A Statistical Exploration of Airline Fare Trends in India

Bianca Otel, Velina Todorova

January 2025

Introduction

One of the most recurring problems that the average traveller is confronted with is the volatility of the price of plane tickets. Designing a model that can predict the price of a plane ticket, can go a long way towards tackling this issue, and that is the main focus of this project, along with analyzing what are the most relevant factors in plane ticket pricing. In our analysis we used a dataset covering flights departing from the top 7 busiest airports in India - a dataset that includes information specific to each individual trip (from the time of departure and arrival, to the number of days prior to the travel that the purchase was made). Alongside these very important factors, research shows that one of the most vital factors in plane ticket pricing is the price of the fuel (kerosene), and that variable was initially missing from our dataset. However, in the spirit of having a more complete analysis, we added a column to our dataset, containing the corresponding monthly average price per gallon of kerosene. Using all this data in a linear regression, we aim to build a model that helps travelers identify the best time and circumstances for purchasing a ticket, in order to save money.

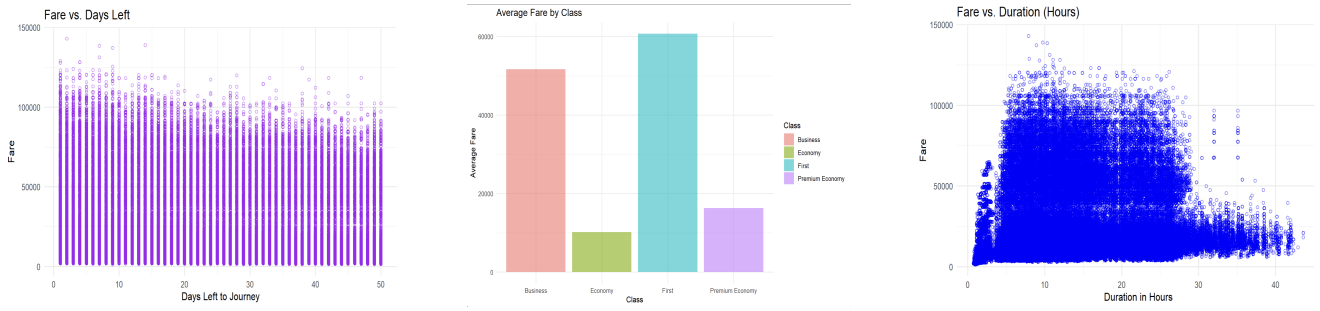
Data Presentation and Visualization

The dataset contains flight fare data collected from the top 7 busiest airports in India over 3 months in 2023. It was collected using a Python script with Beautiful Soup and Selenium libraries through web scraping. The data has been cleaned and processed to remove non-required features and additional features have been extracted to enhance the dataset's usefulness for analysis purposes. We took the dataset from Kaggle and added a column, containing the corresponding monthly average price per gallon of kerosene.

The dataset comprises of 452,088 entries across 14 columns, including features such as date of journey, day of journey, airline, flight code, class, source, departure time, total stops, arrival time, destination, duration in hours, days left before travel, fare and kerosene (added). Flight code and date of journey are irrelevant for our study so we excluded them from our model. To better understand the data, we performed an initial analysis on key variables:

- **Duration_in_hours**: The average flight duration is 12.35 hours, with a standard deviation of 7.43 hours. The shortest flight lasts 45 minutes, while the longest spans over 43.5 hours. This range suggests a wide variation in flight lengths, from short flights to long-haul journeys.
- **Days_left**: On average, flights are booked 25.63 days in advance, with a standard deviation of 14.30 days. The booking window ranges from as little as 1 day to up to 50 days.
- **Fare**: The average fare is INR 22,840.10, with a high standard deviation of INR 20,307.96. Fares range from INR 1,307 to INR 143,019, showing significant variability from budget to premium pricing. Understanding the factors driving fare differences will be crucial for accurate predictions.

We also created visualizations to explore potential trends and relationships between variables. Notably, we identified a negative correlation between the number of days in advance a flight is booked and its fare, as well as an indication that class (unsurprisingly) and flight duration may significantly impact fare prices. However, given the large dataset size, with 452,088 entries, the initial visualizations only provided a little glimpse into the significance of most variables.



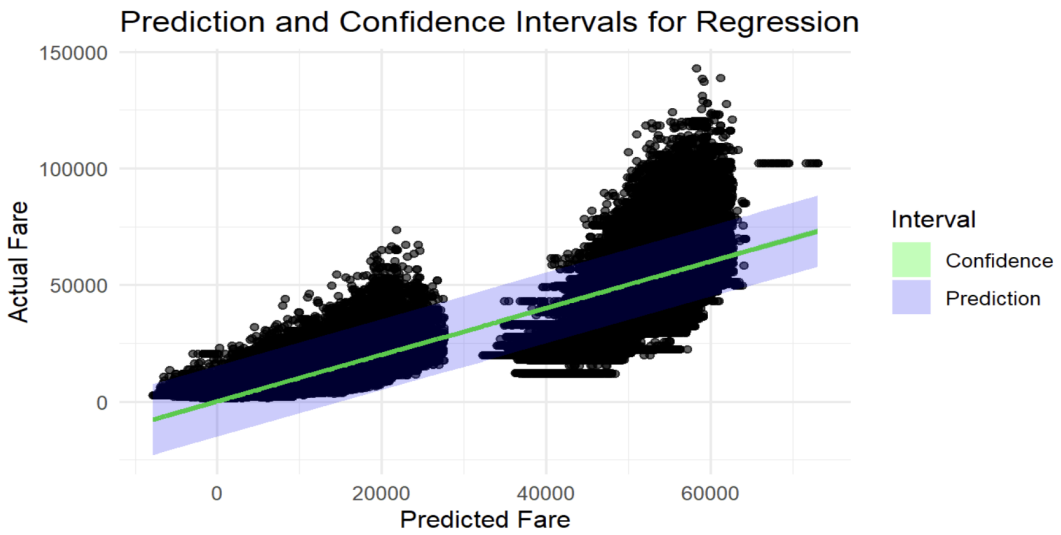
Fare prices vs. Days Left, Class, Flight Duration

Linear Regression and ANOVA Test

As seen above, our model contains a total of three numerical variables (price of kerosene, Duration_in_hours, Days_left), and eight others that we will consider as categorical variables, that are properly encoded at the beginning of the R script. Since our model's goal is to use a linear regression to determine the price of a ticket, we need to make sure our numerical variables are standardized, so that scale discrepancies do not arise when running our model. After standardizing them and running the linear regression, we took a look at the p-values for each variable, since it offers us an understanding of how relevant each factor is for determining the fare price.

Moreover, the value of R^2, also known as the coefficient of determination, highlights the fact that 85.47% of the variability of the dependent variable can be explained by the variables in the model, so we can say, based on this information, that the model provides a good fit to the data, as it explains a large proportion of the variation. Coming back to p-values, we can see that the vast majority of our variables have a small p-value (smaller than 0.05 - as per the significance level we chose), a fact that hints at their significance within the model. However, there are some - within our categorical variables - (for example Monday or Wednesday as the day of journey) that have a high p-value (0.333 and 0.06 respectively), but we hypothesise that this is due, in part, to the correlations between the days of the week, the rest of the days having a small p-value(<2e-16 for Tuesday and Sunday, 3.82e-13 for Thursday, 0.006 for Saturday, hence a high significance), and that overall - in spite of the fact that Monday/Wednesday as singular days don't have a high importance - the day of journey is significant to our model.

Indeed, by running an ANOVA test between our full initial model and a reduced model that doesn't include the day of the journey, we can conclude that this type of variable is indeed significant to our model, as indicated by the small p-value of the ANOVA-test result (< 2.2e-16). Moreover, performing the same test, but this time between the full model, and the model that is lacking the kerosene price information, we can see again that the information about the kerosene is important and significant to our linear model (the p value in ANOVA test is <2.2 e-16), thus supporting our initial decision of adding the price of kerosene per gallon to our dataset.



Model Selection with AIC and Accuracy Testing with Cross Validation

We employed the Akaike Information Criterion (AIC) in order to compare the full model (containing all variables mentioned above) with versions of reduced models that were lacking different variables, and decide whether to keep all variables in our final model. The full, most complex model had the lowest AIC score of 9,379,370, thus we decided in favor of this model since the scores of reduced models that didn't contain: day of the journey (AIC: 9,380,072), departure time (AIC: 9,379,580), number of stops (AIC: 9,422,501), or price of kerosene (AIC: 9,381,689) are higher than the one of the original model.

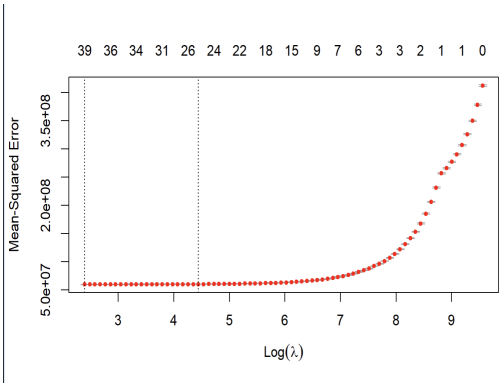
In order to get a better sense of how good our model is for predicting the price of plane tickets, we will use the cross-validation method, and divide our initial dataset into an 80% proportion for training, and 20% for testing. Considering how large our dataset is (452,088 rows), we can be assured that there is enough data in the training set to obtain a good model. The divide was made randomly (so that our model is not influenced by the order the flights were displayed in originally), by setting a seed so that the process can be replicated later on the same sets of data, if needed. We then computed the value of R^2 (0.853), as well as the root mean square error (7,794.269) - both of which suggesting that our model is significant, relevant, since the R^2 is high and the RMSE value is low with respect to the scale of the fare prices that we are trying to predict (tens of thousands).

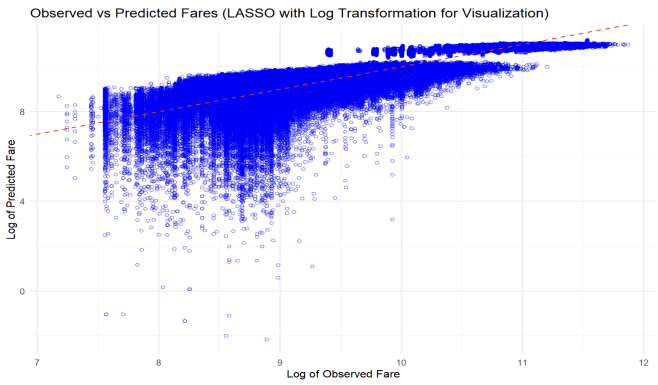
LASSO

Next, we proceeded by performing model selection using the LASSO penalty. The first step is determining the size of the hyperparameter λ. At high λ values, the model is heavily penalized, leading to high bias (underfitting), overlooking potentially valuable models and potentially high MSE. At low λ values, the model is less penalized, leading to overfitting and potentially high MSE. In both cases, the model's predictive performance on unseen data would be diminished. The optimal λ value lies in the *sweet spot* with the lowest MSE, striking a balance between bias and variance.

To select the optimal λ, we conducted 10-fold cross-validation on the training set with MSE loss, using the `cv.glmnet()` function from the glmnet library in R. Using 10 folds provides a good balance between computational efficiency and reliable model evaluation for the size of our dataset. It also aligns with the computational power available in a student setting. Then, we used the computed optimal value to select a model using LASSO penalization.

When visualizing the predicted vs. expected values, due to the large size of our dataset (452,088 entries) the initial scatter plot showed the predictions spread into two distinct "chunks". To improve the model's interpretability, we applied a log transformation to the Fare variable, resulting in a smoother distribution of data points around the line. The LASSO model, however, was trained using the original Fare values to directly evaluate the model's performance on the actual data. It's important to note that this transformation did not affect the model's predictions or fit. The log transformation was used only for visualization purposes to compress the wide range of fares (INR 1,307 to INR 143,019) and make it easier to interpret the relationship between observed and predicted values.





The Mean Squared Error (MSE) of 59,947,725 indicates that the model's predictions have a significant average squared error, which is expected given the wide range of fares (INR 1,307 to INR 143,019). To provide a more interpretable measure, the Root Mean Squared Error (RMSE) is calculated as approxi-

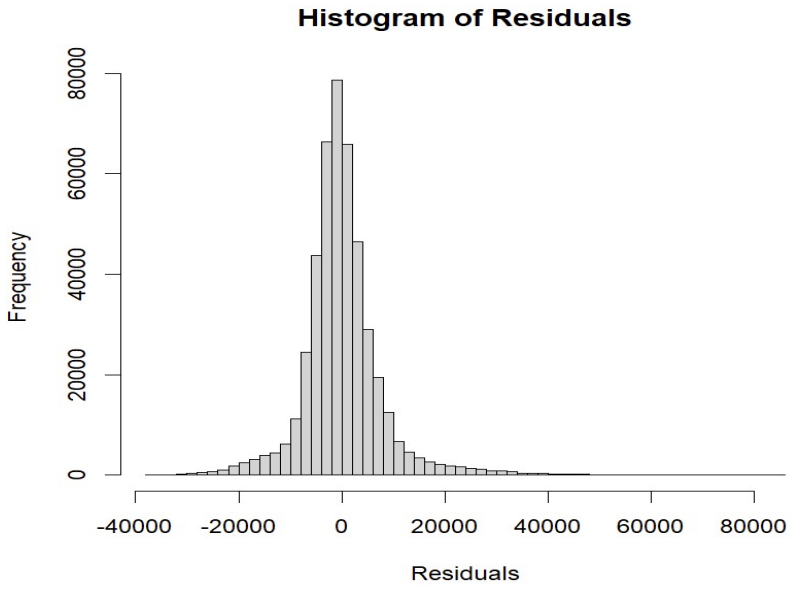
mately INR 7,742.59, meaning that the model's predictions are off by about INR 7,742 on average, in the same unit as the fares. This level of error is normal given the large variety in fare prices across budget and premium options but also potentially indicates the choice of a sub-optimal lambda in the model.

The most significant variables in the model (as intuition confirms) are the `ClassEconomy` and `ClassPremium Economy`, with coefficients of -19,914.00 and -13,695.50, indicating a decrease in fare for these classes compared to the baseline (Business Class). `Total_stopsnon-stop` has a negative impact of -2,839.34, suggesting that non-stop flights tend to be cheaper. `Days_left` also has a significant negative coefficient of -1,350.31, meaning that fares decrease as the booking date is further away.

On the other hand, variables like `AirlineVistara` (+1,927.15) and `DestinationKolkata` (+1,516.03) indicate a slight increase in fares for specific airlines and destinations. `SourceKolkata` (+1,208.91) and `kerosen` (+1,030.64) also show positive relationships with fare prices, albeit less significant than class and stop factors. Less significant variables such as `SourceChennai` (+119.53) or `Journey_dayMonday` (-3.01) have smaller coefficients, suggesting a minor impact on fare price relative to the key factors mentioned above.

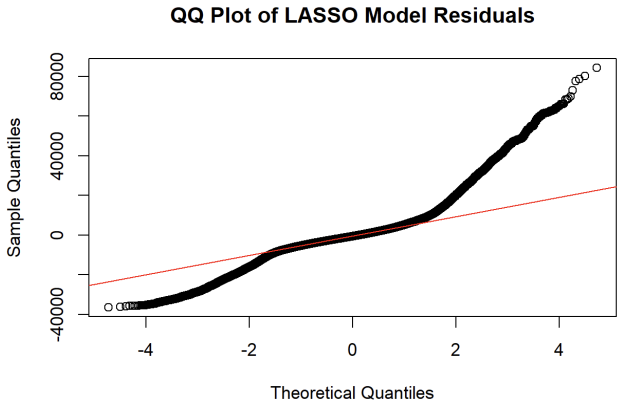
Conclusions and Limitations of the Study

In order to determine whether our linear regression model is indeed statistically relevant, we need to check whether it satisfies the assumptions on the residuals: normality, homoscedasticity, and the mean should be 0. When checking the normality of the residuals from the full model, we plotted the histogram of residuals (see figure below), that is resembling a bell curve, hinting at normality, and has a mean of 0 - both of which are in line with the assumptions that need to be satisfied. However, when we tested for normality of the residuals with the Kolmogorov-Smirnov test, we obtained a p-value smaller than 0.05 (our significance level), which means we reject the null hypothesis (i.e., we reject the hypothesis that the residuals are normally distributed). In hopes of remedying this situation, we took the logarithm of the Fare variable (the one we are trying to predict), in order to reduce its skewness, making it more normally distributed, and we ran again the Kolmogorov-Smirnov test for normality of the residuals of the "new" model (that predicts the log of Fare). Unfortunately, we obtained the same result (p-value <2.2e-16), that the residuals are not normally distributed, and we hypothesized that it might be due to how large our data set is and the existence of outliers. Indeed, when analyzing the distribution of residuals, we can see that it has quite long tails on both sides (see below), which indicate that some predictions deviate significantly from the actual values.



When examining the residuals from the LASSO model, the QQ plot shows (see below) a slight deviation from the normal line as well, in both tails (more in the right). This suggests that the residuals are not perfectly normally distributed, and there may be heavier tails than expected for a normal distribution. The slight asymmetry in the residuals could be attributed to extreme values or outliers in the dataset,

particularly given the large variation in fares (ranging from INR 1,307 to INR 143,019). This departure from normality is consistent with the nature of the LASSO model, which may result in a higher degree of regularization, potentially shrinking certain coefficients and causing a slight shift in the residual distribution. Despite this, the model still performs reasonably well. The R-squared value of 0.8546 indicates that approximately 85.46% of the variability in the fare prices is explained by the model, which suggests a strong fit. This means that the model does a good job of capturing the factors influencing fare prices, though there is still some unexplained variance. While LASSO offers a more regularized model, further refinement of the lambda parameter might help improve normality in the residuals and provide a more balanced model fit. We tried employing 10-fold cross-validation which, in this particular case, might lead to a lambda that is sub-optimal. Even though LASSO performs regularization, it does not inherently correct for non-linearity in the relationship between the predictors and the target variable. The model may still struggle with capturing more complex relationships, leading to residuals that don't follow a normal distribution.



Another potential area for further investigation is determining whether a more optimal lambda value can be identified by employing a more refined approach, such as repeating cross-validation multiple times and averaging the results, or by testing a range of lambda values across multiple model selections and comparing the outcomes. A notable limitation in our study is that the fares were obtained over the period of just 3 months which does not capture the seasonal variations fully and it is limited to a very specific time period. Exploring the impact of seasonality on ticket prices, including trends around holidays or peak travel periods, could also enhance the model's accuracy by accounting for temporal variations in fare patterns as an extension of our study. Another aspect to bring into the picture is the market dynamics - the model doesn't account for competition effects in any way, also it has limited consideration of special events or peak seasons which could be heavily influential over flight fares. Additionally, a very promising research topic for the future is analysing the interaction between variables, specifically how the interaction between different factors impacts the fare prediction.

To conclude, although the large size of our dataset prompted the existence of outliers, our model performed quite well at predicting the price of airplane tickets, as per the results of our cross-validation test. The type of class, as well as the total number of stops and the number of days leading up to the flight appeared to be the most significant variables, determining the ticket fare. Moreover, the Akaike information criterion revealed that the full model, containing all variables mentioned in the beginning, performs better than any of the reduced models (missing one variable), thus hinting at the relevance of each individual variable.

References

- An Introduction to Mathematical Statistics, Bijma et al., Bocconi teaching materials
- https://bookdown.org/ndphillips/YaRrr/comparing-regression-models-with-anova.html
- https://www.kaggle.com/datasets/yashdharme36/airfare-ml-predicting-flight-fares

Appendix

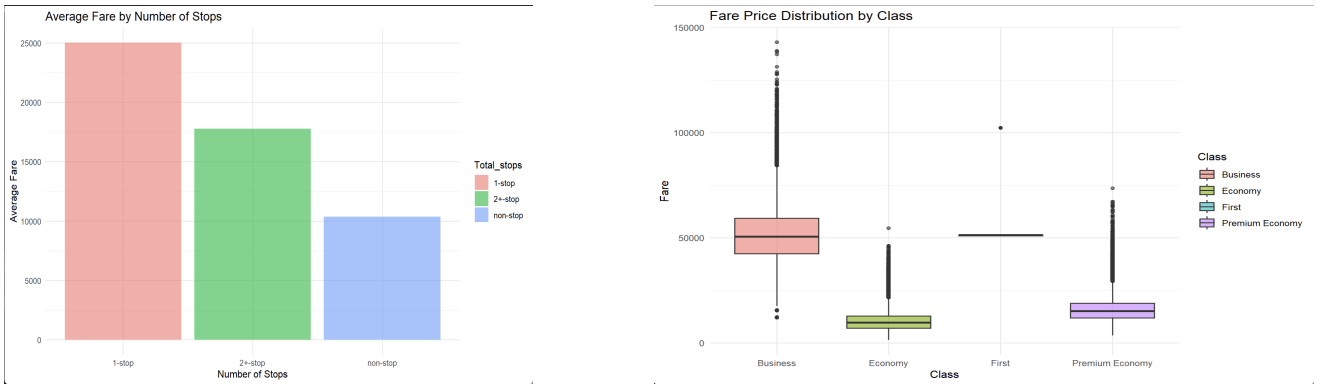


Figure 1: Some more on initial visualizations

```
call:
lm(formula = Fare ~ kerosen + Journey_day + Airline + Class +
    Source + Departure + Total_stops + Arrival + Destination +
    Duration_in_hours + Days_left, data = airline_fares)

Residuals:
   Min     1Q Median     3Q    Max
-36303  -3666   -605   2919  84690

Coefficients:
                          Estimate Std. Error   t value Pr(>|t|)
(Intercept)              48232.69      74.68   645.842  < 2e-16 ***
kerosen                   1038.59      21.53    48.238  < 2e-16 ***
Journey_dayMonday          -40.84      42.20    -0.968  0.33317
Journey_daySaturday        119.46      43.59     2.741  0.00613 **
Journey_daySunday          532.08      43.72    12.170  < 2e-16 ***
Journey_dayThursday       -314.32      43.28    -7.262 3.82e-13 ***
Journey_dayTuesday        -534.42      43.61   -12.253  < 2e-16 ***
Journey_dayWednesday        80.95      43.45     1.863  0.06247 .
AirlineAirAsia           -2022.65      68.56   -29.501  < 2e-16 ***
AirlineAkasaAir           2373.16     170.68    13.904  < 2e-16 ***
AirlineAllianceAir        1736.91     392.91     4.421 9.85e-06 ***
AirlineGO FIRST            609.36      80.45     7.575 3.61e-14 ***
AirlineIndigo              271.10      44.30     6.119 9.41e-10 ***
AirlineSpiceJet            636.62     105.02     6.062 1.34e-09 ***
AirlineStarAir            5497.10     984.34     5.585 2.34e-08 ***
AirlineVistara            3935.82      29.99   131.243  < 2e-16 ***
ClassEconomy            -40160.99      29.91 -1342.552  < 2e-16 ***
ClassFirst               19014.04     646.98    29.389  < 2e-16 ***
ClassPremium Economy    -37290.90      38.66  -964.552  < 2e-16 ***
SourceBangalore             717.37      47.43    15.125  < 2e-16 ***
SourceChennai               550.96      50.15    10.986  < 2e-16 ***
SourceDelhi               -1461.57      46.61   -31.358  < 2e-16 ***
SourceHyderabad            -370.08      49.75    -7.438 1.02e-13 ***
SourceKolkata              3845.27      49.97    76.949  < 2e-16 ***
SourceMumbai                629.31      46.81    13.443  < 2e-16 ***
Departure6 AM - 12 PM       -38.02      29.83    -1.275  0.20240
DepartureAfter 6 PM         213.54      32.70     6.529 6.61e-11 ***
DepartureBefore 6 AM       -586.88      57.20   -10.259  < 2e-16 ***
Total_stops2+-stop         2404.71      48.60    49.483  < 2e-16 ***
Total_stopsnon-stop       -9014.01      43.35  -207.923  < 2e-16 ***
Arrival6 AM - 12 PM       -1760.92      34.17   -51.530  < 2e-16 ***
ArrivalAfter 6 PM            -96.84     30.48    -3.177  0.00149 **
ArrivalBefore 6 AM        -2008.29      51.87   -38.714  < 2e-16 ***
DestinationBangalore       2444.76      47.67    51.290  < 2e-16 ***
DestinationChennai         2856.82      49.65    57.542  < 2e-16 ***
DestinationDelhi            476.22      47.53    10.019  < 2e-16 ***
DestinationHyderabad       1183.11      48.81    24.238  < 2e-16 ***
DestinationKolkata         4968.71      50.07    99.242  < 2e-16 ***
DestinationMumbai          2549.92      46.51    54.828  < 2e-16 ***
Duration_in_hours          -186.39      15.41   -12.092  < 2e-16 ***
Days_left                 -1356.82      21.47   -63.183  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7742 on 452047 degrees of freedom
Multiple R-squared:  0.8547,    Adjusted R-squared:  0.8547
F-statistic: 6.647e+04 on 40 and 452047 DF,  p-value: < 2.2e-16
```

Figure 2: Summary of linear regression

```
Analysis of Variance Table

Model 1: Fare ~ kerosen + Airline + Class + Source + Departure + Total_stops +
    Arrival + Destination + Duration_in_hours + Days_left
Model 2: Fare ~ kerosen + Journey_day + Airline + Class + Source + Departure +
    Total_stops + Arrival + Destination + Duration_in_hours +
    Days_left
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1 452053 2.7137e+13
2 452047 2.7094e+13  6 4.2842e+10 119.13 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3: ANOVA test between full model and reduced model (without journey day)

```
Analysis of Variance Table

Model 1: Fare ~ Journey_day + Airline + Class + Source + Departure + Total_stops +
    Arrival + Destination + Duration_in_hours + Days_left
Model 2: Fare ~ kerosen + Journey_day + Airline + Class + Source + Departure +
    Total_stops + Arrival + Destination + Duration_in_hours +
    Days_left
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1 452048 2.7234e+13
2 452047 2.7094e+13  1 1.3947e+11 2326.9 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 4: ANOVA test between full model and reduced model (without price of kerosene)

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  residuals_full
D = 0.5505, p-value < 2.2e-16
alternative hypothesis: two-sided
```

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  residuals_log_jitter
D = 0.27207, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Figure 5: Kolmogorov-Smirnov test on residuals of the full model and for residuals of the logarithmic full model