Air Flights Price Prediction

Report Outline:

1. Introduction
2. Purpose of the project
3. Data
4. Data Cleaning and manipulation
5. Data Exploration
6. Models
7. Results and Discussion of Results
8. Limitations

Introduction:

- In today's world, Airlines implement dynamic pricing for their tickets and make their pricing decisions based on demand. Because each flight only has a limited number of seats to sell, such a complex method is used so that airlines can regulate the demand.
- In the case where demand is expected to exceed capacity, the airline may increase prices to decrease the rate at which seats fill.
- On the other hand, a seat that goes unsold represents a loss of revenue, and selling that seat for any price above the service cost for a single passenger would have been a better approach.

Purpose of the project:

- The purpose of this project is to predict how airline ticket prices change over time, extract the factors that influence these fluctuations, and describe how they're correlated.

Data:

- Dataset:

Dataset Link from Kaggle: Airlines Fare Prediction | Kaggle

```
> str(data)
'data.frame':   10683 obs. of  11 variables:
 $ Airline        : chr  "IndiGo" "Air India" "Jet Airways" "IndiGo" ...
 $ Date_of_Journey: chr  "24/03/2019" "1/05/2019" "9/06/2019" "12/05/2019" ...
 $ Source         : chr  "Banglore" "Kolkata" "Delhi" "Kolkata" ...
 $ Destination    : chr  "New Delhi" "Banglore" "Cochin" "Banglore" ...
 $ Route          : chr  "BLR ? DEL" "CCU ? IXR ? BBI ? BLR" "DEL ? LKO ? BOM ? COK" "CCU ? NAG ? BLR" ...
 $ Dep_Time       : chr  "22:20" "05:50" "09:25" "18:05" ...
 $ Arrival_Time   : chr  "01:10 22 Mar" "13:15" "04:25 10 Jun" "23:30" ...
 $ Duration       : chr  "2h 50m" "7h 25m" "19h" "5h 25m" ...
 $ Total_Stops    : chr  "non-stop" "2 stops" "2 stops" "1 stop" ...
 $ Additional_Info: chr  "No info" "No info" "No info" "No info" ...
 $ Price          : int  3897 7662 13882 6218 13302 3873 11087 22270 11087 8625 ...
```

Data cleaning and manipulation

As a part of data cleaning and manipulation we have performed the following actions:

- Airlines- We have counted the number of flights per airlines and removed airlines that has small number of observations(count < 20)
- Date of Journey- Separated Date of Journey by Day, Month and Year, transformed 'day' into 'Weekday'
- Total_Stops- Dropped all observations with 4 stops

Data cleaning and manipulation

- Destination- New Delhi, Bangalore and Cochin.
- Departure- Categorized departure time into Morning, Afternoon, Evening and Night.
  - 12 am to 6 am – Night
  - 6 am to 12 pm - Morning
  - 12 pm to 6pm – Afternoon
  - 6pm to 12 am – Evening

	Airline	Day	Month	Source	Destination	Dep_Time	Duration	Total_Stops	Price	Days_of_week
1	IndiGo	24	03	Banglore	New Delhi	22:20	2h 50m	non-stop	3897	Wednesday
2	Air India	1	05	Kolkata	Banglore	05:50	7h 25m	2 stops	7662	Sunday
3	Jet Airways	9	06	Delhi	Cochin	09:25	19h	2 stops	13882	Saturday
4	IndiGo	12	05	Kolkata	Banglore	18:05	5h 25m	1 stop	6218	Sunday
5	IndiGo	01	03	Banglore	New Delhi	16:50	4h 45m	1 stop	13302	Tuesday
6	SpiceJet	24	06	Kolkata	Banglore	09:00	2h 25m	non-stop	3873	Thursday

	Airline	Day	Month	Source	Destination	Duration	Total_Stops	Price	Days_of_week	Departure
1	IndiGo	24	March	Banglore	New Delhi	2h 50m	non-stop	3897	Wednesday	Evening
2	Air India	1	May	Kolkata	Banglore	7h 25m	2 stops	7662	Sunday	Night
3	Jet Airways	9	June	Delhi	Cochin	19h	2 stops	13882	Saturday	Morning
4	IndiGo	12	May	Kolkata	Banglore	5h 25m	1 stop	6218	Sunday	Evening
5	IndiGo	01	March	Banglore	New Delhi	4h 45m	1 stop	13302	Tuesday	Afternoon
6	SpiceJet	24	June	Kolkata	Banglore	2h 25m	non-stop	3873	Thursday	Morning

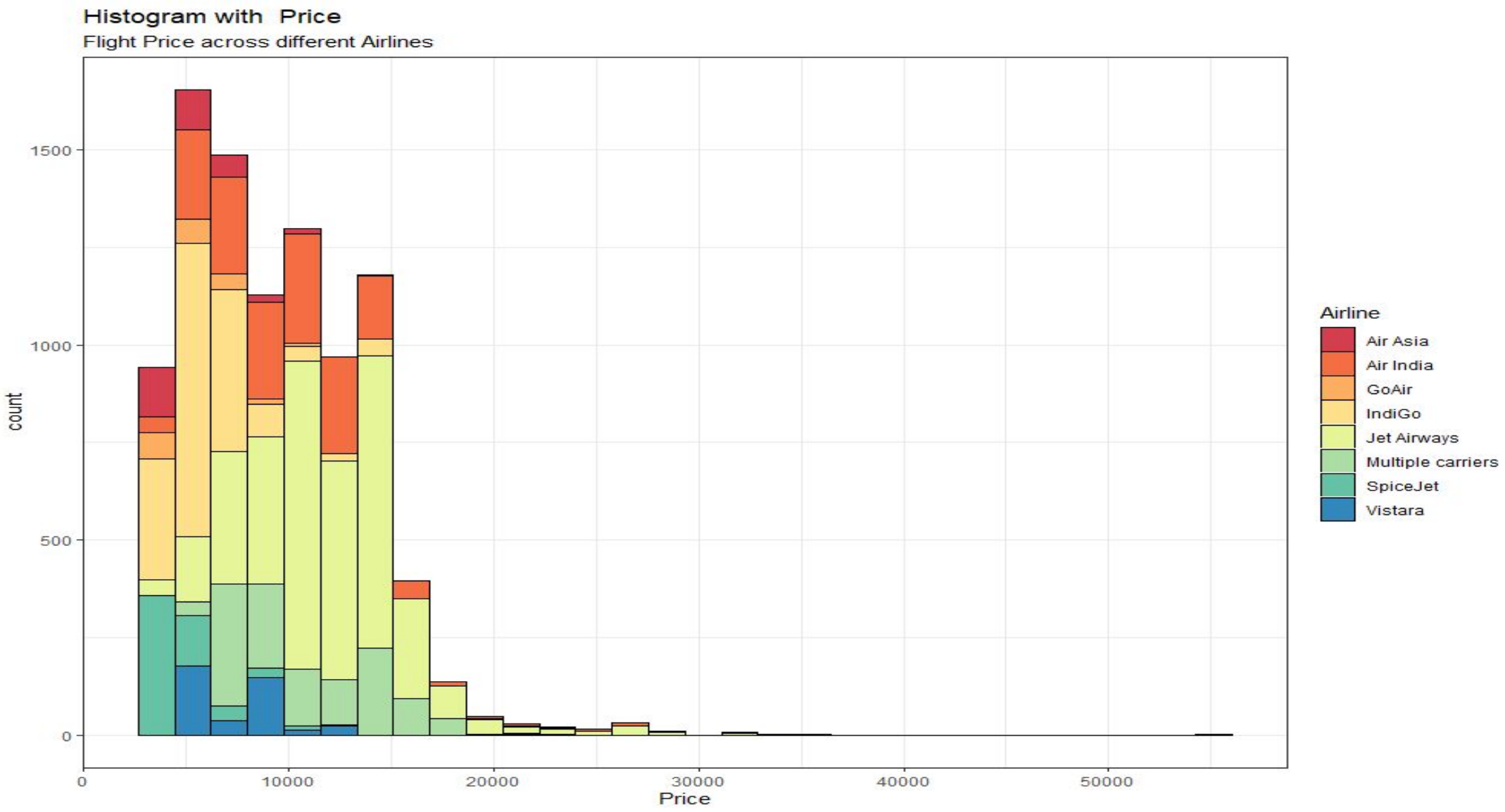
- Duration- Hours and minutes in duration variable has been converted to minutes.

Data cleaning and manipulation

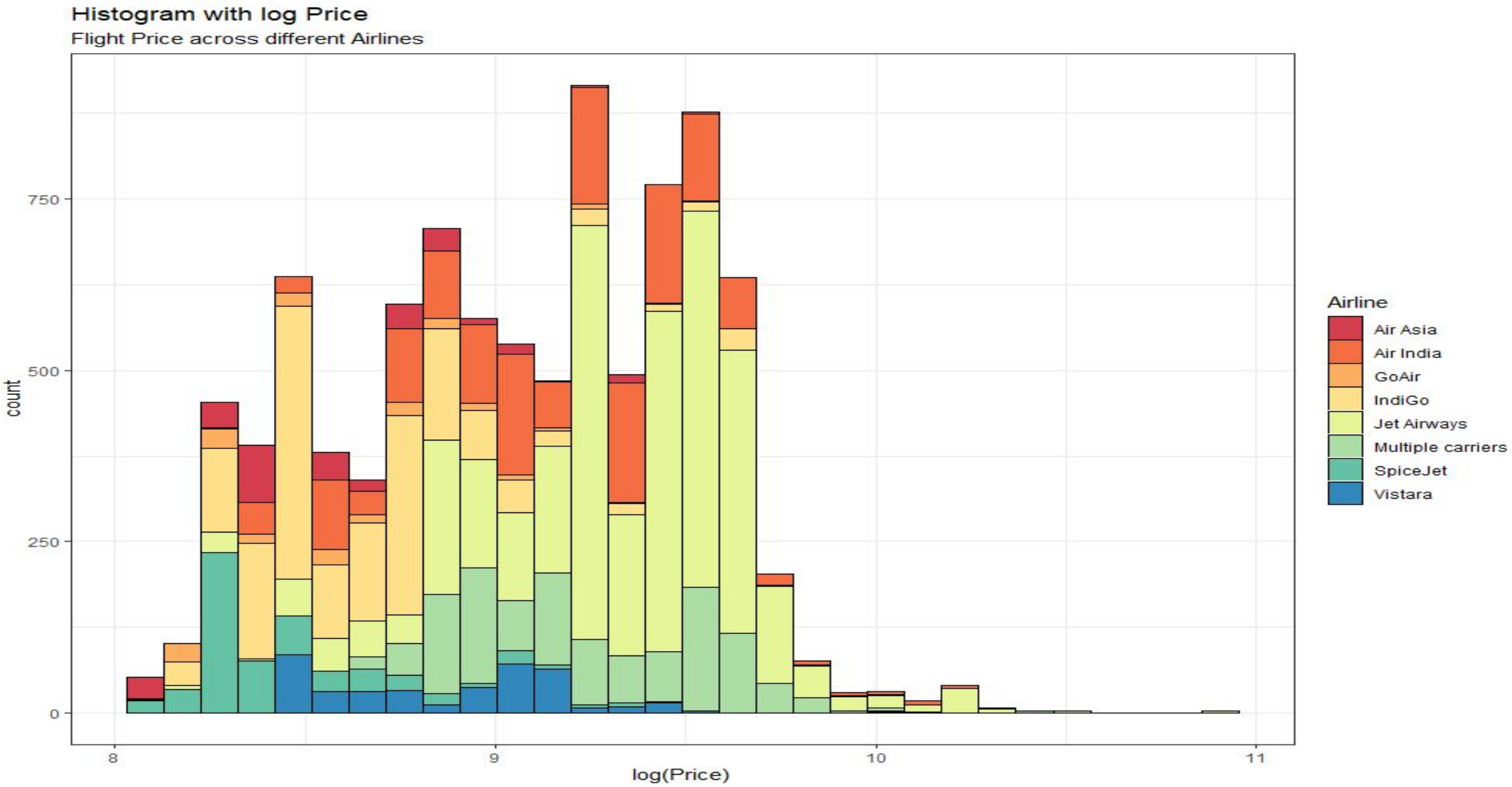
- Between Source and Destination, we chose Destination over Source as the routes were same.
- Another column 'Additional_Info' was dropped since most of its rows had 'No info'.
- Duplicated rows were dropped which were 220 in count.

Data Explorations & Visualizations

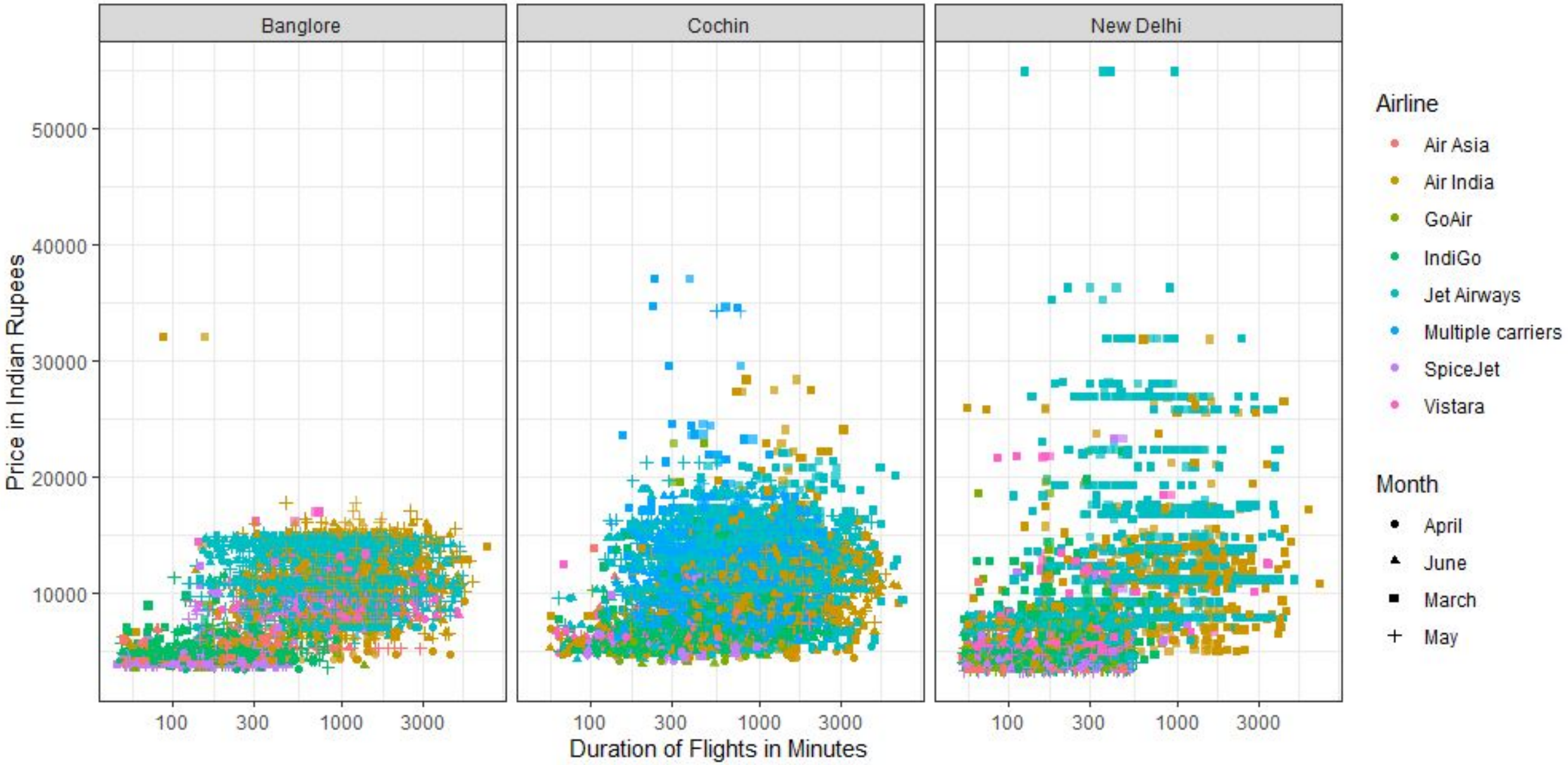
Distribution of Price variable



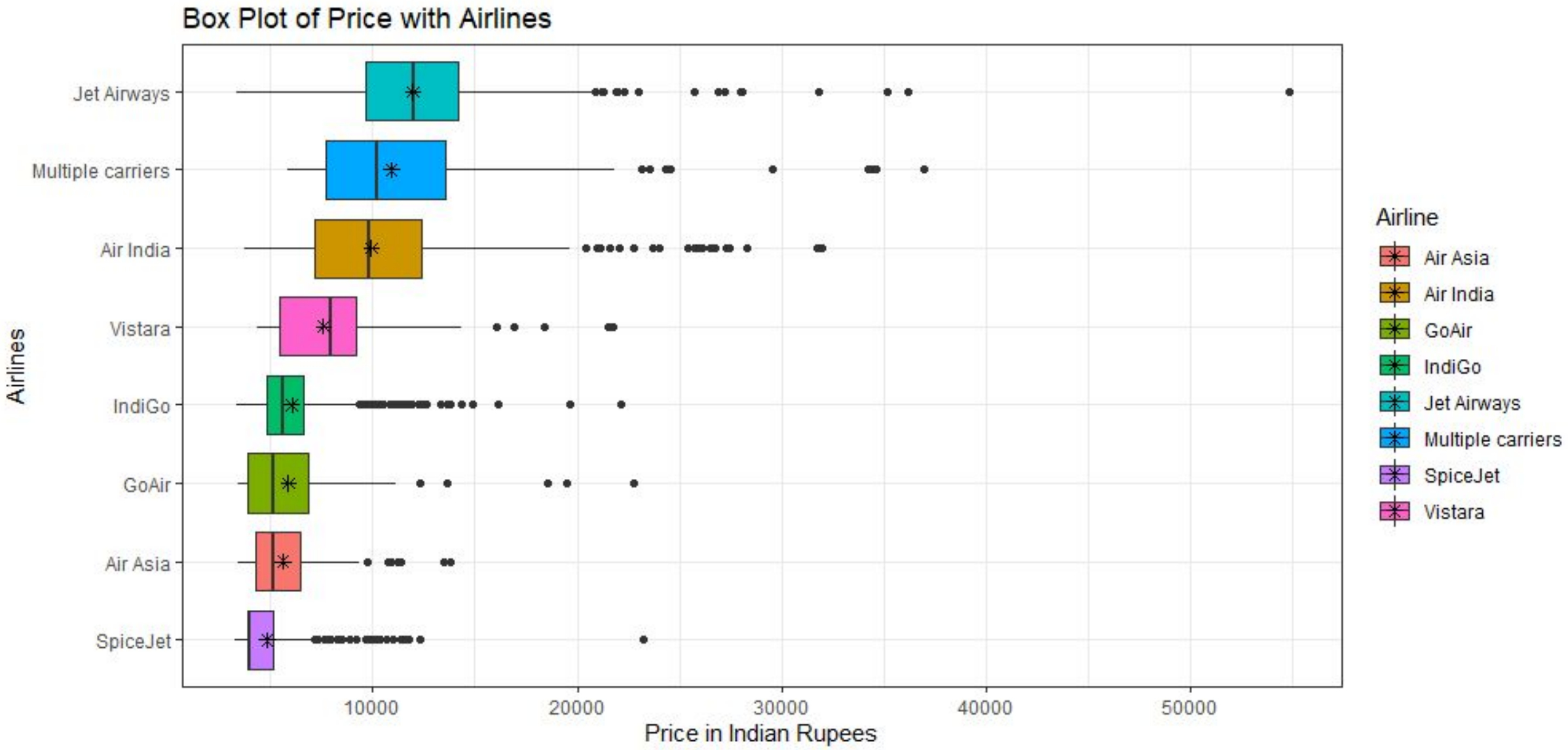
Distribution of log of Price



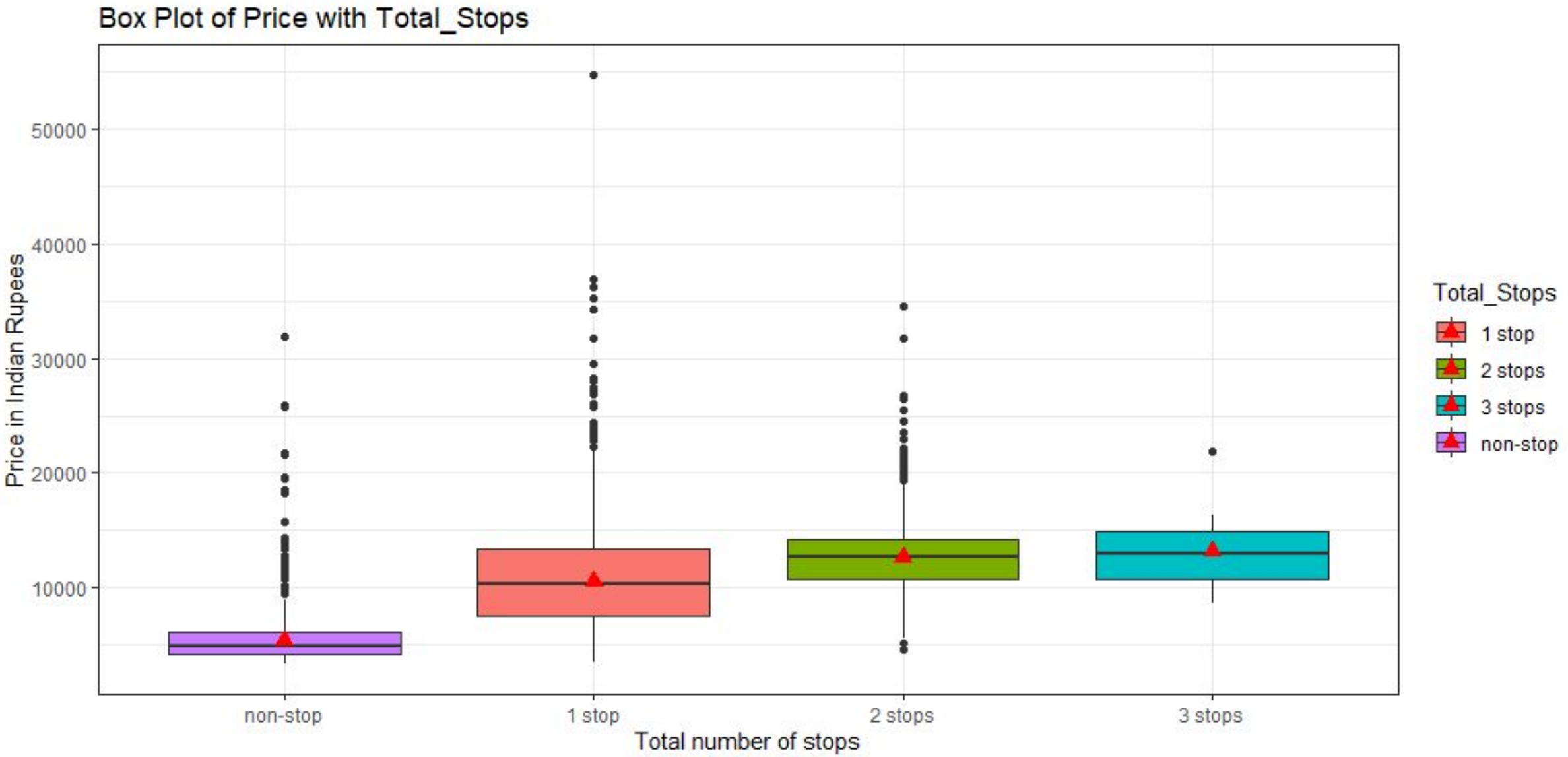
Scatter plot: Price Vs Duration



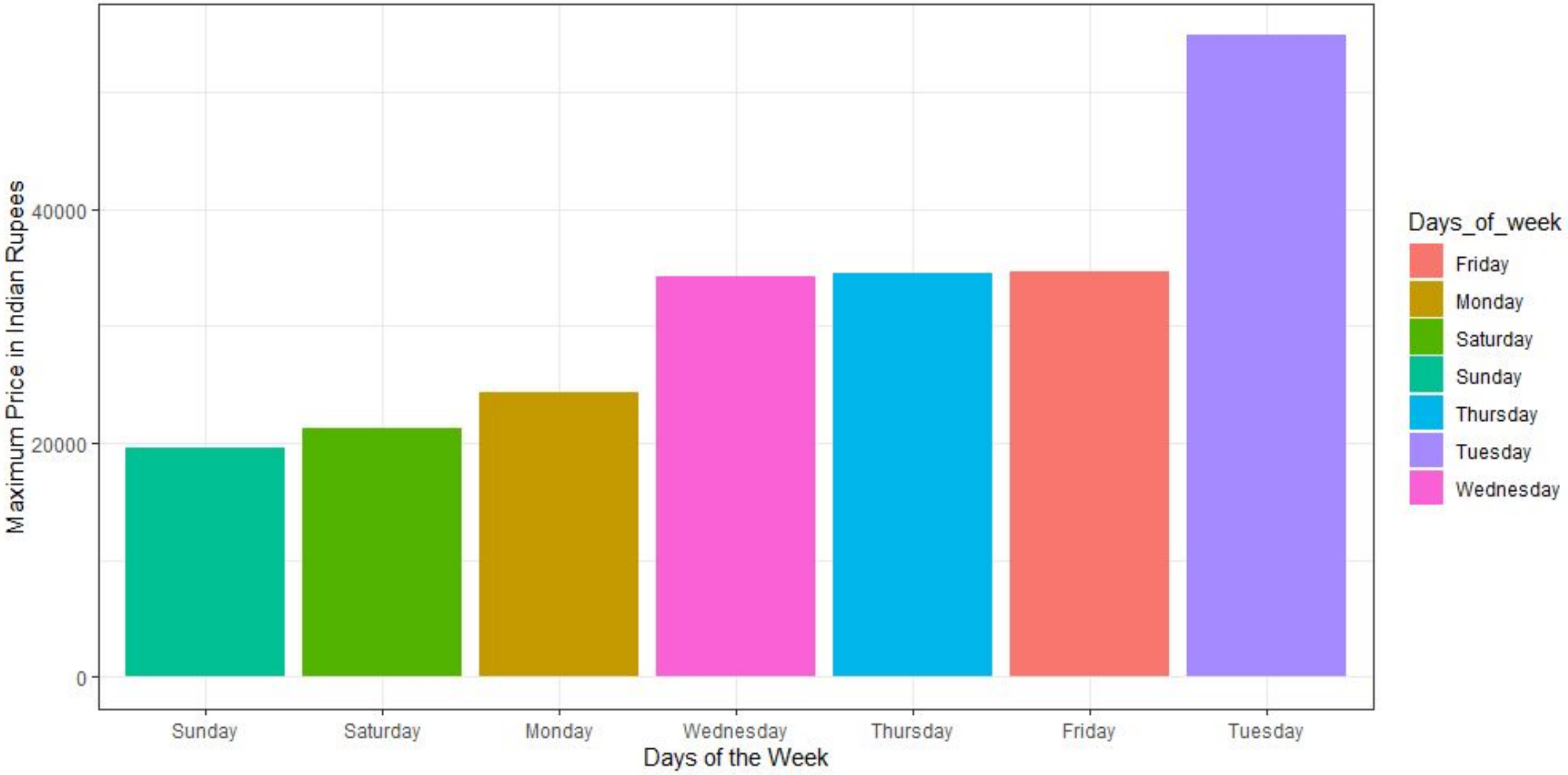
Boxplot of Airline Versus Price



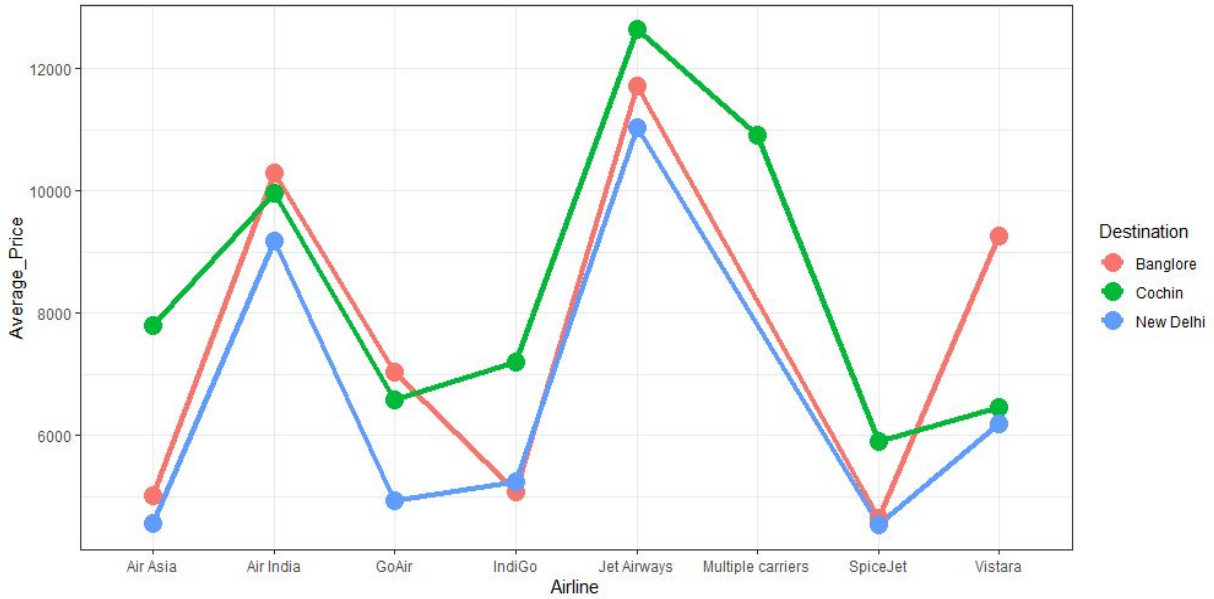
Boxplot for Price Vs Total Stops

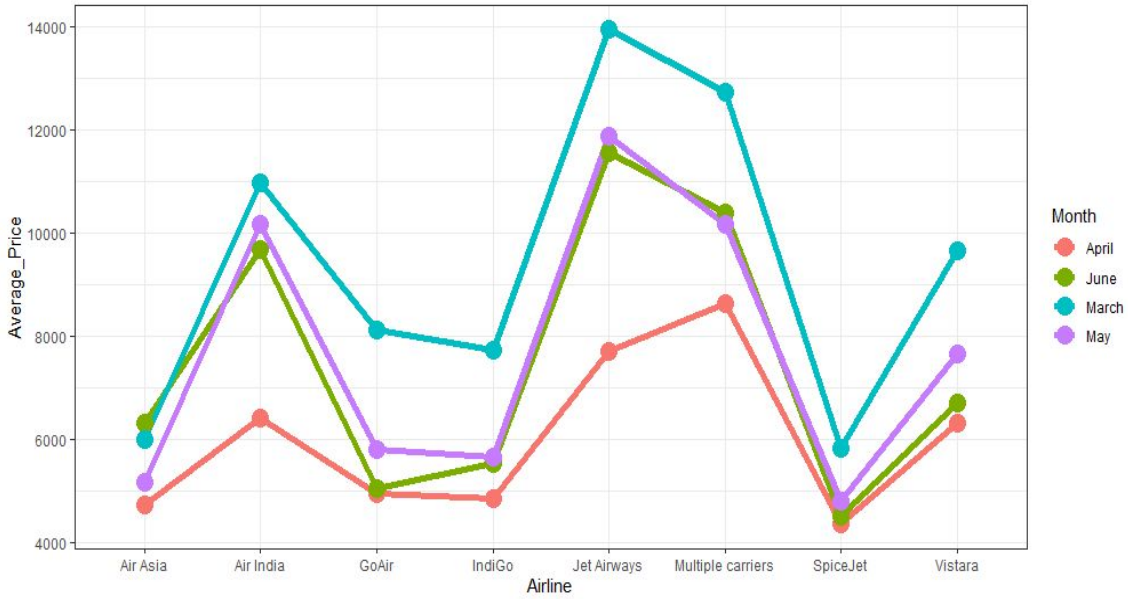


Maximum Price Vs Days of Week



Average Price with respect to Destination & Months





Training and Testing

- We splitted the data into train and test.
- 30 % of data is used in testing and 70 % of data is used for training in all the models.

Models

The models used in the current project are as follows:

1. Linear Regression

2. Regression Tree

3. Random Forest

4. Gradient Boosting

Linear Regression

- R squared: 0.5796
- Adjusted: 0.578
- RMSE for this train model is 2821

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard       2821.
```

- RMSE for this test model is 2791

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard       2791.
```

```
> summary(LR)

Call:
lm(formula = Price ~ Airline + Month + Destination + Total_Stops +
    Days_of_week + Departure + Duration, data = data9_train)

Residuals:
   Min     1Q Median     3Q    Max
 -7679  -1655    -94   1353  39783

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               6140.2485   257.5794  23.838  < 2e-16 ***
AirlineAir India          1967.2941   216.9589   9.068  < 2e-16 ***
AirlineGoAir               287.3881   307.1277   0.936  0.34945
AirlineIndiGo              317.7121   206.2762   1.540  0.12355
AirlineJet Airways        4498.4458   203.4851  22.107  < 2e-16 ***
AirlineMultiple carriers  3505.9469   224.5360  15.614  < 2e-16 ***
AirlineSpiceJet             94.7418   235.4279   0.402  0.68739
AirlineVistara            1983.6105   252.7092   7.849 4.85e-15 ***
MonthJune                 1308.1559   135.3597   9.664  < 2e-16 ***
MonthMarch                3247.5308   136.1088  23.860  < 2e-16 ***
MonthMay                  1416.3265   140.5578  10.076  < 2e-16 ***
DestinationCochin         -256.6925    98.2504  -2.613  0.00901 **
DestinationNew Delhi       -47.2176   118.6614  -0.398  0.69070
Total_Stops2 stops        2394.4354   122.6416  19.524  < 2e-16 ***
Total_Stops3 stops        4177.1099   599.9692   6.962 3.68e-12 ***
Total_Stopsnon-stop      -3375.2171   130.9550 -25.774  < 2e-16 ***
Days_of_weekMonday         998.2427   161.8891   6.166 7.41e-10 ***
Days_of_weekSaturday     -1385.7283   137.5273 -10.076  < 2e-16 ***
Days_of_weekSunday        -100.0013   153.4494  -0.652  0.51462
Days_of_weekThursday       359.1008   147.9896   2.427  0.01527 *
Days_of_weekTuesday       1328.1965   148.3141   8.955  < 2e-16 ***
Days_of_weekWednesday     -209.4882   141.4149  -1.481  0.13856
DepartureEvening           -55.3043   101.3563  -0.546  0.58533
DepartureMorning          -250.4413    90.3996  -2.770  0.00561 **
DepartureNight            -257.0458   136.6941  -1.880  0.06009 .
Duration                    -0.1835     0.1075  -1.708  0.08776 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2826 on 6527 degrees of freedom
Multiple R-squared:  0.5796,    Adjusted R-squared:  0.578
F-statistic: 359.9 on 25 and 6527 DF,  p-value: < 2.2e-16
```

Linear Regression with log(Price)

- More significant variables when log(Price) is taken
- Model significance improved
- Multiple r squared = 0.7092
- Adjusted = 0.708
- RMSE for this train model is 2794

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard       2794.
```

- RMSE for this test model is 2748

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard       2748.
```

```
Call:
lm(formula = log(Price) ~ Airline + Month + Destination + Total_Stops +
    Days_of_week + Departure + Duration, data = data9_train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.81882 -0.15782  0.00806  0.15686  1.53845

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              8.674e+00  2.267e-02 382.687  < 2e-16 ***
AirlineAir India         2.992e-01  1.909e-02  15.670  < 2e-16 ***
AirlineGoAir             1.513e-02  2.703e-02   0.560 0.575707
AirlineIndiGo            6.794e-02  1.815e-02   3.743 0.000183 ***
AirlineJet Airways       5.331e-01  1.791e-02  29.773  < 2e-16 ***
AirlineMultiple carriers 4.169e-01  1.976e-02  21.098  < 2e-16 ***
AirlineSpiceJet         -3.193e-02  2.072e-02  -1.541 0.123285
AirlineVistara           3.065e-01  2.224e-02  13.781  < 2e-16 ***
MonthJune                1.577e-01  1.191e-02  13.236  < 2e-16 ***
MonthMarch               3.244e-01  1.198e-02  27.085  < 2e-16 ***
MonthMay                 1.655e-01  1.237e-02  13.381  < 2e-16 ***
DestinationCochin       -1.215e-02  8.646e-03  -1.405 0.160135
DestinationNew Delhi    -3.432e-02  1.044e-02  -3.287 0.001020 **
Total_Stops2 stops       2.116e-01  1.079e-02  19.608  < 2e-16 ***
Total_Stops3 stops       3.662e-01  5.280e-02   6.937 4.40e-12 ***
Total_Stopsnon-stop     -4.143e-01  1.152e-02 -35.947  < 2e-16 ***
Days_of_weekMonday       6.864e-02  1.425e-02   4.818 1.48e-06 ***
Days_of_weekSaturday    -1.449e-01  1.210e-02 -11.975  < 2e-16 ***
Days_of_weekSunday      -3.087e-02  1.350e-02  -2.286 0.022292 *
Days_of_weekThursday     1.464e-02  1.302e-02   1.124 0.261136
Days_of_weekTuesday      8.394e-02  1.305e-02   6.432 1.35e-10 ***
Days_of_weekWednesday   -3.220e-02  1.244e-02  -2.588 0.009684 **
DepartureEvening        -1.741e-02  8.919e-03  -1.952 0.050926 .
DepartureMorning        -1.960e-02  7.955e-03  -2.464 0.013783 *
DepartureNight          -3.439e-02  1.203e-02  -2.859 0.004266 **
Duration                -4.425e-06  9.458e-06  -0.468 0.639954
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2487 on 6527 degrees of freedom
Multiple R-squared:  0.7092,    Adjusted R-squared:  0.708
F-statistic: 636.6 on 25 and 6527 DF,  p-value: < 2.2e-16
```

Regression Tree

```
> #Decision Tree Train
> Decision_Tree_Train <- rpart(Price~ Airline + Month + Destination + Total_Stops + Days_of_week +
  Departure + Duration,data9_train, method = "anova")
```

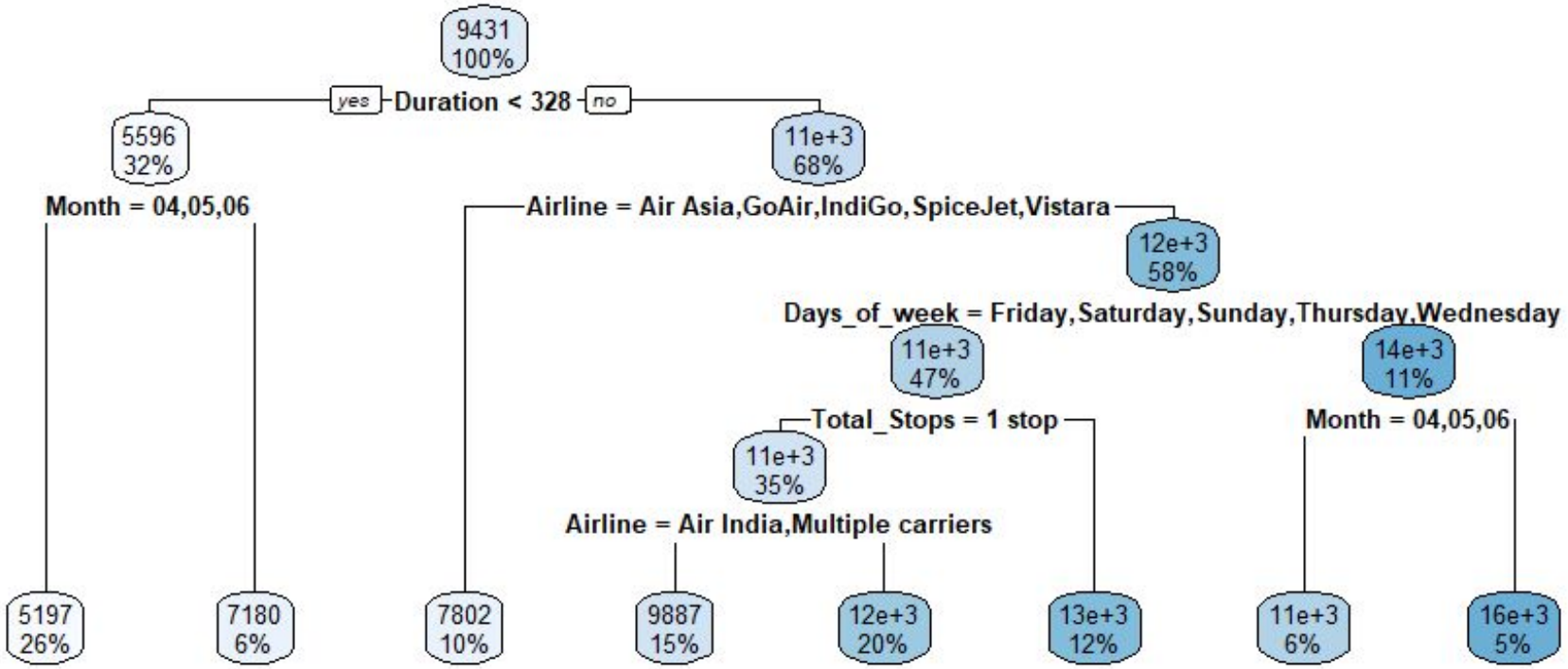
RMSE for train data is 2914

RMSE for test data is 2867

```
Variable importance
    Duration  Total_Stops      Airline  Destination        Month Days_of_week
          30           26           19           14            8            3
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard       2914.
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard       2867.
```



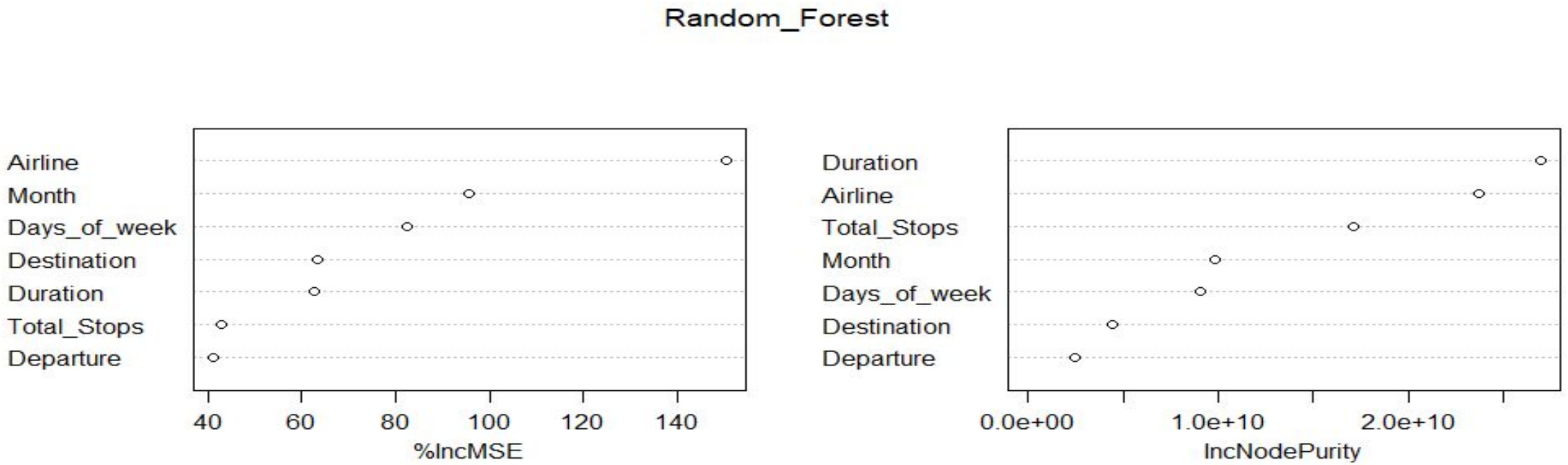
Random Forest

```
> # RANDOM FOREST TRAIN
> Random_Forest <- randomForest(Price~ Airline + Month + Destination + Total_Stops + Days_of_week +
 Departure + Duration,data9_train, ntree = 1000, importance=TRUE)
```

RMSE for train Model is 2383

RMSE for test Model is 2376

```
> rmse(data9_train_forest1, Price, Predicted_Price)
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard       2383.
> rmse(data9_test_forest1, Price, Predicted_Price)
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard       2376.
```



Gradient Boosting

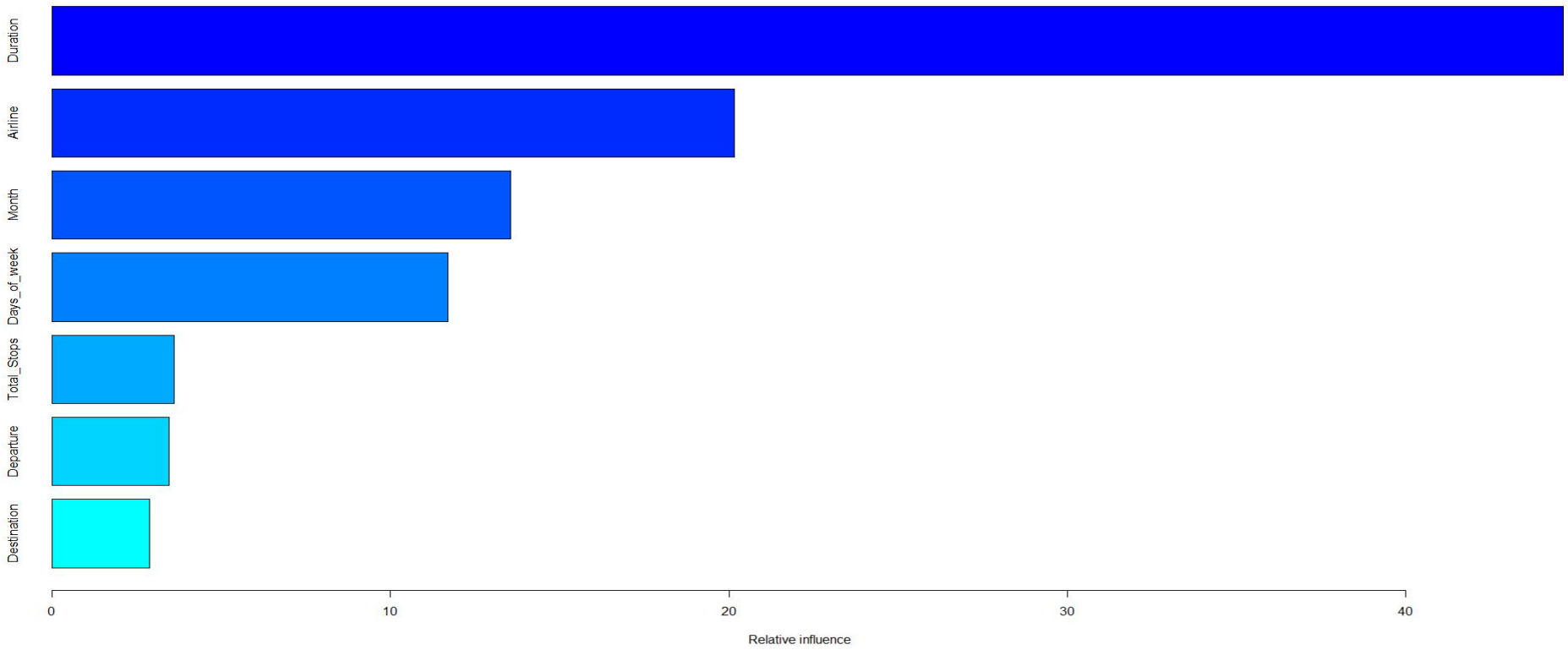
```
> # GRADIENT BOOST MODEL Train
> data9_train_gradient <- gbm(Price~ Airline + Month + Destination + Total_Stops + Days_of_week + Departure +
  Duration,data9_train, distribution = "gaussian",  n.trees = 10000, shrinkage = 0.01, interaction.depth =  4)
> summary(data9_train_gradient)
                      var   rel.inf
Duration         Duration 44.191563
Airline           Airline 20.706609
Month               Month 12.946890
Days_of_week Days_of_week 12.064909
Departure       Departure  3.568378
Total_Stops   Total_Stops  3.536975
Destination   Destination  2.984677
```

RMSE for Train
1985

RMSE for Test
2306

```
> rmse(data9_test_gradient1, Price, Predicted_Price)
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard       2306.
```

```
> Gradient_Train_RMSE
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard       1985.
```



Model Comparison Using RMSE

Model	Train RMSE	Test RMSE
Linear Regression	2821	2791
Linear Regression (log Price)	2794	2748
Decision Tree	2914	2867
Random Forest	2383	2376
Gradient Boosting	1985	**2306**

Discussion of Results

Duration vs. Price:

- Slightly positively correlated
- Higher Prices for March
- Outliers - Banglore and New Delhi

Airline vs. Price:

- Expensive- Jet Airways

Total Stops vs Price:

- Highest price - 1 stop
- Highest average price - 3 stops
- Maximum Price/Days of Week:
  Tuesday (almost double)

Airline vs. Average Price (in terms of Month):

- All Airlines had higher average prices for March

Linear Model:

All variables significant except Duration and Departure

Linear Model (log of Price):

All variables significant except Duration

Regression Tree:

Most significant - Duration

Random Forest:

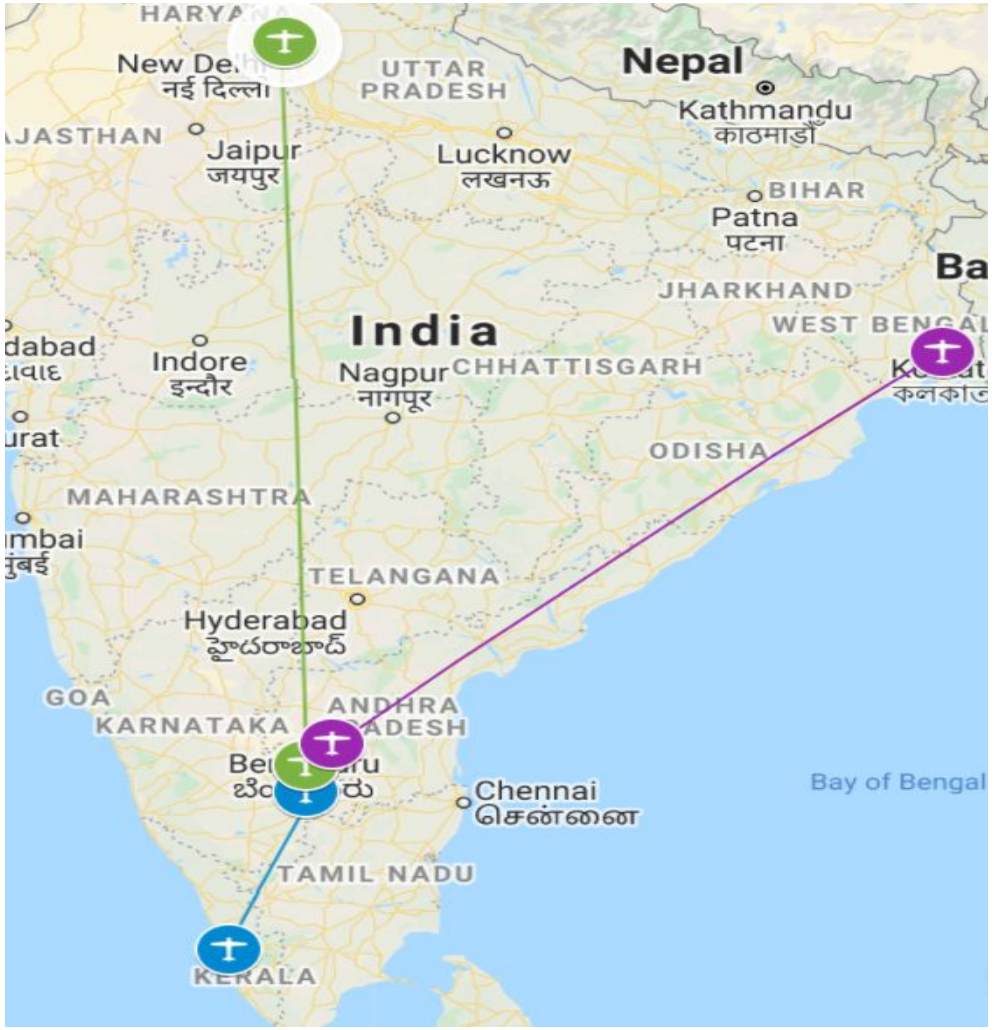
Significant - Airline and Duration

Gradient Boosting:

Most significant - Duration

Limitations:

- Data only for four months of the year 2019
- The finalized dataset contains only three routes
  - Bangalore to New Delhi
  - Kolkata to Bangalore
  - New Delhi to Cochin
- No information about the class (economy/business) by which the passenger is travelling
- No information about the day of purchase.
  - Assumption - the time in between the day of booking and the day of departures could affect the price



Thank you for your time!