Flight Price Prediction and Analysis Using Linear Regression

Saroar Jahan Shuba

Date: 19/11/2025

1. Objectives

The main goal of this project is to understand what factors drive the price of flight tickets and to build models that can predict ticket prices based on those factors. More specifically, this project aims to:

- Explore how flight **duration** affects ticket prices.
- Examine whether the **number of stops** on a journey significantly changes the price.
- Investigate the influence of **additional information** (such as baggage allowance, in-flight meals, or layovers) on pricing differences.
- Test how well these predictors together can explain and predict variations in flight prices.
- Identify whether these relationships are strong, weak, positive, negative, or insignificant.

By doing this, the analysis provides a clear picture of how flight characteristics shape ticket pricing, which can be useful both for airlines (in pricing strategies) and for customers (in making informed choices).

2. List Of Variables

Variable	**Role**	**Nature**
Price	Response Variable	Numerical (continuous)
Duration_Minute	Predictor Variable	Numerical (continuous)
Total_Stops	Predictor Variable	Numerical (0,1,2,. . .)
Additional_Info	Predictor Variable	Categorical (baggage, in flight meal,in-flight services, layovers, No-info)

Table 1: Description of Variables Used in the Regression Model

3. Data Preprocessing

The dataset was obtained from kaggle [1], containing information on ticket prices and flight characteristics such as duration, number of stops, airline, and additional information. Since the dataset included both numerical and categorical variables, several preprocessing steps were necessary before modeling:

- **Missing Values:** The dataset contained two missing values (one in the *Route* column and one in the *Total_Stops* column). Since these were not part of the main predictors used for modeling, no major imputation was required.
- **Categorical Variables:** *Total_Stops* and *Additional_Info* were categorical in nature. *Total_Stops* was primarily treated as a factor (to reflect non-linear price differences between direct and connecting flights), but was also experimented with as a numeric variable using a square-root transformation. *Additional_Info* was treated as a factor, capturing services such as baggage, in-flight meals, and layovers.
- **Transformations:** To address skewness and stabilize variance:
  - *Price* (response variable) was log-transformed.
  - *Duration_Minute* was log-transformed.
  - *Total_Stops* was square-root transformed when treated numerically.
- **Scaling and Standardization:** No standardization or normalization was applied, since regression models in this context do not require predictors on the same scale. The focus was instead on handling skewness through transformations.

In summary, preprocessing was an essential step for this dataset. This flight dataset included both skewed numerical variables and categorical features that required transformation and encoding to ensure valid regression modeling.

4. Exploratory Data Analysis (EDA)

We conducted a brief exploratory data analysis (EDA) prior to model fitting to examine the dataset's key characteristics. This analysis included descriptive statistics and visualizations for both response and predictor variables.

4.1 Descriptive Statistics

Table 2 reports the mean, median, minimum, and maximum values of the main variables used in the regression analysis. The response variable is *Price*, while *Duration_Minute* and *Total_Stops* serve as predictors.

Table 2: Descriptive Statistics of Main Variables

Variable	**Mean**	**Median**	**Min**	**Max**
Price	9,087	8,372	1,759	79,512
Duration_Minute	643	520	5	2,860
Total_Stops	0.824	1	0	4

From Table 2, we observe that **Price** varies widely, with a minimum of 1,759 and a maximum of nearly 80,000. The average ticket price is around 9,087, which highlights a large spread between economy flights and premium international tickets. **Duration_Minute** averages about 643 minutes (approximately 10.7 hours), with most flights being shorter (median = 520 minutes, about 8.6 hours). Some long-haul flights last over 2,800 minutes (around 47 hours).

The variable **Total_Stops** ranges from 0 (non-stop flights) to 4. On average, most flights have around one stop, with non-stop flights generally being priced higher and multi-stop flights tending to be cheaper. This makes *Total_Stops* a strong categorical predictor of ticket price.

4.2 Visualizations

To further motivate the modeling choices, we summarize key patterns from exploratory graphics. The histograms and boxplots are presented *before* and *after* transformations to show how skewness and outliers are handled. A correlation heatmap is also provided.

Summary of observations.

- **Histograms:** *Price* and *Duration_Minute* are right-skewed with long upper tails, while *Total_Stops* is discrete with heavy mass at 0–1 stops. These patterns justify the use of a log transform for *Price* (and *Duration_Minute*) and handling *Total_Stops* as a factor.

- **Boxplots:** Long upper whiskers and many outliers are visible for *Price* and *Duration_Minute*, consistent with premium or very long flights. After log transformation, spread and outliers are reduced.

- **Correlation:** *Price* and *Duration_Minute* show a moderate positive association. Treating *Total_Stops* as numeric yields weak linear correlation with *Price*, reinforcing the choice to model *Total_Stops* as a categorical predictor. VIF checks indicated no serious multicollinearity (VIF < 5).

Graphs Before Transformation

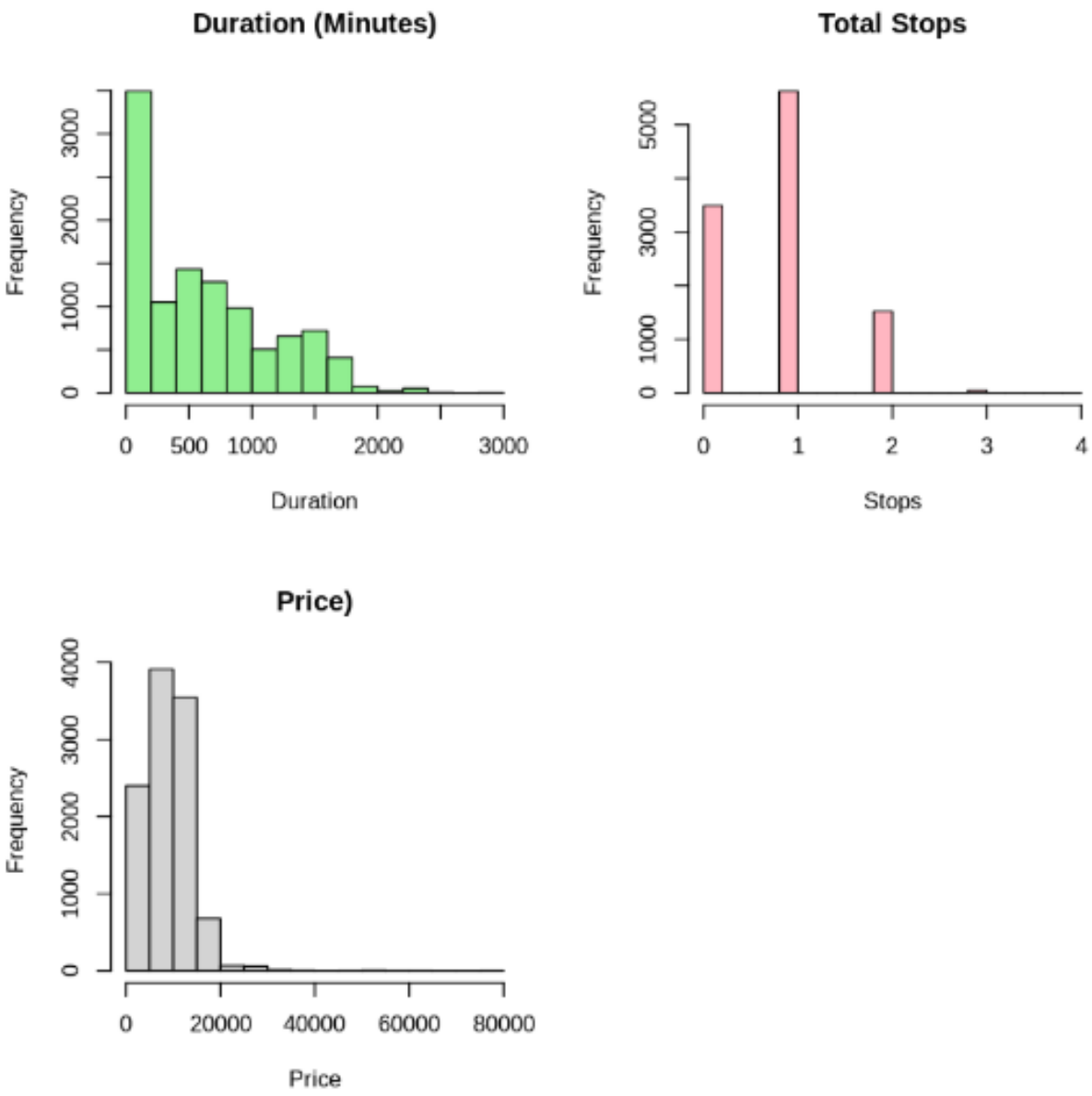

Figure 1: Histograms (before transformation) for *Price*, *Duration_Minute*, and *Total_Stops*.

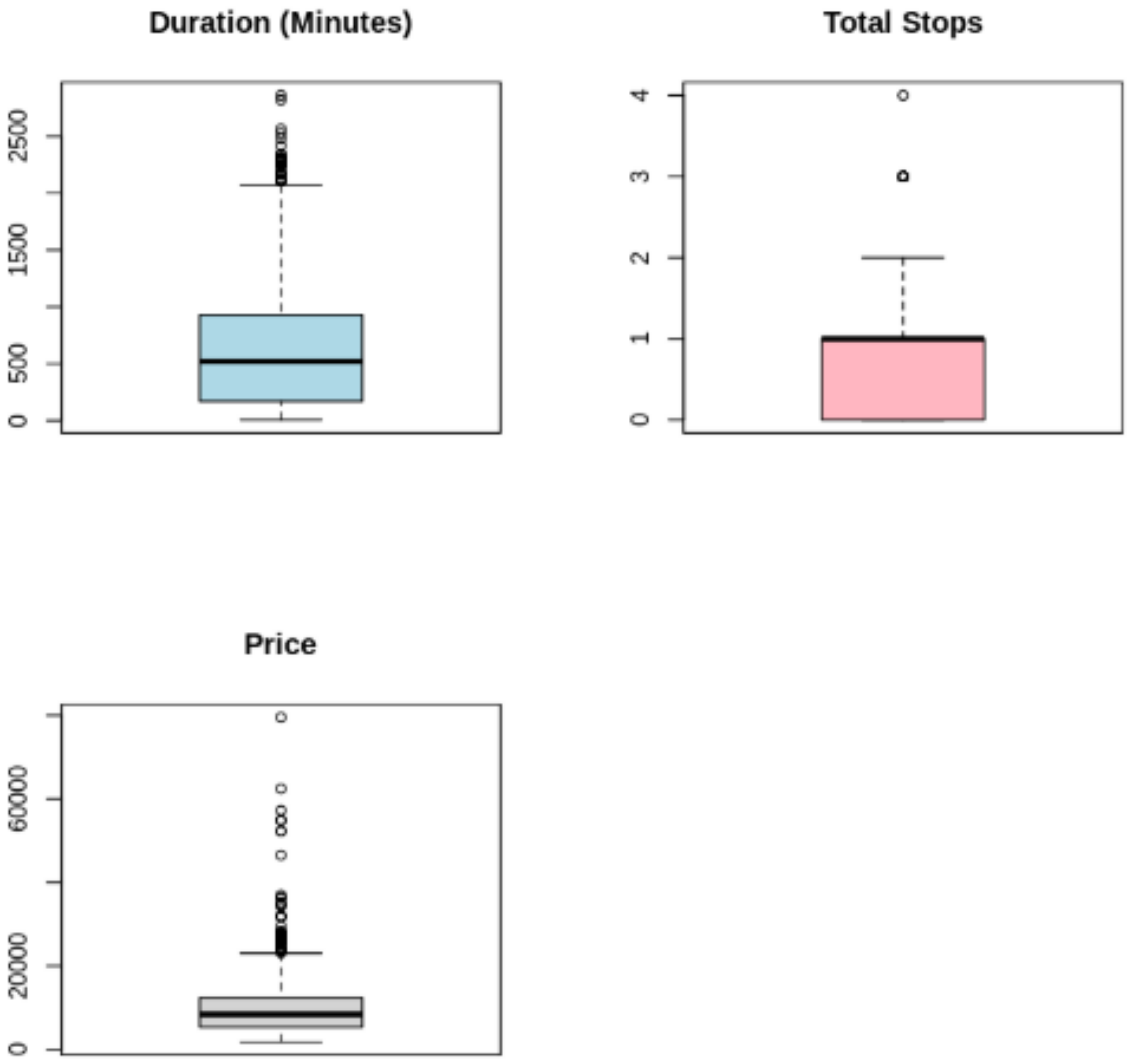

Figure 2: Boxplots (before transformation) for *Price*, *Duration_Minute*, and *Total_Stops*.

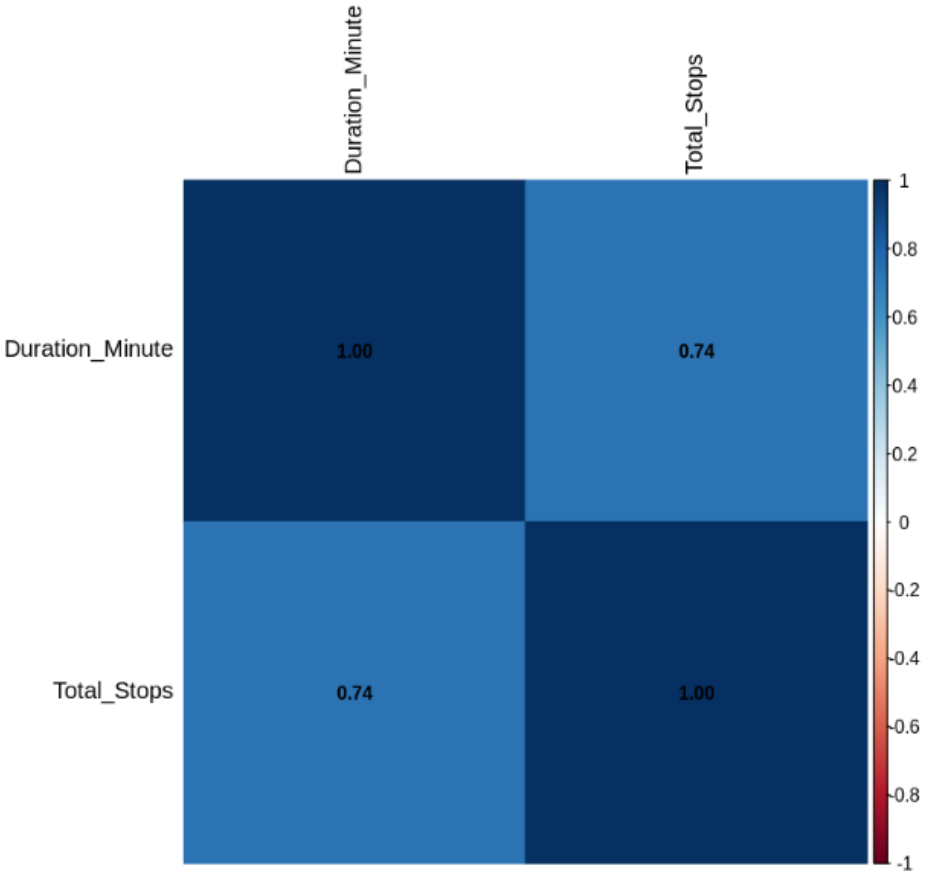

Figure 3: Correlation heatmap of numeric variables (before transformation): *Price*, *Duration_Minute*, and a numeric coding of *Total_Stops*.

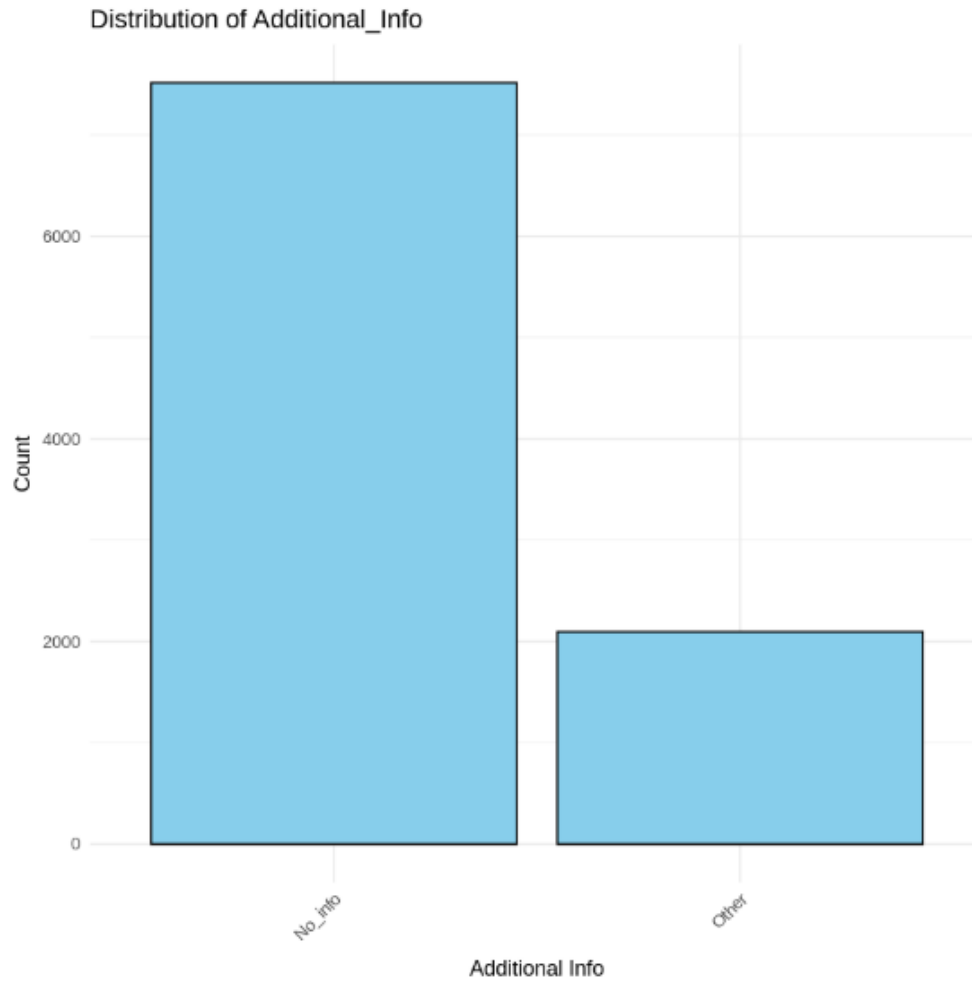

Figure 4: Barplot Of Additional Info

4.2 Summary of EDA Findings

The exploratory data analysis suggests the following key patterns:

- **Price:** Strongly right-skewed with extreme outliers, highlighting the need for a log transformation to stabilize variance and improve model fit.
- **Duration_Minute:** Shows a moderate positive relationship with *Price*, consistent with longer flights tending to be more expensive, though the effect is not strictly linear.
- **Total_Stops:** Exhibits a clear categorical effect. Non-stop flights are generally priced the highest, while flights with more stops tend to be cheaper. This pattern justifies treating *Total_Stops* as a factor rather than a purely numeric variable.
- **Additional_Info:** Contains categorical distinctions (e.g., baggage, in flight meal, in-flight services, layovers, No-info) that can explain part of the variation in ticket pricing. Almost 90% of the data is in-info category. So i create 2 category No-info and other(rest of the category are in this).

Overall, the observed relationships motivate the use of both **simple regression models** (*Price* on *Duration* or *Stops*) to capture direct effects, and **multiple regression**

models that incorporate *Duration*, *Stops*, and *Additional_Info* together to better explain variation in ticket prices.

Graphs After Transformation

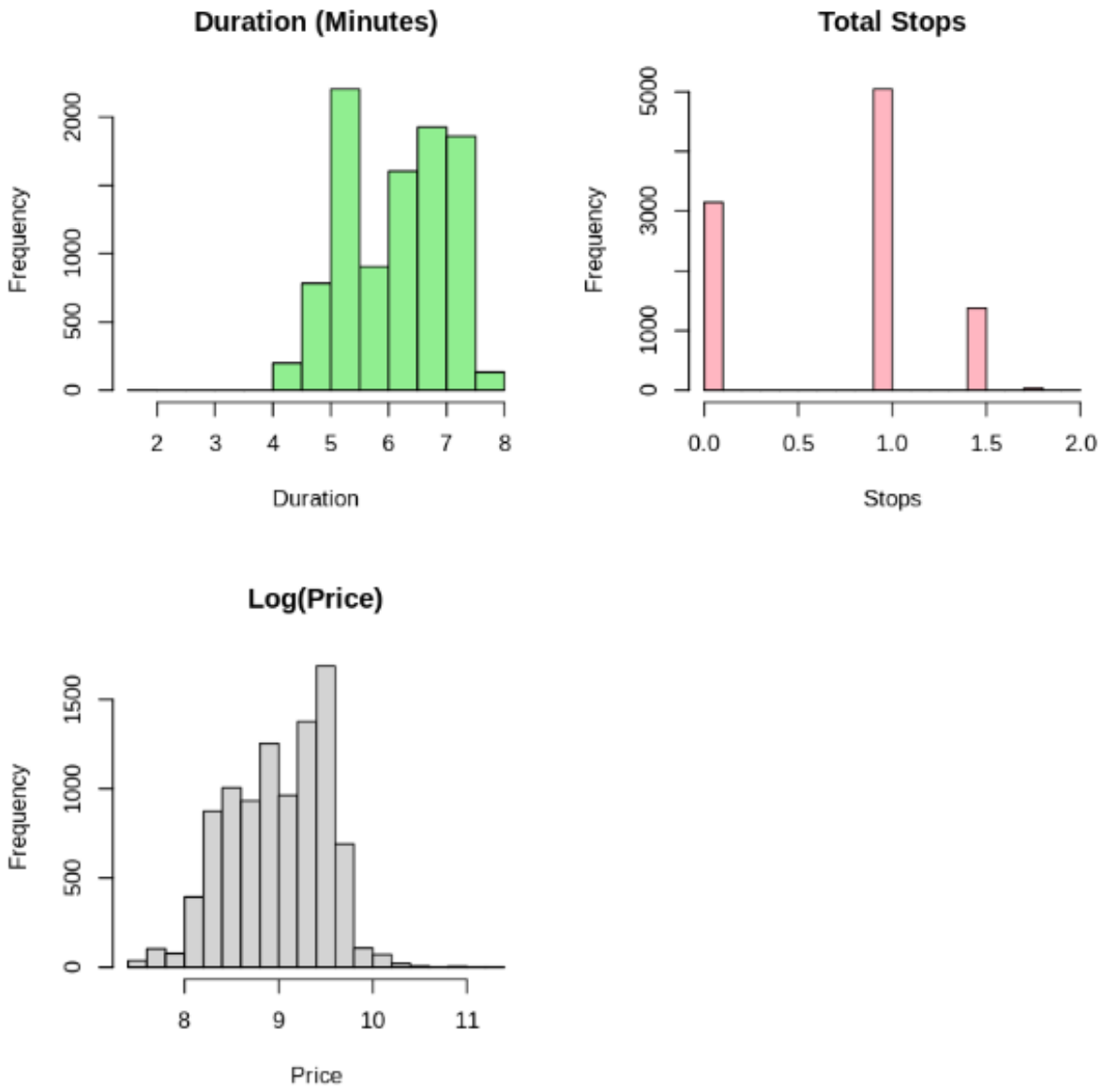

Figure 5: Histograms (after transformation) for log(*Price*), log(*Duration_Minute*), and $\sqrt{Total\_Stops}$.

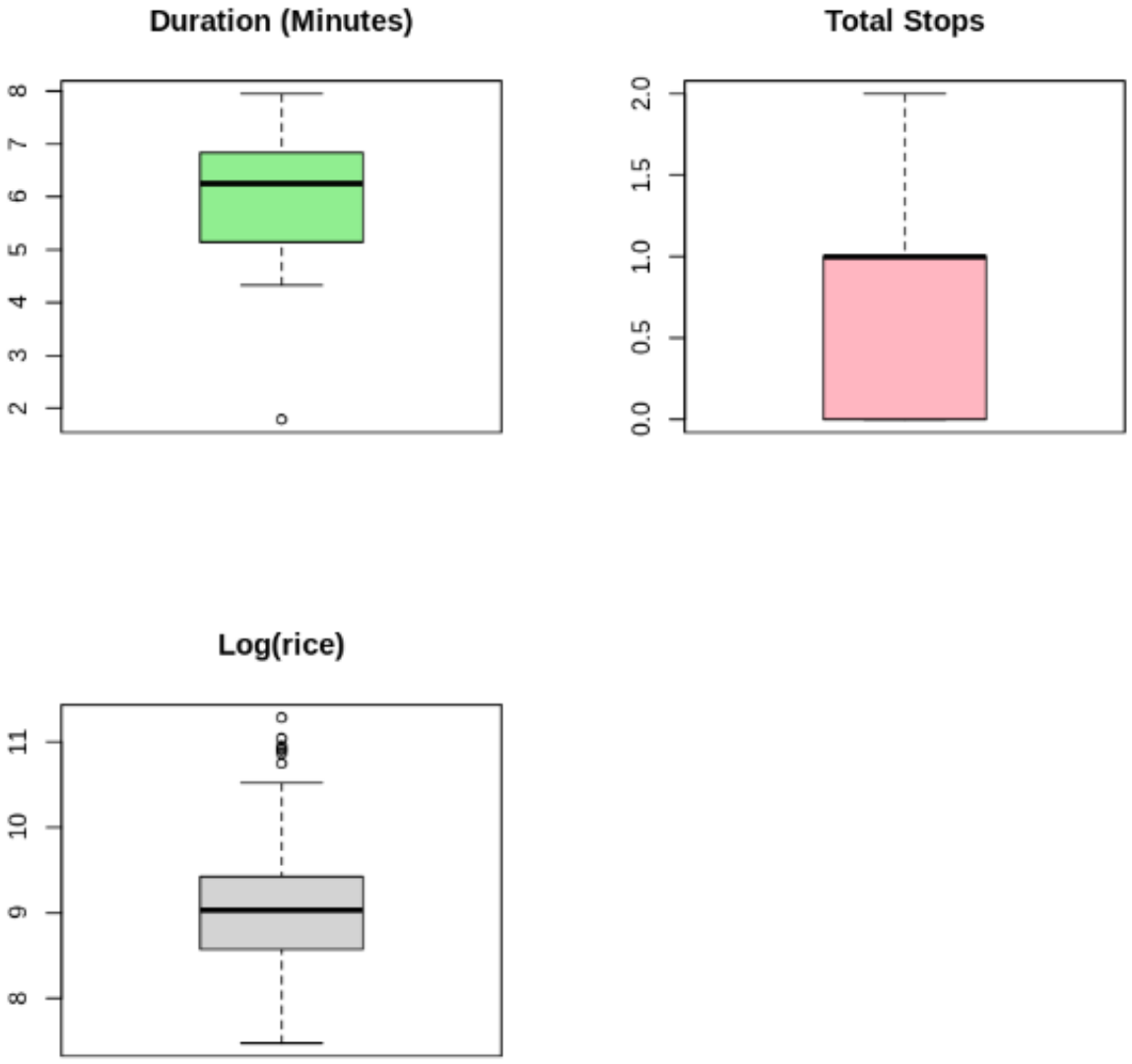

Figure 6: Boxplots (after transformation) for log(*Price*), log(*Duration_Minute*), and *Total_Stops*.

Correlation Matrix

	Duration_Minute	**Total_Stops**
Duration_Minute	1.0000000	0.7380709
Total_Stops	0.7380709	1.0000000

Table 3: Correlation matrix of predictors

5. Data Spliting

After preprocessing and performing exploratory data analysis, the dataset was split into **training** and **test** sets to enable model fitting and evaluation. An **90–10 split** was used, where 90% of the flight records were assigned to the training set and 10% were reserved as the test set.

This split ensured that the regression models were trained on the majority of the data while keeping aside a portion for assessing their predictive performance on unseen flights. To maintain consistency and reproducibility, the split was performed randomly using a fixed random seed.

The training set was used to fit both the **simple regression models** (*Price* on *Duration* or *Stops* individually) and the **multiple regression models** (*Price* on *Duration*, *Stops*, and *Additional_Info*). The test set was then used exclusively for evaluating predictions and assessing the models' ability to generalize beyond the training data.

6. Simple Linear Regression

6.1 General Formula

The general form of a simple linear regression model is expressed as:

$$y = \beta_0 + \beta_1 x + \epsilon$$

where y is the response variable, x is the predictor, β_0 is the intercept, β_1 is the slope coefficient, and ϵ is the error term.

6.2 Application to Flight Price and Duration

In this case:

- Response variable (y): `Price` (ticket price)
- Predictor variable (x): `Duration_Minute` (flight duration in minutes)

6.3 Regression Results

	Estimate	Std. Error	t value	Pr($>\|t\|$)
Intercept	6127.0	65.72	93.23	$< 2e{-}16^{***}$
Duration_Minute	4.616	0.08025	57.53	$< 2e{-}16^{***}$

Table 4: Simple Linear Regression of `Price` on `Duration_Minute`.

Model Statistics:
Residual standard error = 3991 (df = 9611)
Multiple R-squared = 0.2561, Adjusted R-squared = 0.2561
F-statistic = 3309 on 1 and 9611 DF, p-value $< 2.2e{-}16$

6.4 Regression Equation

$$\widehat{Price} = 6127.0 + 4.616 \cdot Duration\_Minute$$

The slope coefficient $\beta_1 = 4.616$ is highly significant ($p < 0.05$), indicating that for every additional minute of flight duration, the ticket price increases on average by about 4.62 units.

The intercept $\beta_0 = 6127$ is also statistically significant ($p < 0.05$), suggesting that when duration is zero, the baseline model predicts a price near 6127 units. Although this has little real-world meaning (since a flight cannot have zero minutes), it reflects the model's adjustment point.

The R^2 value of 25.6% indicates that flight duration explains about a quarter of the variation in ticket prices. This shows that while duration is an important factor, other predictors (such as stops, airline, or booking conditions) also play substantial roles in determining price.

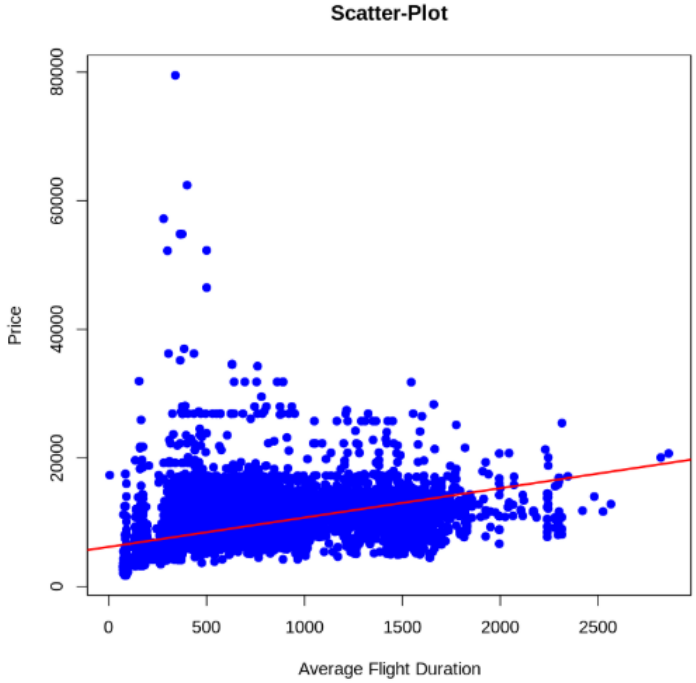

Figure 7: Scatter plot of `Price` vs. `Duration_Minute` with fitted regression line.

The plot shows a weak positive relationship between flight duration and ticket price. While the fitted regression line slopes upward, indicating that longer flights tend to be more expensive, the wide scatter of points around the line suggests that duration alone is not a strong predictor of price. Other factors such as airline, number of stops likely play an important role in explaining price variation.

7. Multiple Linear Regression

7.1 General Formula

The general form of a multiple linear regression model is expressed as:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

where y is the response variable, X_1, X_2 are predictors, β_0 is the intercept, β_1, β_2 are slope coefficients, and ϵ is the error term.

7.2 Application to Flight Prices

In this case:

- Response variable (y): `Price` (ticket price)
- Predictor variables:
  - X_1: `Duration_Minute` (Flight duration in minutes)
  - X_2: `Total_Stops` (Number of stops)

Predictor Symbol	Definition
X_1	Flight Duration (in minutes)
X_2	Number of Stops

Table 5: Table of Predictors and Their Definitions

7.3 Regression Results

	Estimate	Std. Error	t value	Pr($>\|t\|$)
Intercept	5471.80	62.44	87.64	$< 2e{-}16^{***}$
Duration_Minute	1.2296	0.1095	11.23	$< 2e{-}16^{***}$
Total_Stops	3433.32	81.98	41.88	$< 2e{-}16^{***}$

Table 6: Multiple Linear Regression of `Price` on Flight Duration and Stops.

Model Statistics:
Residual standard error = 3670 (df = 9610)
Multiple R-squared = 0.3709, Adjusted R-squared = 0.3708
F-statistic = 2833 on 2 and 9610 DF, p-value $< 2.2e{-}16$

7.4 Regression Equation

$$\widehat{Price} = 5471.80 + 1.2296 \cdot Duration\_Minute + 3433.32 \cdot Total\_Stops$$

The coefficient for `Duration_Minute` ($\beta_1 = 1.2296$) is positive and highly significant ($p < 0.001$), indicating that each additional minute of flight time is associated with an increase in ticket price of about 1.23 units. This shows that longer flights tend to be more expensive.

The coefficient for `Total_Stops` ($\beta_2 = 3433.32$) is also highly significant ($p < 0.05$), indicating that each additional stop increases ticket price on average by about 3433 units. This highlights the strong effect of stops on pricing.

The intercept $\beta_0 = 5471.80$ is also significant, though it mainly represents the baseline price when duration and stops are zero, which has limited real-world meaning.

The R^2 value of 37.1% (Adjusted $R^2 = 37.1\%$) indicates that about one-third of the variation in ticket prices can be explained by flight duration and stops. While meaningful, this also suggests that other factors (such as airline, seasonality, or demand) contribute substantially to price variation.

8. Multiple Linear Regression (Log-Transformed Model)

8.1 General Formula

The general form of a multiple linear regression model is expressed as:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

where $\ln(y)$ is the log of the response variable, X_1, X_2 are predictors, β_0 is the intercept, β_1, β_2 are slope coefficients, and ϵ is the error term.

8.2 Application to Flight Prices

In this case:

- Response variable (y): `Price` (ticket price, log-transformed)
- Predictor variables:
  - X_1: `Duration_Minute` (Flight duration in minutes)
  - X_2: `Total_Stops` (Number of stops)

Predictor Symbol	Definition
X_1	Flight Duration (in minutes)
X_2	Number of Stops

Table 7: Table of Predictors and Their Definitions

8.3 Regression Results

	Estimate	Std. Error	t value	Pr($>\|t\|$)
Intercept	7.4949	0.0399	188.05	$< 2e{-}16^{***}$
Duration_Minute	0.1946	0.0079	24.72	$< 2e{-}16^{***}$
Total_Stops	0.4183	0.0134	31.27	$< 2e{-}16^{***}$

Table 8: Multiple Linear Regression of log(Price) on Flight Duration and Stops.

Model Statistics:
Residual standard error = 0.3375 (df = 9610)
Multiple R-squared = 0.5681, Adjusted R-squared = 0.5680
F-statistic = 6319 on 2 and 9610 DF, p-value $< 2.2e{-}16$

8.4 Regression Equation

$$\widehat{(Price)} = 7.4949 + 0.1946 \cdot Duration\_Minute + 0.4183 \cdot Total\_Stops$$

The coefficient for `Duration_Minute` ($\beta_1 = 0.1946$) is positive and highly significant ($p < 0.05$), indicating that longer flights are associated with higher ticket prices. Since

the model is log-linear, a one-unit increase in duration (scaled appropriately) increases the expected ticket price multiplicatively by $e^{0.1946} \approx 1.215$, or about 21.5%.

The coefficient for `Total_Stops` ($\beta_2 = 0.4183$) is also highly significant ($p < 0.05$). Each additional stop increases expected ticket prices by a factor of $e^{0.4183} \approx 1.52$, or roughly 52% higher compared to flights with fewer stops.

The intercept $\beta_0 = 7.4949$ corresponds to the baseline log-price when both predictors are zero. While not interpretable in real-world terms (as a flight cannot have zero minutes and zero stops), it serves as the model's baseline adjustment.

The R^2 value of 56.8% (Adjusted R^2 = 56.8%) indicates that the log-linear model explains over half of the variation in flight prices, substantially improving fit compared to the simple duration-only model.

9. Multiple Linear Regression (with `Additional_Info`)

9.1 General Formula

The multiple linear regression with a log-transformed response is

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon,$$

where y is ticket price, X_1 is flight duration, X_2 encodes the number of stops, and X_3 encodes additional information.

9.2 Application to Flight Prices

In this case:

- Response variable (y): `Price` (log-transformed).
- Predictor variables:
  - X_1: `Duration_Minute` (flight duration in minutes).
  - X_2: `Total_Stops` (baseline = 0; levels: 1, 2, 3 , 4).
  - X_3: `Additional_Info` (categorical; baseline = `No_info`; level used here: `Other`).

Predictor Symbol	**Definition**
X_1	Flight Duration (minutes)
X_2	Number of Stops (0, 1 , 2 , 3 , 4)
X_3	Additional Info (`No_info` vs `Other`)

Table 9: Table of Predictors and Their Definitions

9.3 Regression Results

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
Intercept	7.4691	0.0400	186.950	$< 2e{-}16^{***}$
Duration_Minute	0.2011	0.0079	25.323	$< 2e{-}16^{***}$
Total_Stops: 1 stop	0.4362	0.0144	30.288	$< 2e{-}16^{***}$
Total_Stops: 2 stops	0.5461	0.0195	27.970	$< 2e{-}16^{***}$
Total_Stops: 3 stops	0.5231	0.0562	9.303	$< 2e{-}16^{***}$
Total_Stops: 4 stops	0.8795	0.3354	2.622	0.00875^{**}
Additional_Info: Other	-0.0721	0.0083	-8.637	$< 2e{-}16^{***}$

Table 10: Multiple Linear Regression of ln(`Price`) on Duration, Stops, and Additional Info. Baselines: non–stop, `No_info`.

Model Statistics:
Residual standard error = 0.3347 (df = 9604)
Multiple $R^2 = 0.5732$, Adjusted $R^2 = 0.5729$
F-statistic = 2149 on 6 and 9604 DF, p-value $< 2.2e{-}16$

Regression Equation

$$\widehat{(\text{Price})} = 7.4691 + 0.2011 \cdot X_1 + 0.4362 \cdot \mathbb{1}\{\text{1 stop}\} + 0.5461 \cdot \mathbb{1}\{\text{2 stops}\} + 0.5231 \cdot \mathbb{1}\{\text{3 stops}\} + 0.8795 \cdot \mathbb{1}\{\text{4 sto}$$

Interpretation

Because the model is on the log scale, coefficients can be expressed as approximate percentage changes:

- A one–unit increase in `Duration_Minute` is associated with about a **20% increase** in expected price.
- Relative to non–stop flights, expected prices are higher by approximately:
  - **44%** for 1 stop,
  - **55%** for 2 stops,
  - **52%** for 3 stops,
  - **88%** for 4 stops.
- Flights with **Additional_Info = Other** are about **7% cheaper** than those with `No_info`.

The inclusion of `Additional_Info` slightly improves model fit ($R^2 = 0.5732$). The negative coefficient indicates that flights with explicit ancillary details (such as baggage or meals) tend to be slightly less expensive, after controlling for duration and stops. Meanwhile, the number of stops remains the strongest determinant of ticket price on the log scale.

10. Model Selection using AIC

To evaluate competing regression models, the Akaike Information Criterion (AIC) was used. AIC balances model fit against model complexity, with lower values indicating a better trade-off between explanatory power and parsimony.

Model	AIC
Model_new2: Duration + Total_Stops + Additional_Info	6245.87
Model_log: Duration + Total_Stops	6354.66

Table 11: Comparison of competing models using AIC. Lower AIC indicates preferred model.

The results show that `model_new2` has a substantially lower AIC value (6245.87) compared to `model_log` (6354.66). The difference of approximately 109 points is well above the usual cutoff of 10, providing strong evidence that `model_new2` offers a better fit to the data despite having more parameters.
Conclusion: The model including `Additional_Info` should be preferred, while the simpler model without it can be rejected.

11. Model Evaluation on Test Data

11.1 Performance Metrics

To assess the predictive performance of the multiple linear regression model, the fitted model was applied to the test dataset. Two key evaluation metrics were computed: the coefficient of determination (R^2) and the root mean squared error (RMSE).

- **Test** R^2: -13.4328, indicating that the model performs very poorly on the test data. The negative value suggests that the model fits the data worse than a simple baseline (predicting the mean of the response variable).
- **Test RMSE**: 0.6507, representing the average magnitude of prediction errors. A higher RMSE indicates substantial deviations between predicted and actual prices.

The negative R^2 and relatively high RMSE demonstrate that the model does not generalize well to unseen data and fails to capture the main patterns in flight prices. This highlights the need for model refinement, such as incorporating additional predictors, feature engineering, or considering alternative modeling approaches.

12. Model Validation: Regression Assumptions

Before interpreting the results of the linear regression model, it is crucial to validate that the underlying statistical assumptions are met. This ensures the model's coefficients and p-values are reliable and that the conclusions drawn from the analysis are statistically sound. For the flight price model, diagnostic checks were performed on the residuals, with the following findings:

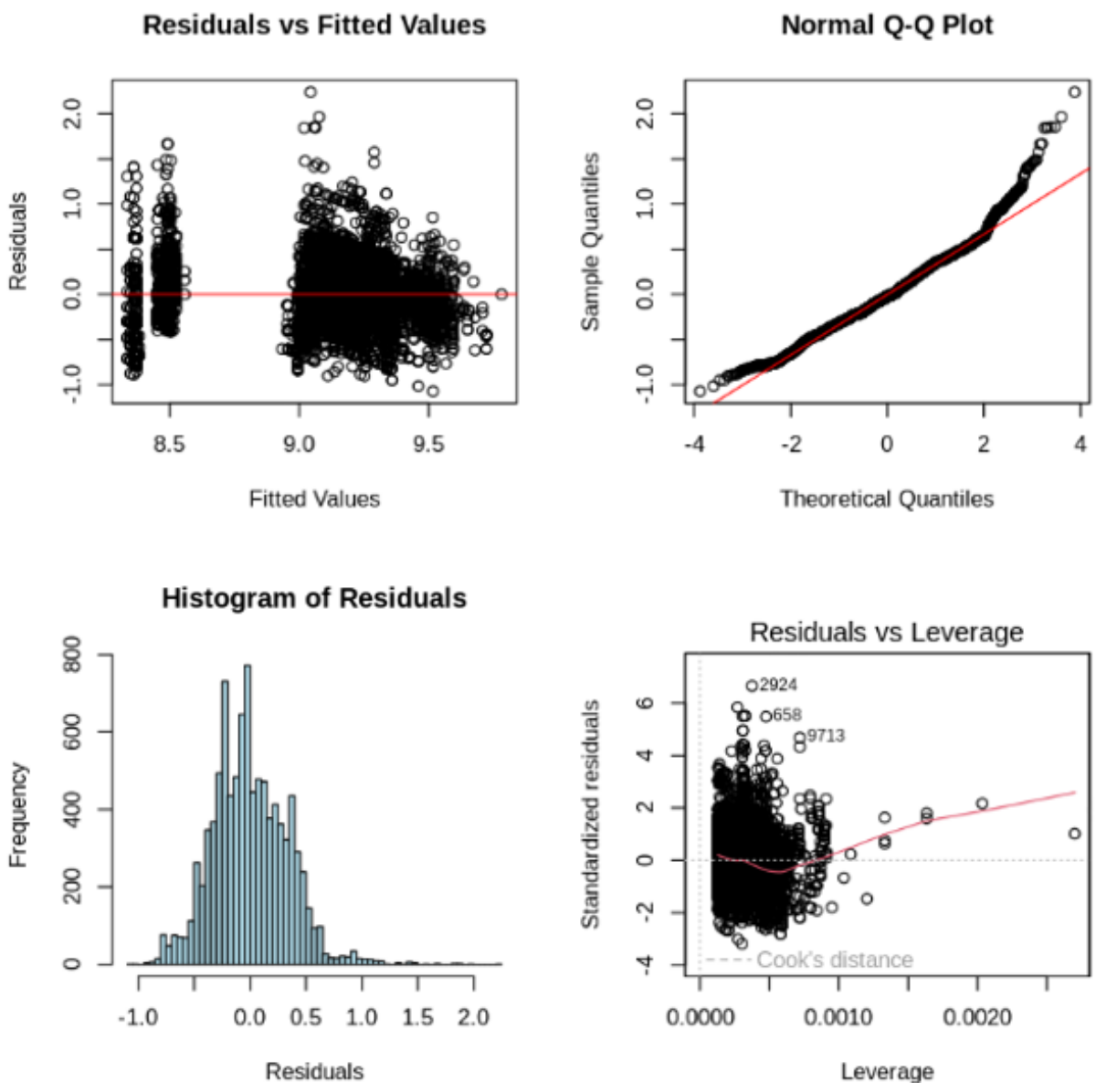

Figure 8: Enter Caption

- **Linearity:** The Residuals vs. Fitted plot shows clear patterns rather than a random scatter, indicating that the linearity assumption is violated. The model does not fully capture the relationship between predictors and the log of price.
- **Normality of Residuals:** The Normal Q–Q plot reveals systematic deviations from the diagonal line, particularly at the tails. This suggests that the residuals are not normally distributed, violating the normality assumption.
- **Homoscedasticity:** The Residuals vs. Fitted plot also indicates non-constant variance, with clusters of points showing different spreads. This heteroscedasticity undermines the constant-variance assumption.
- **Influential Points:** The Residuals vs. Leverage plot highlights several influential observations (e.g., cases 2924, 6568, and 9713). These points may disproportionately affect the regression coefficients and model stability.
- **Independence:** Since the dataset consists of individual flight records, there is no evidence of temporal correlation, and the independence assumption is reasonably satisfied.

The diagnostics reveal multiple violations of regression assumptions, including non-linearity, non-normal residuals, heteroscedasticity, and the presence of influential points. These issues contribute to the poor test performance (negative R^2 and high RMSE) and indicate that the linear regression model does not generalize well. Model refinement, such as transformations, removal of outliers, or use of more flexible non-linear approaches, will be necessary to improve predictive performance.

13. Independence of Errors

The independence of errors assumption was also checked using the Box–Ljung test [2]. This test evaluates whether residuals are autocorrelated across multiple lags.

Test Statistic	**Result**
Chi-squared (X^2)	13.855
Degrees of freedom	10
p-value	0.1797

Table 12: Box–Ljung test results for residual autocorrelation.

The p-value of 0.1797 is greater than the conventional significance level of 0.05, indicating that we fail to reject the null hypothesis of no autocorrelation. Therefore, the assumption that the errors are independent is satisfied.

14. Assumptions and Limitations

While this study applies linear regression models to analyze the relationship between flight ticket prices and their predictors (Duration, Total Stops, and Additional Info), several limitations should be considered:

- **Violation of Linear Regression Assumptions:** Diagnostic checks revealed violations of key assumptions, including non-linearity, non-normal residuals, and heteroscedasticity. These issues limit the reliability of coefficient estimates and p-values, reducing confidence in the model's explanatory power. Although the independence assumption was satisfied (as confirmed by the Durbin–Watson test), other assumption violations highlight that a linear model may not be the best fit.

- **Simplification of Predictors:** Important real-world drivers of ticket price, such as airline, seasonality, demand surges, and booking time, were not included in the model. This omission means that the regression only partially explains price variation. For example, the model's R^2 value of about 0.57 suggests that nearly half of the variation in flight prices remains unexplained.

- **Categorical Encoding Choices:** `Additional_Info` were simplified into categorical factors. While necessary for regression analysis, this process may lose some nuances (e.g., grouping diverse "Other" flight information into a single category), potentially obscuring subtle effects.

- **Presence of Outliers and Influential Observations:** Several flights with extreme durations or prices exerted high influence on the regression fit, as identified by leverage and Cook's distance diagnostics. These influential points may distort coefficient estimates and inflate error measures such as RMSE.

Despite these limitations, the regression framework provides a useful starting point for identifying the main relationships between flight duration, number of stops, additional flight information, and prices. The findings highlight the need for more sophisticated modeling approaches—such as non-linear methods, robust regression, or machine learning algorithms—to better capture the complex structure of flight pricing.

Conclusion

This study explored the main factors that drive flight ticket prices, focusing on flight duration, number of stops, and additional flight information. The results showed that longer flights and flights with more stops tend to cost more, while some categories of additional information were linked to slightly lower prices. Including the `Additional_Info` variable improved the model compared to a simpler version, which means it adds useful predictive value.

Even though the model explained about 57% of the variation in prices, its performance on test data was weak, with a negative R^2 and higher errors. The diagnostic checks also revealed issues such as non-linearity, unequal variance in residuals, and outliers, although the independence of errors assumption was satisfied.

Overall, linear regression provided a good starting point and helped identify important price drivers, but it is not flexible enough to fully capture the complexity of airline pricing. A more robust approach, possibly using non-linear or machine learning methods and additional predictors like airline, booking time, and seasonality, would likely give more accurate and reliable results.

References

1. Imran, M. B. (n.d.). *Flight Price Prediction Dataset*. Kaggle. Retrieved from https://www.kaggle.com/datasets/muhammadbinimran/flight-price-prediction

2. Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2), 297–303. https://doi.org/10.1093/biomet/65.2.297