American University of Sharjah
1997
الجامعة الأمريكيّة في الشارقة

STA 301

Final Project Report

Name	ID	Sections completed by each member
Baraa Abed	b00088000	1, 2
Mohamed Elmohandes	b00083108	3, 4, 5
Jawad Zaabalawi	b00086528	3, 4, 5

Instructor: Dr. Hana Sulieman

Submission Date: May 16th, 2024

Abstract

In this project, we statistically analyze the AirFare ML dataset to predict the ticket fare for various flights in India. We first processed the dataset to simplify the analysis, then looked at some explanatory and inferential statistics of the data. This includes mean and variance tests, correlation test, chi-square tests, and ANOVA tests. Following that, we built multiple linear regression models to predict the fare. This includes simple linear regression models, where the best adjusted R-squared was found to be 15.89%. We also worked on multiple linear regression models, where we added interaction and polynomial terms and tested their significance. The best adjusted R-squared for these models was found to be 86.67%. Finally, we also used a binary logistic regression model and an ordinal logistic regression model to predict the class of the ticket. The best accuracy for the binary regression model was found to be 90.1% while for the ordinal model, the best accuracy was found to be 63.8%.

Section 1: Data Set Information

The AirFare ML dataset [1] contains the flight fare data of 9 different Indian airlines from the top 7 busiest airports in India during the time period of January to March 2023. The 9 airlines are Vistara, Air India, Indigo, AirAsia, GO FIRST, SpiceJet, AkasaAir, AllianceAir, and StarAir. The data was collected from the EaseMyTrip website using web scraping techniques, with the goal of providing information that will aid people in making decisions on where and when to purchase flight tickets. The dataset consists of 452,088 samples with 13 features each. These features are summarized in Table 1.

Table 1 - Summary of the Airfare ML dataset features

Feature	Type	Values
Date_of_journey	datetime	2023-01-16 – 2023-03-06
Journey_day	categorical	Sunday – Saturday
Airline	categorical	Vistara, Air India, … (9 airlines)
Flight_code	categorical	UK-936, UK-918, … (1405 flight codes)
Class	categorical	Economy, Premium Economy, Business, First
Source	categorical	Dheli, Mumbai, … (7 cities)
Departure	categorical	Before 6 AM , 6 AM - 12 PM, 12 PM - 6 PM, After 6 PM
Total_stops	categorical	non-stop, 1-stop, 2+-stop
Arrival	categorical	Before 6 AM , 6 AM - 12 PM, 12 PM - 6 PM, After 6 PM
Destination	categorical	Dheli, Mumbai, … (7 cities)
Duration_in_hours	numeric	0.750000 – 43.583300
Days_left	numeric	1 – 50
Fare	numeric	1307 - 143019

To simplify the analysis, multiple changes were applied to the data. First, the "Flight_code" feature was dropped to avoid overfitting the model. Next, the "Date_of_journey" feature was aggregated into 4 periods: Jan_16-31, Feb_1-15, Feb_16-29, and Mar_1-16. This makes it easier to gain a more general outlook about which time periods affect the flight tickets' prices. Similarly, "Journey_day" was aggregated to two types: Weekday and Weekend. This simplifies the analysis and reduces the number of categories. As a result of these changes, the dataset is left with 12 features, with "period" and "day_type" replacing "Date_of_journey" and "Journey_day" respectively.

Section 2: Exploratory Data Analysis and Inferential Statistics

Subsection 2.1: *Exploratory Data Analysis: Graphs and Summary Statistics*

The dataset has 3 quantitative variables and 9 qualitative variables. The fare is considered the main response variable of this dataset. What follows is an analysis of both the numeric and categorical variables, and their relationship with the response variable.

Quantitative variables

Table 2 - Summary statistics for quantitative features

Feature	Mean	Std	CV	Min	Q1	Median	Q3	Max
Duration_in_hours	12.35	7.43	0.602	0.75	6.58	11.33	16.50	43.58
Days_left	25.63	14.30	0.558	1	13	26	38	50
Fare	22840	20308	0.889	1307	8763	13407	35587	143019

The data consists of 3 quantitative variables: “Duration_in_hours”, “Days_left”, and “Fare”. Table 2 - Summary statistics for quantitative featuresTable 2 provides the summary statistics for these three variables. From the table, it can be seen that the “Fare” variable has the highest variability out of the three variables, due to its higher coefficient of variation (CV). On the other hand, “Days_left” has the lowest CV, signifying that it has the lowest variability.

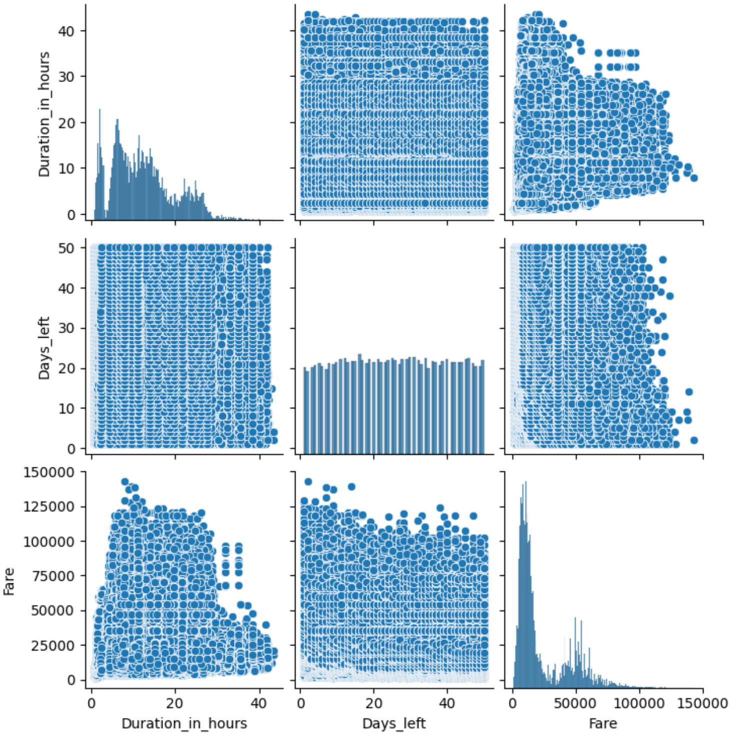


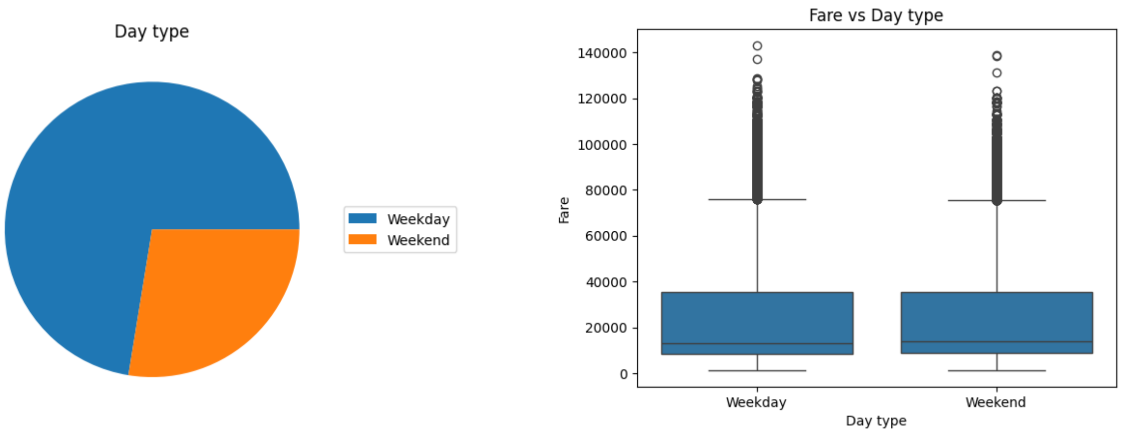
Figure 1 - Pair plot between the three quantitative variables

Figure 1 explores the possible relationships between the three qualitative variables. The diagonal plots represent the distribution of the three quantitative variables. From it we can see that the number of days left has an almost uniform distribution, while the duration of the flight as well as the fare both have right skewed distributions, with the fare having a more severe skew. Looking at the rest of the plots, it can be seen that the number of days left and the duration of the flight do not show any correlation. On the other hand, there is an apparent relationship between the fare and the duration of the flight, where after the duration of the flight increases past approximately 28 hours, the price of the ticket falls significantly. On the other hand, the fare and the number of days left show a much weaker relationship, where the fare increases slightly when the number of days left is 15 or less. To further explore these relationships, the correlation coefficient between each of these variables is explored in section 2.2.

Qualitative variables

The data has 9 qualitative variables. What follows is a closer look at the frequency of the categories for each variable as well as the relationship between each variable and the fare.

Day Type:

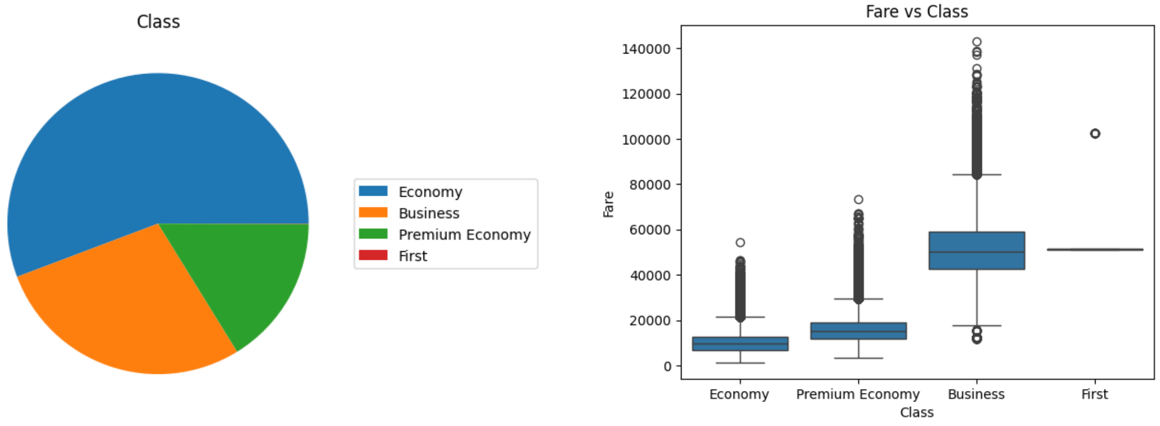


(a) (b)

Figure 2 – (a) A pie chart for the Day Type variable (b) A boxplot showing the relationship between the Fare and the Day Type

From Figure 2, it can be seen that there are more tickets recorded for flights on weekdays in general than on weekends. This makes sense since weekdays consist of five days while weekends only consist of two days. At the same time, from the boxplot, it can be seen that the average fare is about equal whether it was a weekday or a weekend. Further analysis on this will be done in the ANOVA section.

Class:

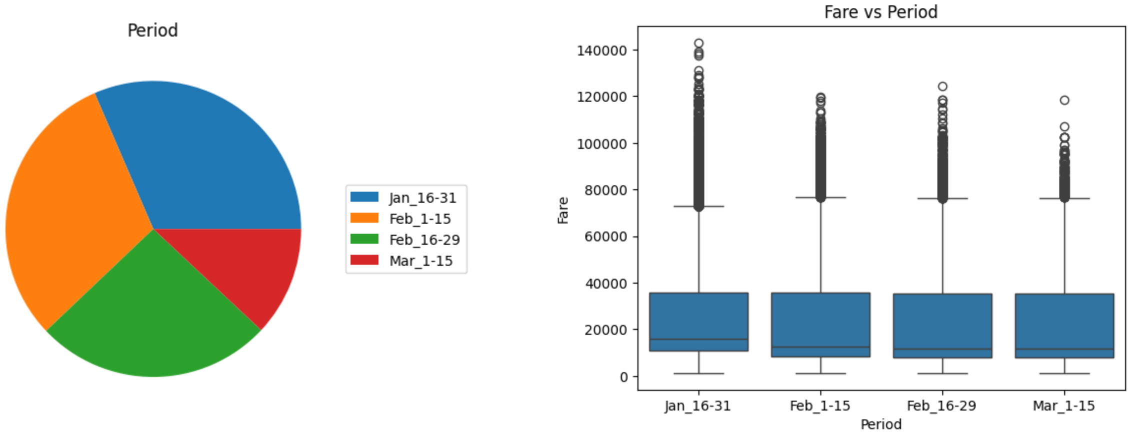


(a) (b)

Figure 3 – (a) A pie chart for the Class variable (b) A boxplot showing the relationship between the Fare and the Class variable

From Figure 3, it is apparent that most recorded tickets are economy class tickets. At the same time, it can be seen that there is a negligible number of recorded first class tickets. Furthermore, from the boxplot, the economy tickets appear to have the lowest mean fares, followed closely by the premium economy tickets. On the other hand, the business tickets have the greatest range of fares, as well as the highest mean fare, alongside the first-class tickets.

Period:



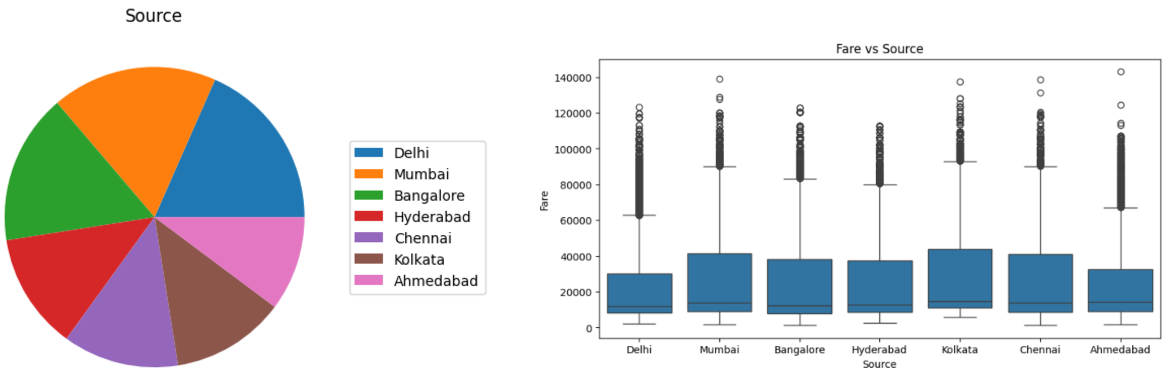
(a) (b)

Figure 4 – (a) A pie chart for the Period variable (b) A boxplot showing the relationship between the Fare and the Period

From Figure 4, it can be seen that most tickets were from January 16th to February 15th. The number of tickets from February 16th to February 29th are noticeably less than the earlier two periods. Finally, there were the least number of tickets between March 1st to March 15th. However, this can be contributed to the fact that the last data samples were recorded on the 6th of March. India's republic day happens to be on the 26th of January, which could be the reason for the period from January 16th to January 31st to have the largest number of samples recorded.

Furthermore, from the boxplot, it can be seen this period has the highest mean fare, as well as the highest range of fares, which might also be attributed to this holiday. On the other hand, the rest of the periods have roughly the same mean fares, with March 1st to March 15th having a relatively smaller range than the other two.

Source:

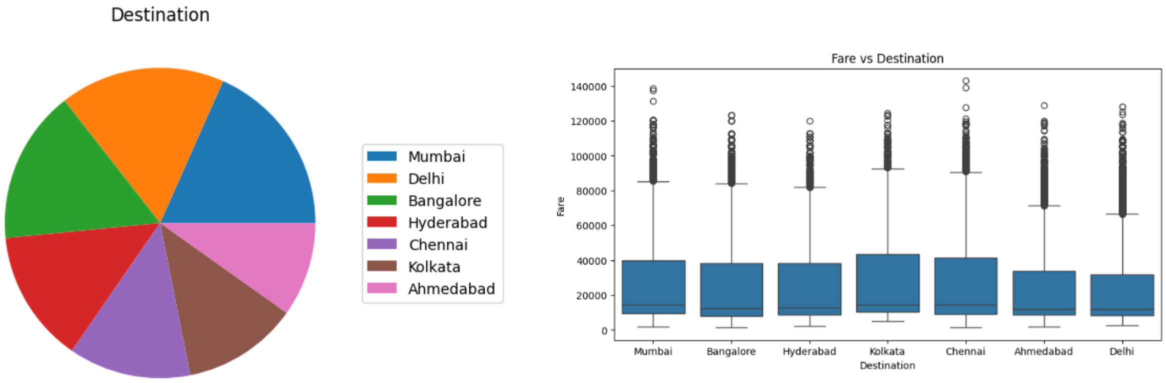


(a) (b)

Figure 5 – (a) A pie chart for the Source variable (b) A boxplot showing the relationship between the Fare and the Source

From Figure 5, it can be seen that most tickets were for flights from either Delhi, Mumbai, or Bangalore. The rest of the cities have an almost equal distribution of tickets. Looking at the boxplot, it can be seen that Kolkata has the highest mean fare out of all of the other cities.

Destination:

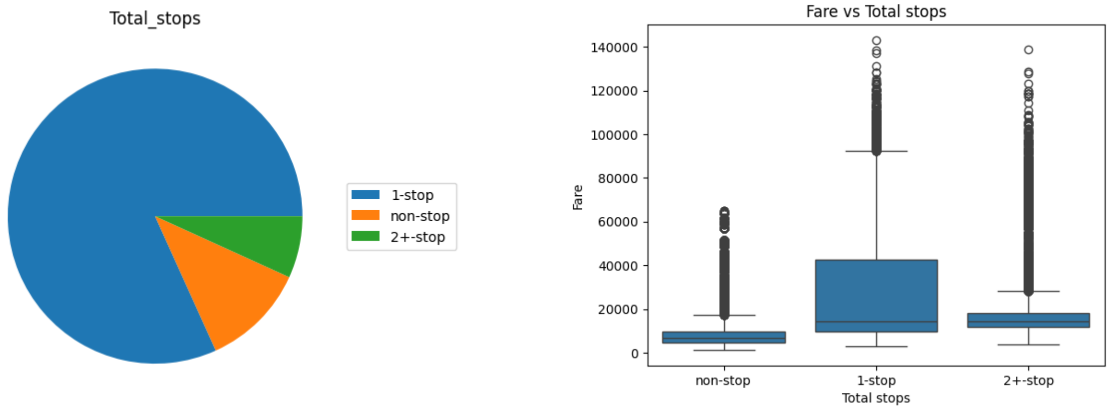


(a) (b)

Figure 6 – (a) A pie chart for the Destination variable (b) A boxplot showing the relationship between the Fare and the Destination

From Figure 6, it can be seen that most tickets were for flights to either Delhi, Mumbai, or Bangalore. The rest of the cities have an almost equal distribution of tickets. Looking at the boxplot, it can be seen that Kolkata still has the highest mean fare out of all of the other cities.

Total Stops:

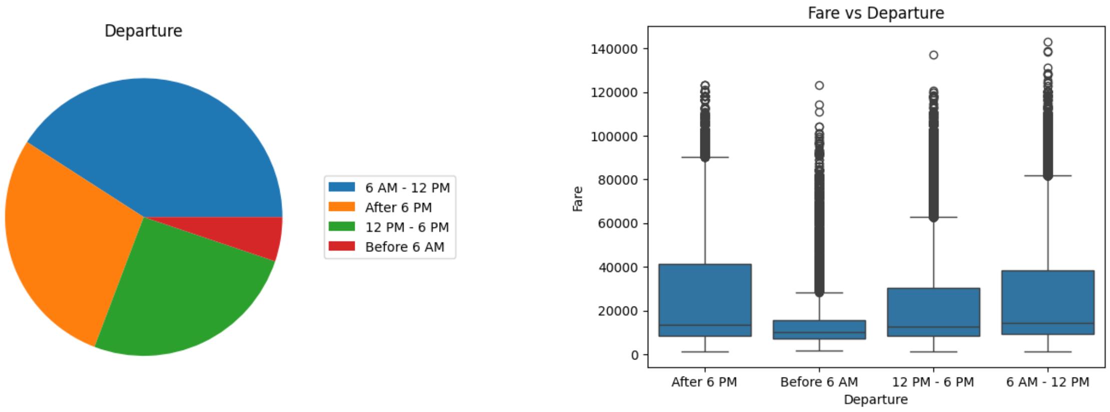


(a) (b)

Figure 7 – (a) A pie chart for the Total Stops variable (b) A boxplot showing the relationship between the Fare and the total number of stops

From Figure 7, it can be seen that most tickets were for flights that have exactly one stop. Furthermore, looking at the boxplot, it can be seen that the tickets for trips with one stop have the highest variability, even though the average fare is still very close to the trips with two or more stops. On the other hand, trips with no stops appear to be the much cheaper than the rest.

Departure:

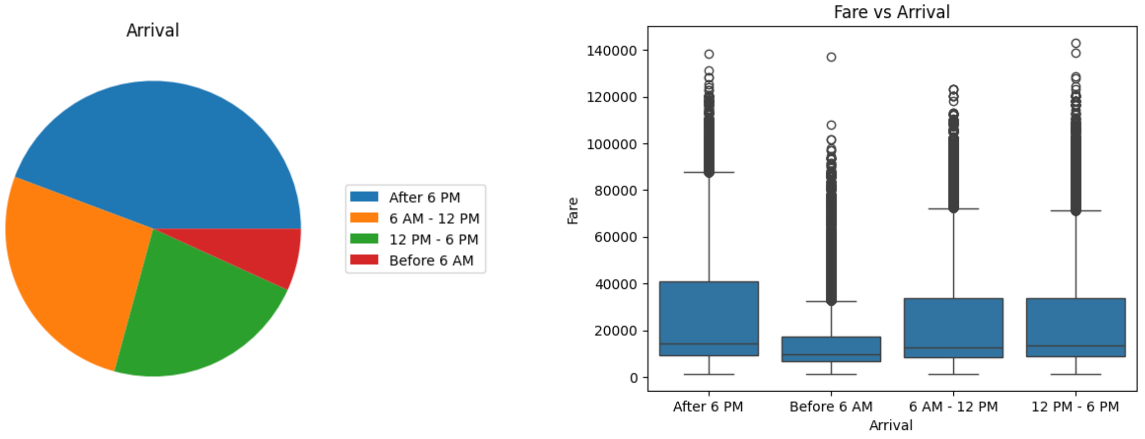


(a) (b)

Figure 8 – (a) A pie chart for the Departure variable (b) A boxplot showing the relationship between the Fare and the departure time

From Figure 8, it can be seen that most tickets were for flights that depart from 6 AM to 12 AM. On the other hand, before 6 AM, a lot less flights depart. Looking at the boxplot, it can be seen that the tickets for trips from 6 AM to 12 AM and for trips after 6 PM have the highest average prices. On the other hand, flights before 6 AM have the lowest average fare as well as the lowest variability.

Arrival:

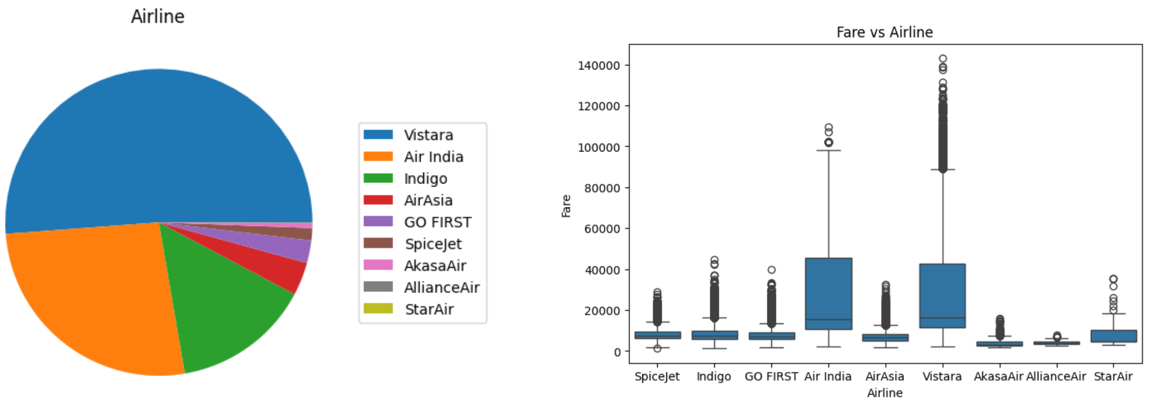


(a) (b)

Figure 9 – (a) A pie chart for the Arrival variable (b) A boxplot showing the relationship between the Fare and the arrival time

From Figure 9, it can be seen that most tickets were for flights that arrive at their destination after 6 PM. On the other hand, before 6 AM, a lot less flights arrive. Looking at the boxplot, it can be seen that the tickets for trips that arrive before 6 AM have the lowest average fare as well as the lowest variability. On the other hand, flights after 6 PM have the highest variability as well as the highest average fare.

Airline:



(a) (b)

Figure 10 – (a) A pie chart for the Airline variable (b) A boxplot showing the relationship between the Fare and various airlines

From Figure 10, it can be seen that most tickets were from the Vistara airline, followed by Air India. Furthermore, looking at the boxplot, Air India and Vistara airlines both have high variability and high mean fares, as opposed to the other airlines. In addition, Vistara airlines appears to have a much higher range of fares than Air India and the rest, which could indicate that perhaps the Business Class tickets that reached approximately 140,000 rupees from Figure 3 (b) are from this airline.

Subsection 2.2: *Inferential Statistics*

In this section, the relationship between the variables will be further explored through hypothesis tests, to gain a better understanding of possible associations between variables.

Pearson Correlation Tests

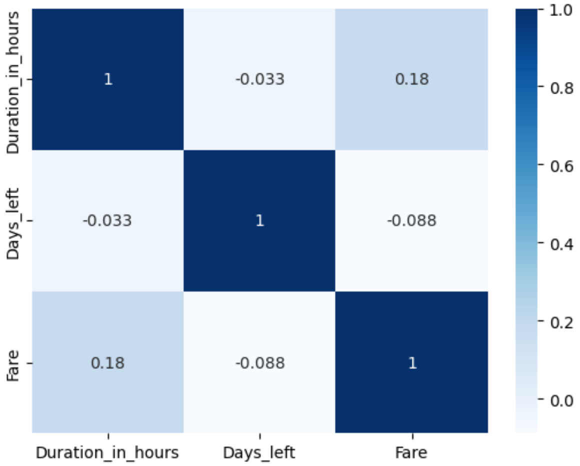


Figure 11 - Correlation matrix for the three quantitative variables

Table 3 - Results for the Pearson correlation tests

Feature 1	Feature 2	r	p-value	Confidence Interval
Duration_in_hours	Days_left	-0.033	2.39 x10^{-108}	[-0.036, -0.030]
Duration_in_hours	Fare	0.180	0.0	[0.177, 0.183]
Days_left	Fare	-0.088	0.0	[-0.091, -0.085]

In Table 3, it can be seen that the number of days left has a very week negative linear relationship with both the duration of the flight and the fare. On the other hand, the fare has a relatively stronger, but still weak, positive linear relationship with the duration of the flight. To test the significance of these correlations, the following hypothesis test is used:

$$H_0: The\ correlation\ is\ not\ significant$$

$$H_a: The\ correlation\ is\ significant$$

At a 5% significance, these correlations are significant since the confidence intervals do not include 0. The p-values are also less than 0.05, further emphasizes the significance of the relationship.

Mean and Variance Tests

The three numeric variables in our data have different units and different scales; therefore, their means should not be equal. To confirm this, a t-test is performed with the following hypotheses:

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

The results are displayed in Table 4. Note that none of the confidence intervals include 0, and the p-values for all tests were less than 0.05. From this, we can conclude that at a 5% significance level, there is evidence that the means of all three variables are significantly different.

Table 4 - Results for difference in means test

Mean 1	Mean 2	p-value	Confidence Interval
Days_left	Duration_in_hours	0.0	[13.232, 13.326]
Days_left	Fare	0.0	[-22874, -22755]
Fare	Duration_in_hours	0.0	[22769, 22887]

Next, the variances of the three variables are compared through an F-test. The hypotheses used are:

$$H_0: \frac{\sigma_1}{\sigma_2} = 1$$

$$H_a: \frac{\sigma_1}{\sigma_2} \neq 1$$

The results are shown in Table 5. As seen in the table, the confidence intervals for the three tests do not include 1, and the p-values are all less than 0.05. From this, it can be concluded that at a 5% significance level, there is evidence that the variances of the three variables are significantly different. Furthermore, from the confidence intervals, it is apparent that the fare variable has by far the largest variance, with several hundred thousands of times the variance of the other two variables. On the other hand, the variance of the days left variable is approximately 3.7 times the variance of the duration variable, indicating that days left has a higher variability.

Table 5 - Results for the variance ratio test

Variance 1	Variance 2	p-value	Confidence Interval
Days_left	Duration_in_hours	0.0	[3.6816, 3.7248]
Fare	Days_left	0.0	[7424213, 7511285]
Fare	Duration_in_hours	0.0	[2004829, 2028342]

Chi-Square Tests

To understand the relationship between the different categorical features, chi-square tests were performed for all pairs of categorical features. Figure 12 summarizes the results. The hypotheses for chi-square tests are:

$$H_0: The\ two\ features\ are\ independent$$

$$H_a: The\ two\ features\ are\ associated\ (dependent)$$

From the heatmap, it can be observed that at a 5% significance level, most categorical variables are associated, since the p-values are less than 0.05. For example, the chi-square test between class and period shows that they are strongly associated, which may indicate that in different periods, the distribution of the classes of the tickets bought are not equal. On the other hand, there are two chi-square tests that have p-values larger than 0.05, indicating that at a 5% significance level, the variables are independent. The first one is the test between the airline and the day type. The fact that they appear independent indicates that whether it is a weekday or a weekend, the distribution of airlines of the tickets are equal. The second one is the test between the class and the day type. Their independence can mean that the class of the ticket is not affected by whether the plane departs in a weekday or a weekend.

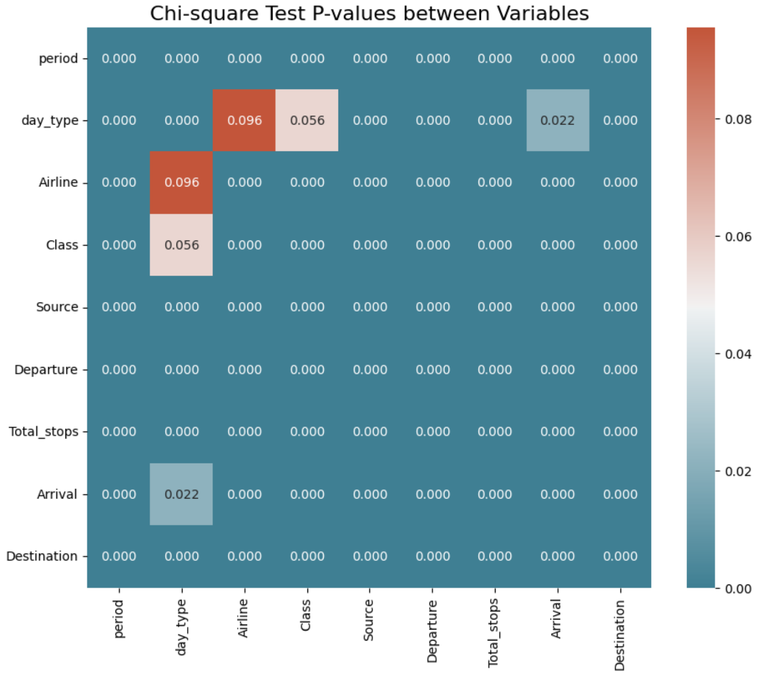


Figure 12 – A heatmap summarizing the results of the Chi-Square test

ANOVA Tests

One-way ANOVA was performed on all categorical variables, with the fare as the response. On the other hand, two-way was performed twice, one for the airline and class variable, and one between the class and the day type variable. Bonferroni's multiple comparison was used to identify which means are significantly different, since the number of samples (n) for each category is not constant, making it unsuitable for using Tukey's method.

One-way ANOVA:

Table 6 - One-way ANOVA test results

Feature	period	day type	Airline	Class	Source	Departure	Total stops	Arrival	Destination
p-value	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 6 shows the p-values from the ANOVA test for all categorical variables. As seen above, at a 5% significance level, all categorical variables have at least one group that is significantly different from the rest. Let's take the period variable as an example. To test if any of the groups are different, the following hypotheses are used:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_a: Atleast\ one\ \mu_i\ is\ significantly\ different$$

At a 5% significance level, there is evidence to reject H_0, meaning that the mean fare for at least one of the periods is significantly different from the rest. To identify which periods are different, Bonferroni's LSD was calculated, and the confidence intervals for the different periods were plotted in Figure 13. From the figure, it can be seen that other than the periods Feb_16-29 and Mar_1-15, which are very close to the 0 line, all other periods are significantly different. It appears that Jan_16-31 has the highest mean fare out of all of the periods, while Mar_16-31 generally has the lowest mean fare. This conforms to the results from the box plot in Section 2.1.

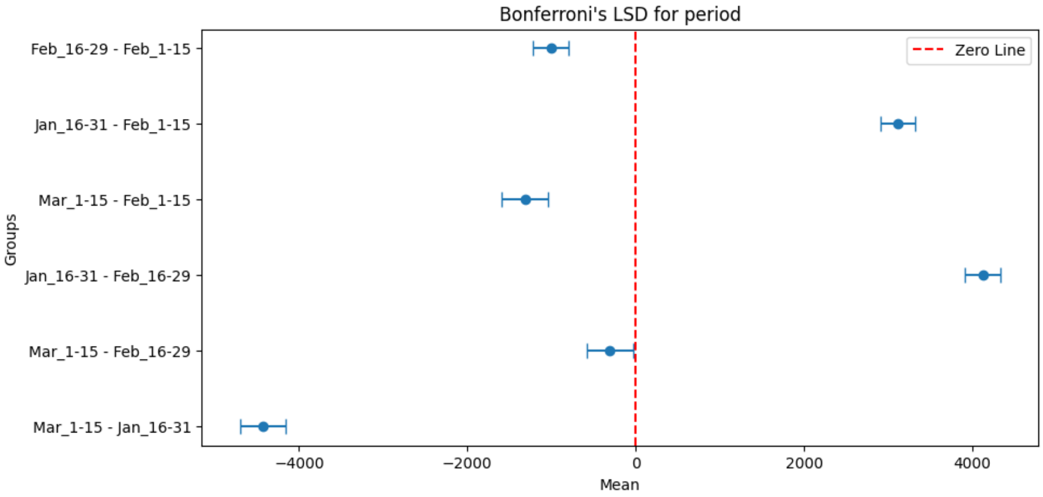


Figure 13 - Bonferroni's LSD for period

Two-way ANOVA:

Table 7 - Two-way ANOVA table for class and airline

	SS	df	F	p-value
Class	$1.49\text{x}10^{14}$	3	$5.99\text{x}10^{5}$	0.00
day_type	$1.05\text{x}10^{9}$	1	12.69	0.00
Class*day_type	$2.14\text{x}10^{10}$	3	86.22	0.00
Error	$3.75\text{x}10^{13}$	452080	-	-
Total	$1.86\text{x}10^{14}$	452087	-	-

Table 7 summarizes the results of the Two-way ANOVA test. To test for the effect of Factor A (Class) on the response, the following hypotheses are used:

$$H_0: \mu_{.1} = \mu_{.2} = \mu_{.3} = \mu_{.4}$$

$$H_a: Atleast\ one\ \mu_{.j}\ is\ significantly\ different$$

At a 5% significance level, there is evidence to reject H_0 ($p < 0.05$), meaning that at least one class has a significant effect on the response (fare). To find which means are different, Bonferroni's LSD was calculated, and the confidence intervals for each difference of mean are shown in Figure 14. As seen in the figure, all confidence intervals do not intersect with the dotted red line (0 mean), meaning that all means are significantly different. Moreover, from the diagram, it can be seen that first-class tickets have a higher mean fare than all other tickets.

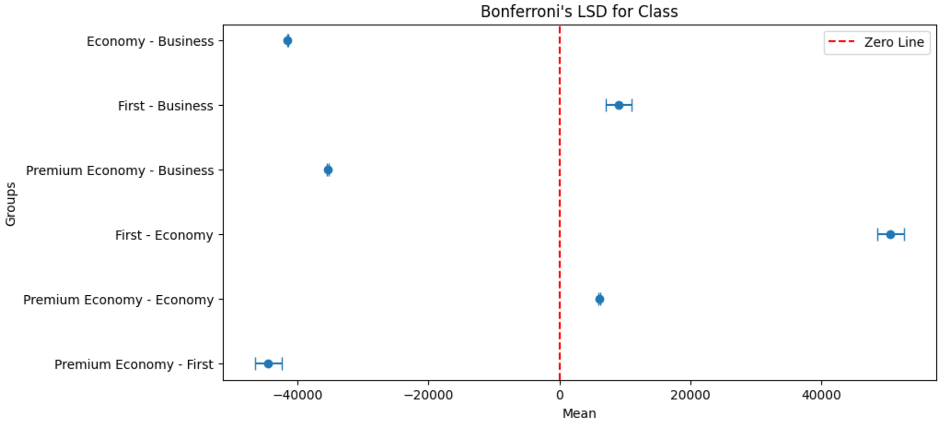


Figure 14 - Bonferroni's LSD for Factor A

To test for the effect of Factor B (day_type), the following hypotheses are used:

$$H_0: \mu_{1.} = \mu_{2.}$$

$$H_a: Atleast\ one\ \mu_{i.}\ is\ significantly\ different$$

At a 5% significance level, there is evidence to reject H_0 ($p < 0.05$), meaning that either weekdays or weekends have a significant effect on the response (fare). Bonferroni's LSD was calculated to find which means are different. As seen in the Figure 15, the confidence interval for the difference between the mean fare for weekdays and weekends does not include 0, meaning that they are significantly different. In particular, the mean fare on weekends is greater than the mean fare on weekdays.

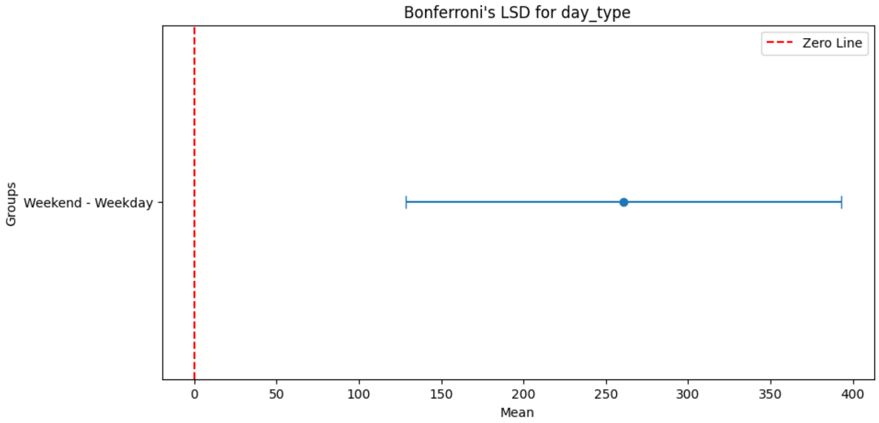


Figure 15 - Bonferroni's LSD for Factor B

To test for the effect of the interaction, the following hypotheses are used:

$$H_0: \mu_{11} = \ \mu_{12} = \cdots = \mu_{23} = \ \mu_{24}$$

$$H_a: Atleast\ one\ \mu_{ij}\ is\ significantly\ different$$

At a 5% significance level, there is evidence to reject H_0 (p < 0.05), meaning that at least one interaction term has a significant effect on the response (fare). Figure 16 shows the confidence intervals of the difference of means after calculating Bonferroni's LSD. As seen in the figure, the majority of the confidence intervals do not intersect with the dotted red line (0 mean), meaning that most of the means are significantly different.

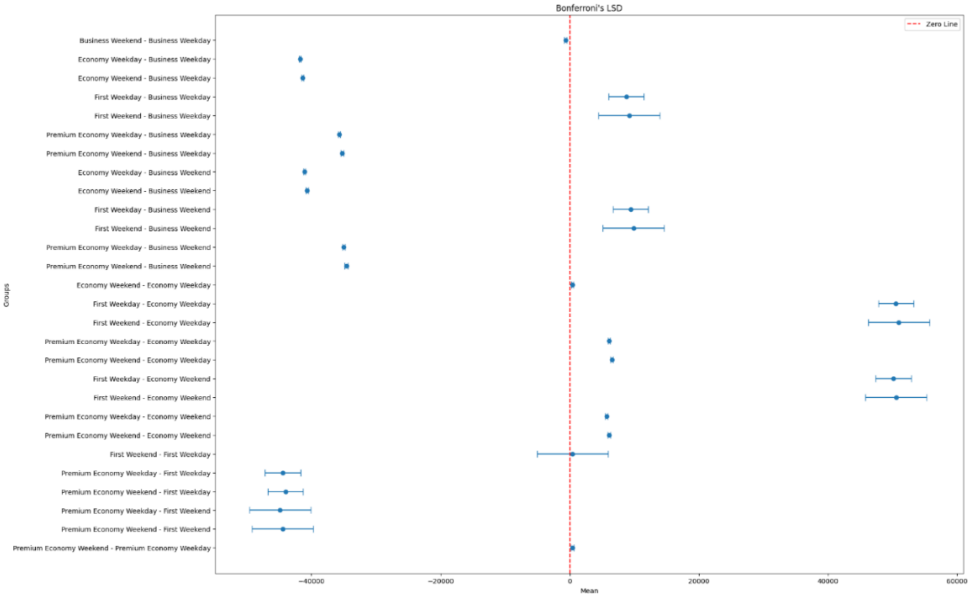


Figure 16 - Bonferroni's LSD for the interaction term

Section 3: Model Building

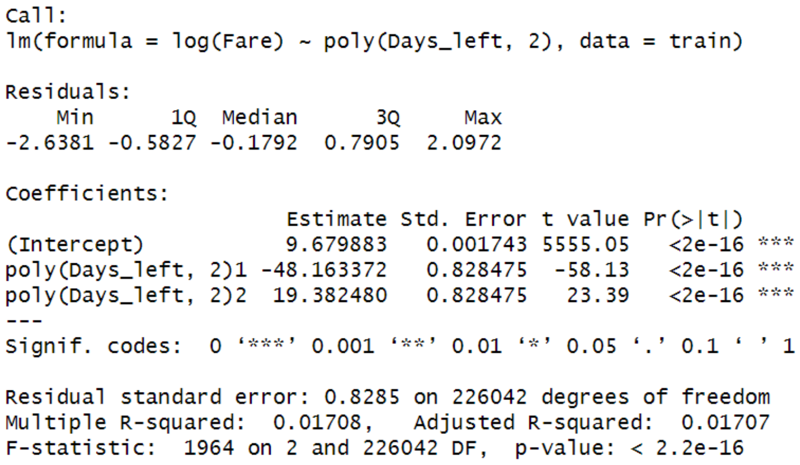
In this section, the model building process is discussed for the creation of models using four methods: simple linear regression, multiple linear regression, binary logistic regression, and ordinal logistic regression. It is worth noting that, for all four models, the same train and test data was used.

Simple Linear Regression

We tested out different models that utilize just one predictor. We tried two variables. One model had Days_left as the predictor and Fare as the response. The other model had Duration_in_hours and Fare as the response. After the implementation of the Days_left model and looking at the resid plots we have tested out different models to find out the best model to use.

1. Days_left as the predictor and Fare as the response.

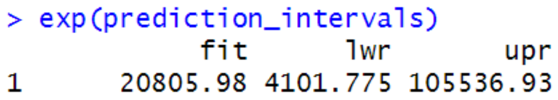
We used boxcox to find lambda which gave us -0.1818182, hence we do log transform as it is close to 0. We also tested out polynomial of degrees 2 and 3. The best model summary is shown in the figure below.



```
Call:
lm(formula = log(Fare) ~ poly(Days_left, 2), data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.6381 -0.5827 -0.1792  0.7905  2.0972 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           9.679883   0.001743 5555.05   <2e-16 ***
poly(Days_left, 2)1 -48.163372   0.828475  -58.13   <2e-16 ***
poly(Days_left, 2)2  19.382480   0.828475   23.39   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8285 on 226042 degrees of freedom
Multiple R-squared:  0.01708,	Adjusted R-squared:  0.01707 
F-statistic:  1964 on 2 and 226042 DF,  p-value: < 2.2e-16
```

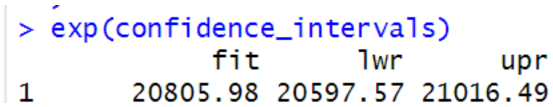
Figure 17 Best model with Days_left as predictor and Fare as response

We have also tried to look at the prediction and confidence intervals at 95% for a datapoint for the best model. The results are shown in the two figures below.



```
> exp(prediction_intervals)
       fit      lwr       upr
1 20805.98 4101.775 105536.93
```

Figure 18 PI intervals for a datapoint



```
> exp(confidence_intervals)
       fit      lwr      upr
1 20805.98 20597.57 21016.49
```

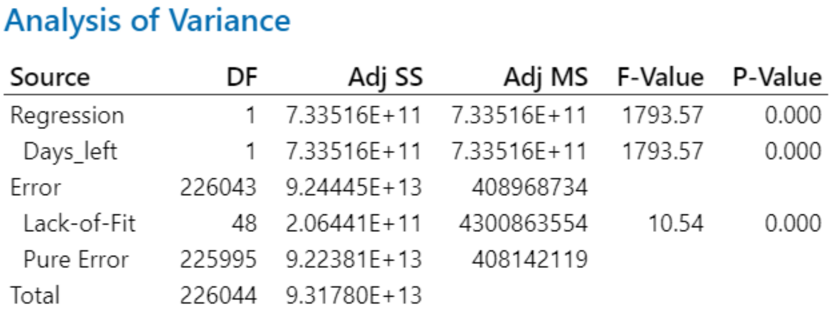
Figure 19 CI intervals for a datapoint

Finally, we did the lack of fit test using Minitab.

$$H_0: There\ is\ no\ lack\ of\ fit$$

$$H_a: There\ is\ a\ lack\ of\ fit$$

At a 5% significance level we have enough evidence to reject the null hypothesis (p<0.05). Hence, we can conclude that there is a lack of fit.



Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	7.33516E+11	7.33516E+11	1793.57	0.000
Days_left	1	7.33516E+11	7.33516E+11	1793.57	0.000
Error	226043	9.24445E+13	408968734		
Lack-of-Fit	48	2.06441E+11	4300863554	10.54	0.000
Pure Error	225995	9.22381E+13	408142119		
Total	226044	9.31780E+13			

Figure 20 Simple regression lack of fit test

The comparison results of different models are in the validation section below.

2. Model diagnostics and remedial measures.

These are the diagnostic plots before any transformations.

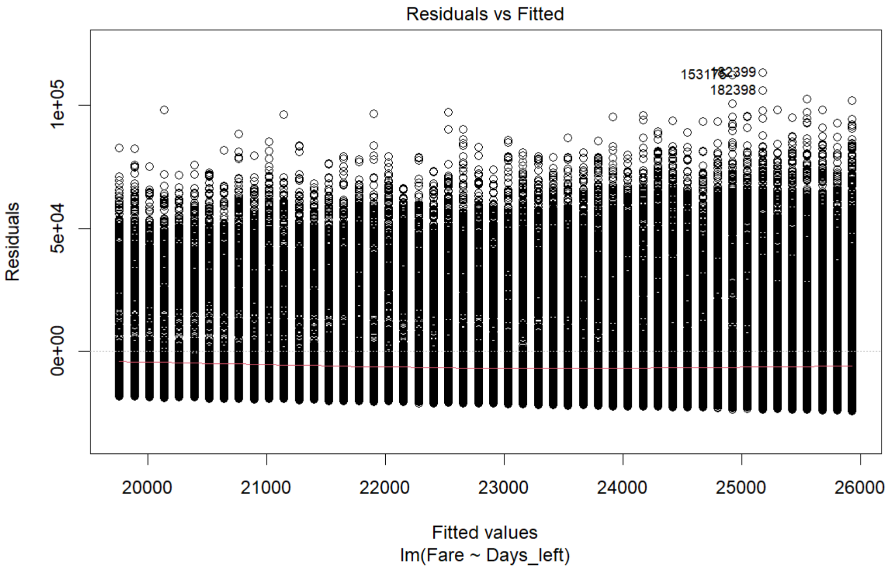


Figure 21 Simple Regression Residuals vs Fitted plot before remedy

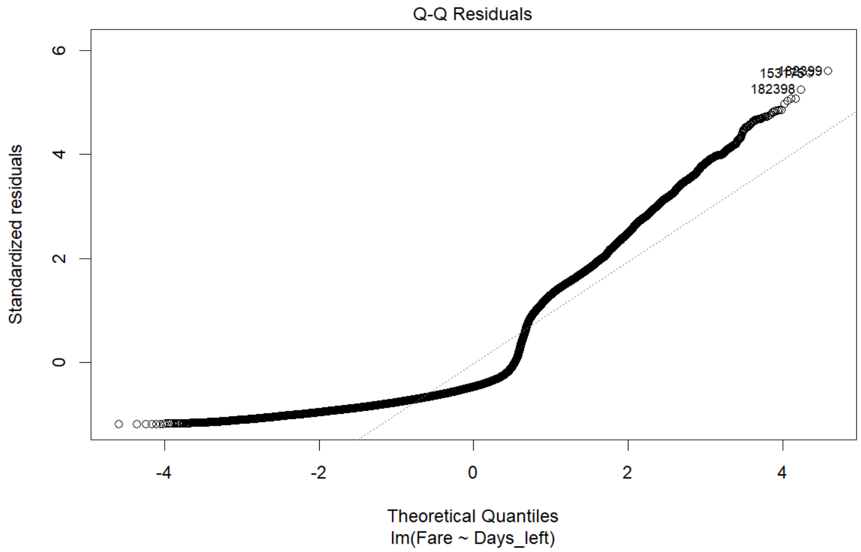


Figure 22 Simple Regression Q-Q plot before remedy

These are the diagnostic plots after the transformations and for the best model.

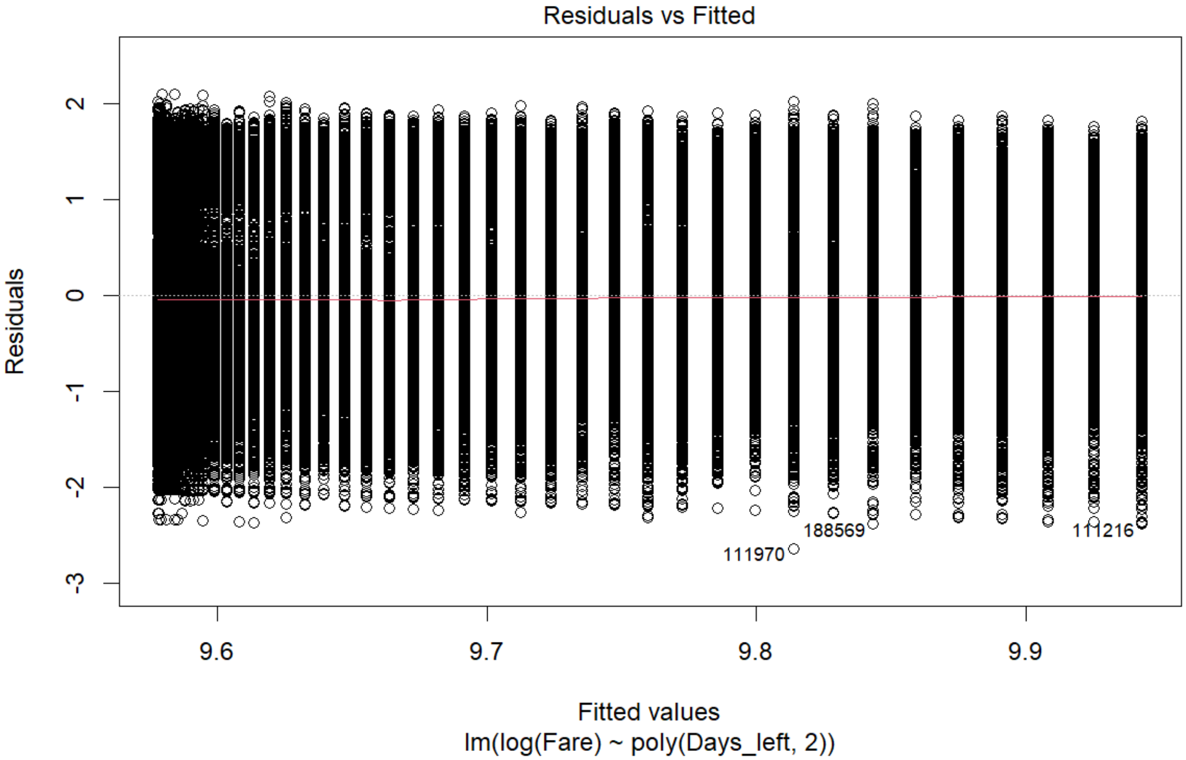


Figure 23 Simple Regression Residuals vs Fitted plot after remedy

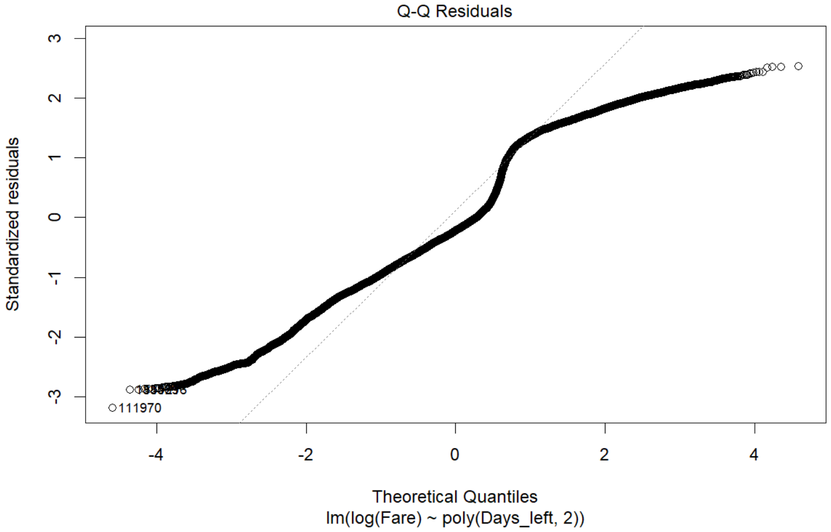
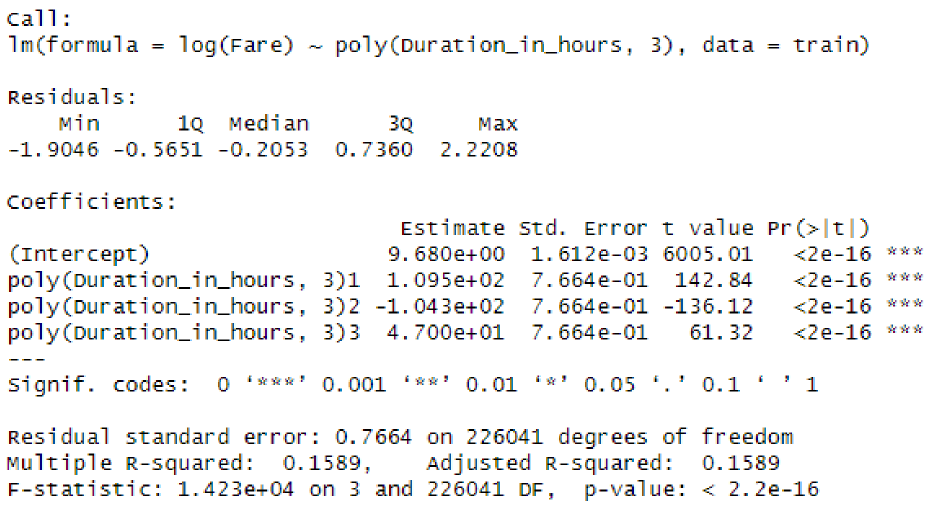


Figure 24 Simple Regression Q-Q plot after remedy

As can be seen the normality plot shows that we were able to somewhat fix the normality assumption but there are still heavy tails. If we look at the residual vs fitted plot, we can notice that initially the zero mean assumption was not satisfied but after transformations we can see that it is having zero mean. Furthermore, we can notice that there was some sort of fanning out hence not const variance but after the transformations it has const variance. We can also see that it is independent before and after transformations. Moreover, as seen in Figure 1 there is no clear relationship between Days_left and Fare.

3. Duration_in_hours as the predictor and Fare as the response.

Our main goal was to test our different predictors against the response. As requested, we implemented a simple linear regression with Days_left as the predictor. But we decided to test Duration_in_hours as a predictor. We used this predictor as it showed a high correlation with Fare. The model did yield results better than the first model with Days_left. We also realized that the model is significant when we add higher order terms of the predictor but that makes the model complicated and introduces overfitting. The best model summary is shown in the figure below.



```
Call:
lm(formula = log(Fare) ~ poly(Duration_in_hours, 3), data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.9046 -0.5651 -0.2053  0.7360  2.2208 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  9.680e+00  1.612e-03 6005.01   <2e-16 ***
poly(Duration_in_hours, 3)1  1.095e+02  7.664e-01  142.84   <2e-16 ***
poly(Duration_in_hours, 3)2 -1.043e+02  7.664e-01 -136.12   <2e-16 ***
poly(Duration_in_hours, 3)3  4.700e+01  7.664e-01   61.32   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7664 on 226041 degrees of freedom
Multiple R-squared:  0.1589,	Adjusted R-squared:  0.1589 
F-statistic: 1.423e+04 on 3 and 226041 DF,  p-value: < 2.2e-16
```

Figure 25 Best model with Duration_in_hours as predictor and Fare as response

The comparison results of different models are in Section 4.

Multiple Linear Regression

To build the best multiple linear regression to predict the quantitative response, Fare, we performed several steps: importing the data, encoding variables, sampling for train/test split, addition of interaction and polynomial terms, variable selection (including best subset, stepwise, forward, and backward selection), model diagnostics for the best model, remedial measures, and model performance analysis (for this specific step, see Section 4).

1. Importing the data and encoding variables

The first step was to import the data into R from the CSV file. The data has 452,088 observations and 12 variables. Next, for the purpose of model analysis, we dropped the Flight_code variable and we encoded the Date_of_journey and Journey_day variables as described in Section 1. The encoded Date_of_journey variable was named as “period”, and the encoded Journey_day variable was named as “day_type”. Once this was done, we encoded all of the categorical variables using one-hot encoding, which were Airline, Class, Source, Destination, Departure, Arrival, Total_stops, period, and day_type.

2. Sampling for train/test split

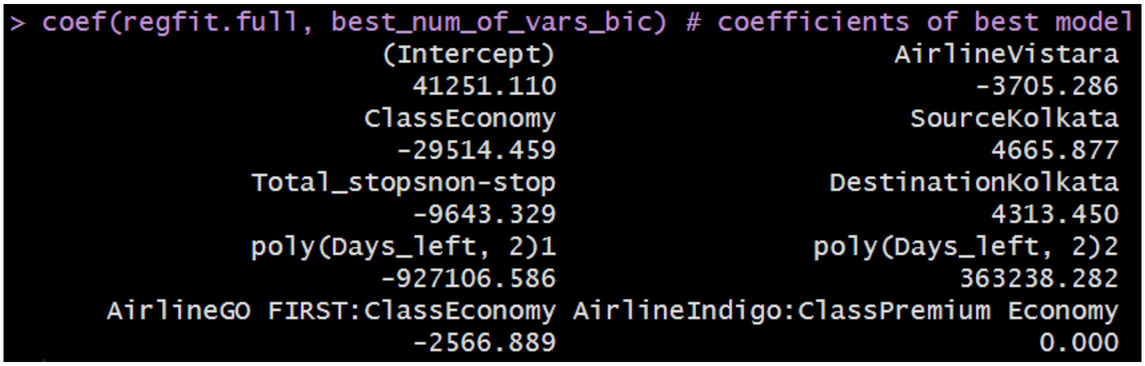
Once the variables were encoded and the data was ready for analysis, we first performed sampling to split the data into train and test data. We chose to split the data into 50% train and 50% test. The justification for this is that the dataset is large, meaning that 50% train data would have about 225,000 observations, which is sufficient for training. This would leave 225,000 observations for testing, which is essential for our analysis of the performance of the model. It is also worth noting that once the whole model building and validation process was done, we tried it again with an 80% train and 20% test split, but this split resulted in slightly lower adjusted R-squared and much higher BIC, therefore we stayed with the 50% train/test split.

3. Variable selection with interaction and polynomial terms

Once the train and test data were prepared, we performed variable selection using best subset, stepwise, forward, and backward selection. It is worth noting that for all four methods, we included two interaction terms, being Airline*Class and day_type*Class, as well as two quadratic terms, being Days_left^2 and Duration_in_hours^2. It is also worth noting that, due to the large number of dummy variables and computational limitations, we limited the size of the model in variable selection to a maximum of 9 variables, including dummy variables. The table below displays a summary of the results of variable selection using the four methods, and the figures following display the coefficients and variables of each method's best model, where the best model is determined according to the lowest BIC value. The final results show that, if we consider a dummy variable included as the whole variable included, then all four methods agree on the same best variables to include, which are Airline, Class, Source, Total_stops, Destination, Days_left, Days_left^2, and Airline*Class.

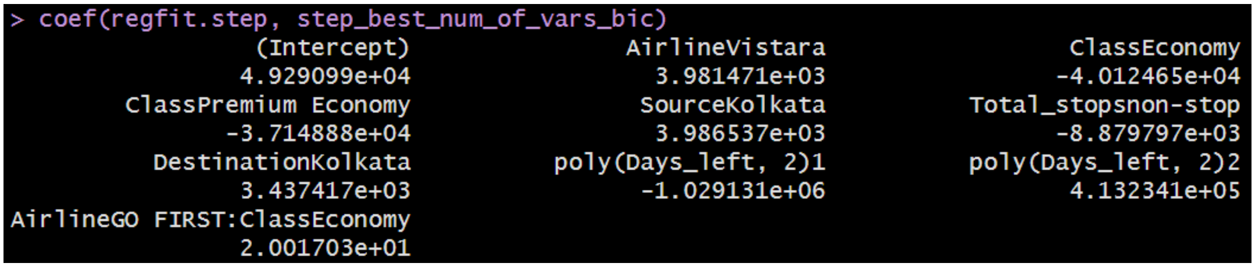
Table 8: Variable Selection Results

Method	No. of Variables	R^2	Adj. R^2	Cp	BIC
Best Subset	9	0.8533786	0.8533728	13311.82	-433860.9
Stepwise	9	0.8533786	0.8533728	13311.82	-433860.9
Forward	9	0.8533786	0.8533728	13311.82	-433860.9
Backward	9	0.8533786	0.8533728	13311.82	-433860.9



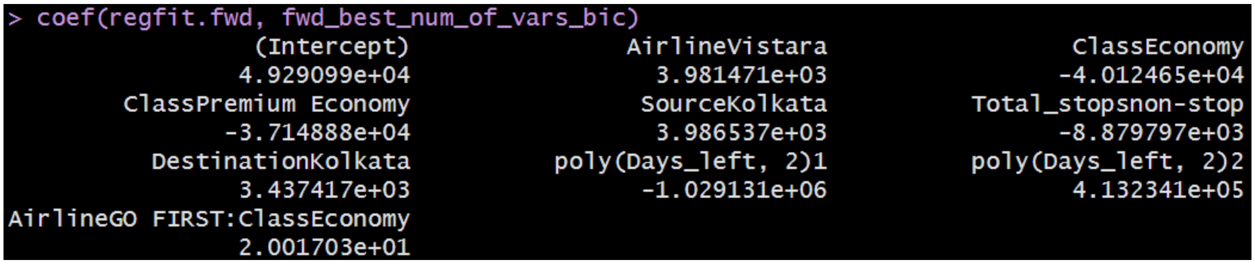
```
> coef(regfit.full, best_num_of_vars_bic) # coefficients of best model
                      (Intercept)                     AirlineVistara
                        41251.110                          -3705.286
                     ClassEconomy                       SourceKolkata
                       -29514.459                           4665.877
             Total_stopsnon-stop                   DestinationKolkata
                        -9643.329                           4313.450
              poly(Days_left, 2)1                  poly(Days_left, 2)2
                      -927106.586                         363238.282
        AirlineGO FIRST:ClassEconomy AirlineIndigo:ClassPremium Economy
                        -2566.889                              0.000
```

Figure 26: Best Subsets' Best Model's Variables and Coefficients



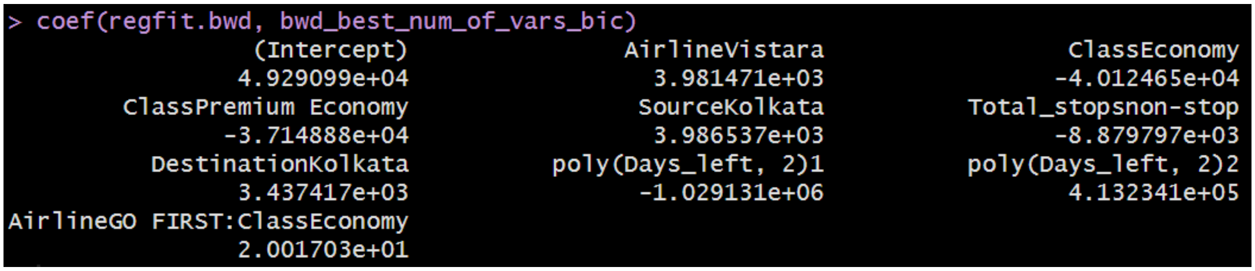
```
> coef(regfit.step, step_best_num_of_vars_bic)
                 (Intercept)              AirlineVistara                ClassEconomy
                4.929099e+04                3.981471e+03               -4.012465e+04
       ClassPremium Economy               SourceKolkata         Total_stopsnon-stop
               -3.714888e+04                3.986537e+03               -8.879797e+03
          DestinationKolkata         poly(Days_left, 2)1         poly(Days_left, 2)2
                3.437417e+03               -1.029131e+06                4.132341e+05
AirlineGO FIRST:ClassEconomy
                2.001703e+01
```

Figure 27: Stepwise's Best Model's Variables and Coefficients



```
> coef(regfit.fwd, fwd_best_num_of_vars_bic)
                (Intercept)            AirlineVistara              ClassEconomy
               4.929099e+04              3.981471e+03             -4.012465e+04
      ClassPremium Economy            SourceKolkata      Total_stopsnon-stop
              -3.714888e+04              3.986537e+03             -8.879797e+03
         DestinationKolkata      poly(Days_left, 2)1      poly(Days_left, 2)2
               3.437417e+03             -1.029131e+06              4.132341e+05
AirlineGO FIRST:ClassEconomy
               2.001703e+01
```

Figure 28: Forward's Best Model's Variables and Coefficients



```
> coef(regfit.bwd, bwd_best_num_of_vars_bic)
                (Intercept)            AirlineVistara              ClassEconomy
               4.929099e+04              3.981471e+03             -4.012465e+04
      ClassPremium Economy            SourceKolkata      Total_stopsnon-stop
              -3.714888e+04              3.986537e+03             -8.879797e+03
         DestinationKolkata      poly(Days_left, 2)1      poly(Days_left, 2)2
               3.437417e+03             -1.029131e+06              4.132341e+05
AirlineGO FIRST:ClassEconomy
               2.001703e+01
```

Figure 29: Backward's Best Model's Variables and Coefficients

4. Model diagnostics and remedial measures

Once we identified the best model, i.e. the best variables to include, we built this model and used diagnostic plots (Q-Q plot and residuals vs fitted plot) to assess the validity of the regression assumptions. The figures below display the two plots. As seen from the Q-Q plot, the residuals are not normal, so normality is not satisfied. As for the residuals vs fitted plot, we can clearly see that the residual variance is not constant. We can also deduce that independence is not satisfied, zero mean is not satisfied, and linearity is not satisfied.

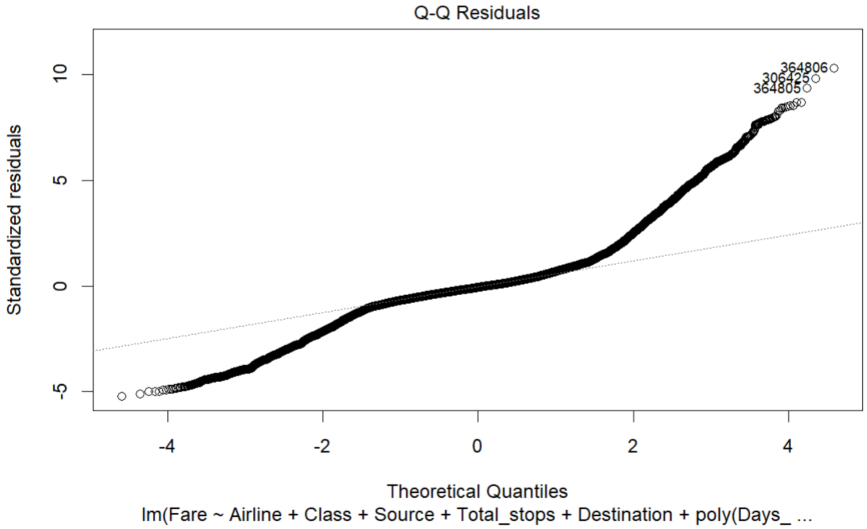


Figure 30: Best Model's Q-Q Plot

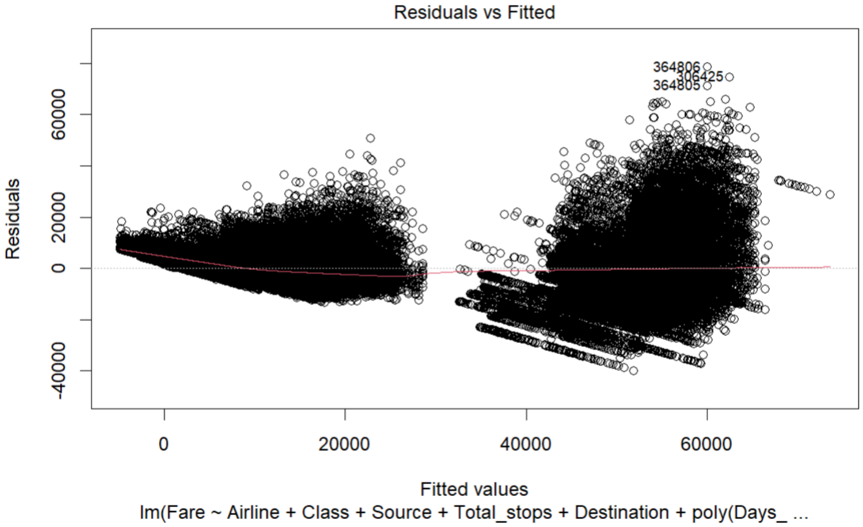


Figure 31: Best Model's Residuals vs Fitted Plot

To remedy the normality assumption, we performed Box-Cox transformation on the response, Fare. The figure below shows the optimal value of lambda found, which turned out to be 0.14. This value is very close to zero, as a result, we applied the log transformation on the response.

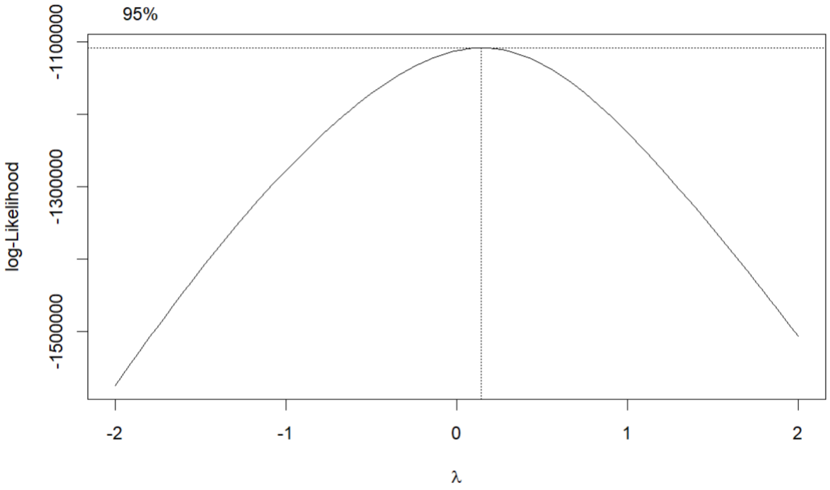


Figure 32: Optimal Lambda Value of Box-Cox Transformation

Once we performed the log transformation on the response and re-fitted the model, we got the diagnostic plots below. As seen in the Q-Q plot, the normality assumption is approximately satisfied. The reason why it is not fully satisfied is due to the fact that the transformed response, log(Fare), has a bimodal distribution (as seen in the subsequent histogram figure) due to the ticket class variable (Economy and Premium Economy contribute to the first mode, and Business and First contribute to the second mode). Also, by inspecting the residuals vs fitted plot, we can observe that all of the other regression assumptions are also now approximately satisfied. It is also worth noting that we decided not to eliminate outliers, as we did not find any that very clearly stand out from the rest of the data points. Therefore, we have now achieved the best multiple regression model for predicting the response, Fare, whose coefficients can be seen in the figure following the diagnostic figures. In Section 4, we use test data to evaluate the performance of this model.

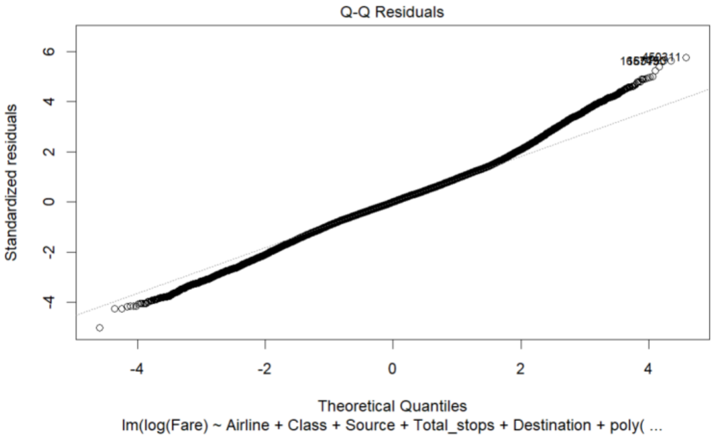


Figure 33: Q-Q Plot of Transformed Model

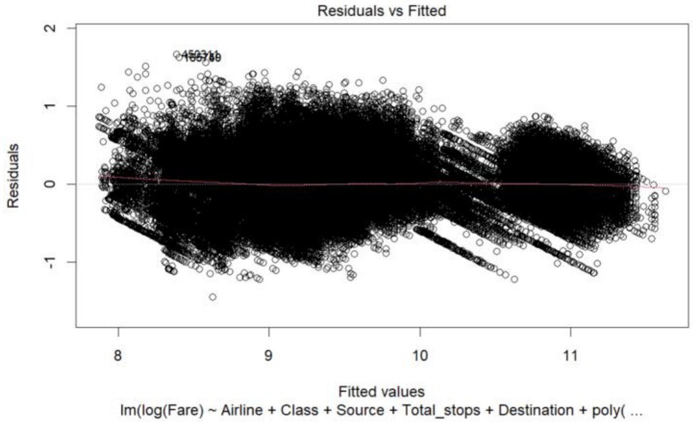


Figure 34: Residuals vs Fitted Plot of Transformed Model

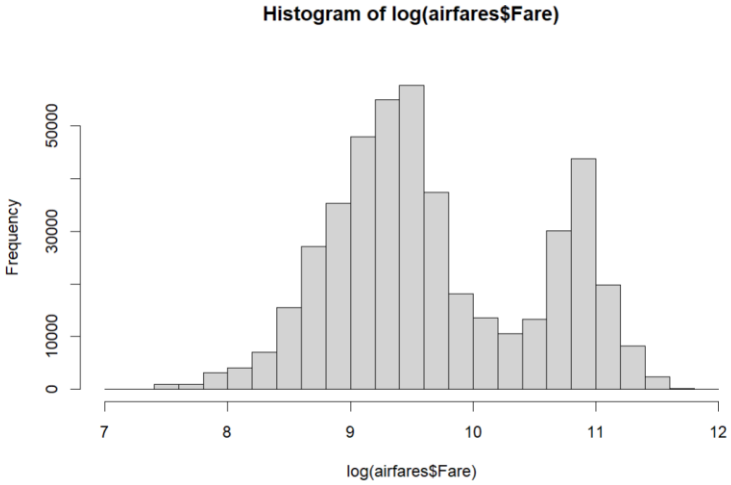
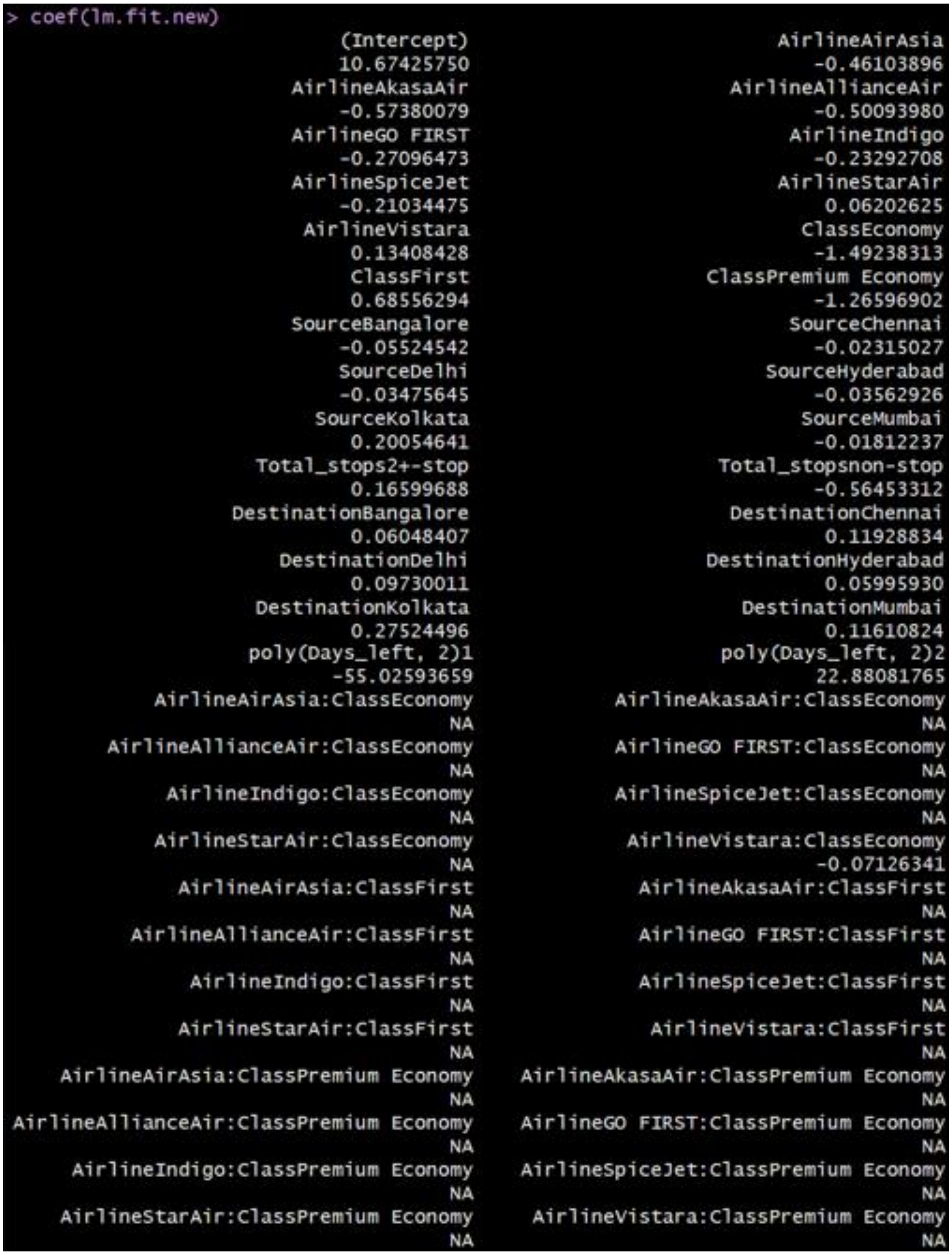


Figure 35: Histogram of Transformed Response

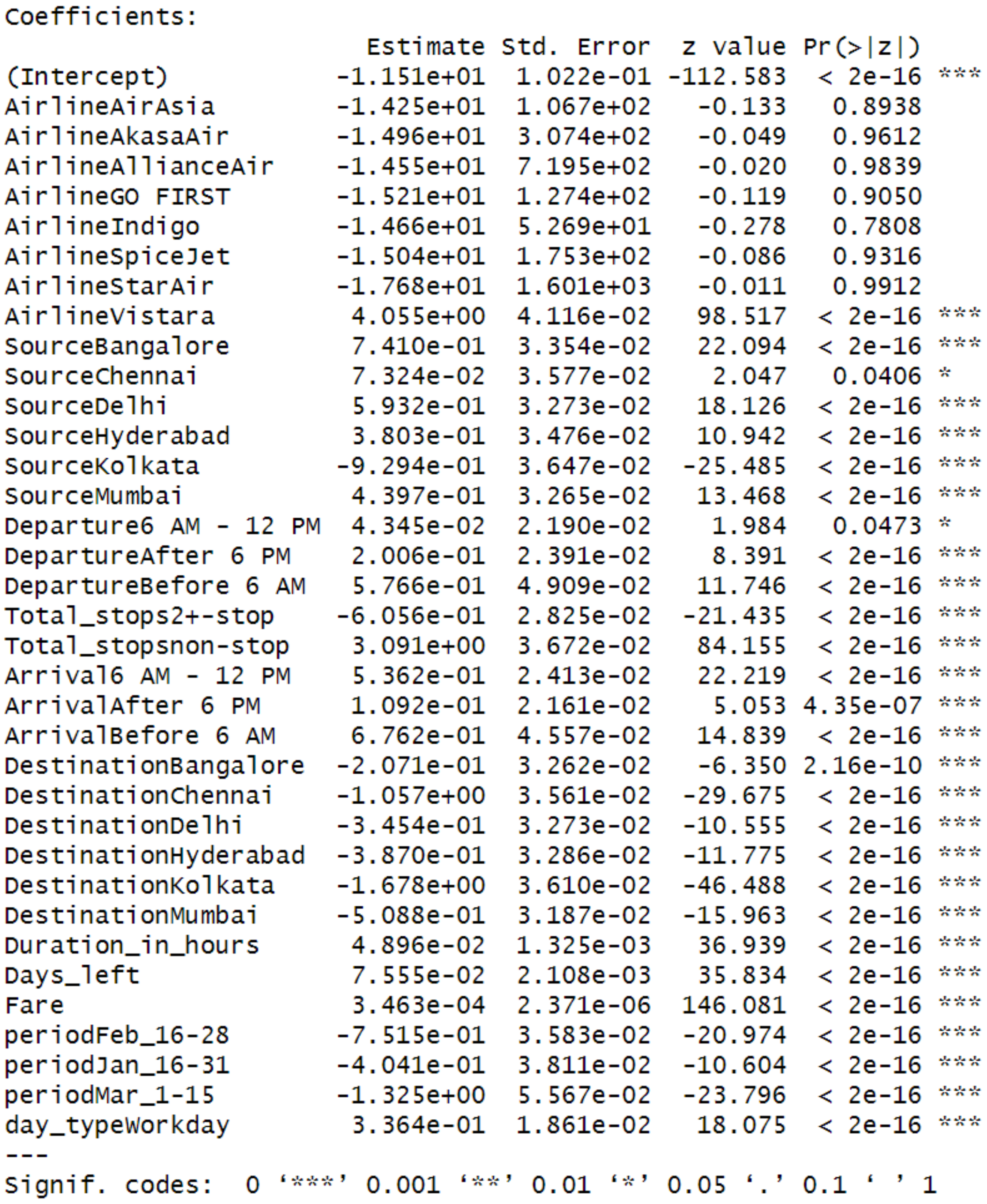


```
> coef(lm.fit.new)
                            (Intercept)                          AirlineAirAsia
                            10.67425750                             -0.46103896
                        AirlineAkasaAir                      AirlineAllianceAir
                            -0.57380079                             -0.50093980
                        AirlineGO FIRST                           AirlineIndigo
                            -0.27096473                             -0.23292708
                        AirlineSpiceJet                          AirlineStarAir
                            -0.21034475                              0.06202625
                         AirlineVistara                            ClassEconomy
                             0.13408428                             -1.49238313
                             ClassFirst                    ClassPremium Economy
                             0.68556294                             -1.26596902
                        SourceBangalore                           SourceChennai
                            -0.05524542                             -0.02315027
                            SourceDelhi                         SourceHyderabad
                            -0.03475645                             -0.03562926
                          SourceKolkata                            SourceMumbai
                             0.20054641                             -0.01812237
                     Total_stops2+-stop                     Total_stopsnon-stop
                             0.16599688                             -0.56453312
                   DestinationBangalore                      DestinationChennai
                             0.06048407                              0.11928834
                       DestinationDelhi                    DestinationHyderabad
                             0.09730011                              0.05995930
                     DestinationKolkata                       DestinationMumbai
                             0.27524496                              0.11610824
                    poly(Days_left, 2)1                     poly(Days_left, 2)2
                           -55.02593659                             22.88081765
            AirlineAirAsia:ClassEconomy            AirlineAkasaAir:ClassEconomy
                                     NA                                      NA
        AirlineAllianceAir:ClassEconomy            AirlineGO FIRST:ClassEconomy
                                     NA                                      NA
             AirlineIndigo:ClassEconomy            AirlineSpiceJet:ClassEconomy
                                     NA                                      NA
            AirlineStarAir:ClassEconomy             AirlineVistara:ClassEconomy
                                     NA                             -0.07126341
              AirlineAirAsia:ClassFirst              AirlineAkasaAir:ClassFirst
                                     NA                                      NA
          AirlineAllianceAir:ClassFirst              AirlineGO FIRST:ClassFirst
                                     NA                                      NA
               AirlineIndigo:ClassFirst              AirlineSpiceJet:ClassFirst
                                     NA                                      NA
              AirlineStarAir:ClassFirst               AirlineVistara:ClassFirst
                                     NA                                      NA
    AirlineAirAsia:ClassPremium Economy    AirlineAkasaAir:ClassPremium Economy
                                     NA                                      NA
AirlineAllianceAir:ClassPremium Economy    AirlineGO FIRST:ClassPremium Economy
                                     NA                                      NA
     AirlineIndigo:ClassPremium Economy    AirlineSpiceJet:ClassPremium Economy
                                     NA                                      NA
    AirlineStarAir:ClassPremium Economy     AirlineVistara:ClassPremium Economy
                                     NA                                      NA
```

Figure 36: Coefficients of Transformed Model

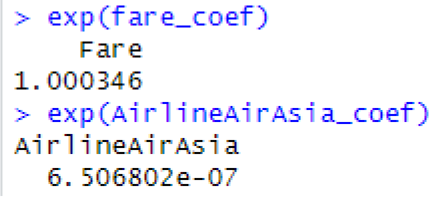
Binary Logistic Regression

The goal is to create a model that can predict the class of the ticket based on the remaining predictors. The difference in the model predictors is in the class variable. We have merged Premium Economy, Business, and First class into one class called Premium and kept the Economy class as is. Hence, we now have a binary response we can use in our model for prediction. The figure below shows the summary of the data:



```
Coefficients:
                         Estimate Std. Error  z value Pr(>|z|)    
(Intercept)            -1.151e+01  1.022e-01 -112.583  < 2e-16 ***
AirlineAirAsia         -1.425e+01  1.067e+02   -0.133   0.8938    
AirlineAkasaAir        -1.496e+01  3.074e+02   -0.049   0.9612    
AirlineAllianceAir     -1.455e+01  7.195e+02   -0.020   0.9839    
AirlineGO FIRST        -1.521e+01  1.274e+02   -0.119   0.9050    
AirlineIndigo          -1.466e+01  5.269e+01   -0.278   0.7808    
AirlineSpiceJet        -1.504e+01  1.753e+02   -0.086   0.9316    
AirlineStarAir         -1.768e+01  1.601e+03   -0.011   0.9912    
AirlineVistara          4.055e+00  4.116e-02   98.517  < 2e-16 ***
SourceBangalore         7.410e-01  3.354e-02   22.094  < 2e-16 ***
SourceChennai           7.324e-02  3.577e-02    2.047   0.0406 *  
SourceDelhi             5.932e-01  3.273e-02   18.126  < 2e-16 ***
SourceHyderabad         3.803e-01  3.476e-02   10.942  < 2e-16 ***
SourceKolkata          -9.294e-01  3.647e-02  -25.485  < 2e-16 ***
SourceMumbai            4.397e-01  3.265e-02   13.468  < 2e-16 ***
Departure6 AM - 12 PM   4.345e-02  2.190e-02    1.984   0.0473 *  
DepartureAfter 6 PM     2.006e-01  2.391e-02    8.391  < 2e-16 ***
DepartureBefore 6 AM    5.766e-01  4.909e-02   11.746  < 2e-16 ***
Total_stops2+-stop     -6.056e-01  2.825e-02  -21.435  < 2e-16 ***
Total_stopsnon-stop     3.091e+00  3.672e-02   84.155  < 2e-16 ***
Arrival6 AM - 12 PM     5.362e-01  2.413e-02   22.219  < 2e-16 ***
ArrivalAfter 6 PM       1.092e-01  2.161e-02    5.053 4.35e-07 ***
ArrivalBefore 6 AM      6.762e-01  4.557e-02   14.839  < 2e-16 ***
DestinationBangalore   -2.071e-01  3.262e-02   -6.350 2.16e-10 ***
DestinationChennai     -1.057e+00  3.561e-02  -29.675  < 2e-16 ***
DestinationDelhi       -3.454e-01  3.273e-02  -10.555  < 2e-16 ***
DestinationHyderabad   -3.870e-01  3.286e-02  -11.775  < 2e-16 ***
DestinationKolkata     -1.678e+00  3.610e-02  -46.488  < 2e-16 ***
DestinationMumbai      -5.088e-01  3.187e-02  -15.963  < 2e-16 ***
Duration_in_hours       4.896e-02  1.325e-03   36.939  < 2e-16 ***
Days_left               7.555e-02  2.108e-03   35.834  < 2e-16 ***
Fare                    3.463e-04  2.371e-06  146.081  < 2e-16 ***
periodFeb_16-28        -7.515e-01  3.583e-02  -20.974  < 2e-16 ***
periodJan_16-31        -4.041e-01  3.811e-02  -10.604  < 2e-16 ***
periodMar_1-15         -1.325e+00  5.567e-02  -23.796  < 2e-16 ***
day_typeWorkday         3.364e-01  1.861e-02   18.075  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 37 Coefficients of the Binary Logistic Model



```
> exp(fare_coef)
    Fare 
1.000346 
> exp(AirlineAirAsia_coef)
AirlineAirAsia 
  6.506802e-07 
```

Figure 38 Odds ratio of Fare and one of the factors

We have done variable selection, and we got that all the coefficients should be in the model. The confusion matrix and ROC curve are in Section 4.

Ordinal Logistic Regression

The goal is to create a model that can predict the class of the ticket based on the remaining predictors. The difference in the model predictors is in the class variable. We have merged Business and First class into one class called Premium and kept the Economy, and Premium Economy classes as is. Hence, we now have 3 classes in the response which we can use in our model for prediction. Hence, we used the ordinal logistic regression model as our model. We have also decided to use all the variables for the model. The confusion matrix is in the validation section below.

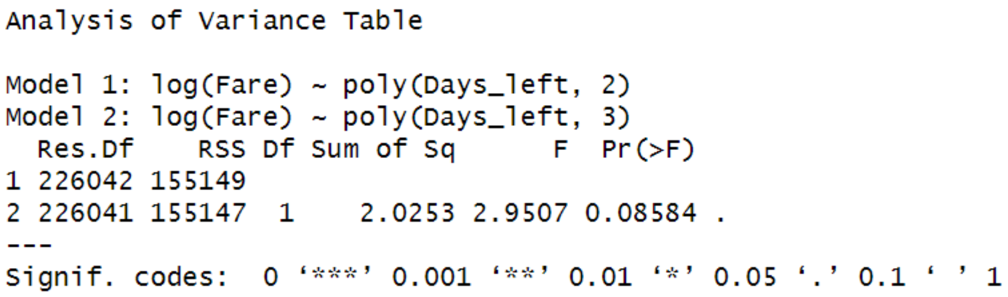
Section 4: Model Validation

Simple Linear Regression

1. Results for using Days_left as the predictor and Fare as the response.

Table 9 Simple Linear Regression results 1

Predictor	Response	R^2	Adj. R^2
Days_left	Fare	0.007872198	0.007867809
Days_left	log(Fare)	0.01469623	0.01469187
Poly(Days_left,2)	log(Fare)	0.0170763	0.01706761
Poly(Days_left,3)	log(Fare)	0.01708913	0.01707609



```
Analysis of Variance Table

Model 1: log(Fare) ~ poly(Days_left, 2)
Model 2: log(Fare) ~ poly(Days_left, 3)
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1 226042 155149
2 226041 155147  1    2.0253 2.9507 0.08584 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

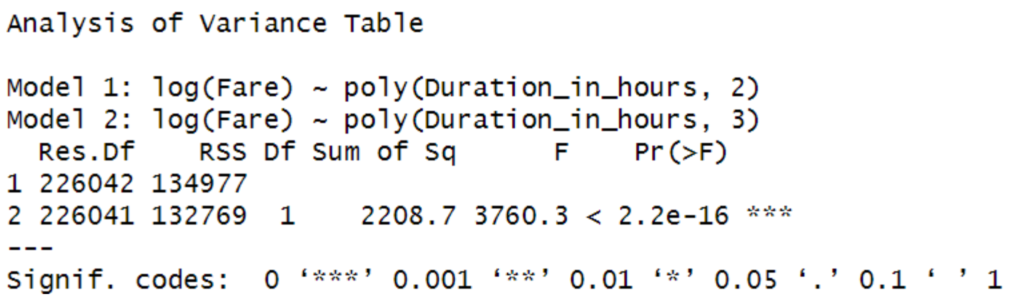
Figure 39 Simple Linear Regression Partial F-test 1

In conclusion after using the transformation and using a polynomial of order 2 we have achieved the best results. Adding a third order polynomial makes the model insignificant according to the partial F-test.

2. Results for using Duration_in_hours as the predictor and Fare as the response.

Table 10 Simple Linear Regression results 2

Predictor	Response	R^2	Adj. R^2
poly(Duration_in_hours,2)	log(Fare)	0.1448685	0.144861
poly(Duration_in_hours,3)	log(Fare)	0.1588614	0.1588503



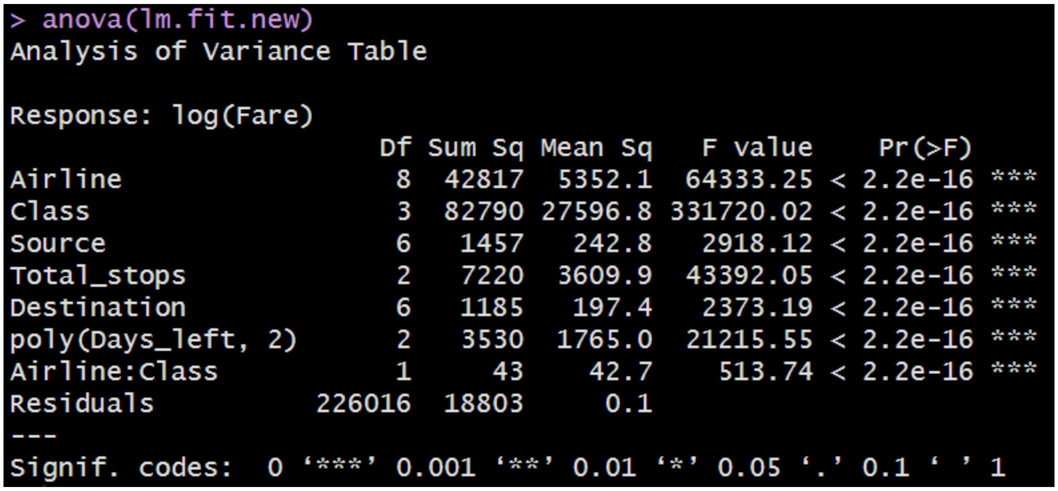
```
Analysis of Variance Table

Model 1: log(Fare) ~ poly(Duration_in_hours, 2)
Model 2: log(Fare) ~ poly(Duration_in_hours, 3)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1 226042 134977                                  
2 226041 132769  1    2208.7 3760.3 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 40 Simple Linear Regression Partial F-test 2

In conclusion after using the transformation and using a polynomial of order 3 we have achieved the best results. Adding a third order polynomial makes the model significant according to the partial F-test. We also tested with higher order polynomials but that leads to overfitting and a more complicated model.

Multiple Linear Regression

Once we built the best multiple linear regression model for predicting the Fare, we evaluated the performance of this model using test data by finding its R-squared and adjusted R-squared values. To do this, we predicted the Fare using our model for all the test data observations, and from that we calculated SST and SSE, which allowed us to calculate R-squared and adjusted R-squared. The value of R-squared is 0.8667395, and the value of adjusted R-squared is 0.8667094. These values are considered high, which signifies the strong predictive ability of the model. Additionally, we viewed the ANOVA table results of this model, which can be seen in the figure below, which shows that all of the variables in the model are significant.

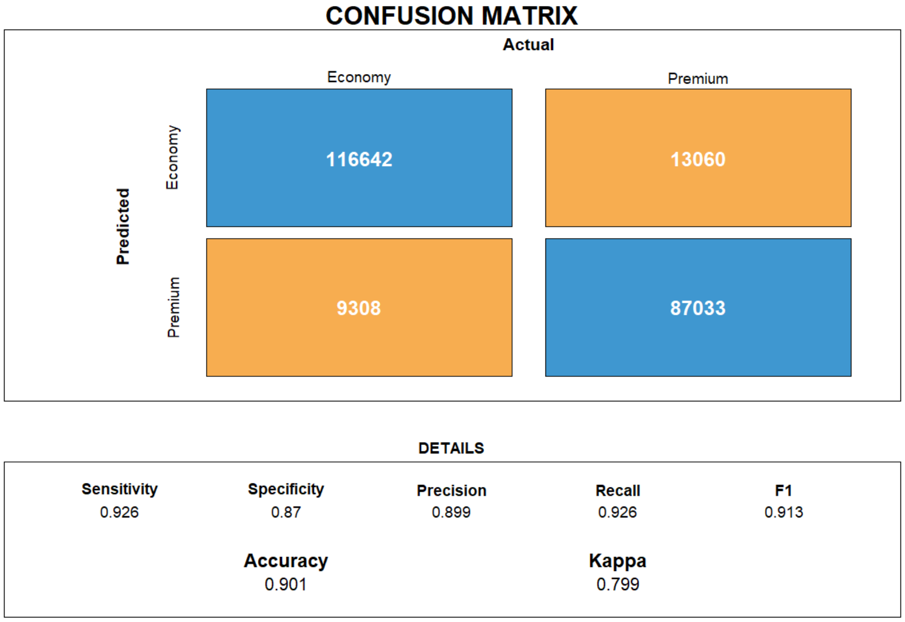


```
> anova(lm.fit.new)
Analysis of Variance Table

Response: log(Fare)
                      Df Sum Sq Mean Sq   F value    Pr(>F)    
Airline                8  42817  5352.1  64333.25 < 2.2e-16 ***
Class                  3  82790 27596.8 331720.02 < 2.2e-16 ***
Source                 6   1457   242.8   2918.12 < 2.2e-16 ***
Total_stops            2   7220  3609.9  43392.05 < 2.2e-16 ***
Destination            6   1185   197.4   2373.19 < 2.2e-16 ***
poly(Days_left, 2)     2   3530  1765.0  21215.55 < 2.2e-16 ***
Airline:Class          1     43    42.7    513.74 < 2.2e-16 ***
Residuals         226016  18803     0.1                        
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 41: ANOVA Results of Final Model

Binary Logistic Regression

Below are the confusion matrix and the ROC curve for the binary logistic regression model. As seen in the figures, the sensitivity and specificity are high because the model is good in predicting both classes. The results are discussed in detail in Section 5.



CONFUSION MATRIX

Predicted	Actual: Economy	Actual: Premium
Economy	116642	13060
Premium	9308	87033

DETAILS

Sensitivity	Specificity	Precision	Recall	F1
0.926	0.87	0.899	0.926	0.913

Accuracy	Kappa
0.901	0.799

Figure 42 Binary Logistic model confusion matrix and details

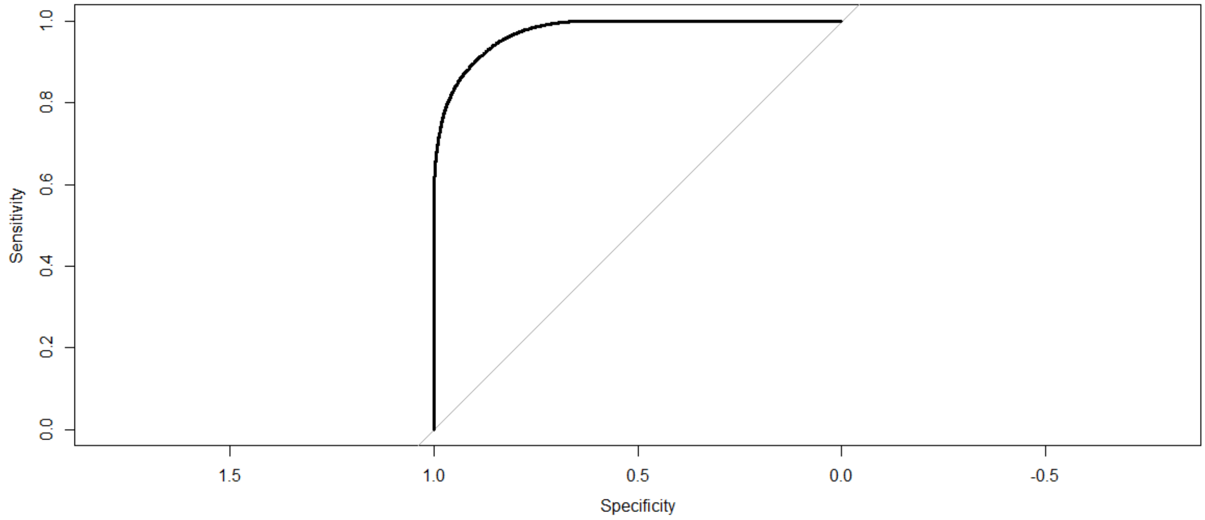
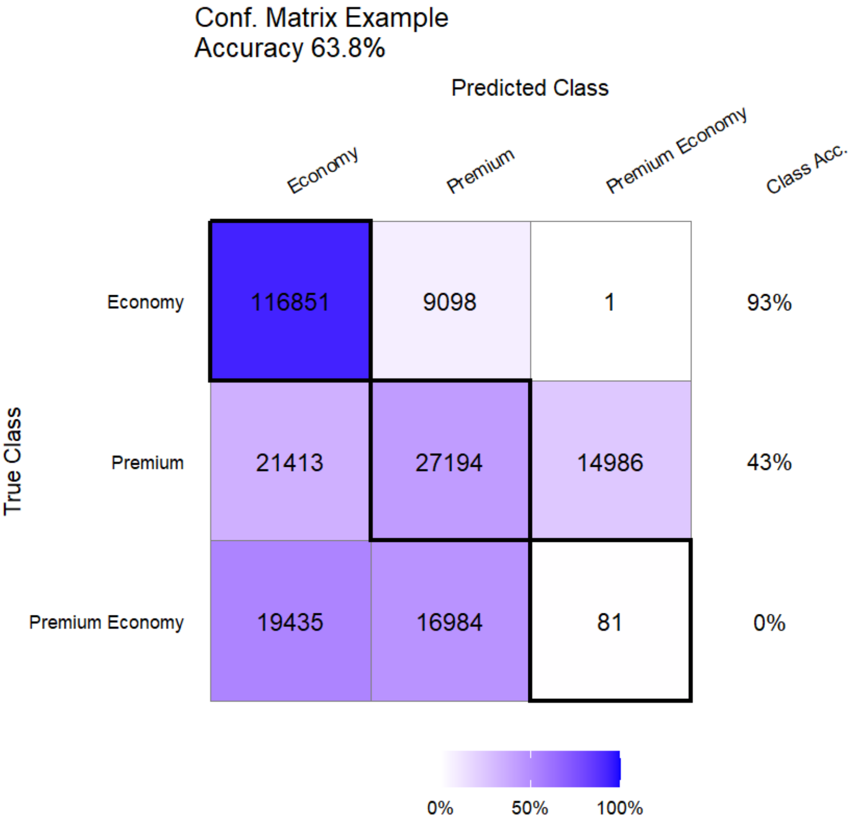


Figure 43 - ROC curve

Ordinal Logistic Regression

Below are the confusion matrix and the ROC curve for the ordinal logistic regression model. As shown below, the model has a high class accuracy for the Economy class. The Premium Economy class has an accuracy of 0%, which may indicate that there are missing variables that could explain the response. The results are discussed in detail in Section 5.



	Economy	Premium	Premium Economy	Class Acc.
Economy	116851	9098	1	93%
Premium	21413	27194	14986	43%
Premium Economy	19435	16984	81	0%

Figure 44 Ordinal Logistic model confusion matrix and details

Section 5: Discussion and Conclusions

We have tried a simple linear regression model with fare as the response and found that Duration in hours has a better performance than just using Days left. This is due to the high correlation between the Fare and the Duration in hours. We can also notice that after applying the remedies we were able to some extent meet all the assumptions. The transformations used was the log of the response and the best model for Days left had was the 2nd order polynomial. The best model had an adjusted R-squared of 0.01706761, or 1.706761%. When we used the Duration in hours, we realized that polynomial of order 3 was significant. We also tested with higher orders but that would make the model complicated and can cause overfitting. The best model at polynomial of order 3 with Duration in hours had adjusted R-squared of: 0.1588503, or 15.88503%.

Next, we tried multiple linear regression. The best model that we achieved has an adjusted R-squared value of 0.8667094, or 86.67094%, which signifies the strong predictive ability of the model. The model includes the variables Airline, Class, Source, Total_stops, Destination, Days_left, Days_left^2, and the interaction Airline*Class. By inspecting the model coefficients shown in Figure 36, we can find some insightful interpretations. For the Airline, Vistara, Air India, and StarAir have positive coefficients, which indicates that their flights are more expensive on average. For the Class, Business and First have positive coefficients, which indicates that they are more expensive than Economy and Premium Economy, which aligns with our expectations. For the Source, only Kolkata has a positive coefficient, which indicates that it is the most expensive city to travel from. For the Total_stops, 1 stop and 2+ stops have positive coefficients, which indicates their flights are more expensive compared to non-stop, which aligns with our expectations. For the Destination, Ahmedabad is the only negative coefficient, which indicates it is the cheapest city to fly to. For Days_left, its terms have the coefficients 22.9*Days_left^2 -55.0*Days_left, which indicates that the Fare has a quadratic relationship with the number of days left since booking. Finally, for the interaction of Airline*Class, we can see that if the ticket is an Economy ticket on Vistara, then the Fare is lower on average.

Furthermore, we tried logistic regression to predict the Class. We did two models, binary and ordinal. The ordinal model had three classes. In binary we combined the four classes which are Economy, Premium Economy, Business and First to just Economy and the remaining classes were grouped into Premium. For ordinal we combined First and Business to Premium and kept Economy and Premium Economy as is. We got an accuracy of 90.1% in the binary model. In the ordinal model we got a lower accuracy of 63.8%.

Future work includes trying various other supervised learning techniques such as neural networks, lasso, and ridge regression, SVM, KNN, random forest, bagging, and gradient boosting, as well as unsupervised techniques like clustering to view patterns in the data. For three classes we can also investigate different balancing techniques.

References

[1] “Airfare ML: Predicting Flight Fares.” Accessed: May 16, 2024. [Online]. Available: https://www.kaggle.com/datasets/yashdharme36/airfare-ml-predicting-flight-fares