



**Carnegie Mellon University**

# **Predicting Flight Fares**

## **Final Report**

19-433 Data Science for Technology, Innovation and Policy

May 2023

### **Team 1**

Aarohi Kapadia, Aadit Javeri, Prince Jain, Farhan Ahmad, Fagun Parikh

[GitHub Repository Link](#)

## **Introduction**

The airline industry is subject to numerous internal and external factors that affect the prices of airline tickets. These factors include fuel costs, demand, route availability, time of travel, seasonal trends, and airline pricing policies, among others. These factors make airline prices highly volatile and difficult to predict accurately. As a result, pricing forecasting and optimization have become critical factors for airlines to remain competitive and profitable while providing customers with the best possible value for their money. For customers, finding the best price for a flight can be a time-consuming and frustrating process.

The goal is to develop a model that can accurately forecast airline prices based on a range of factors, such as departure and arrival locations, travel dates, airline carriers, seat class, and number of stops. To address this problem, several approaches have been proposed in the past, including regression-based models, machine learning algorithms, and time-series forecasting techniques. While significant progress has been made in enhancing the accuracy and scalability of current technologies, there remain opportunities for further improvements in these areas.

The study also aims to develop a Graphical User Interface (GUI) that can provide an easy-to-use interface for customers to predict future flight tickets and compare prices across different airlines. The GUI is designed to be user-friendly and intuitive, allowing customers to enter their travel information and receive real-time pricing information for different flight options. The dataset used in this report contains flight fare information for the aviation industry in India. It is collected from EaseMyTrip [1] using web scraping techniques.

## **Research Questions**

The goal of the study is to analyze the flight booking dataset received from the "Ease My Trip" [1] and run various statistical hypothesis tests to analyze the following research questions:

- How does price vary across Airlines?
- Does ticket price change depending on the departure time and arrival time?
- How does the ticket price vary closer to the departure dates?
- Is the ticket price influenced by Source and Destination?

## Literature Review

There have been innumerable studies related to prediction of flight fares through machine learning approach. Based on that, many tools are available online such as Hopper [2], AirHint [3], Alternative Airlines [4], etc. We will be discussing the following 2 papers based on this field of study:

### 1) *Flight Fare Prediction System Using Machine Learning* [5]

The paper "Flight Fare Prediction Using Machine Learning" presents a study on the use of machine learning algorithms for predicting flight fares. The authors used a dataset that included flight details and prices of different airlines and applied various regression algorithms such as Linear Regression, Random Forest, and Gradient Boosting Regression to the data. They also performed feature engineering to extract useful features from the dataset. The authors explain the data collection process, which involved scraping flight data from multiple websites, and then preprocessing the data to extract relevant features. They extracted relevant features such as flight origin, destination, flight duration, departure time, and airline name, and converted them into numerical data for analysis. They then describe the modeling and evaluation process, which involves training and testing various regression models, including linear regression, decision trees, random forests, and gradient boosting algorithms.

The study found that the Gradient Boosting Regression algorithm performed the best in terms of accuracy, with an R-squared value of 0.81, indicating that the model can explain 81% of the variability in the data. The authors also evaluated the feature importance of the dataset and found that the most important features were the airline name, the number of stops, and the departure time. The authors evaluate the performance of each algorithm using metrics such as mean absolute error, mean squared error, and root mean squared error. They find that the random forest algorithm outperforms the other algorithms, achieving a mean absolute error of 681.67 and a root mean squared error of 1174.77. The authors also perform feature importance analysis to identify the most significant factors affecting ticket prices. They find that the airline, the day of the week, and the time of day have the most significant impact on ticket prices.

Overall, the study demonstrates the effectiveness of machine learning algorithms in predicting flight fares and provides insights into the important features that affect the pricing of airline

tickets. The findings of this study could be useful for airlines and travel agencies in pricing their tickets more accurately and for consumers in making informed decisions when purchasing airline tickets.

## 2) *A Framework for Airfare Price Prediction: A Machine Learning Approach* [6]

The paper "A Framework for Airfare Price Prediction: A Machine Learning Approach" proposes a machine learning-based framework for predicting airfare prices. Airfare pricing is a complex task that depends on various factors such as the time of the year, the route, the airline, the day of the week, and demand. Traditional methods of airfare prediction involve manual analysis and forecasting based on historical data, which can be time-consuming and error-prone. To address these challenges, the authors proposed a machine learning-based framework that can accurately predict airfare prices. The framework consists of four stages: data collection, data preprocessing, feature selection, and model development.

In the data collection stage, the authors collected airfare data from various sources such as airline websites, travel agencies, and third-party providers. The data include information such as the departure and arrival airports, the airline, the date of travel, the price, and other relevant attributes. In the data preprocessing stage, the authors performed various operations such as outlier detection, missing value imputation, and data normalization. Outlier detection involves identifying and removing data points that are significantly different from other data points in the dataset. Missing value imputation involves filling in missing values with plausible values based on the other attributes. Data normalization involves scaling the data to a common range to avoid bias towards attributes with a larger range.

In the feature selection stage, the authors used various techniques to identify the most important features for airfare prediction. Feature selection is an important step in machine learning as it helps to reduce the number of features and avoid overfitting. The authors used techniques such as correlation analysis, mutual information, and principal component analysis to select the most relevant features. In the model development stage, the authors experimented with different machine learning algorithms such as linear regression, decision trees, and random forests. They evaluated the performance of each model using different performance metrics such as mean squared error and mean absolute error. The authors found that the random forest algorithm

outperformed other algorithms in terms of prediction accuracy. Overall, the paper presents a comprehensive framework for airfare price prediction using machine learning techniques. The framework can be extended to predict airfare trends and help airlines in pricing strategies. The paper provides a valuable contribution to the field of airfare pricing and demonstrates the potential of machine learning in this domain.

## Data Sources

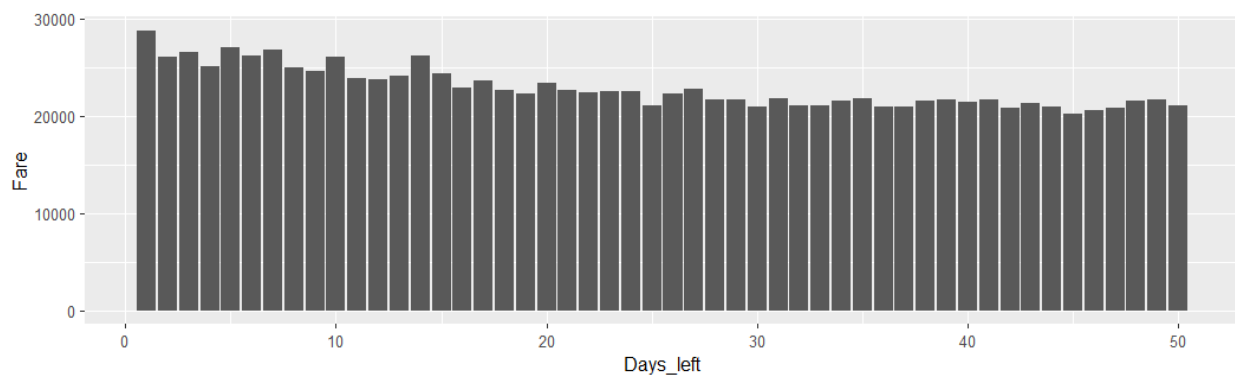
The features that we have in the data are Date of Journey, Journey\_day, Airline, Flight code, Class, Source city, Departure time, Total\_stops, Arrival city, Flight Duration, Days Left to Travel and Fare of the Ticket. We will be using this to predict the fare of the flight tickets. Table 1 shows the unique values for the main features and range of these features.

*Table 1. Exploratory Data Analysis*

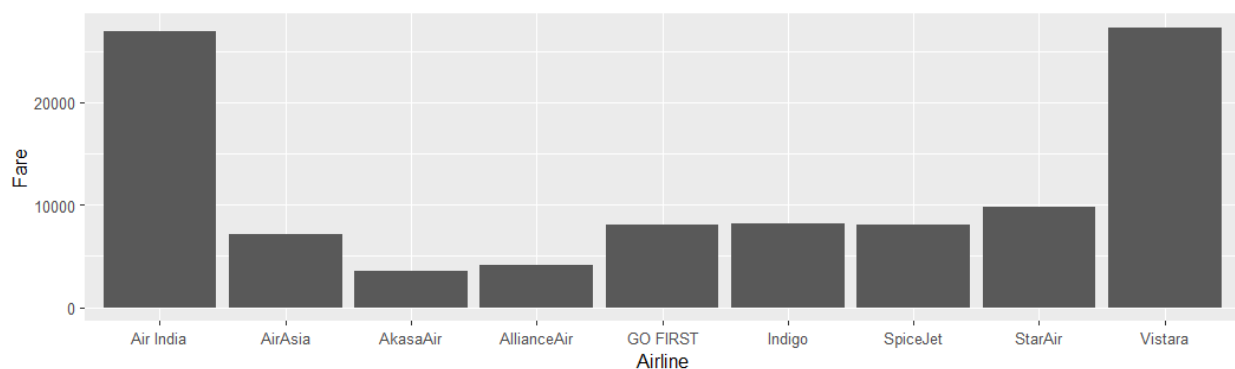
Features	Count/Range
Number of Observations (flights)	452088
Date of Journey (range)	2023-01-16 - 2023-03-06 (50 days)
Airlines	SpiceJet, Indigo, GO FIRST, Air India, AirAsia, Vistara, AkasaAir, AllianceAir, StarAir
Times of Departure and Arrival	After 6 PM, Before 6 AM, 12 PM - 6 PM, 6 AM - 12 PM
Cities	Mumbai, Bangalore, Hyderabad, Kolkata, Chennai, Ahmedabad, Delhi
Ticket Class	Economy, Premium Economy, Business and First class
Number of Stops	0,1,2
Flight Duration	0.75 - 43.58 hours
Days left to book (range)	1 - 50 days
Flight Fares	INR 1307 - 143019
Unique values of First Class	144
Max Departures from City	Delhi, 83153 departures

It has also been observed that we don't have a uniform number of datapoints on all airlines and different classes (See Figures A1, A2 and A3 in Appendix). However this might be because some airlines operate more flights and most of the tickets sold are either economy class or premium economy. Hence the prediction for airlines with less data and first class tickets might have higher uncertainty. It is also seen that the number of flights is more or less the same on each day of the week. Hence the prediction might have similar uncertainty any day of the week.

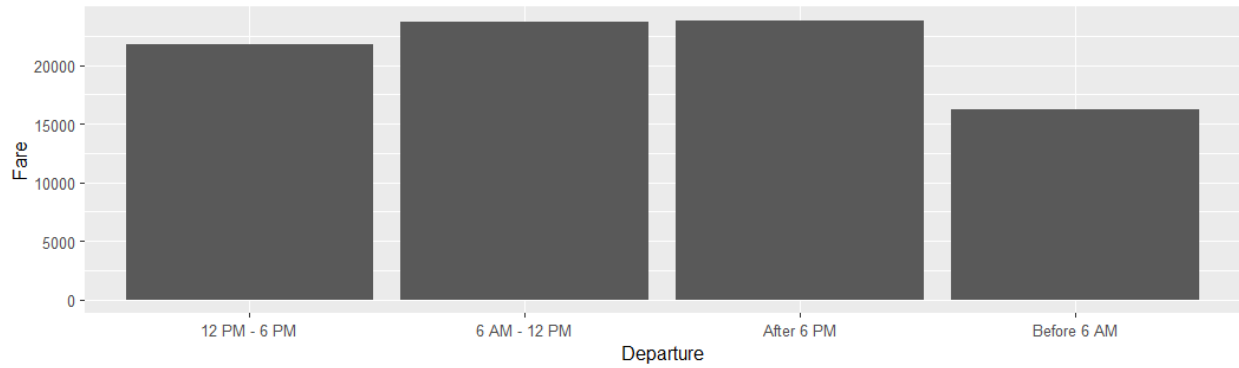
The relation of the ticket fare vs multiple variables in the data was examined to understand the dependence of the fare on these variables. From the plots below, it can be seen that the fare of the ticket depends to a high extent on the days left to the flight, number of stops, the airline, ticket class (economy, business etc.) and departure time. Hence these should be the main elements of the model. Other variables such as, journey day (day of the week), sector, source and destination cities don't have much effect on the fare (see Figures A4, A5, A6 and A7 in Appendix).



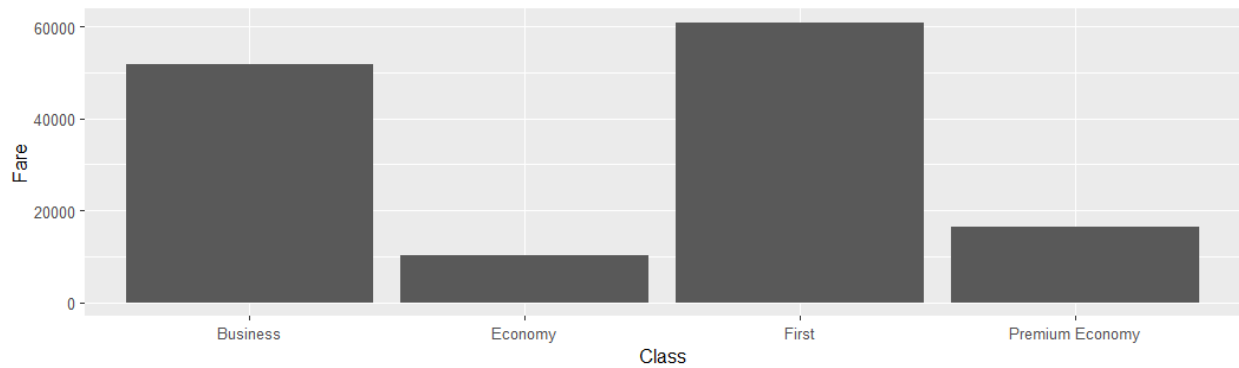
*Figure 1. Plot of Fare vs. Days left*



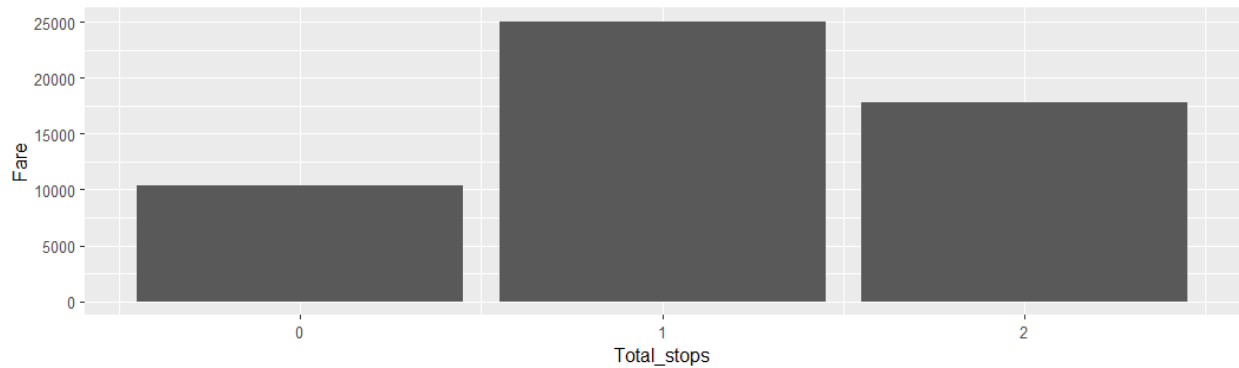
*Figure 2. Plot of Fare vs. Airline*



*Figure 3. Plot of Fare vs. Departure*



*Figure 4. Plot of Fare vs. Class*



*Figure 5. Plot of Fare vs. Total Stops*

Another thing to note is that even though the departure and the destination city look highly correlated (See figures A5 and A6 in Appendix), the destination and the source cities will never be the same and hence both (destination and source cities) need to be considered for flight price

prediction. The same also applies to the arrival and departure times (See Figure 3 above and A9 in Appendix).

## Analysis

### *Model fitting*

We trained a Random Forest model with 25 trees to predict the price of air tickets based on different features such as departure and arrival locations, departure and arrival times, travel dates, airline carriers, seat class, and number of stops. Initially, we tested a Lasso model to predict the target variable. While the model had a relatively high R-squared value of 0.851, we observed that the plot of residuals displayed a pattern of uneven variance, indicating the presence of outliers or error terms that could potentially impact the model's predictive variability (Figure 2). Despite our efforts to rectify these issues, we could not obtain satisfactory results.

To overcome these limitations, we switched to a Random Forest model. The R-squared value of the Random Forest model of 0.929 was slightly higher than that of the Lasso model and demonstrated a statistically significant and high predictive ability (Figure 3). This suggests that the Random Forest model is better suited for our data and can provide more reliable predictions.

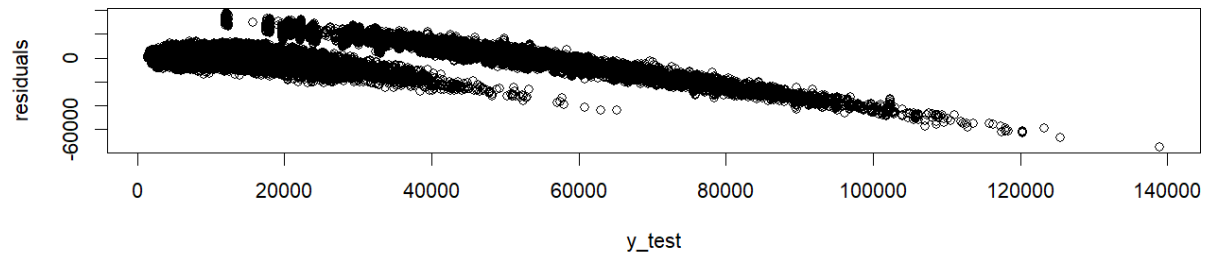
```
Call:
  randomForest(x = x_train, y = y_train, ntree = 25, do.trace = TRUE)
      Type of random forest: regression
      Number of trees: 25
No. of variables tried at each split: 12

      Mean of squared residuals: 30335071
      % Var explained: 92.63

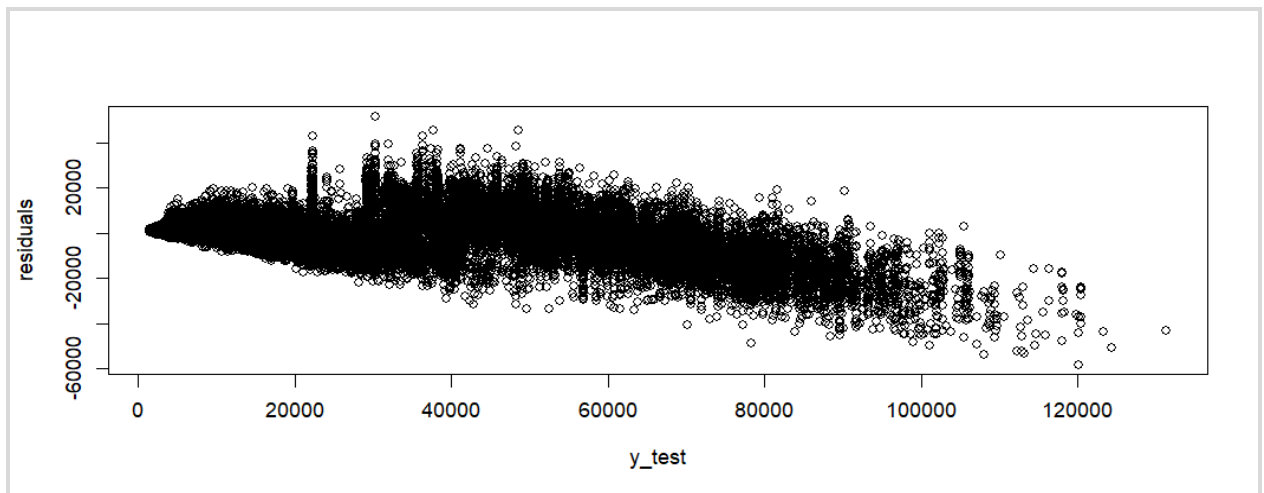
      RMSE      Rsquared      MAE
5428.6725860    0.9289331 2991.8254202
```

*Figure 6. Input parameters and Results of the Random Forest model*





*Figure 7. Plot of residuals vs the fare for optimal Lasso Model*

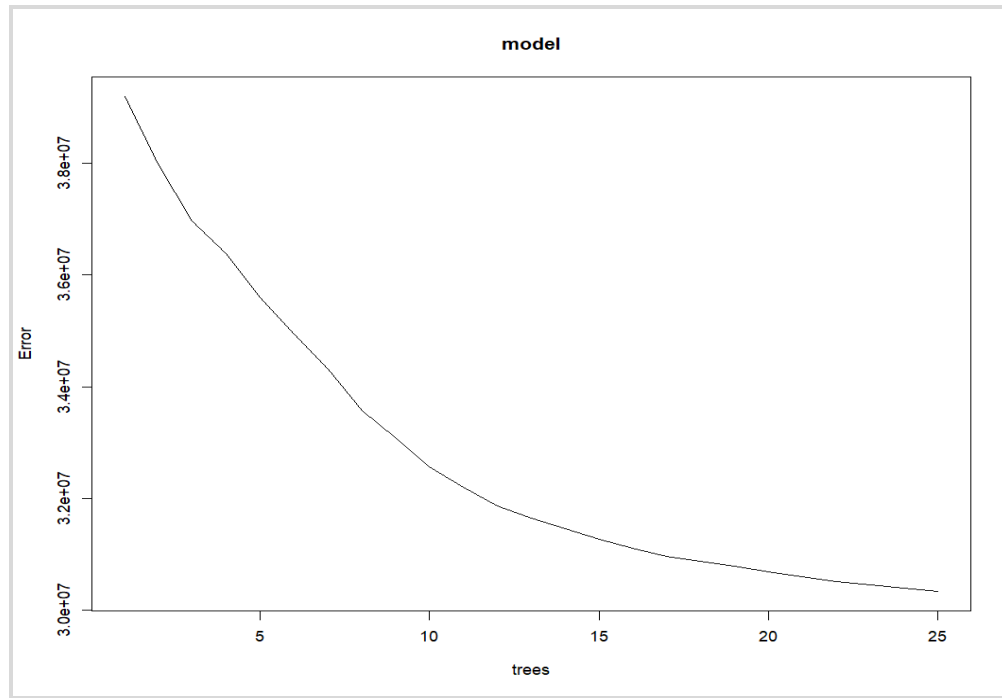


*Figure 8. Plot of residuals vs the fare for optimal Random Forest model*

### *Model validation*

To validate the performance of our model, we employed an 80-20 train-test split, where 80% of the data was used for training and the remaining 20% was held out for testing. This approach allowed us to evaluate how well our model generalizes to unseen data.

## Model tuning



*Figure 9. Error vs the no. of trees used for the model*

To enhance the predictive power of our model, we adjusted the features using the following techniques:

- One-hot encoding: We used one-hot encoding to transform the "stops" feature into multiple binary features, which allowed the model to capture more information about the data.
- Converting dates to weekdays: We converted dates to weekdays to see the price distribution based on the day of the week. This helped us identify any patterns or trends in the data that were specific to certain days of the week.
- Random forest parameters: We also adjusted the parameters of the Random Forest model, such as the number of trees, to optimize its performance. We allowed the model to select the optimal depth of trees. Specifically, we chose 25 trees for our model as a tradeoff between performance and computational efficiency.
- Removing sector: We initially thought that adding sector of flight, eg. Mumbai-Delhi etc. as categorical variables. However it increased the complexity of the model by increasing the number of model coefficients to 47. Thus, we just added the destination and source cities and reduced the number of variables to 39.

- Removing Duration: We have removed duration from our model. It would have been very cumbersome for the user to input the flight duration. Also the duration is a factor of the source and the destination city as well as the aircraft used. Hence, the airline, number of stops, source city and destination city will cover this and thus it was removed. We also compared the R squared error with and without the duration value in the model and it was about the same and thus it did not have any major effect on the prediction.

Due to limitations on the available computing power, multiple variations of the model could not be tried to obtain optimal tuning. However, our current model still had a R squared value of 0.929 on the withheld data set and hence is accurate enough to give a reasonable prediction.

### *Extending the analysis*

To extend our analysis, we developed a graphical user interface (GUI) that allows users to input their travel information and receive a prediction on the lowest expected price of the ticket and on what day of the week. Additionally, the GUI displays a plot of ticket price variation for the number of days left to flight, which can help inform their decision. The GUI was built using the Random Forest model that we trained on our dataset, which takes into account several features such as the travel dates, the number of stops, and the day of the week.

By developing this GUI, we aimed to provide users with a tool that can help them make informed decisions about their travel plans using our Random Forest model. Overall, this GUI serves as a practical application of our machine learning analysis and can benefit users who want to save money on their travel expenses [7].

To create this application, we used the Shiny package in R which is an easy library to make interactive applications. The structure of any shiny application consists of three components -

- 1) User Interface (UI) - this part consists of the layout and the appearance of the applications. This consisted of the inputs from the user to determine which day and type of ticket the user is looking for.
- 2) Server function - this consists of instructions to the application on what the computer is supposed to display and analyze. The function which consists of the model is called in this part of the application.

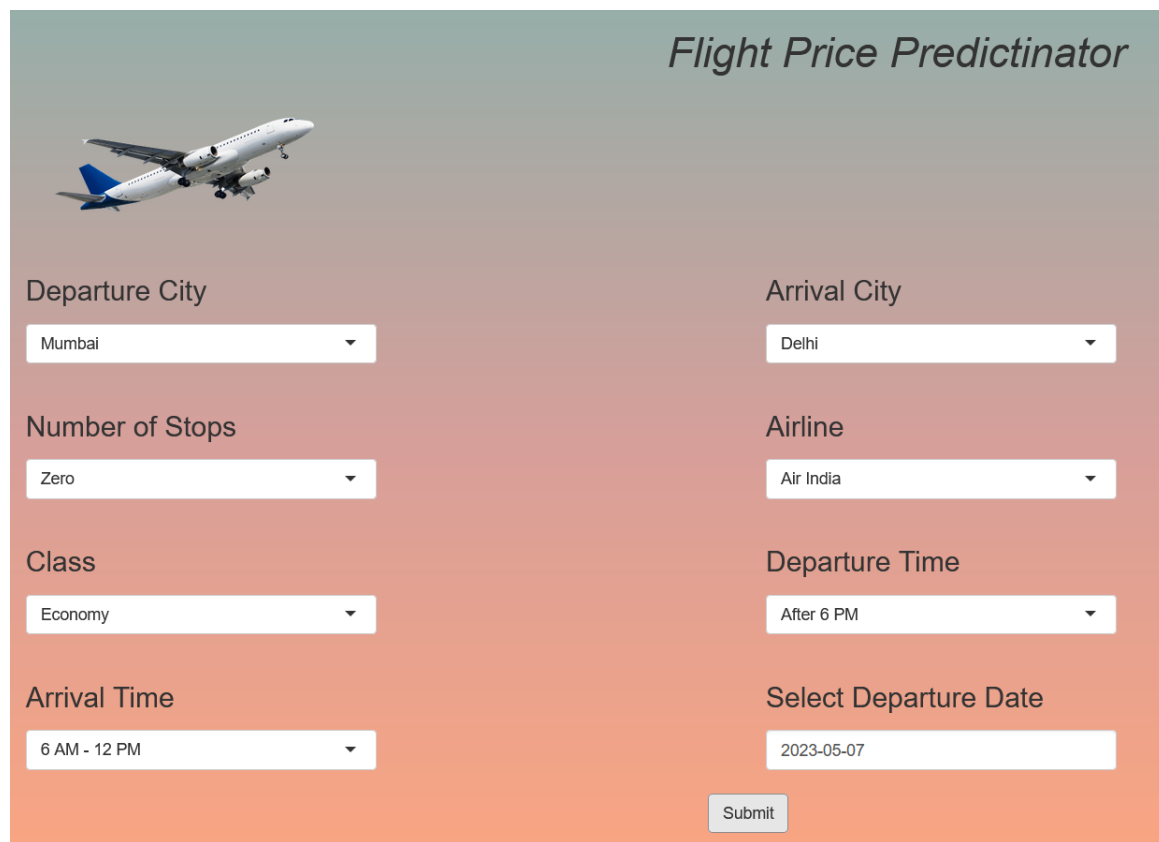
- 3) Shinyapp Call Function - this creates the shiny app objects from the user interface and server pair [7].

### *GUI Output*

The inputs that the user had to give were departure, arrival, number of stops, airlines, class of travel, departure time, arrival time and the date of departure. The screenshot below shows the inputs taken from the user in the form of dropdown menus and calendar inputs.

The output after the submit button is as shown below for an example run. This includes a graph of the prices predicted on that particular route at all the days ranging from today to the date of departure and when the price is predicted to be at the lowest. A line graph has also been included to show the prices on all days and to allow the user to make an informed decision.

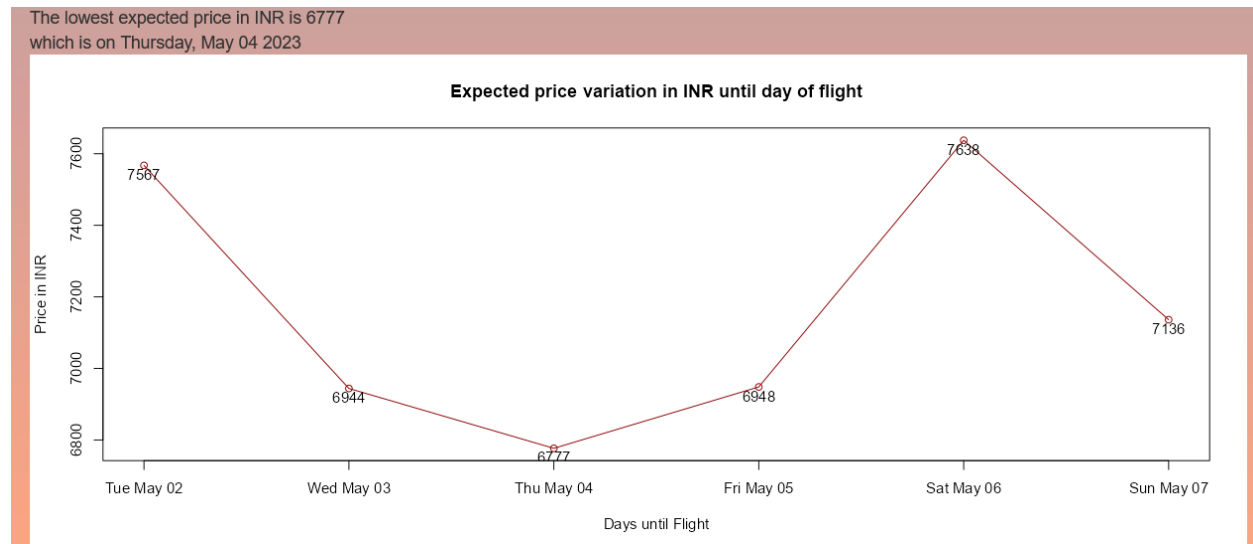
User Input Example:



The screenshot displays the 'Flight Price Predictinator' web application. At the top left is an image of a white and blue airplane. The title 'Flight Price Predictinator' is at the top right. The interface features two columns of input fields. The left column includes 'Departure City' (Mumbai), 'Number of Stops' (Zero), 'Class' (Economy), and 'Arrival Time' (6 AM - 12 PM). The right column includes 'Arrival City' (Delhi), 'Airline' (Air India), 'Departure Time' (After 6 PM), and 'Select Departure Date' (2023-05-07). A 'Submit' button is located at the bottom center.

*Figure 10. Screenshot of the sample GUI*

## GUI Output Example:



*Figure 11. Output of the GUI showing variations in price to the days until flight*

## Interpretation

Since the model used was a Random Forest model, what we could derive from the model was the importance of the different categorical variables. From the plot below it can be clearly seen that the class Economy or Premium Economy were the most statistically significant. This is quite intuitive as selecting an Economy vs a Premium Economy class ticket can have a huge impact on price. The other major factors that affect the price of the ticket are the Number of Stops, the Airline and the Number of Days Left to the Flight. The Price is not significantly impacted by other features.

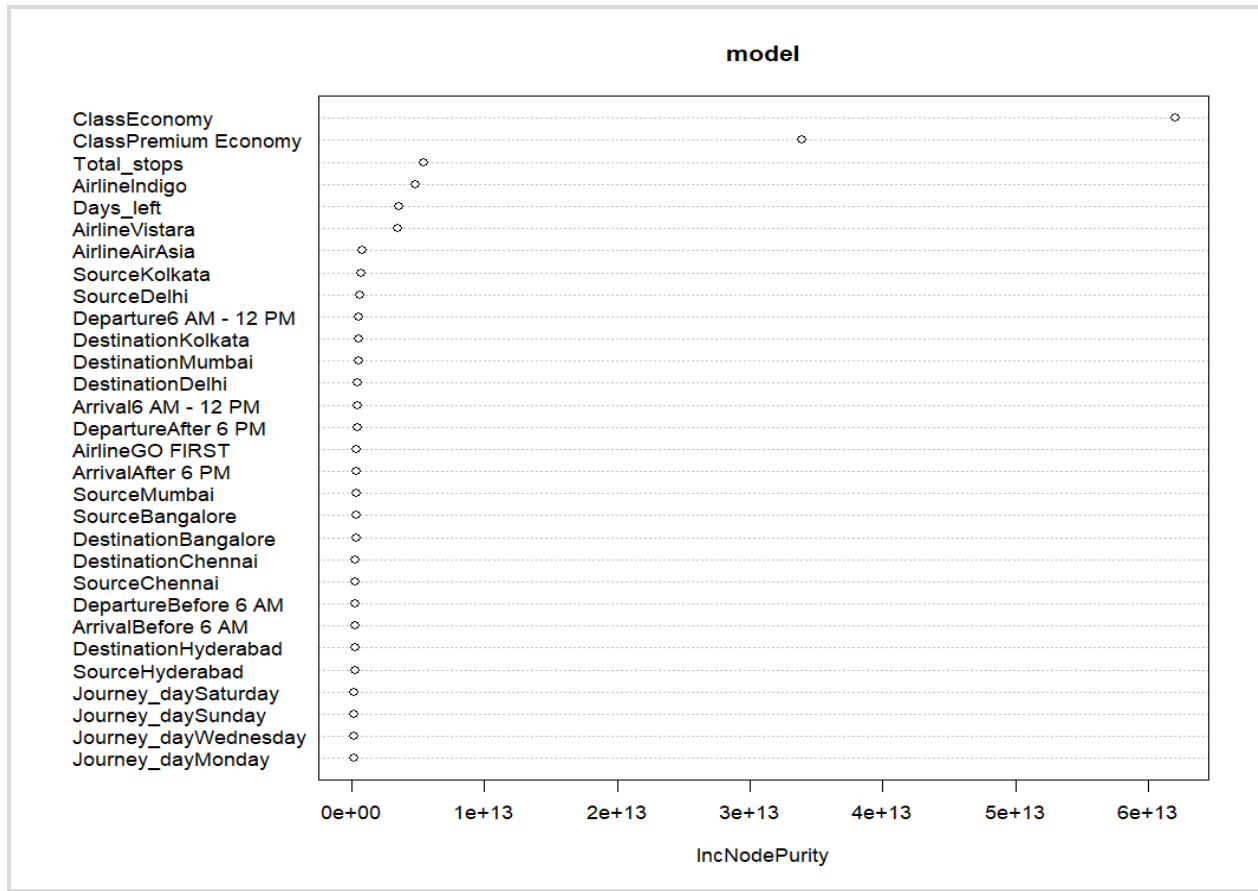


Figure 12. Importance of the Variables

As seen in Fig. 13 the R squared value is 0.929 for withheld data and 0.926 on the cross validated training data. This also means that model is not over-fitting as the R squared value for the withheld data is higher than on the training dataset.

```
Call:
randomForest(x = x_train, y = y_train, ntree = 25, do.trace = TRUE)
Type of random forest: regression
Number of trees: 25
No. of variables tried at each split: 12

Mean of squared residuals: 30335071
% Var explained: 92.63
RMSE      Rsquared      MAE
5428.6725860  0.9289331 2991.8254202
```

Figure 13. R squared value for the Random Forest model with training and withheld dataset

## Limitations

The limitations of the project are several from the data source and the model. The data available currently is very minimal to be able to predict the future prices as there are various factors apart from just the historical prices considered in the pricing of an airline ticket. A couple of variables being:

- 1) *Fuel Prices*: A major component of the ticket pricing strategy is the price of the aviation fuel and this can fluctuate at a global level causing the ticket price to vary. This has not been taken into consideration and can cause the model to not predict accurately at the time when these prices fluctuate at a global level.
- 2) *Consumer Demand*: The consumer demand isn't taken into account here showing the seasonal travel times and taking into account the daily happenings which can impact air travel.
- 3) *Competitive Pricing*: This is the strategy that airlines use to have a competitive edge over the other competitors in the market to manage and provide the most efficient performance at the most optimal prices such as RyanAir and Spirit Airlines in the USA.
- 4) *Limited Data*: This includes data only from 7 major airports in India and includes 9 airlines that operate domestically. This doesn't include the local airlines that run only in a particular region, the missing data of other destinations and a lack of equal observations for each airline and class of fare impacts the predictions.

In addition to the limitations arising from the data sources, there are a few limitations associated with using the Random Forest model.:

- 1) *Computational power*: Random Forest models can be computationally expensive and time-consuming to train, especially if there are a large number of features and data points.
- 2) *Bimodal distribution*: We see in the residual distribution that a majority of the data can in figure 2 that through the Lasso model, the entire data wasn't being explained by the data and causing high residuals. This problem was addressed by using a Random Forest model which gave a better result, however still around 8-10% of the data remains unexplained and predicts with a high value of residuals. This problem can be traced back to the class of travel as there are only 144 observations of the first class travel.

## Conclusion

Based on the above analysis, the predictions are highly accurate. In figure 13 we can see that the r-squared value for the model is 0.928, indicating that the model is able to explain ~93% of the dependent variable variance by the variance of the independent variable. However, the  $RMSE = 5428.67$  shows that the average difference between the predicted values and the actual values is 5428.672 pertaining to the residuals coming in from ~8% of unexplained data of the first class.

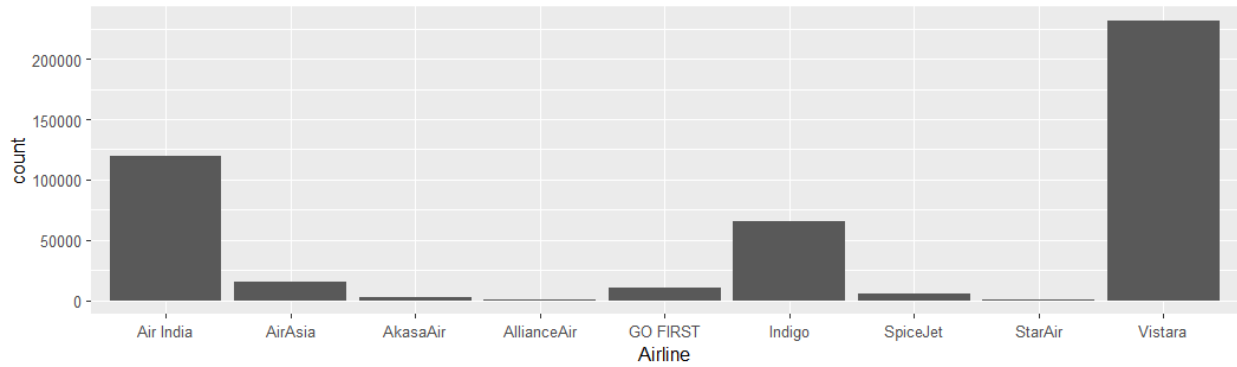
The analysis reveals that the model is highly precise for flight tickets in economy, premium economy, and business class, priced between INR 0-20000. Although the model's accuracy is limited by the available data, it performs well for most of the data. To further enhance the model's accuracy, future steps could involve segregating ticket classes and increasing the data coverage for each class and airline. This would facilitate better price predictions.



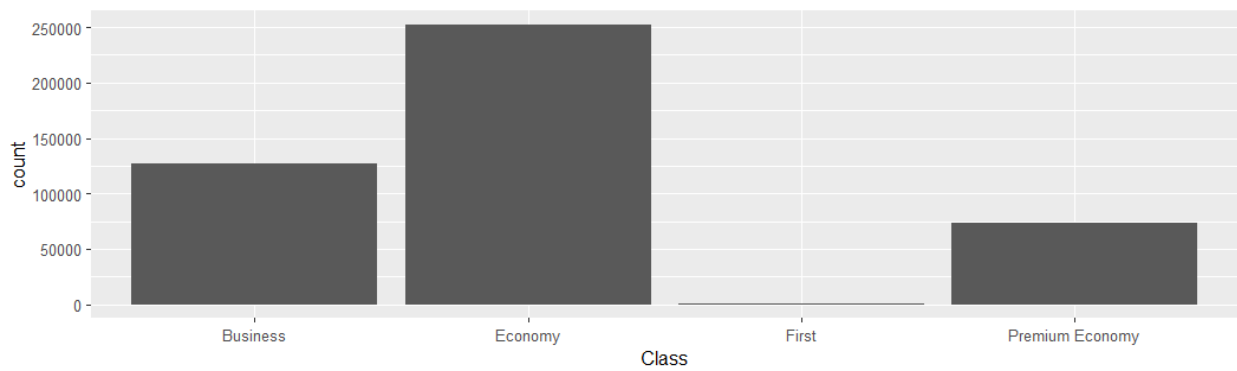
## References

- [1] “Airfare ML : Predicting Flight Fares.”  
<https://www.kaggle.com/datasets/yashdharme36/airfare-ml-predicting-flight-fares> (accessed May 01, 2023).
- [2] “Price Prediction,” *Hopper*. <https://hopper.com/product/price-prediction> (accessed Apr. 30, 2023).
- [3] “AirHint Flight Price Predictor - when to book cheap flight tickets.”  
<https://www.airhint.com/> (accessed Apr. 30, 2023).
- [4] “Flight Price Predictor | Flight Price Predictor.”  
<https://www.alternativeairlines.com/flight-price-predictor> (accessed Apr. 30, 2023).
- [5] P. Sarao and P. Samanta, “Flight Fare Prediction Using Machine Learning.” Rochester, NY, Oct. 20, 2022. doi: 10.2139/ssrn.4269263.
- [6] T. Wang *et al.*, “A Framework for Airfare Price Prediction: A Machine Learning Approach,” in *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*, Jul. 2019, pp. 200–207. doi: 10.1109/IRI.2019.00041.
- [7] “Build a user interface.” <https://shiny.rstudio.com/tutorial/written-tutorial/lesson2/> (accessed May 01, 2023).

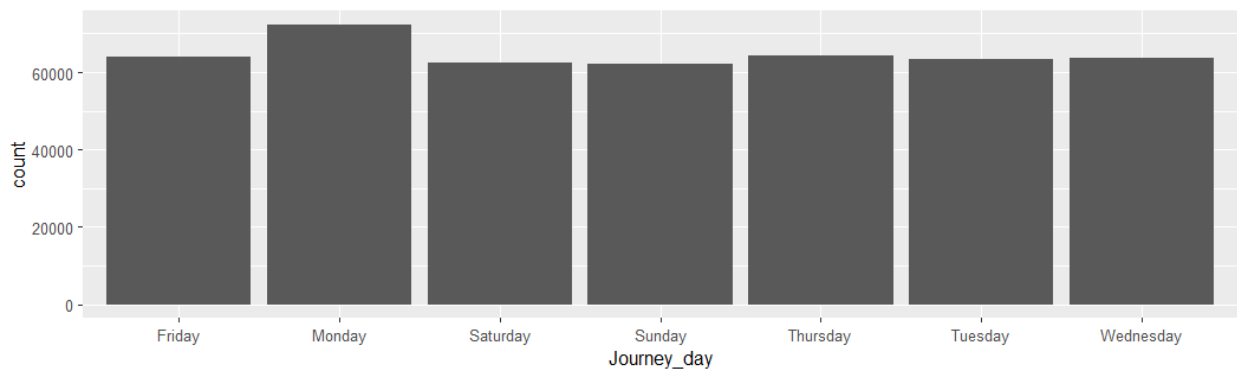
## Appendix



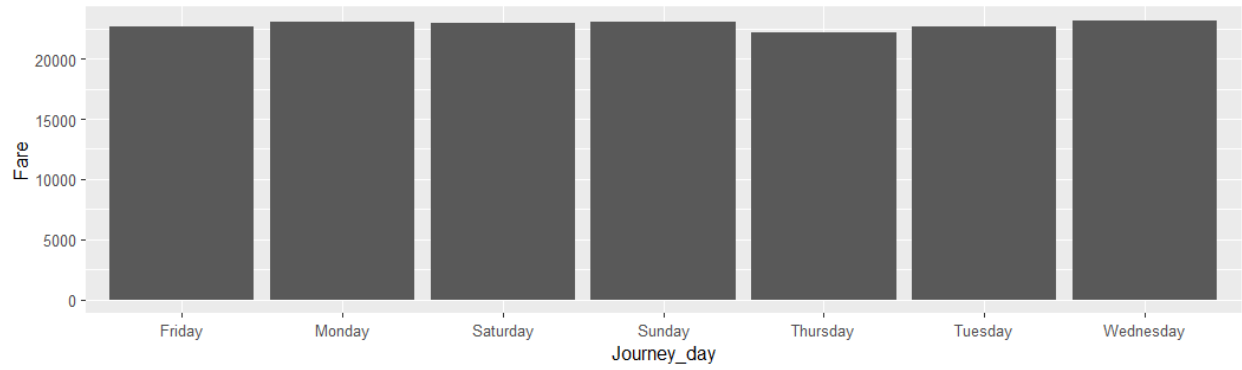
*Figure A.1 Count vs. Airline*



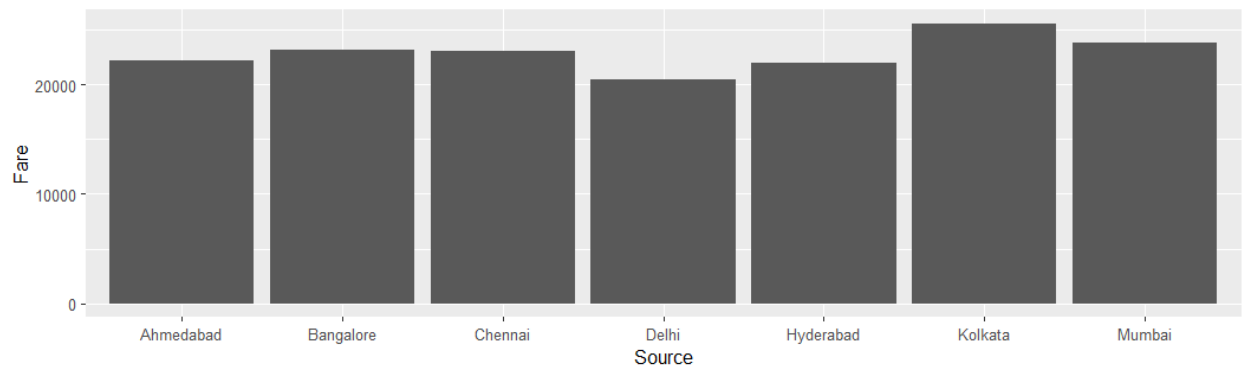
*Figure A.2 Count vs. Class*



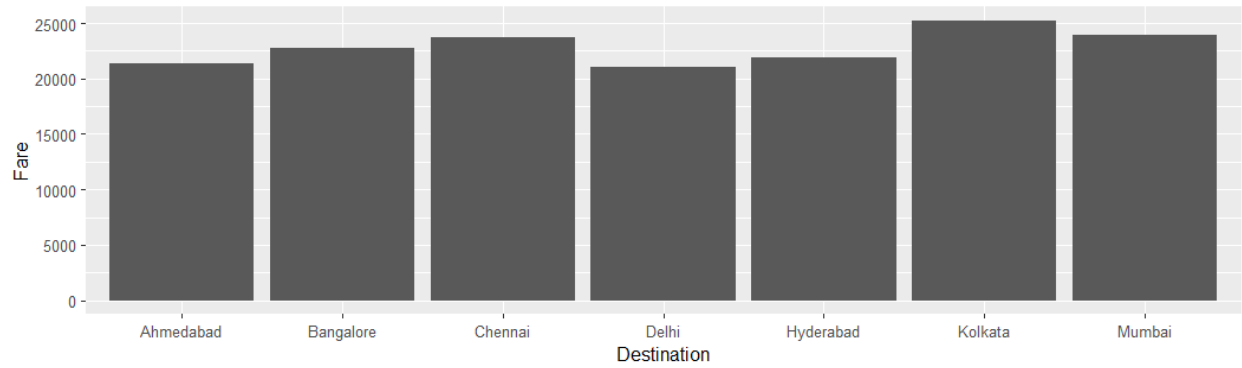
*Figure A.3 Count vs. Journey Day*



*Figure A.4 Fare vs. Journey Day*



*Figure A.5 Fare vs. Location*



*Figure A.6 Fare vs. Destination*

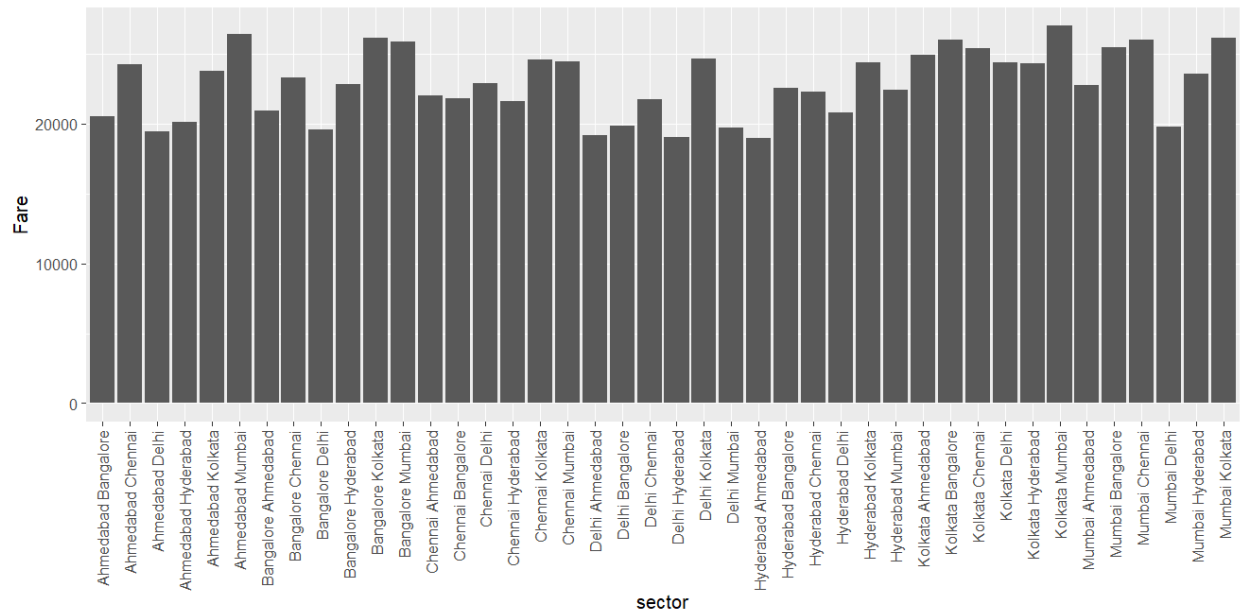


Figure A.7 Fare vs. Sector

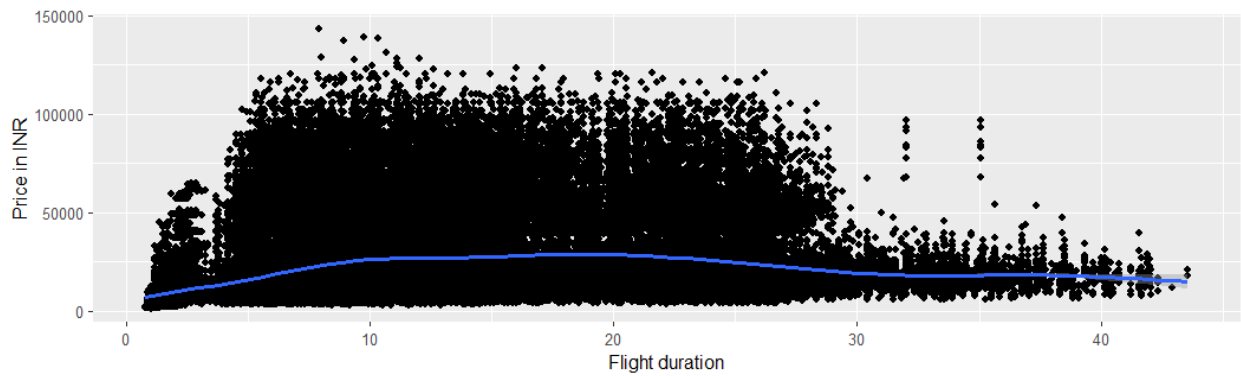


Figure A.8 Price vs. Flight Duration

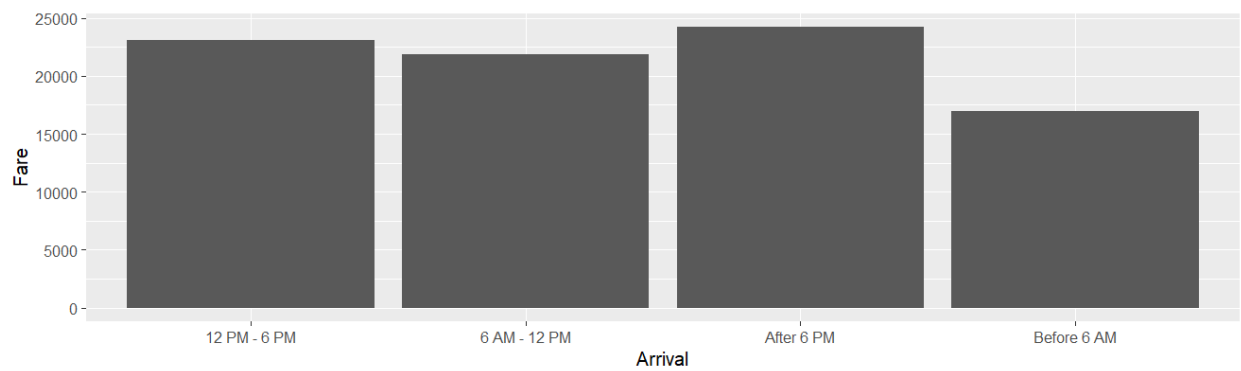


Figure A.9 Price vs. Arrival Time