DATA ANALYTICS 17:610:561:C1
Assignment - 3

Gayathri Ravipati
NetID: gr485

Introduction:

Air Travel has gained popularity among individuals, due to its time-saving benefits for both long-distance and local trips. The ability to quickly reach one's destination has made flying an attractive option for individuals with hectic schedules or limited time. However, alongside the convenience, travelers are also keenly aware of the expenses associated with air travel and make careful plans to ensure their trips fit within their budget constraints.

Despite the convenience and efficiency of flying, the cost factor cannot be ignored. Airfare can vary significantly based on several factors, such as the distance between the source and destination, the chosen airline, the time of travel, and other variables. To ensure well-informed decisions regarding air travel, individuals must carefully consider these factors and plan accordingly.

While we can make rough estimates of the factors that influence flight costs, I am determined to gain a more comprehensive understanding by delving into the analysis and visual representation of various parameters. I hope to uncover valuable insights about the interrelationships among these factors and their impact on flight prices. Analyzing the data will enable me to identify patterns, trends, and correlations that can further enhance my understanding of the pricing dynamics of air travel.

Problem Statement:

My objective is to predict the price of travel for a trip considering information like distance between source and destination, Airlines, number of stops, duration, departure_hour, and day_of_travel.

Data Sources:

This dataset is taken from Kaggle. There is already a different dataset for training and testing.

Training dataset:
https://www.kaggle.com/datasets/absin7/airlines-fare-prediction?select=Data_Train.xlsx
It has a total of 10683 rows and 11 columns

Testing dataset:

https://www.kaggle.com/datasets/absin7/airlines-fare-prediction?select=Test_set.xlsx
It has a total of 2671 rows and 10 columns
The dataset used for analysis and prediction is sourced solely from the provided dataset, without any additional data collection or scraping. This dataset comprises various trip details that are instrumental in predicting the flight journey price.

Unit of Analysis:

The unit of analysis for this dataset is the details of individual flight journeys. Each row in the dataset represents a specific flight journey, including details such as the airline, date of the journey, source, destination, route, departure time, arrival time, duration, total stops, additional information, and price. By examining and analyzing these individual flight journeys, patterns, trends, and relationships can be identified to understand the factors influencing flight prices.

Model Description:

I aim to use a multi-linear regression model to predict the price of air travel. Through multi-linear regression, we establish a linear relationship between the target variable (price) and several predictor variables (airlines, distance between source and destination, total stops, duration of travel, day_of_journey, departure_hour).

Independent Predictor Variables:

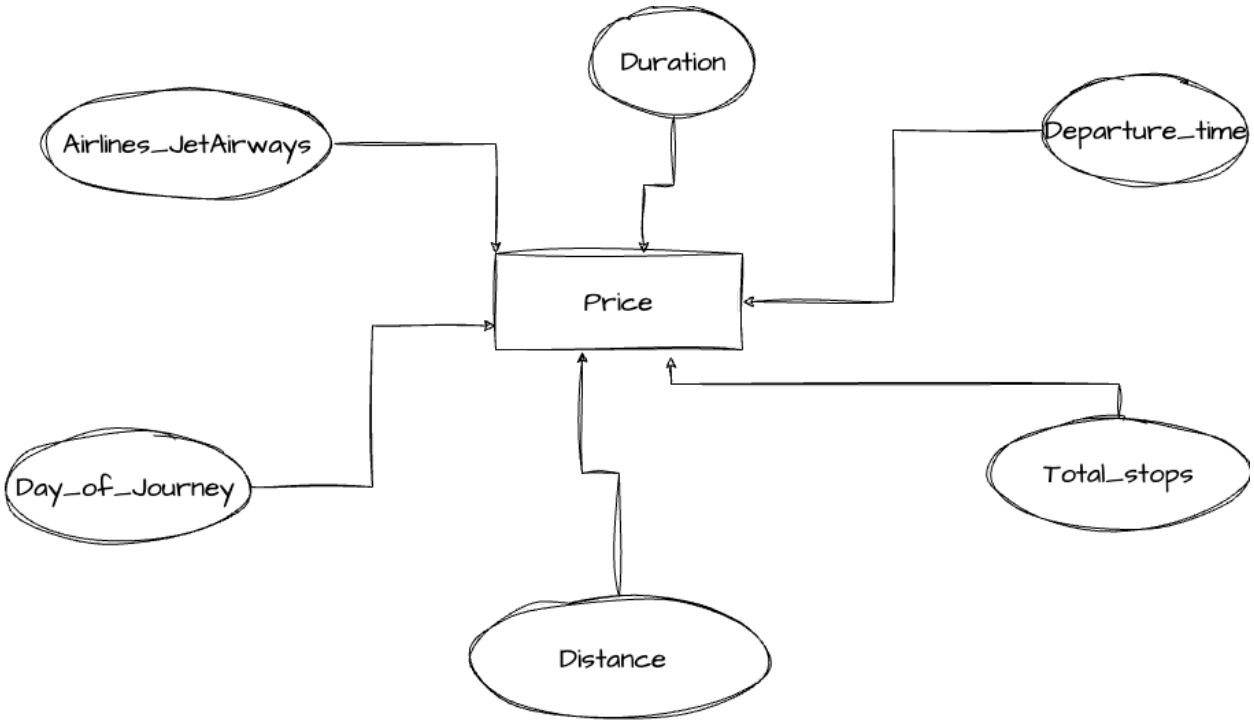
- **Airline_JetAirways -** The airline names, such as Indigo, Jet Airways, Air India, etc., which are currently in string format, will be converted to a binary value that takes the value 1 if the airline is Jet Airways or 0 if the airline name is anything else.
- **Day_of_journey** - This is obtained when we find the week_day of the date_of_journey from the dataset. Considering Monday is 0, Tuesday is 1, .. and Sunday is 6. Furthermore, it is considered to be a binary value that takes 0 if it is a weekday and 1 if it's either on Saturday or Sunday.
- **Distance**(in km) - This is the calculated distance from the given source to the destination.
- **Total stops** - Total number of stops between the source and destination. Non-stop is considered as zero, one-stop is considered as 1, etc.
- **Duration(mins)** - The given duration will be converted into minutes.
- **Departure_time** - We extract the hour component from the given departure time, disregarding the minutes. The resulting format is in numerical form, representing the hour of the departure time. Further, it is considered 0 if it is from 9 AM - 10 PM else it is taken as 1.

Target Dependent Variable:

Price is the target dependent variable which is the focus of our prediction. Using the above independent variables we use multi-linear regression to predict the price value. The regression equation can be given as

Price = β0 + β1 ∗ (airline_jetairways) + β2 ∗ (distance) + β3 ∗ (Day_of_journey) + β4 ∗ (total_stops) + β5 ∗ (duration) + β6 ∗ (departure_time)

Visualization of the model:



Data Curation:

Sample of Initial Dataset:

	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662
2	Jet Airways	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops	No info	13882
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218
4	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302

During the data curation process, multiple steps were undertaken to refine the dataset. Categorical variables were converted into binary values to make them more amenable to

analysis. Unnecessary columns were removed to streamline the dataset. The date of the journey was transformed into weekdays and then categorized as either weekday or weekend, providing a more informative representation. Additionally, time values were converted into an hour format, and a binary value was assigned based on whether the departure time falls within the 9 AM - 10 PM window or the 10 PM - 9 AM window. These transformations and adjustments were implemented to enhance the data, making it more suitable for subsequent analysis and processing.

1. **Computation of distance from source to destination:**
   In order to calculate the distance between the source and destination for each record, I performed a computation and assigned the resulting distance value to the "Distance" column in the dataset. This process allowed us to obtain a numerical variable representing the distance, effectively replacing the source and destination information with a more meaningful and quantifiable measure.

   The distance values have been rounded to the nearest integer. This rounding operation converts the floating-point values representing the distances to whole numbers, providing a more concise and standardized representation for analysis purposes.

2. **Dropping unnecessary columns:**
   Based on my analysis, I have identified certain columns that do not contribute significantly to the analysis. Therefore, I have removed these columns from the dataset. The dropped columns include 'Route', 'Arrival_Time', and 'Additional_Info'. By removing these columns, we have simplified the dataset and focused on the relevant information for further analysis.

	Airline	Date_of_Journey	Dep_Time	Duration	Total_Stops	Price	Distance
0	IndiGo	24/03/2019	22:20	2h 50m	non-stop	3897	1732
1	Air India	1/05/2019	05:50	7h 25m	2 stops	7662	1559
2	Jet Airways	9/06/2019	09:25	19h	2 stops	13882	2080
3	IndiGo	12/05/2019	18:05	5h 25m	1 stop	6218	1559
4	IndiGo	01/03/2019	16:50	4h 45m	1 stop	13302	1732

3. **Total_stops conversion:**
   The values in the "Total_Stops" column have been assigned proper numerical values based on their corresponding stop counts. For example, "non-stop" is assigned a value of 0, "1 stop" is assigned a value of 1, "2 stops" is assigned a value of 2, and so on. This conversion allows us to represent the total number of stops as a numerical variable.

	Airline	Date_of_Journey	Dep_Time	Duration	Total_Stops	Price	Distance
0	IndiGo	24/03/2019	22:20	2h 50m	0	3897	1732
1	Air India	1/05/2019	05:50	7h 25m	2	7662	1559
2	Jet Airways	9/06/2019	09:25	19h	2	13882	2080
3	IndiGo	12/05/2019	18:05	5h 25m	1	6218	1559
4	IndiGo	01/03/2019	16:50	4h 45m	1	13302	1732

4. **Conversion of duration to minutes:**
The duration of the trip is initially provided in the format of hours and minutes. To ensure consistency and comparability across all records, this duration is converted into a single unit of minutes. By converting it to a numerical value expressed solely in minutes, we achieve a standardized representation of the trip duration that allows for easier analysis and comparisons among different records.

	Airline	Date_of_Journey	Dep_Time	Duration	Total_Stops	Price	Distance
0	IndiGo	24/03/2019	22:20	170	0	3897	1732
1	Air India	1/05/2019	05:50	445	2	7662	1559
2	Jet Airways	9/06/2019	09:25	1140	2	13882	2080
3	IndiGo	12/05/2019	18:05	325	1	6218	1559
4	IndiGo	01/03/2019	16:50	285	1	13302	1732

5. **Date_of_Journey to numerical:**
The process involved two steps. Firstly, the day of travel was determined by extracting the day component from the "Date_of_Journey". Secondly, a numerical value was assigned to each day, starting from 0 for Monday and incrementing by 1 for each subsequent day of the week until Sunday, which was assigned a value of 6. Additionally, Day_of_Journey is considered as 0 if it is from Monday to Friday and considered as 1 if it is on Saturday or Sunday.

	Airline	Day_of_Journey	Dep_Time	Duration	Total_Stops	Price	Distance
0	IndiGo	1	22:20	170	0	3897	1732
1	Air India	0	05:50	445	2	7662	1559
2	Jet Airways	1	09:25	1140	2	13882	2080
3	IndiGo	1	18:05	325	1	6218	1559
4	IndiGo	0	16:50	285	1	13302	1732

6. **Converting Airline name which is categorical to numerical:**
The airline names, initially in string format, contain different airline names like JetAirways, Indigo, Air India, etc. Airline is given a binary value, it is taken as 1 if the airline is Jet Airways and it is taken as 0 if it is any other Airline name.

	Airline	Day_of_Journey	Dep_Time	Duration	Total_Stops	Price	Distance
0	0	1	22:20	170	0	3897	1732
1	0	0	05:50	445	2	7662	1559
2	1	1	09:25	1140	2	13882	2080
3	0	1	18:05	325	1	6218	1559
4	0	0	16:50	285	1	13302	1732

7. **Consider only Departure_hour from Departure_time:**
The departure time provided in the dataset is in the format of HH: MM, which is not inherently numerical. To simplify the analysis, I have extracted only the hour component (HH) from the departure time. Further, it takes only binary value 0 if the departure_hour is between 9 AM to 10 PM else 0.

	Airline	Day_of_Journey	Dep_Time	Duration	Total_Stops	Price	Distance
0	0	1	0	170	0	3897	1732
1	0	0	1	445	2	7662	1559
2	1	1	0	1140	2	13882	2080
3	0	1	0	325	1	6218	1559
4	0	0	0	285	1	13302	1732

The data curation process for the dataset has been successfully completed with the necessary preprocessing steps carried out, the dataset is now in a suitable state for further analysis, modeling, and making predictions based on the curated data. Below is the snippet of the final

curated data.

Airline_JetAirways	Distance	Day_of_Journey	Dep_Time	Duration	Total_Stops	Price
0	1732	1	0	170	0	3897
0	1559	0	1	445	2	7662
1	2080	1	0	1140	2	13882
0	1559	1	0	325	1	6218
0	1732	0	0	285	1	13302
0	1559	0	0	145	0	3873
1	1732	0	0	930	1	11087
1	1732	0	1	1265	1	22270
1	1732	0	1	1530	1	11087
0	2080	0	0	470	1	8625
0	2080	1	0	795	1	8907
0	1559	0	0	155	0	4174
0	1355	0	0	135	0	4667
1	1559	0	0	730	1	9663
0	1559	0	0	155	0	4804

Informed Estimate of Findings:

There are certain factors that play a crucial role in determining the price of the air travel journey. One factor is the distance of travel. It measures the distance between the source and the destination. If the distance value is higher, it implies that the source and destination are considerably far and there is a greater probability for the air travel price to be higher. The value of the distance is measured in km.

The second factor which can directly impact the price of the ticket is the total stops in the travel. If the number of stops is zero it is expected that the price of the ticket will be higher. As the number of stops increases the ticket cost decreases. This independent variable value will be 0,1,2,... where 0 indicates no intermediate stops between the source and destination.

The third factor which can have a significant impact on the price is the duration of the travel. Sometimes, for the same distance, the duration of the travel impacts the price of the air ticket. By considering this independent variable we can find its impact on the price of the ticket for different locations. The duration value is the total minutes taken to reach from source to destination.

Prices of the flight can be cheaper if their departure time is from 10 PM - 9 AM which are not the business hours of the day. By this independent variable, departure_hour we can analyze if time of departure can really have an impact on the prices of the flight ticket.

The independent variable day_of_journey can also impact the price of the ticket. For example, if the date of travel is on a weekend or holiday season and if the demand is higher for any unprecedented reason then the price of the ticket will be higher. But in the dataset, we can only try to find the impact considering on which day is the passenger traveling. Hence, we will try to analyze if the price is higher on weekends when compared to weekdays.

The independent variable, airlines may have a slight impact on the price of the ticket. Even if all other variables remain constant, different airlines may offer varying prices for the same route. This can be attributed to several factors, including the standards, reputation, and branding of the airlines. In this analysis, we are considering Airlines Jet Airways vs all the other airlines.

SPSS Statistical Analysis:

Correlations

		Price	Airline_JetAirways	Day_of_Journey	Dep_Time	Duration	Total_Stops	Distance
Price	Pearson Correlation	1	.428**	.020*	-.024*	.506**	.604**	.317**
	Sig. (2-tailed)		.000	.036	.013	.000	.000	<.001
	N	10682	10682	10682	10682	10682	10682	10682
Airline_JetAirways	Pearson Correlation	.428**	1	-.023*	-.060**	.305**	.216**	.016
	Sig. (2-tailed)	.000		.018	<.001	<.001	<.001	.091
	N	10682	10682	10682	10682	10682	10682	10682
Day_of_Journey	Pearson Correlation	.020*	-.023*	1	.008	-.004	-.020*	.036**
	Sig. (2-tailed)	.036	.018		.413	.669	.043	<.001
	N	10682	10682	10682	10682	10682	10682	10682
Dep_Time	Pearson Correlation	-.024*	-.060**	.008	1	-.003	.026**	.047**
	Sig. (2-tailed)	.013	<.001	.413		.766	.007	<.001
	N	10682	10682	10682	10682	10682	10682	10682
Duration	Pearson Correlation	.506**	.305**	-.004	-.003	1	.738**	.310**
	Sig. (2-tailed)	.000	<.001	.669	.766		.000	<.001
	N	10682	10682	10682	10682	10682	10682	10682
Total_Stops	Pearson Correlation	.604**	.216**	-.020*	.026**	.738**	1	.442**
	Sig. (2-tailed)	.000	<.001	.043	.007	.000		.000
	N	10682	10682	10682	10682	10682	10682	10682
Distance	Pearson Correlation	.317**	.016	.036**	.047**	.310**	.442**	1
	Sig. (2-tailed)	<.001	.091	<.001	<.001	<.001	.000	
	N	10682	10682	10682	10682	10682	10682	10682

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Correlation Scores:

- Airline_JetAirways and Price => r = 0.428
- Day_of_Journey and Price => r = 0.020
- Dep_Time and Price => r = -0.024
- Duration and Price => r = 0.506
- Total_Stops and Price => r = 0.604
- Distance and Price => r = 0.317

Significance Values:

- Airline_JetAirways and Price => p = 0.00, statistically significant($p < 0.05$)
- Day_of_Journey and Price => p = 0.036, statistically significant($p < 0.05$)
- Dep_Time and Price => p = 0.13, statistically not significant($p > 0.05$)
- Duration and Price => p = 0.00, statistically significant($p < 0.05$)
- Total_Stops and Price => p = 0.00, statistically significant($p < 0.05$)
- Distance and Price => p = <0.001, statistically significant($p < 0.05$)

Correlation Direction:

- Airline_JetAirways and Price => positive direction
- Day_of_Journey and Price => positive direction
- Dep_Time and Price => negative direction
- Duration and Price => positive direction
- Total_Stops and Price => positive direction
- Distance and Price => positive direction

Correlations: Interpretation and Analysis

Considering the r-values for each of the independent variables with the dependent variable price, it can be said that there is a moderate correlation between the Airline_JetAirways and Price, Duration and Price, Distance and Price. Also, there is a weak correlation between Day_of_Journey and Price, Dep_Time and Price. And there is a considerably strong positive correlation between Total_Stops and Price.

The correlation(r) value between **Airline_JetAirways** and **Price**(r = 0.428) indicates that there is a moderate positive correlation. This suggests that flights with Jet Airways as an airline tend to have moderately higher prices compared to other airlines. The scatter plot of these two variables would show a general upward trend.

The correlation(r) value between **Day_of_Journey** and **Price**(r = 0.020) indicates that there is a very weak positive correlation. This indicates that the day of the journey may have a slight influence on the flight prices, but the correlation is too weak to draw any significant conclusions. The scatter plot of these two variables would show a minimal or no linear trend.

The correlation(r) value between **Dep_Time** and **Price**(r = -0.024) indicates that there is a very weak negative correlation. This suggests that the departure time may have a minor impact on flight prices, but the correlation is negligible. The scatter plot of these two variables would show minimal or no linear trend, indicating that the departure time has little to no impact on flight prices.

The correlation(r) value between **Duration** and **Price**(r = 0.506) indicates that there is a moderate positive correlation. The scatter plot of these two variables would show a general upward trend, suggesting that longer flight durations have moderately higher prices.

The correlation(r) value between **Total_Stops** and **Price**(r = 0.604) indicates that there is a relatively strong positive correlation between the variable Total_Stops and the Price. The scatter plot of these two variables would show a clear upward trend, indicating that flights with less stops have higher prices.

The correlation(r) value between **Distance** and **Price**(r = 0.317) indicates that there is a moderate positive correlation. The scatter plot of these two variables would show a general upward trend, suggesting that longer flight distances have moderately higher prices.

p-values:

Considering the p-values for each of the independent and dependent variable relationships, it is evident that only one relationship produced a p-value higher than the 0.05 critical value. Thus, five of the six independent variables statistically possess genuine correlations with the dependent variable(price).

The p-value calculated from the relationship between **Airline_JetAirways** and **Price(**p = 0.00**)** is below the 0.05 critical value. This implies statistical significance and a statistically genuine correlation, with a lower chance of coincidence.

The p-value calculated from the relationship between **Day_of_Journey** and **Price(**p = 0.036**)** is below the 0.05 critical value. This implies that there is a greater likelihood of this value being a genuine correlation, rather than a coincidence.

The p-value calculated from the relationship between **Dep_Time** and **Price(**p = 0.13**)** is above the 0.05 critical value. This was not statistically significant, implying a greater chance of this result being coincidental, rather than statistically genuine.

The p-value calculated from the relationship between **Duration** and **Price(**p = 0.00**)**is below the 0.05 critical value. This implies statistical significance and a statistically genuine correlation, with a lower chance of coincidence.

The p-value calculated from the relationship between **Total_stops** and **Price(**p = 0.00**)**is below the 0.05 critical value. This implies statistical significance and a statistically genuine correlation, with a lower chance of coincidence.

The p-value calculated from the relationship between **Distance** and **Price(**p = 0.00**)**is below the 0.05 critical value. This implies that there is a greater likelihood of this value being a genuine correlation, rather than a coincidence.

Thus, through bivariate correlations, there are independent variables exhibiting all types of linear relationships. The p-value is not statistically significant for only one of the independent variables.

Multiple Linear Regression:

The multi-linear regression model is as follows:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.683[a]	.467	.467	3367.504

a. Predictors: (Constant), Distance, Airline_JetAirways, Day_of_Journey, Dep_Time, Duration, Total_Stops

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.061E+11	6	1.768E+10	1559.231	.000[b]
	Residual	1.211E+11	10675	11340080.99		
	Total	2.271E+11	10681			

a. Dependent Variable: Price

b. Predictors: (Constant), Distance, Airline_JetAirways, Day_of_Journey, Dep_Time, Duration, Total_Stops

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	3150.994	161.334		19.531	<.001
	Airline_JetAirways	3017.372	71.642	.314	42.117	.000
	Day_of_Journey	340.172	71.621	.034	4.750	<.001
	Dep_Time	-216.239	69.622	-.022	-3.106	.002
	Duration	.314	.098	.035	3.216	.001
	Total_Stops	3215.376	75.933	.471	42.345	.000
	Distance	1.124	.096	.093	11.706	<.001

a. Dependent Variable: Price

Percentage of Variance Explained by the Created Model

R – Square = 0.467 = 46.7%

The created model explained **46.7%** of the variance.

Significant variables and coefficients:

Significance:

- Airline_JetAirways and Price => p = 0.00, statistically significant($p < 0.05$)
- Day_of_Journey and Price => p = <0.001, statistically significant($p < 0.05$)
- Dep_Time and Price => p = 0.002, statistically significant($p < 0.05$)
- Duration and Price => p = 0.001, statistically significant($p < 0.05$)
- Total_Stops and Price => p = 0.00, statistically significant($p < 0.05$)
- Distance and Price => p = <0.001, statistically significant($p < 0.05$)

Unstandardized Coefficients:
Unstandardized coefficients in multiple linear regression represent the direct impact of each independent variable on the dependent variable without any normalization or scaling of the data. They are valuable for interpreting how practical changes in the independent variables affect the dependent variable. For instance, if an unstandardized coefficient for an independent variable is 0.5, it indicates that a one-unit increase in that independent variable corresponds to a 0.5-unit increase in the dependent variable while keeping all other variables constant.

Airline_JetAirways: 3017.3(positive)
For every 1 unit of change in Airline_JetAirways, the dependent variable (*price*), increases by 3017.3 which is very significant.

Day_of_Journey: 340.17(positive)
For every 1 unit of change in Day_of_Journey, the dependent variable (*price*), increases by 340.17

Dep_Time: -216.239(negative)
For every 1 unit of change in Dep_Time, the dependent variable (*price*), decreases by 216.23

Duration: 0.314(positive)
For every 1 unit of change in Duration, the dependent variable (*price*), increases by 0.314 which is very minor.

Total_Stops:3215.3(positive)
For every 1 unit of change in Total_Stops, the dependent variable (*price*), increases by 3215.3 which is very significant.

Distance: 1.124(positive)
For every 1 unit of change in Distance, the dependent variable (*price*), increases by 1.124 which is very minor.

The **unstandardized coefficients** enable us to calculate the estimated equation to predict price, which is as follows:

Calculating the estimated equation to predict price, using standardized coefficients is as follows:

$\hat{y} = 3150.994 + 3017.3 * (\text{Airline\_JetAirways}) + 340.17 * (\text{Day\_of\_Journey}) - 216.239 * (\text{Dep\_Time}) + 0.314(\text{Duration}) + 3215.3(\text{Total\_Stops}) + 1.124(\text{Distance})$

Standardized coefficients:
Standardized coefficients, transform the variables to have a mean of 0 and a standard deviation of 1. The standardized coefficients represent the change in the dependent variable (Y) corresponding to a one-standard deviation change in the corresponding independent variable (X).

Airline_JetAirways: 0.314(positive)
For every 1 standard deviation of change in Airline_JetAirways, the dependent variable (*price*), increases by 0.314.

Day_of_Journey: 0.34(positive)
For every 1 standard deviation of change in Day_of_Journey, the dependent variable (*price*), increases by 0.34.

Dep_Time: -0.22(negative)
For every 1 standard deviation of change in Dep_Time, the dependent variable (*price*), decreases by 0.22.

Duration: 0.35(positive)
For every 1 standard deviation of change in Duration, the dependent variable (*price*), increases by 0.35.

Total_Stops: 0.471(positive)
For every 1 standard deviation of change in Total_Stops, the dependent variable (*price*), increases by 0.471.

Distance: 0.093(positive)
For every 1 standard deviation of change in Distance, the dependent variable (*price*), increases by 0.093.

Multi-Linear Regression Interpretation and Analysis:

R-squared values range from 0% to 100%. An R^2 of 0% indicates that the independent variables have no explanatory power in predicting the dependent variable, while an R^2 of 100% indicates that the independent variables perfectly explain the variation in the dependent variable.
Our model has given the R-squared value of **46.7%** which indicates a moderate linear relationship between the variables. There is a discernible pattern and some association between the variables, and the fit of the data to the regression line is not extremely tight. The scatter plot exhibits a more dispersed pattern rather than a concentrated cluster of points around the regression line.

Bivariate Results:
The p-value calculated between Airline_JetAirways, Day_of_Journey, Dep_Time, Duration, Total_Stops, and Distance was below the 0.05 critical value. Hence they show a statistical significance and a stronger chance of a genuine result. Thus, **all** the independent variables possess a genuine predictive relationship with the outcome (dependent) ***price*** variable.

Both standardized and unstandardized coefficients indicate a negative relationship between the independent variable (Dep_Time) and the dependent variable (Price). As the value of Dep_Time increases, the Price is expected to decrease, regardless of whether the coefficients are standardized or unstandardized.

On the other hand, the independent variables Airline_JetAirways, Day_of_Journey, Duration, Total_Stops, and Distance demonstrated the behavior where the dependent price variable increases in the index for the unstandardized and standardized coefficients for each independent variable increase in these relationships.

As all the variables are statistically significant, standardized coefficients provide a way to rank the independent variables based on their influence on the dependent variable, irrespective of the original units of measurement. So, based on the standardized coefficient values we can say that independent variables impact the dependent variable in the following order

1. Total_stops - 0.471
2. Airline_JetAirways - 0.314
3. Distance - 0.093
4. Duration - 0.035
5. Day_of_Journey - 0.034
6. Dep_Time - -0.22

Considering the results and above_sequence it is evident that independent variables, Total_stops make a significant impact on the dependent variable price.

Aspects to the Policymaker:

1. If the Policymaker is an **End-User:**
   - **Considering Layovers:** As an end-user, if you have sufficient time and enjoy exploring new places, opting for flights with layovers can be a rewarding choice. Layovers provide opportunities to discover new destinations and break up long journeys, making the travel experience more enjoyable.
   - **Consider Airline Options:** It is advisable to compare flight options from different airlines, particularly when it comes to Jet Airways. Unless there is a specific reason for choosing Jet Airways, exploring other airlines may lead to cost savings as other carriers may offer more competitive prices.

2. If the Policymaker is the **CEO:**

- Dynamic Pricing: As the CEO of an airline, considering dynamic pricing based on departure times can be a lucrative strategy. Charging slightly higher fares for flights departing between 9 AM - 10 PM, when demand is usually higher, can help maximize revenue.

- Encouraging Layovers: Promoting layovers on certain routes, especially in attractive layover cities, can benefit both the airline and the passengers. This approach can increase the appeal of the travel package and potentially lead to increased ticket sales.

Ethical Implication and Challenges/Limitations:

Ethical Implication:

The data analysis project on prediction of airline prices obtains all its data from Kaggle, and it is crucial to note that it does not include any personal information of individual users. All the data used is solely related to travel patterns, flight routes and other non-sensitive aspects of airline travel. Given these circumstances, the study is devoid of any direct ethical implications related to privacy, informed consent, or data protection.

Challenges:

- One challenge is the dynamic nature of flight prices. Airline ticket prices are subject to various factors that can fluctuate over time, such as fuel costs, demand-supply dynamics, seasonal variations, economic conditions, and airline pricing strategies.
- Second, this dataset is based on 2019 records which might not accurately reflect the current pricing landscape. Prices may have changed due to market conditions, inflation, or other external factors. Therefore, the predictions based on the dataset may not align with the actual prices in the present-day real world. To overcome this challenge, it is essential to regularly update the dataset with the most recent data or employ real-time data sources for accurate predictions.

More time and data:

If I had access to more data and additional time, I would have invested the effort to collect recent data that is not readily available online. Additionally, I would have taken the opportunity to analyze the relationship between the dependent variable (flight ticket price) and other relevant variables, such as the class of travel and the number of days before the ticket was booked.

Acquiring more recent data would have allowed for a more up-to-date analysis, potentially providing more accurate insights into the current market trends and pricing patterns. Such information could be crucial in understanding how flight prices have evolved over time and identifying any recent fluctuations or seasonality.

By delving into these additional variables like analyzing the relationship between the class of travel and ticket prices, studying the correlation between the number of days before the ticket booking and the ticket price the analysis would have gained depth and comprehensive insights, enabling a more robust understanding of the factors influencing flight ticket prices. Although limited by the available data and time constraints, exploring these aspects could enhance the

overall understanding of the relationships in the data and offer more comprehensive conclusions.

For future research or analysis, considering these aspects and expanding the dataset could be beneficial in uncovering more nuances and refining the predictions or conclusions related to flight ticket pricing.

References:

- Prediction of Flight Fares using Machine Learning
  https://ieeexplore.ieee.org/document/10048801

- Flight Price Prediction Using Machine Learning Techniques
  https://www.ijcseonline.org/pub_paper/3-IJCSE-08981.pdf