration	134.2209495	29.035	4.62273	4.05066E-06
Booking_day	-170.2860447	8.650045	-19.6861	1.90664E-78
Departure_Morning	961.5425265	352.6774	2.726408	0.006463519
Departure_Evening	663.4942835	339.1849	1.956143	0.050599403
Arrival_Early_Morning	-1255.341487	683.5986	-1.83637	0.066463886

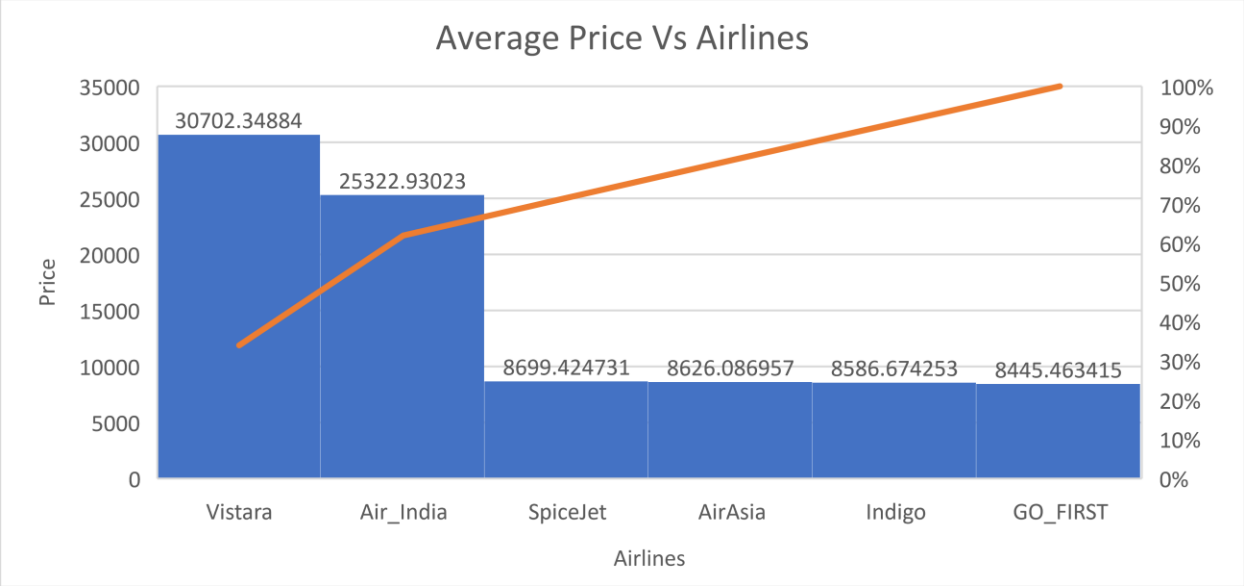
We were successful in adding one more input variable to the model while maintaining our P-values below 0.07.

Note the R Square for this model is higher than the previous model and the Standard Error went down.

Appendix D – Data Analysis

Input Variable – Airlines

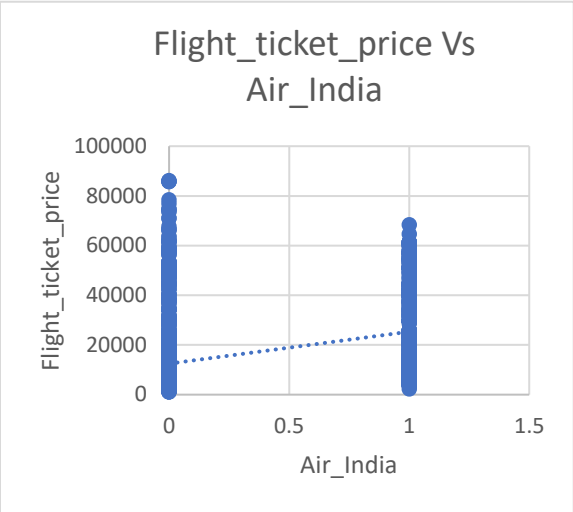
Histogram

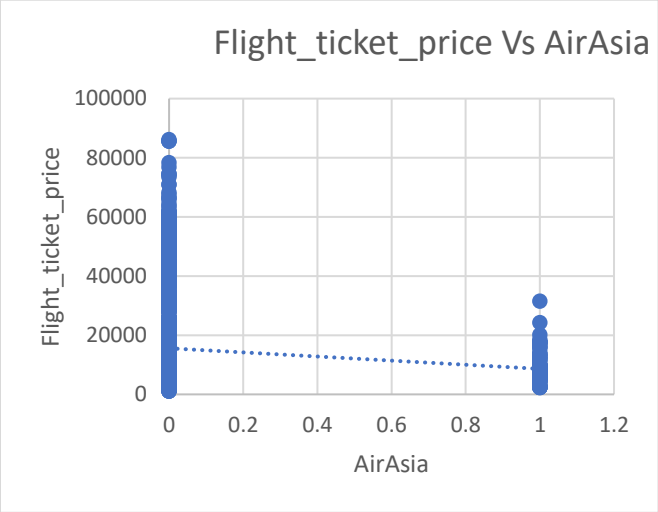


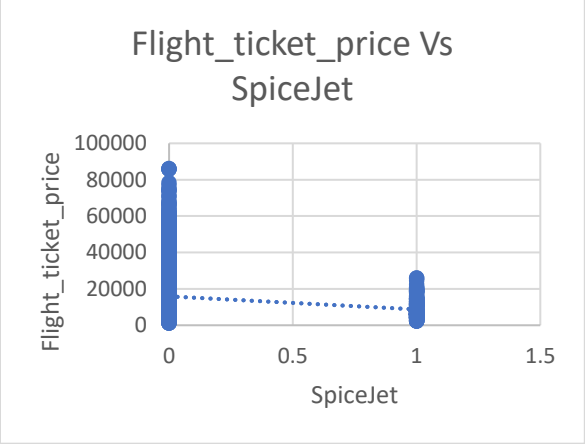
Descriptive Statistics

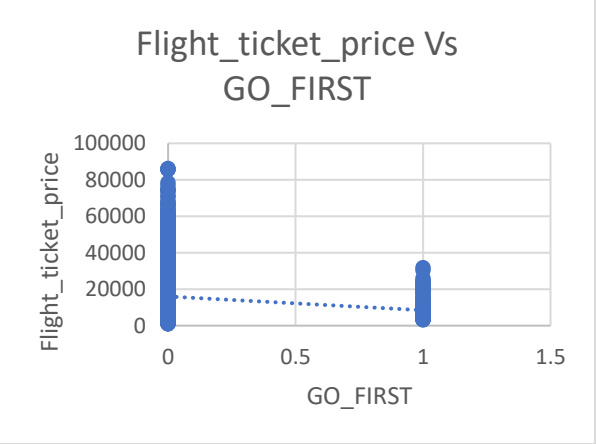
	Air_India	*AirAsia*	*GO_FIRST*	*Indigo*	*SpiceJet*
Mean	0.208738	0.062028	0.110572	0.37918	0.100324
Standard Error	0.009441	0.005603	0.007285	0.011271	0.006979
Median	0	0	0	0	0
Mode	0	0	0	0	0
Standard Deviation	0.406516	0.241272	0.313686	0.485314	0.300512
Sample Variance	0.165256	0.058212	0.098399	0.23553	0.090308
Kurtosis	0.057892	11.22134	4.182726	-1.75345	5.096221
Skewness	1.434514	3.634452	2.485602	0.498442	2.662841
Range	1	1	1	1	1
Minimum	0	0	0	0	0
Maximum	1	1	1	1	1
Sum	387	115	205	703	186
Count	1854	1854	1854	1854	1854

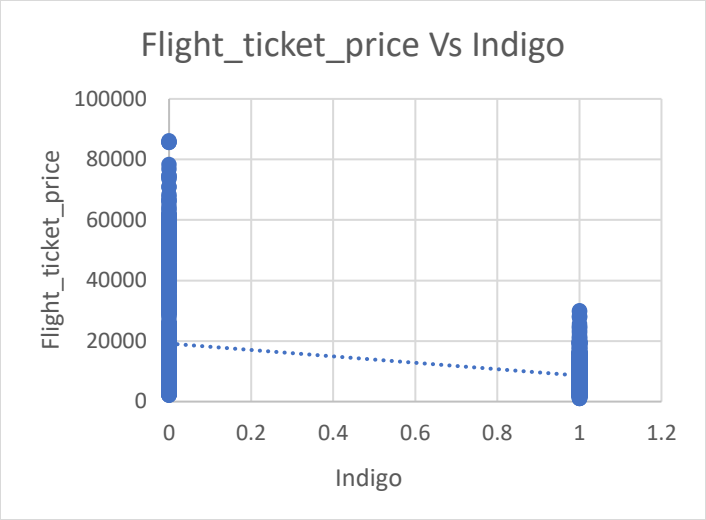
Scatter Plots









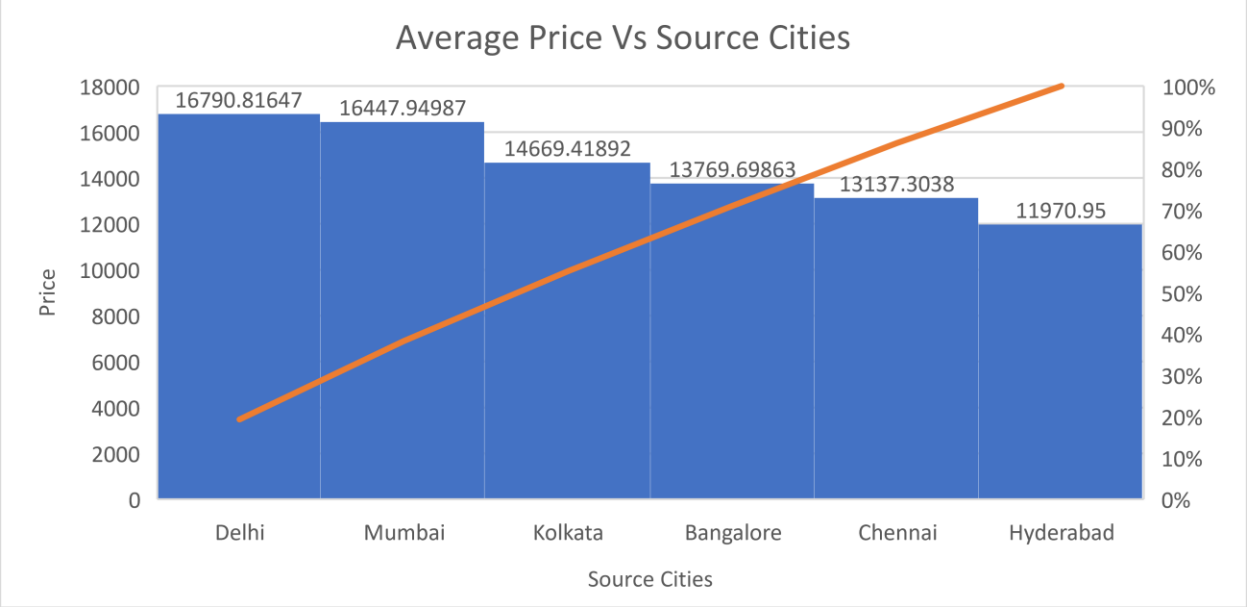


Observations:

Looks like prices do vary depending on the airline.
Vistara sells the most expensive flight tickets in comparison to the other airlines.
Indigo Airlines has the maximum number of flight count in this dataset

Input Variable – Source Cities

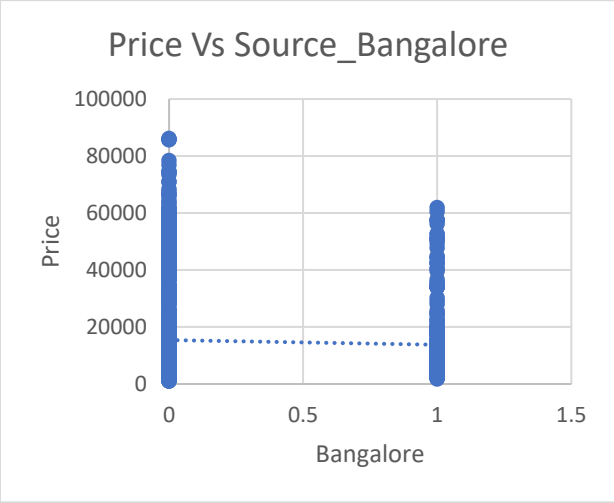
Histogram

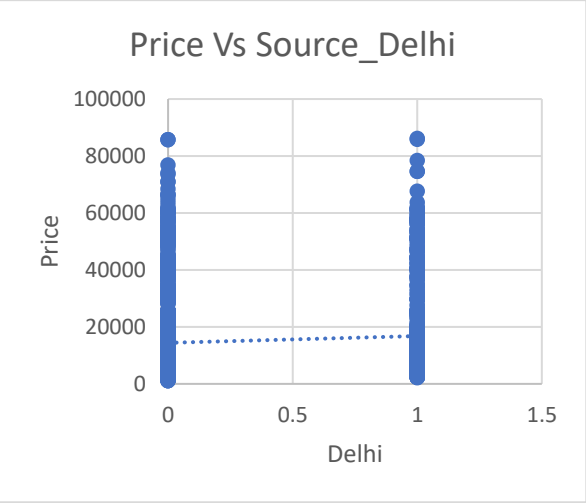


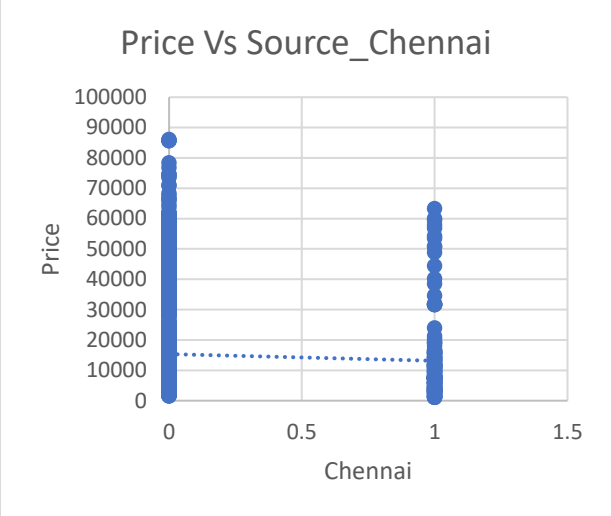
Descriptive Statistics

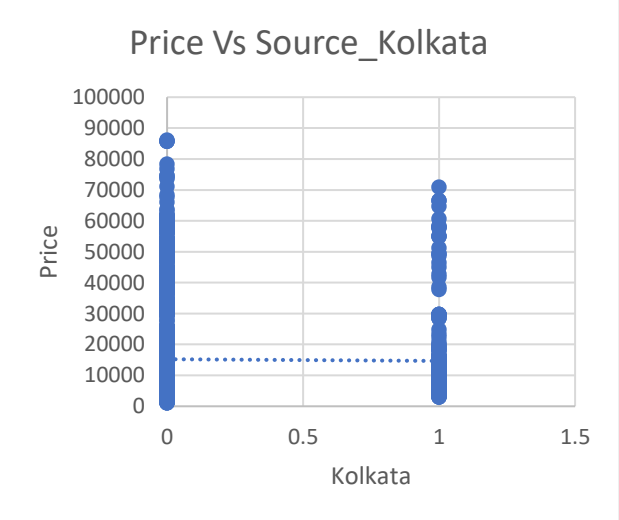
	Source_Bangalore	*Source Delhi*	*Source_Chennai*	*Source_Kolkata*	*Source_Mumbai*
Mean	0.157497303	0.314455232	0.085221143	0.1197411	0.215210356
Standard Error	0.008462224	0.01078598	0.006486255	0.007542045	0.009547079
Median	0	0	0	0	0
Mode	0	0	0	0	0
Standard Deviation	0.364367276	0.464423811	0.279285824	0.324746134	0.411079074
Sample Variance	0.132763512	0.215689476	0.078000572	0.105460052	0.168986005
Kurtosis	1.543649272	-1.361638032	6.849029104	3.500047318	-0.076127304
Skewness	1.882016266	0.799894992	2.973489849	2.344413105	1.387067404
Range	1	1	1	1	1
Minimum	0	0	0	0	0
Maximum	1	1	1	1	1
Sum	292	583	158	222	399
Count	1854	1854	1854	1854	1854

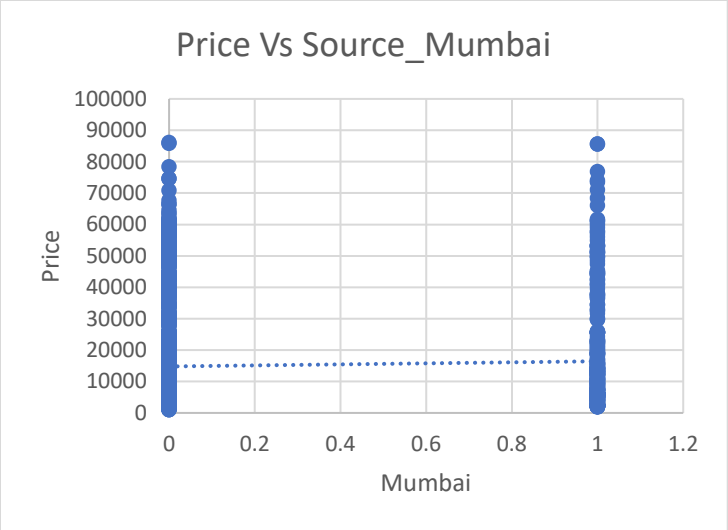
Scatter Plots









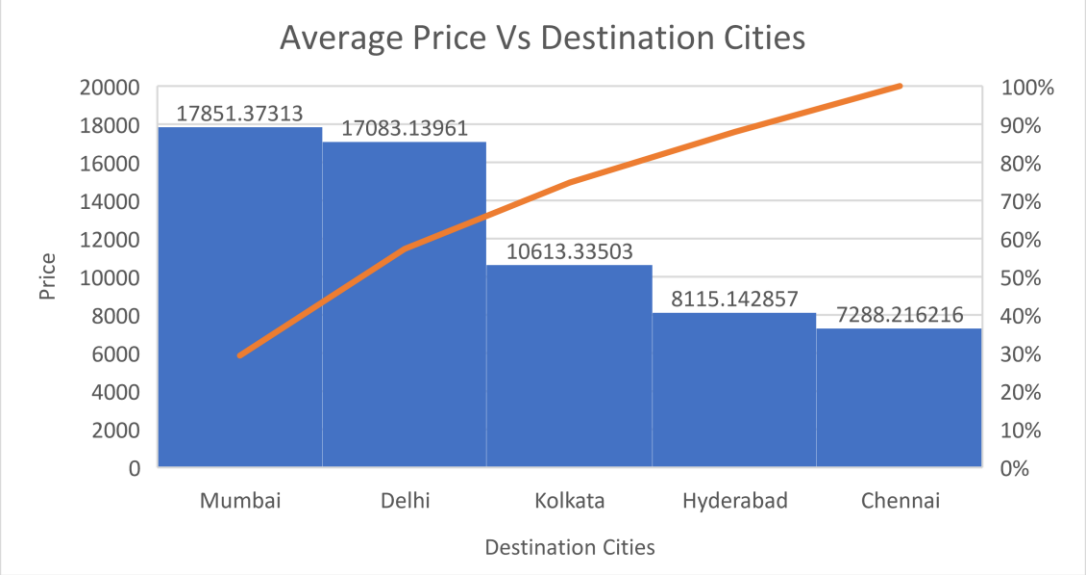


Observations:

Flights starting from Delhi seem to be more expensive.

Input Variable – Destination Cities

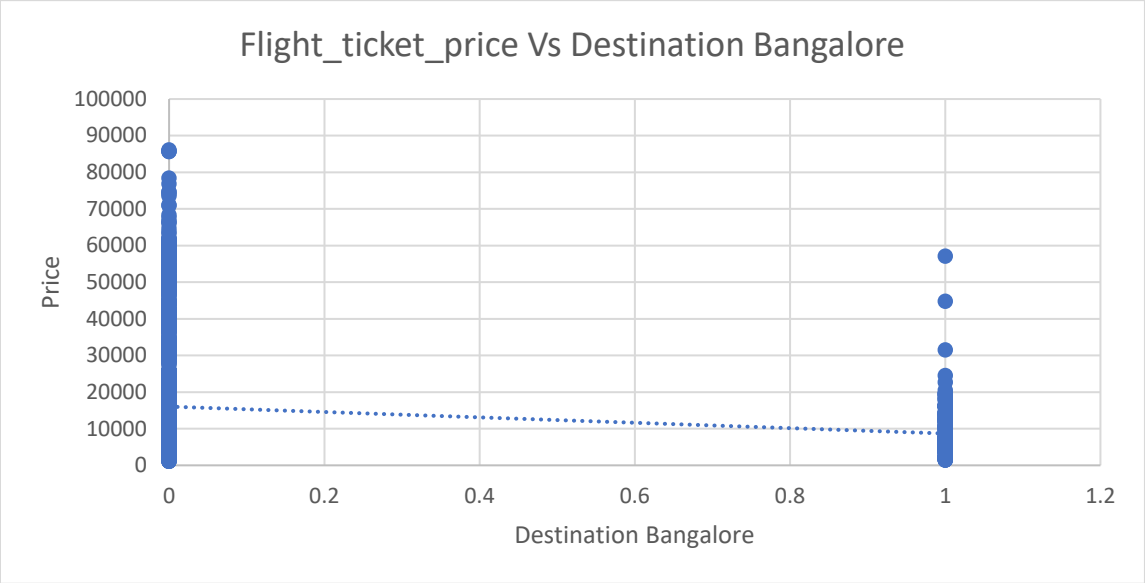
Histogram

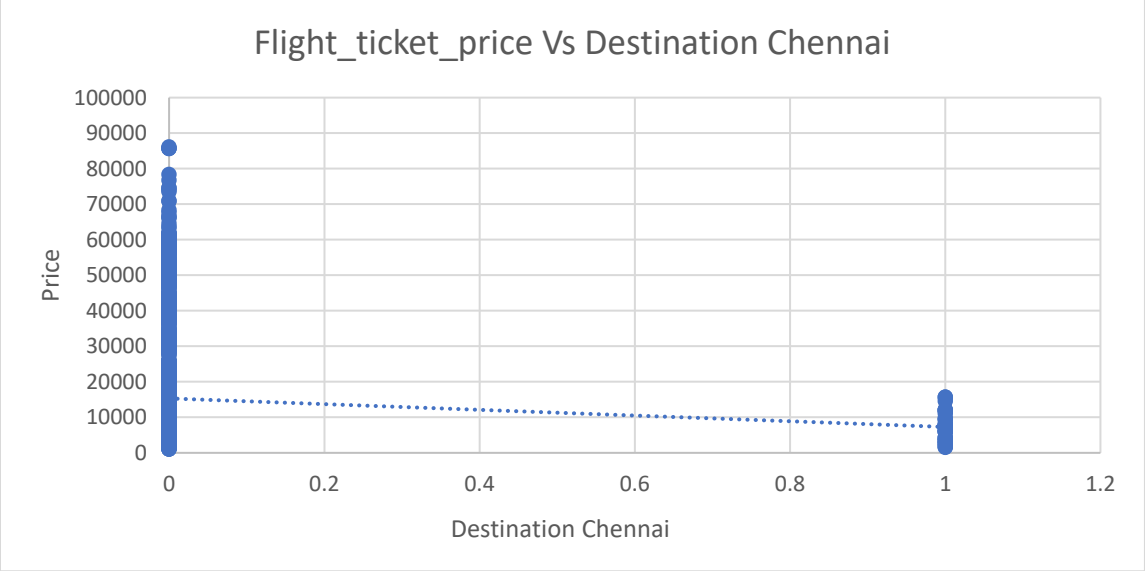


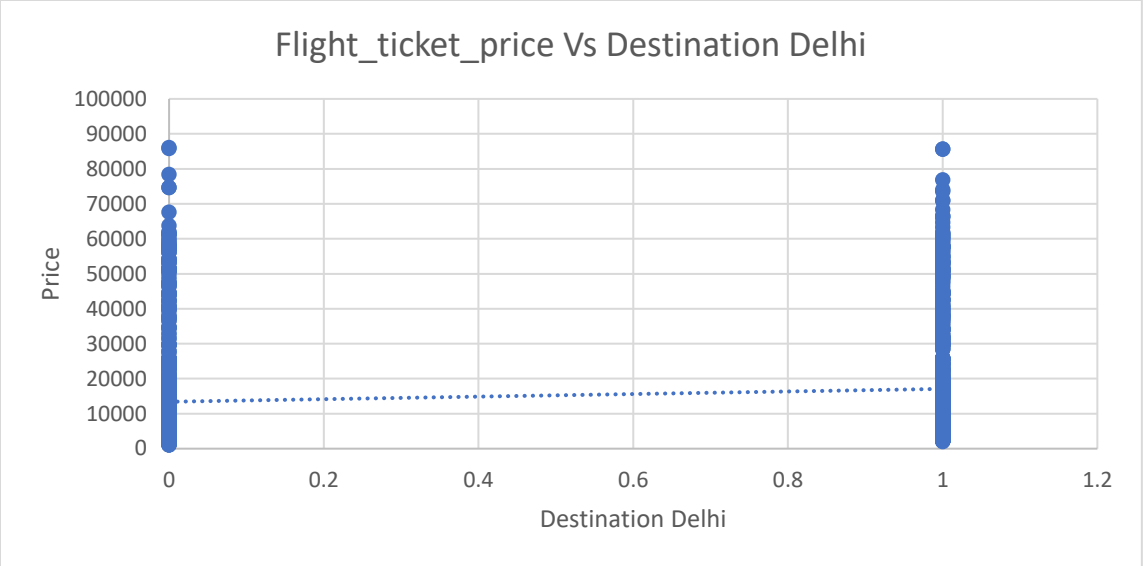
Descriptive Statistics

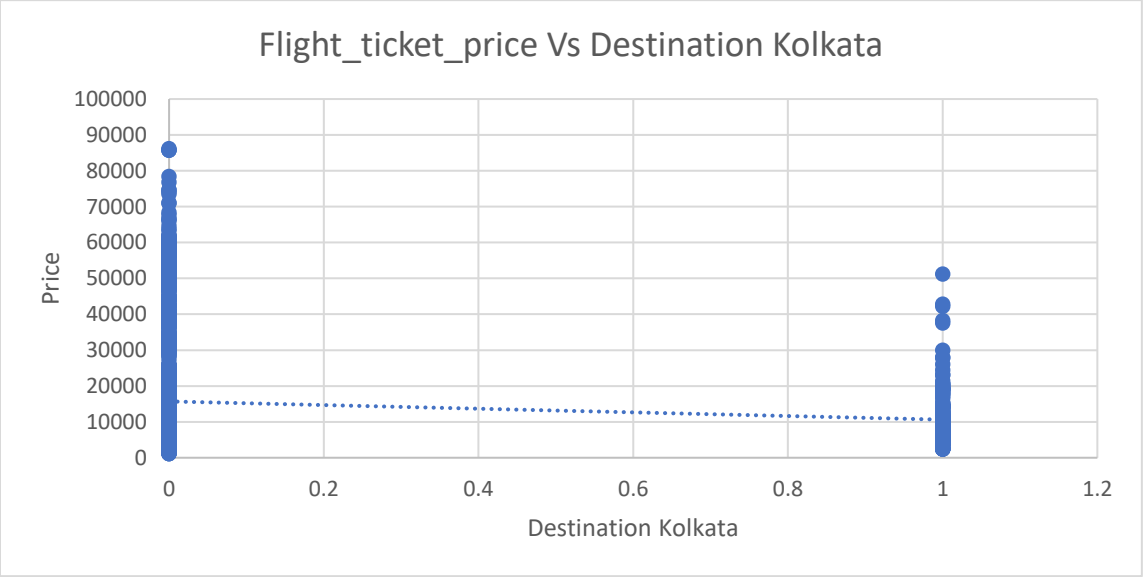
	Destination_Bangalore	*Destination_Delhi*	*Destination_Chennai*	*Destination_Kolkata*	*Destination_Mumbai*
Mean	0.119201726	0.475188781	0.01995685	0.106256742	0.252966559
Standard Error	0.007527344	0.01160104	0.003248859	0.00715891	0.010098663
Median	0	0	0	0	0
Mode	0	0	0	0	0
Standard Deviation	0.324113152	0.499518756	0.139889704	0.308249083	0.434829233
Sample Variance	0.105049335	0.249518988	0.019569129	0.095017497	0.189076462
Kurtosis	3.537240798	-1.992260455	45.25365703	4.545541061	-0.706950772
Skewness	2.352323568	0.099447769	6.870577912	2.557467248	1.137459017
Range	1	1	1	1	1
Minimum	0	0	0	0	0
Maximum	1	1	1	1	1
Sum	221	881	37	197	469
Count	1854	1854	1854	1854	1854

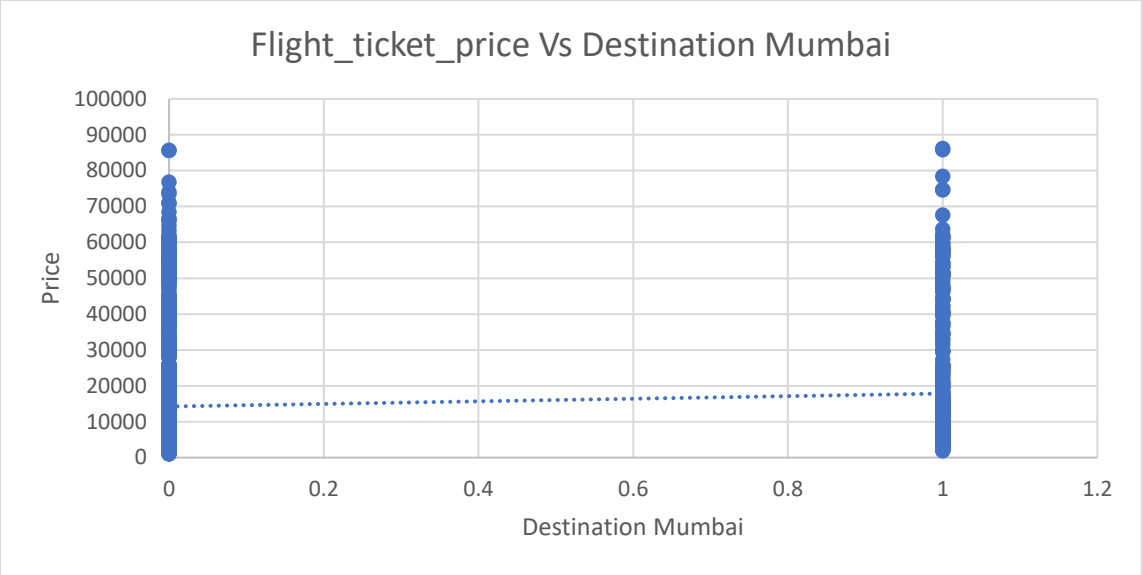
Scatter Plots



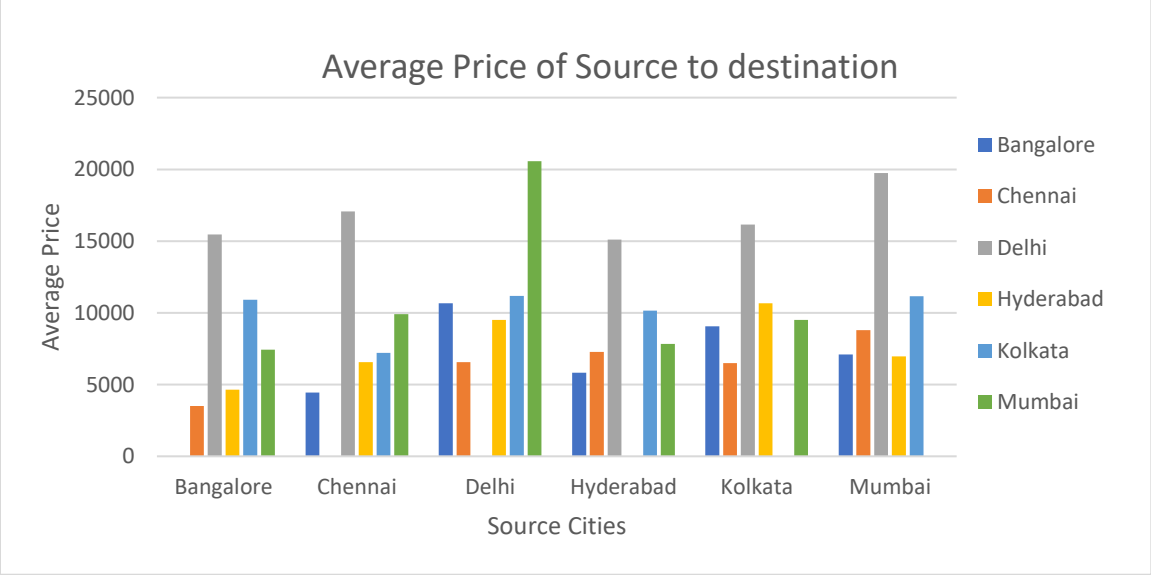



Flight_ticket_price Vs Destination Delhi
Price
100000
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
0
0.2
0.4
0.6
0.8
1
1.2
Destination Delhi


Flight_ticket_price Vs Destination Kolkata
Price
100000
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
0
0.2
0.4
0.6
0.8
1
1.2
Destination Kolkata



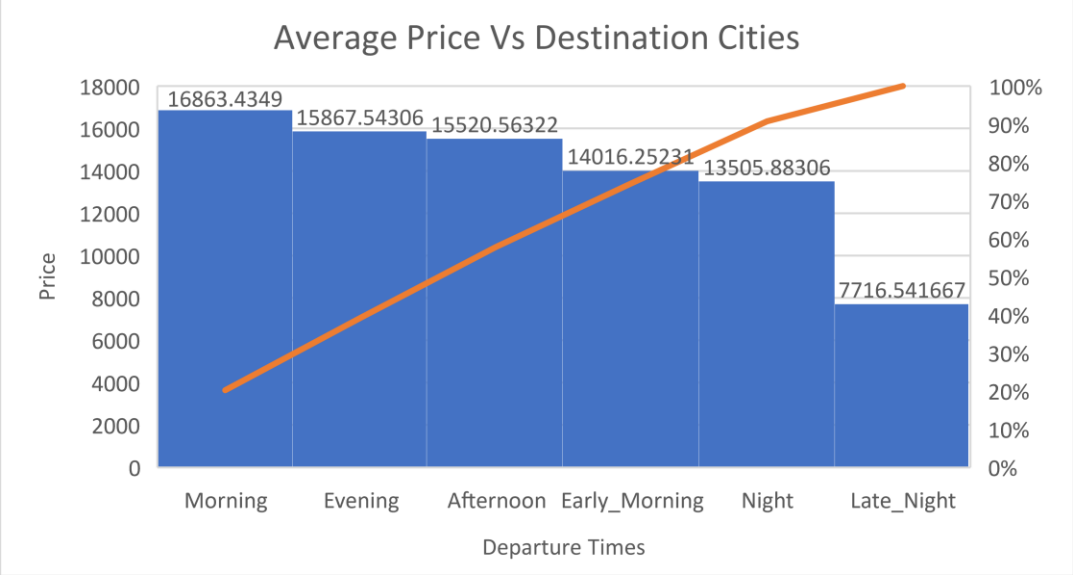
Observations:



Flights from all other cities to Delhi seem to be expensive.
For Delhi, flights from Delhi to Mumbai are the most expensive.
Bangalore to Chennai seems to be the cheapest route.

Input Variable – Departure Times

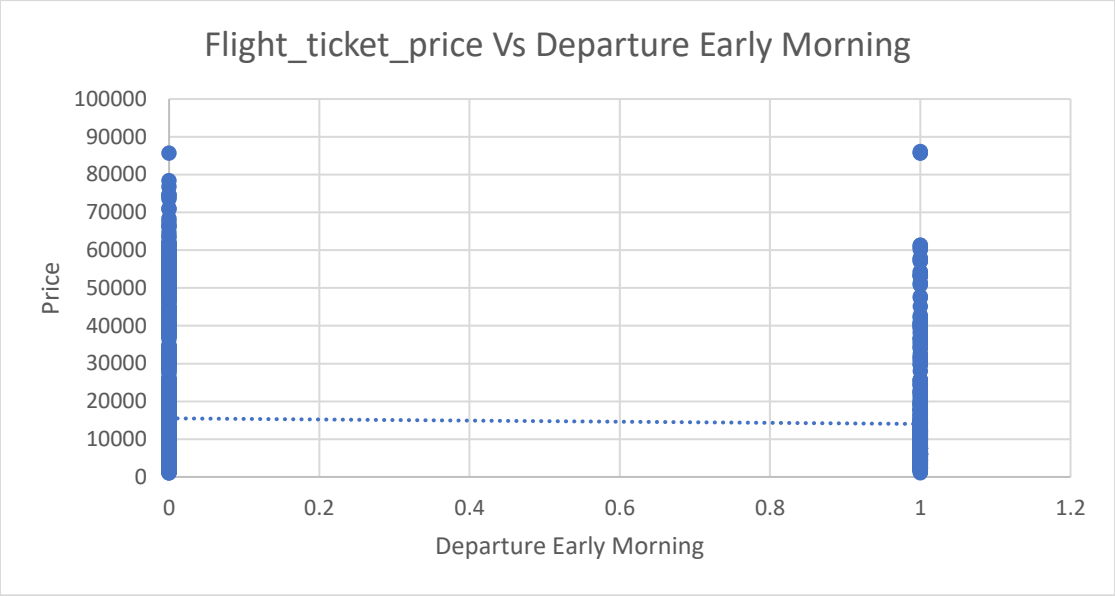
Histogram

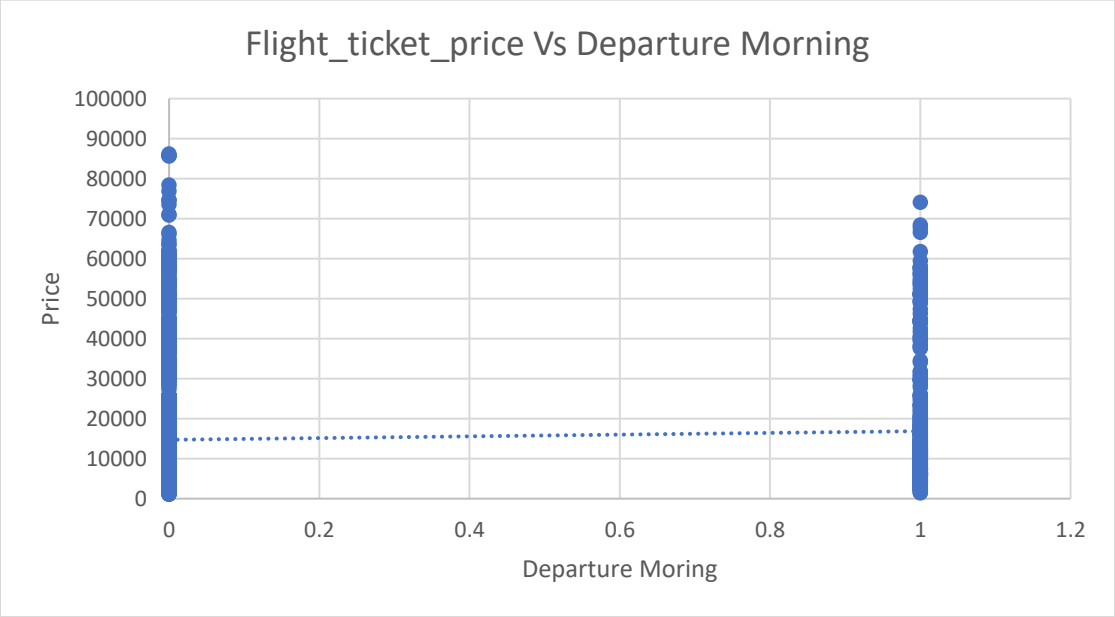


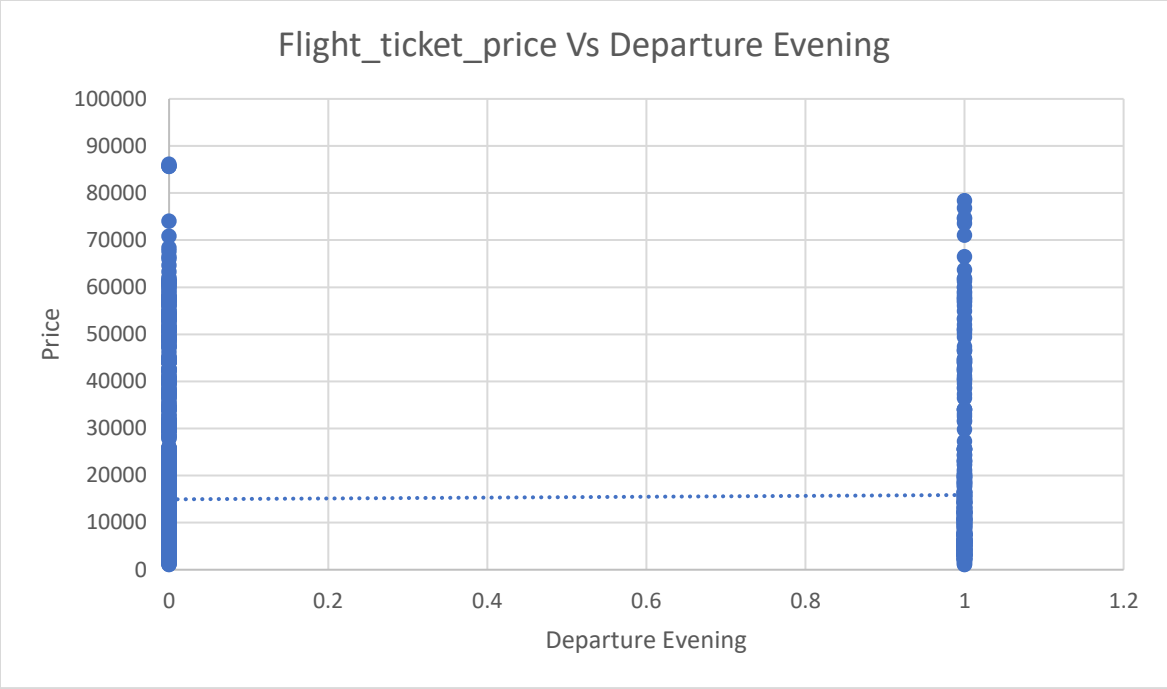
Descriptive Statistics

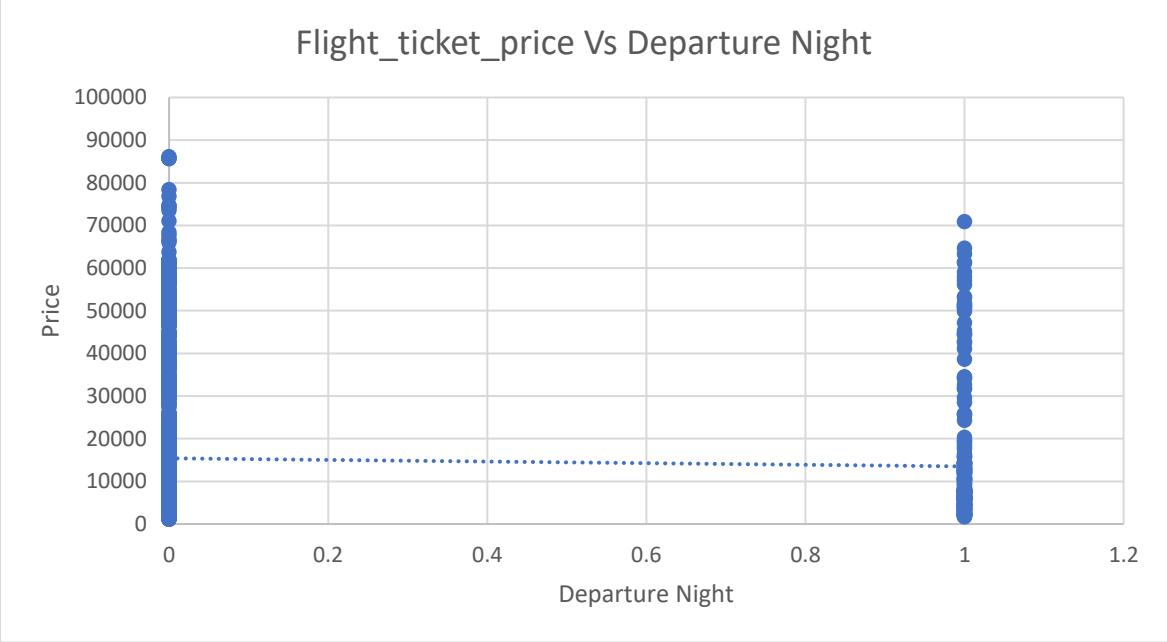
	Departure_Early_Morning	*Departure_Morning*	*Departure_Evening*	*Departure_Night*	*Departure_Late_Night*
Mean	0.233009709	0.207119741	0.225458468	0.133764833	0.012944984
Standard Error	0.00982074	0.009414057	0.009707735	0.007907718	0.002625934
Median	0	0	0	0	0
Mode	0	0	0	0	0
Standard Deviation	0.422862426	0.405351426	0.417996636	0.340491313	0.113067709
Sample Variance	0.178812631	0.164309778	0.174721187	0.115934334	0.012784307
Kurtosis	-0.402385727	0.092834164	-0.271002757	2.640577798	72.46162144
Skewness	1.26413983	1.446628903	1.315025	2.15353905	8.624584319
Range	1	1	1	1	1
Minimum	0	0	0	0	0
Maximum	1	1	1	1	1
Sum	432	384	418	248	24
Count	1854	1854	1854	1854	1854

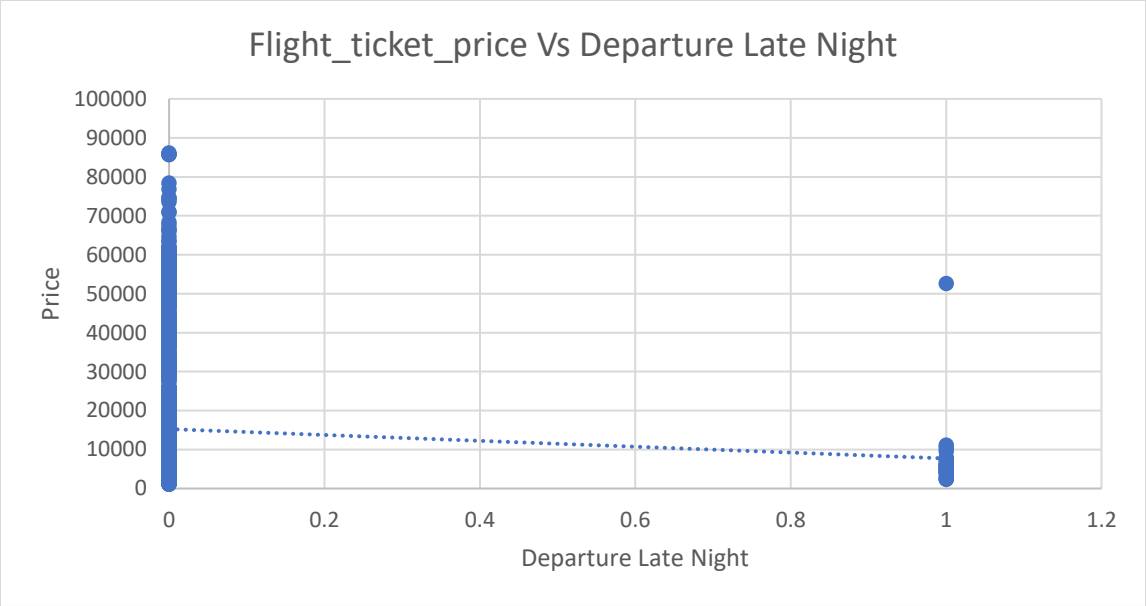
Scatter Plots






Flight_ticket_price Vs Departure Evening
Price
100000
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
0
0.2
0.4
0.6
0.8
1
1.2
Departure Evening


Flight_ticket_price Vs Departure Night
Price
100000
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
0
0.2
0.4
0.6
0.8
1
1.2
Departure Night

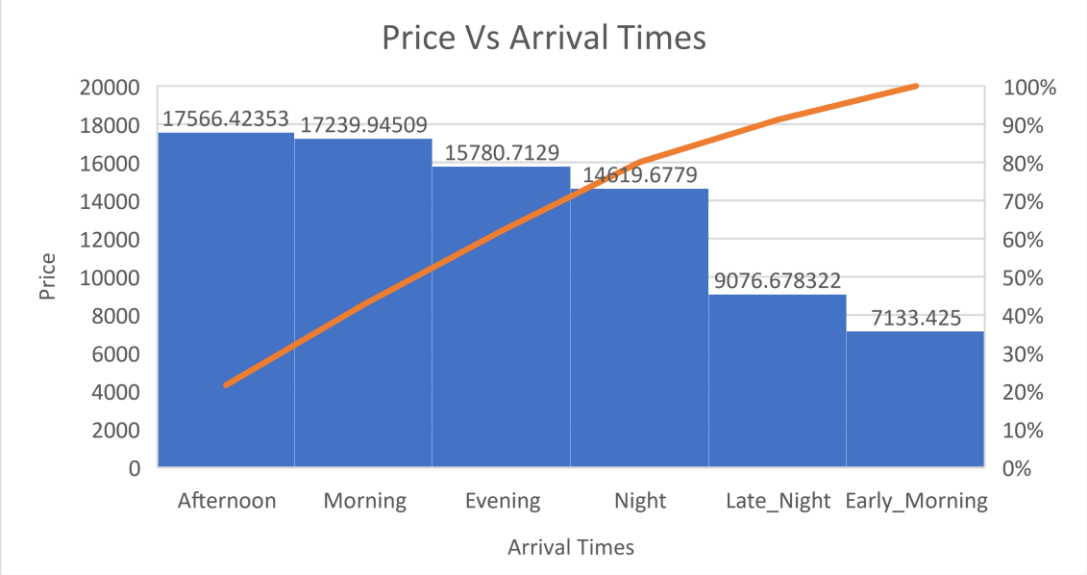


Observations:

Late Night flight seems to be the cheapest.
Morning flights are expensive.

Input Variable – Arrival Times

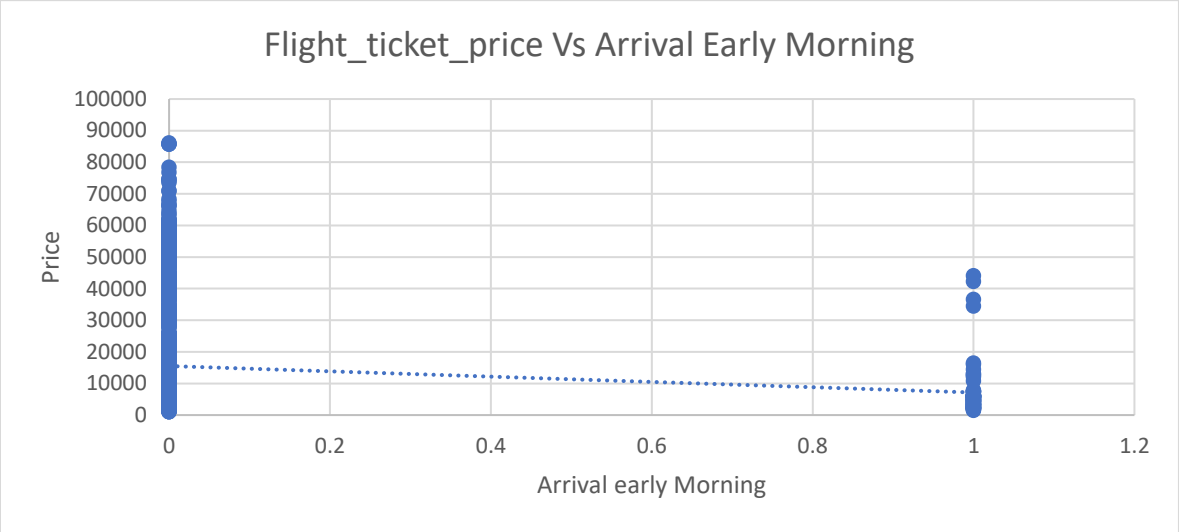
Histogram

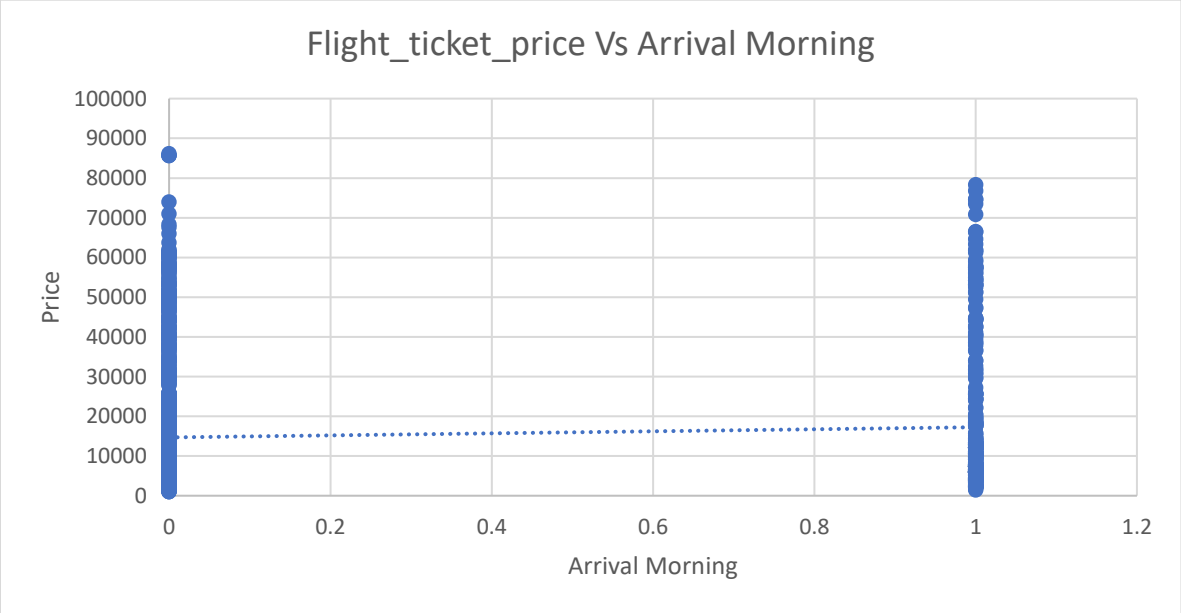


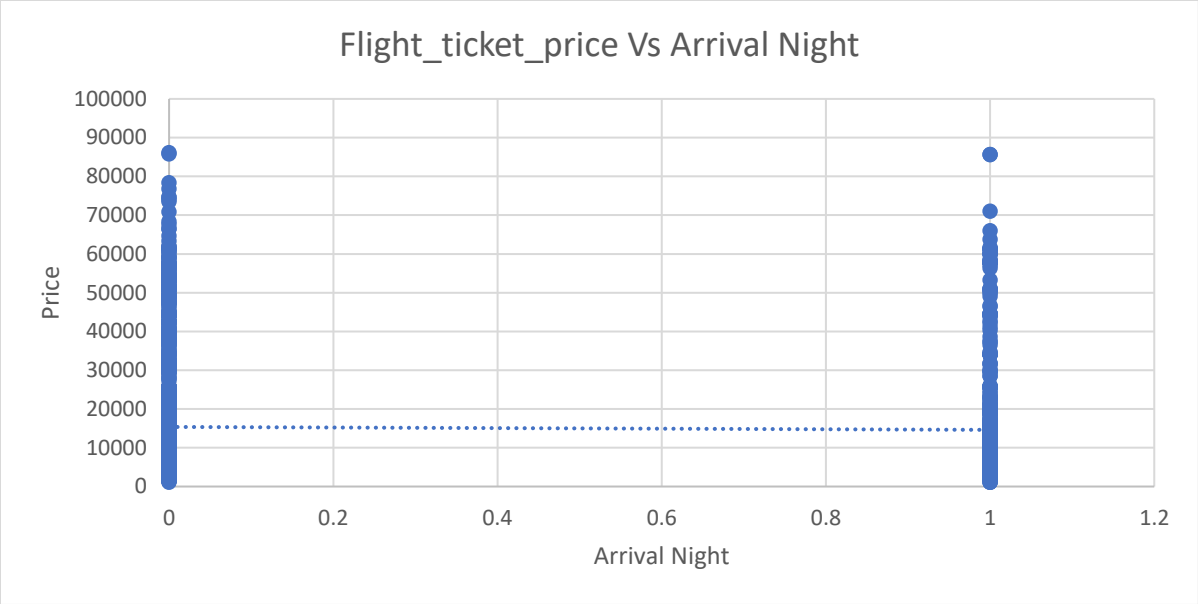
Descriptive Statistics

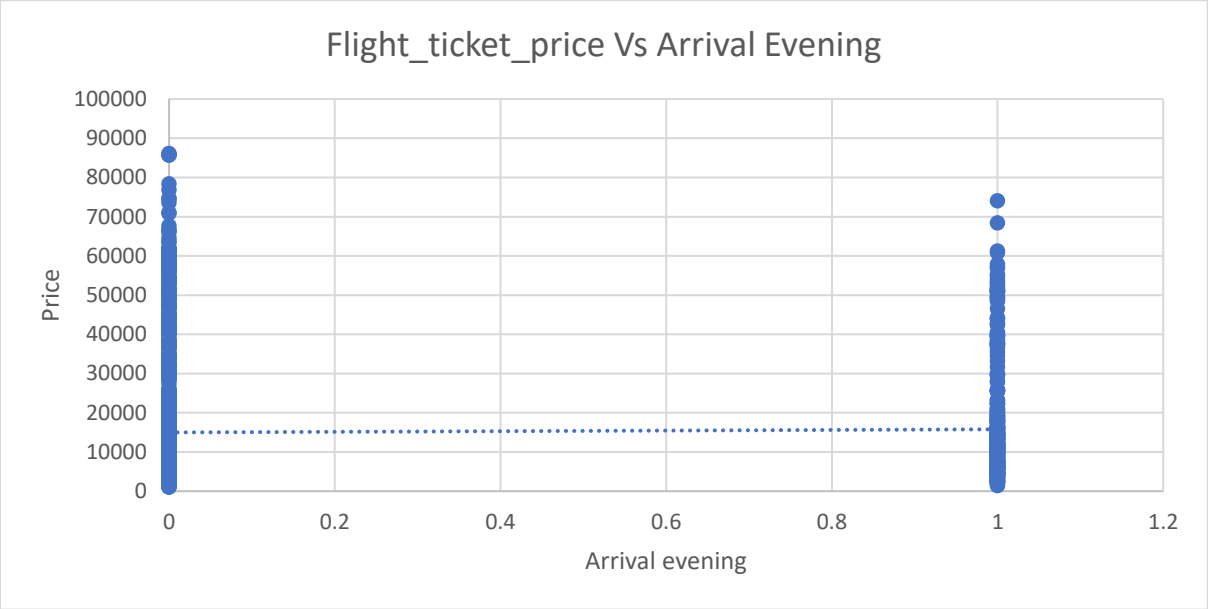
	Arrival_Early_Morning	*Arrival_Morning*	*Arrival_Evening*	*Arrival_Night*	*Arrival_Late_Night*
Mean	0.043149946	0.186623517	0.221682848	0.28802589	0.077130529
Standard Error	0.004720352	0.00905089	0.009649541	0.010519865	0.006197915
Median	0	0	0	0	0
Mode	0	0	0	0	0
Standard Deviation	0.203249379	0.389714124	0.415490886	0.45296539	0.266870426
Sample Variance	0.04131031	0.151877098	0.172632676	0.205177644	0.071219824
Kurtosis	18.27257191	0.592656086	-0.201536416	-1.123337141	8.073603359
Skewness	4.500317939	1.60997451	1.341149565	0.936950277	3.172521888
Range	1	1	1	1	1
Minimum	0	0	0	0	0
Maximum	1	1	1	1	1
Sum	80	346	411	534	143
Count	1854	1854	1854	1854	1854

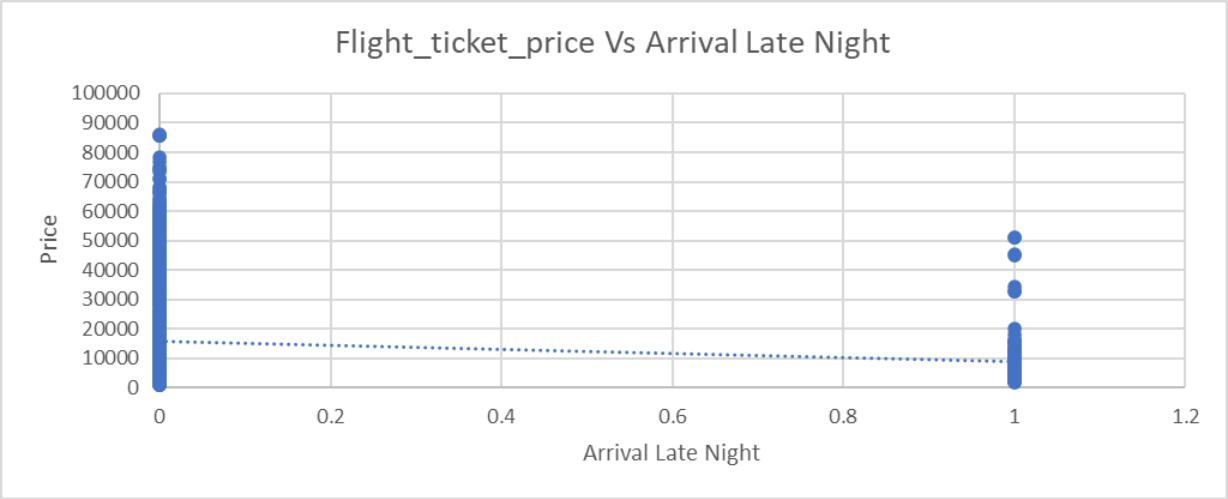
Scatter Plots






Flight_ticket_price Vs Arrival Night
Price
100000
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
0
0.2
0.4
0.6
0.8
1
1.2
Arrival Night


Flight_ticket_price Vs Arrival Evening
Price
100000
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
0
0.2
0.4
0.6
0.8
1
1.2
Arrival evening

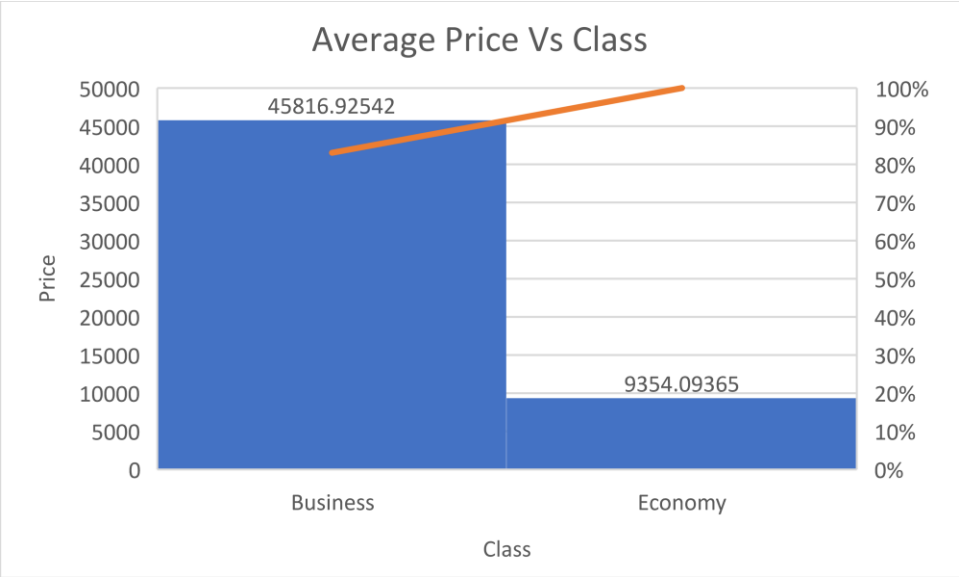


Observations:

I feel that Departure time drives the arrival time of flights, so arrival time should have less influence over flight ticket prices.
However, the graphs show that flights that reach early morning and late-night are cheaper.

Input Variable – Class (Economy, Business)

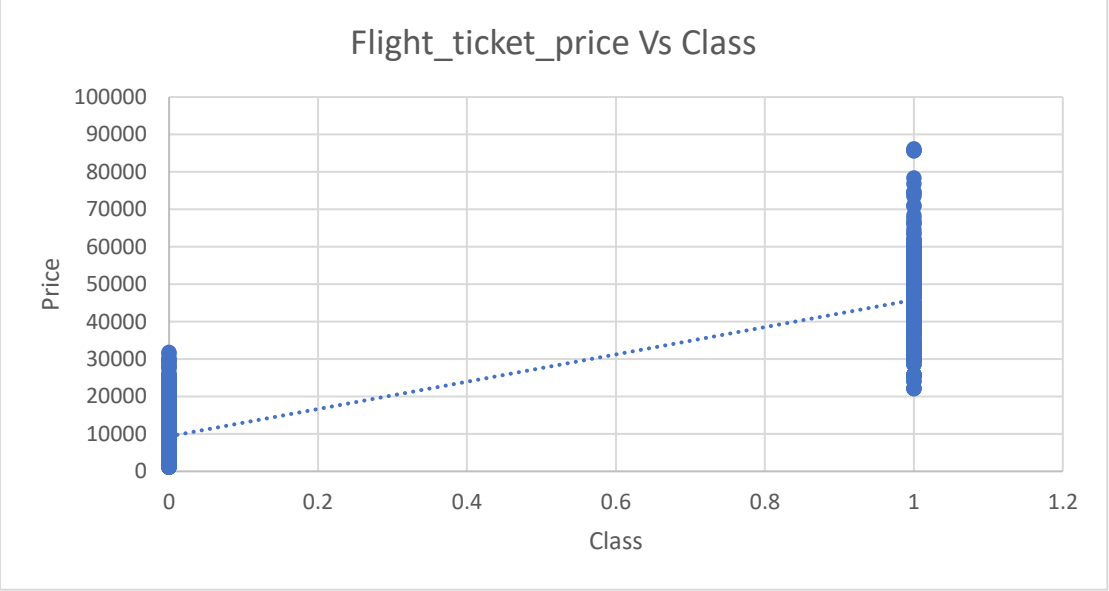
Histogram



Descriptive Statistics

Class	
Mean	0.159115
Standard Error	0.008497
Median	0
Mode	0
Standard Deviation	0.365882
Sample Variance	0.13387
Kurtosis	1.481196
Skewness	1.865368
Range	1
Minimum	0
Maximum	1
Sum	295
Count	1854

Scatter Plot

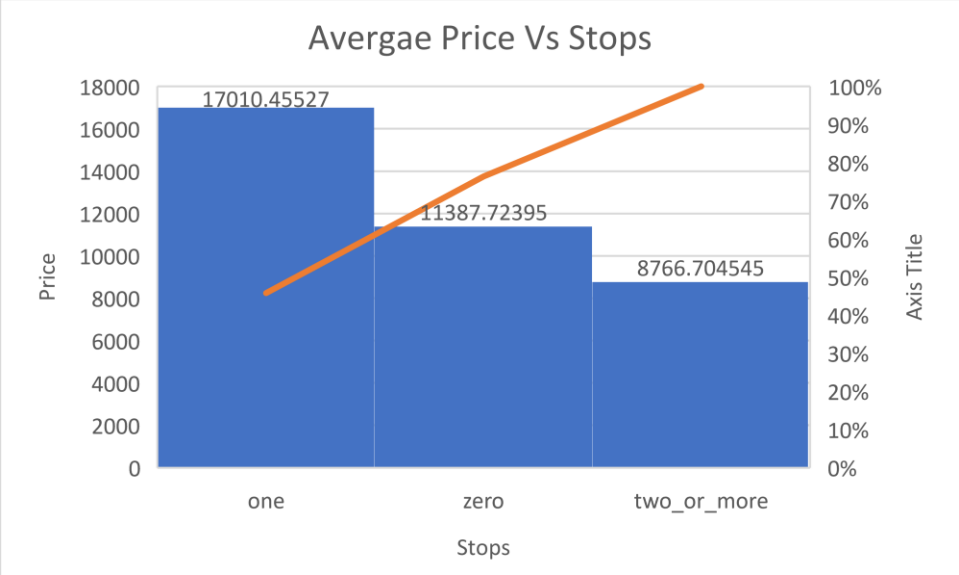


Observation:

Clearly Business Class is more expensive than the Economy. Economy seats are roughly 20% cheaper than the Business Class seats.

Input Variable – Stops (0,1, 2+)

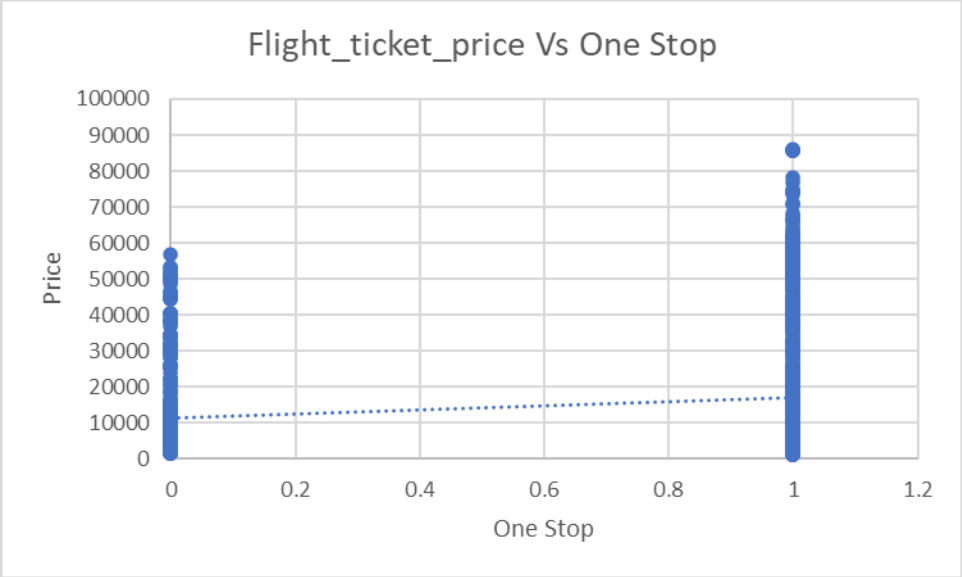
Histogram



Descriptive Statistics

	Stops_one	*Stops_two_or_more*
Mean	0.681229773	0.02373247
Standard Error	0.0108255	0.003536048
Median	1	0
Mode	1	0
Standard Deviation	0.466125477	0.152255521
Sample Variance	0.217272961	0.023181744
Kurtosis	-1.395536912	37.26432878
Skewness	-0.778440548	6.262917138
Range	1	1
Minimum	0	0
Maximum	1	1
Sum	1263	44
Count	1854	1854

Scatter Plot

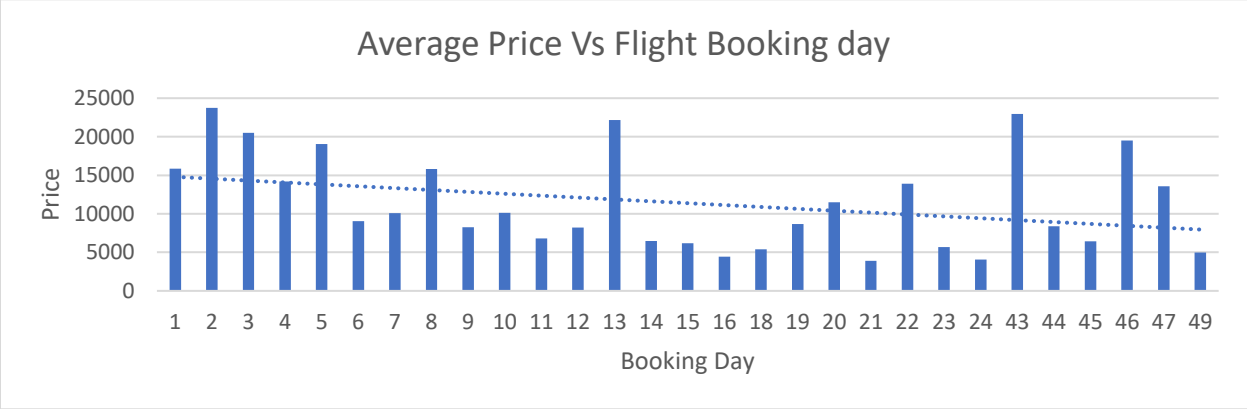


Observation:

Flight ticket price is lower when stops between source and destination are two or more.

Input Variable – Booking Days

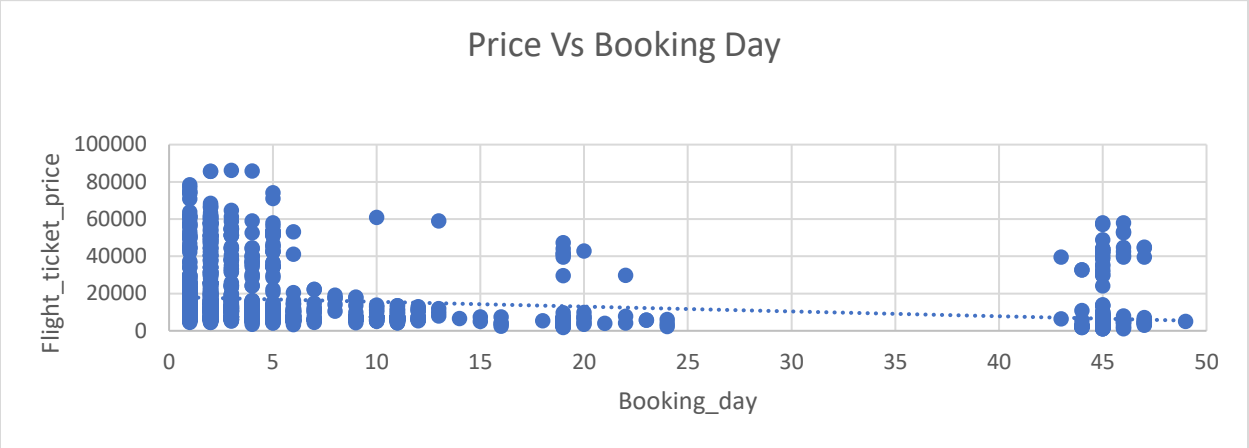
Histogram



Descriptive Statistics

Booking_day	
Mean	11.6429342
Standard Error	0.37992987
Median	3
Mode	1
Standard Deviation	16.35905842
Sample Variance	267.6187925
Kurtosis	0.294532722
Skewness	1.439679934
Range	48
Minimum	1
Maximum	49
Sum	21586
Count	1854

Scatter Plot

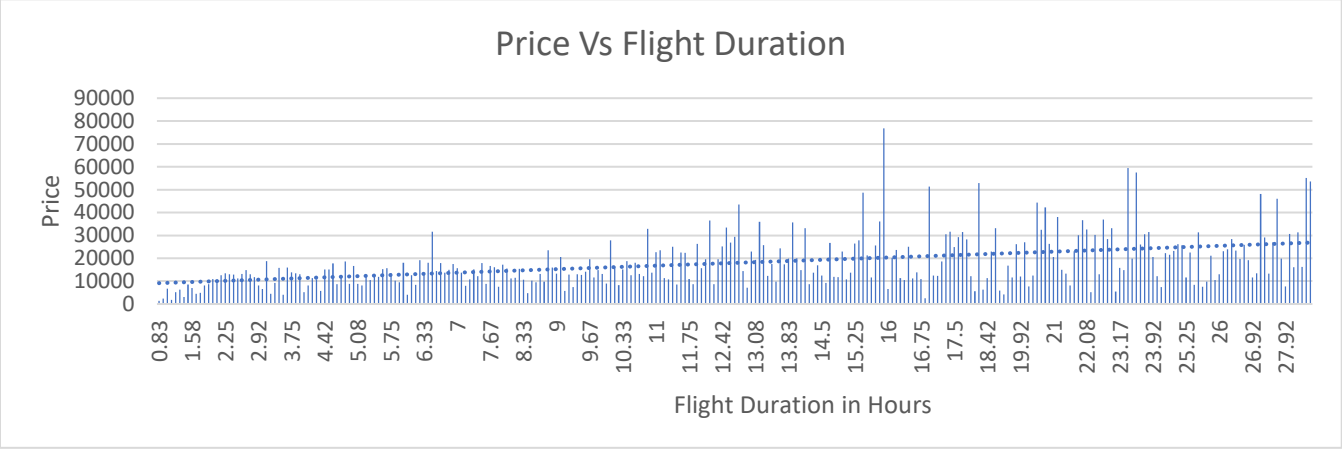


Observation:

Flight ticket price is cheaper when you book them in advance.
Flight ticket prices are the highest when you book them the flight takes off.
One day before the flight takes off the flight ticket price slightly drops.

Input Variable – Flight Duration (In Hours)

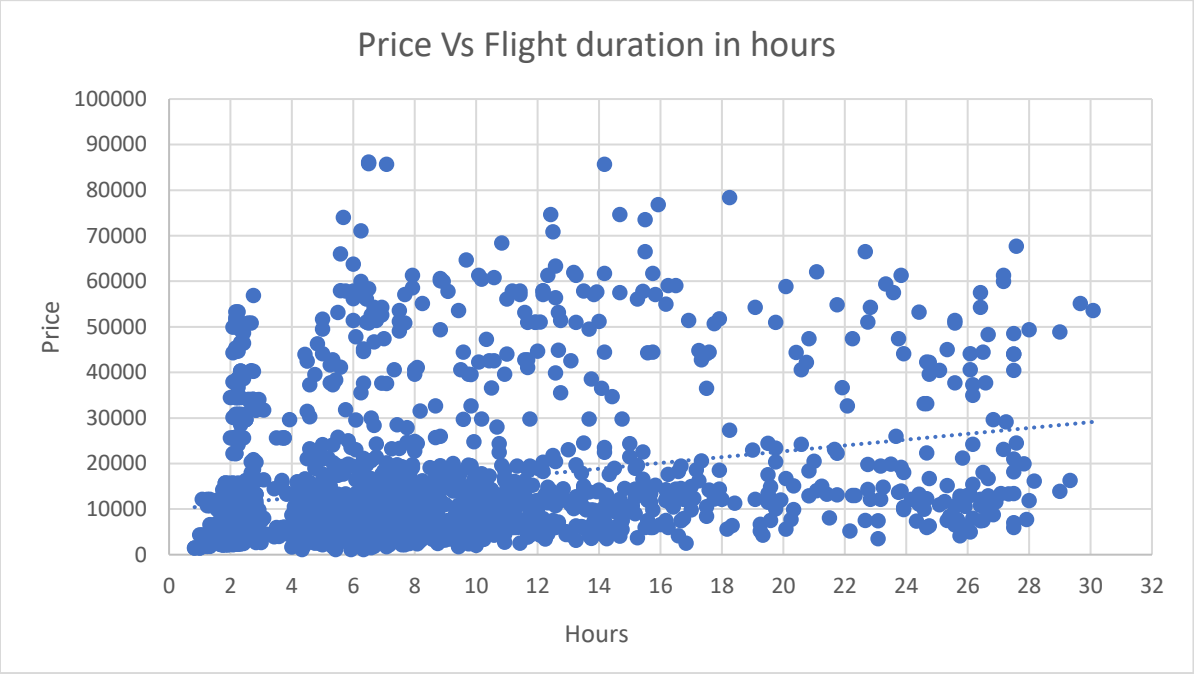
Histogram



Descriptive Statistics

duration	
Mean	8.221936
Standard Error	0.149868
Median	6.5
Mode	2.17
Standard Deviation	6.453038
Sample Variance	41.6417
Kurtosis	1.380642
Skewness	1.374335
Range	29.25
Minimum	0.83
Maximum	30.08
Sum	15243.47
Count	1854

Scatter Plot

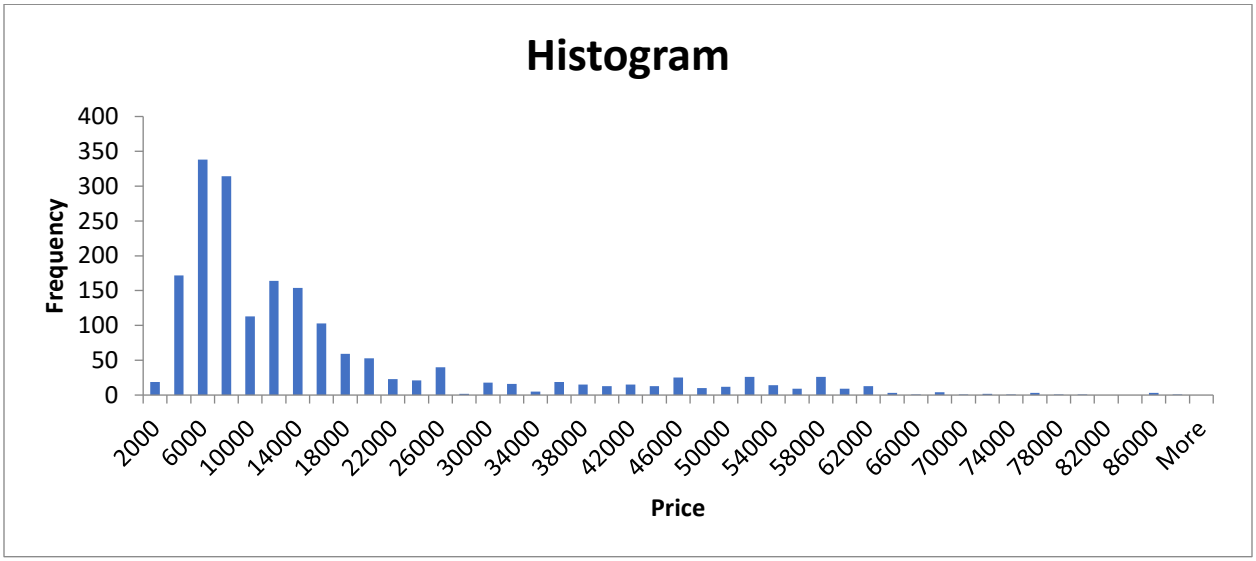


Observation:

Ticket price increases as fight duration increases.

Output Variable – Price

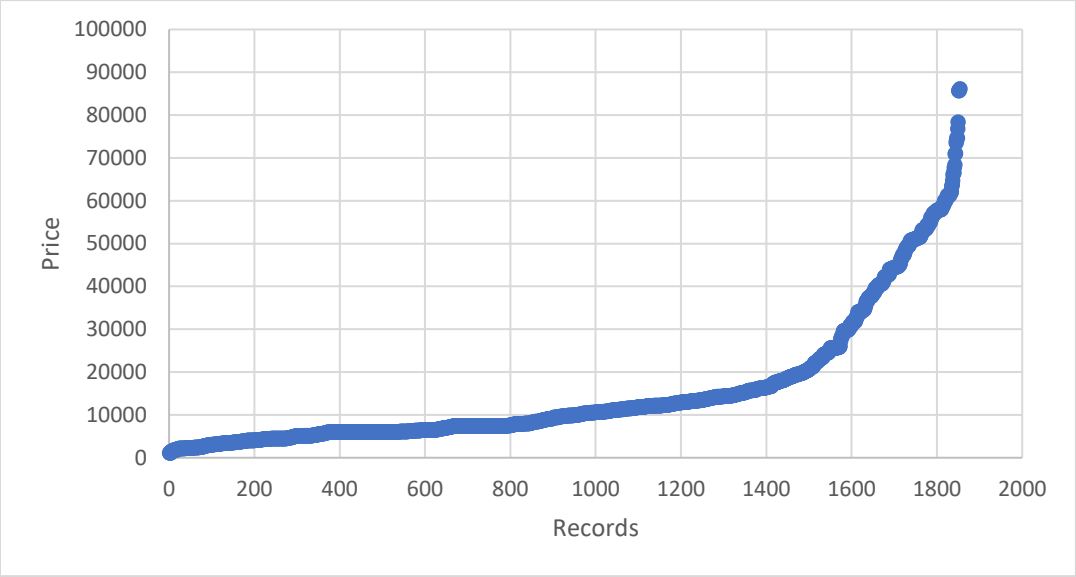
Histogram



Descriptive Statistics

Flight price	
Mean	15155.89
Standard Error	351.0887
Median	9735.5
Mode	5955
Standard Deviation	15117.21
Sample Variance	2.29E+08
Kurtosis	3.280507
Skewness	1.94988
Range	85032
Minimum	1105
Maximum	86137
Sum	28099025
Count	1854

Scatter Plot



Observation:

The output Variable has some outliers and may need transformation as its histogram is skewed towards the right.

Comment added
Please note that we have 1854 records and close to two dozen outliers. It was noted that the skewness for the Output variable is high (above 1) but it was assumed that is not so high that we need an output transformation.

Correlation Matrix

	Air_India	*AirAsia*	*GO_FIRST*	*Indigo*	*SpiceJet*	*Source_Bangalore*	*Source_Delhi*	*Source_Chennai*	*Source_Kolkata*	*Source_Mumbai*
Air_India	1									
AirAsia	-0.132080759	1								
GO_FIRST	-0.181095237	-0.090670442	1							
Indigo	-0.401402827	-0.200973655	-0.275553926	1						
SpiceJet	-0.171513723	-0.085873186	-0.117740277	-0.260974726	1					
Source_Bangalore	-0.054474186	0.165057214	0.003366761	-0.017458397	-0.001451464	1				
Source_Delhi	0.069477232	-0.029679081	0.009396815	-0.093528106	0.013577488	-0.292827873	1			
Source_Chennai	-0.037934321	-0.022428204	-0.070657307	0.107858532	0.007387264	-0.131967361	-0.206717442	1		
Source_Kolkata	0.006786745	0.015358019	-0.024088167	0.023359846	0.059325816	-0.159465714	-0.249791647	-0.112572427	1	
Source_Mumbai	0.011992658	-0.074811819	0.053911781	-0.05759846	-0.021969988	-0.226415348	-0.354663465	-0.159834516	-0.193139615	1
Departure_Early_Morning	0.012008991	0.011657569	0.009084926	-0.028415851	0.05376553	0.04189475	-0.007817023	0.009982099	-0.042160577	0.012508666
Departure_Morning	0.048616666	-0.059698667	-0.010438866	0.034002466	-0.051055725	-0.012711696	-0.022219047	0.010845281	0.008278945	-0.002075268
Departure_Evening	0.01507806	-0.0263689	-0.029712097	-0.019945003	0.021759341	0.000588645	0.032129889	-0.002877392	-0.016108446	0.003272833
Departure_Night	-0.041979272	0.063176276	0.00292152	-0.02298071	0.032276741	-0.004607984	-0.04090142	0.004909805	0.021007178	0.010131896
Departure_Late_Night	-0.035337229	0.010115293	-0.025162496	0.097360823	-0.038241818	0.029081196	0.004656301	-0.000774295	-0.012842432	-0.001916339
Stops_one	0.163374319	0.031951921	0.060337725	-0.071340473	-0.126015395	-0.006097876	0.06691739	-0.039938634	0.013429884	-0.016365275
Stops_two_or_more	-0.062642353	0.22433974	0.103218599	-0.048815805	-0.052064955	0.04932082	-0.029276494	-0.009514976	0.051641213	-0.021290812
Arrival_Early_Morning	-0.089476127	0.011420485	-0.015623234	0.118534039	0.017442342	0.024777784	0.010540063	0.011240284	-0.012912571	-0.03369586
Arrival_Morning	0.023084404	-0.008389419	-0.02321077	-0.106134311	0.093488094	-0.01327915	0.033390538	-0.002412272	-0.027420403	0.02539008
Arrival_Evening	0.023032673	-0.040340683	0.002298127	0.027183413	-0.039906574	0.00095751	-0.025844747	-0.028024339	0.007144958	0.011212123
Arrival_Night	-0.001365795	-0.030235425	0.003626005	0.030730158	0.013587361	-0.042845922	-0.002357789	0.036225636	0.018557355	0.01761454
Arrival_Late_Night	-0.108689592	0.185480901	0.052786163	0.032406102	-0.022517713	0.07480085	-0.069524144	-0.015832491	0.024142291	0.006025681
Destination_Bangalore	-0.164373074	0.036519543	-0.018240264	0.192819371	0.059997543	-0.159057428	0.120123554	-0.05266598	-0.022881776	0.046331041
Destination_Delhi	0.14112541	0.091138372	0.050236947	-0.240550245	-0.069691402	0.246826011	**-0.644455651**	0.065452867	0.224587001	0.237582412
Destination_Chennai	-0.063803341	-0.036696321	-0.03801586	0.071305776	0.106396663	-0.051110811	0.019646456	-0.0435551	-0.040751371	0.019118363
Destination_Kolkata	-0.138337122	-0.059643537	-0.088086132	0.246394123	0.077112662	-0.03376381	0.086900632	0.020131354	-0.127171053	0.002570501
Destination_Mumbai	0.104113487	-0.062196719	0.036169751	-0.104428942	-0.070422858	-0.115354197	**0.578615926**	-0.070962182	-0.145832398	-0.304730894
Class	0.393390494	-0.111863159	-0.153374992	-0.339960102	-0.14526012	-0.046397323	0.061090933	-0.02714672	-0.033263288	0.055661124
duration	0.295051858	0.04328157	-0.009367631	-0.341862205	0.051728599	-0.016313246	0.124689165	-0.079081189	0.03387206	-0.013338015
Booking_day	-0.068637897	-0.036498046	0.047976155	0.178501185	0.019805015	0.009349159	-0.025843484	0.070565841	-0.029939785	-0.027247083

	Departure_Early_Morning	*Departure_Morning*	*Departure_Evening*	*Departure_Night*	*Departure_Late_Night*	*Stops_one*	*Stops_two_or_more*
Air_India							
AirAsia							
GO_FIRST							
Indigo							
SpiceJet							
Source_Bangalore							
Source_Delhi							
Source_Chennai							
Source_Kolkata							
Source_Mumbai							
Departure_Early_Morning	1						
Departure_Morning	-0.281707896	1					
Departure_Evening	-0.297374001	-0.275751489	1				
Departure_Night	-0.216593363	-0.200844533	-0.212013732	1			
Departure_Late_Night	-0.06312075	-0.058531145	-0.06178613	-0.045002138	1		
Stops_one	-0.033652639	0.086850948	-0.040865806	-0.122223297	-0.065016513	1	
Stops_two_or_more	0.06494087	-0.018478832	-0.04172142	-0.050859031	0.013492925	-0.227926586	1
Arrival_Early_Morning	0.090162762	-0.101985992	-0.057402572	0.111503709	0.233995752	-0.076890772	-0.033109699
Arrival_Morning	0.233747843	-0.074007198	-0.056347543	0.072056241	0.030875843	-0.079337102	-0.047398291
Arrival_Evening	-0.060716193	0.1630142	-0.067315879	-0.198276455	-0.061117815	0.09478184	0.044752218
Arrival_Night	-0.212514399	-0.13109364	0.238297755	0.008991156	-0.07283896	-0.022433029	0.018207777
Arrival_Late_Night	-0.14499726	-0.122813855	0.081079456	0.260556466	0.056317051	-0.058202244	0.034615314
Destination_Bangalore	0.005925477	-0.027823193	0.016625478	-0.007638546	0.016775425	-0.048408478	0.062937511
Destination_Delhi	-0.018603631	0.041384989	-0.024887263	0.048080679	-0.013420399	0.022798885	0.000650637
Destination_Chennai	0.030823486	-0.034865434	0.015302422	0.023234599	-0.016341935	-0.084463747	0.00308862
Destination_Kolkata	0.004542186	0.00949079	-0.010116362	-0.058368239	0.068901459	-0.008271601	0.003733675
Destination_Mumbai	0.002108634	-0.027982103	0.027494323	-0.017261741	-0.03371126	0.019977982	-0.049972459
Class	-0.016525956	0.01418992	0.01937149	-0.006327264	-0.036770893	0.034925274	-0.058135226
duration	-0.049888866	0.00130317	0.02729002	-0.043397711	-0.068096261	**0.561722434**	0.0615046
Booking_day	0.03052279	-0.058261128	0.006728306	0.048690196	0.046264467	0.075795323	-0.010896001

	Arrival_Early_Morning	*Arrival_Morning*	*Arrival_Evening*	*Arrival_Night*	*Arrival_Late_Night*	*Destination_Bangalore*	*Destination_Delhi*	*Destination_Chennai*	*Destination_Kolkata*	*Destination_Mumbai*	*Class*	*duration*	*Booking_day*
Air_India													
AirAsia													
GO_FIRST													
Indigo													
SpiceJet													
Source_Bangalore													
Source_Delhi													
Source_Chennai													
Source_Kolkata													
Source_Mumbai													
Departure_Early_Morning													
Departure_Morning													
Departure_Evening													
Departure_Night													
Departure_Late_Night													
Stops_one													
Stops_two_or_more													
Arrival_Early_Morning	1												
Arrival_Morning	-0.101719816	1											
Arrival_Evening	-0.113332882	-0.255637831	1										
Arrival_Night	-0.135067807	-0.304663931	-0.339446557	1									
Arrival_Late_Night	-0.061391918	-0.138477876	-0.154287506	-0.18387669	1								
Destination_Bangalore	0.028376529	0.007503371	-0.056071659	0.012300562	0.055866569	1							
Destination_Delhi	-0.090443659	0.037659635	0.038216606	-0.02802693	0.060918786	-0.350053624	1						
Destination_Chennai	0.007657755	-0.008959317	0.007406888	0.054022092	-0.026798358	-0.052496028	-0.13578598	1					
Destination_Kolkata	-0.01292531	-0.021405447	0.001384104	0.043516476	-0.014397938	-0.126845453	-0.328097856	-0.049203416	1				
Destination_Mumbai	0.084038731	-0.011230391	-0.026791729	-0.024889992	-0.084520319	-0.214074323	**-0.553723645**	-0.08303954	-0.200647334	1			
Class	-0.063347576	0.071706068	-0.004957285	-0.029200892	-0.087068259	-0.150924592	0.123483278	-0.062074231	-0.126064174	0.103033562	1		
duration	-0.062737832	0.047508602	0.058643526	-0.086057256	-0.09821915	-0.072117772	0.032828487	-0.072630831	-0.078366984	0.105932477	0.150586081	1	
Booking_day	0.119387336	-0.029665569	-0.023600474	-0.013060103	0.062803115	0.09821032	-0.066399402	0.026461694	0.00677885	-0.037063313	-0.105640017	-0.063116774	1

Appendix E – Data

Dataset Details:

Flight Price Dataset initially had ten columns and 300154 records.

The Flight Number column was used strategically as a primary key to filter out flights such that we have only one flight number record for each of the classes (Economy and Business), which left us with 1854 records.

The Flight Number column was then deleted because it does not contribute to our target variable Price.

After splitting variables with categorical data, we now have 30 Input variables and 1 output variable-Price.

Variables Description:

Airline: The name of the airline company is stored in the airline column. It is a categorical feature having 6 different airlines.

Flight No.: It was assumed to be a primary key and was discarded later as it had no impact on the target variable.

Source City: City from which the flight takes off. It is a categorical feature having 6 unique cities.

Departure Time: This is a categorical variable obtained by grouping time periods. It stores information about the departure time and has 6 unique time labels.
Stops: A categorical feature with 3 distinct values that store the number of stops between the source and destination cities.

Arrival Time: This is a derived categorical feature created by grouping time intervals. It has six distinct time labels and keeps the information about the arrival time.

Destination City: City where the flight will land. It is a categorical feature having 6 unique cities.

Class: A binary variable that contains information on seat class; it has two distinct values: Business and Economy.

Duration: Captures the amount of time it takes to travel between cities in hours.

Days Left (Booking day): This variable tells us how early the flight was booked in comparison to the travel date.

Price: Target variable stores information about the ticket price.