FLIGHT PRICE PREDICTION

Project report submitted in partial fulfillment of the requirement for the degree of Bachelor of Technology

in

Computer Science and Engineering

By

Pranav Verma (181333)

Under the supervision
of
Dr. Kapil Sharma

Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology Waknaghat, Solan-173234,Himachal Pradesh

CERTIFICATE OF ORIGINALITY

This is to certify that the work which is being presented in the project report titled Steganography of images in completion of the requirements for the degree of B.Tech in Computer Science And Engineering and submitted to the Department of Computer Science And Engineering, Jaypee University of Information Technology, Waknaghat is an authentic record of work carried out by Pranav during the period from July 2021 to December 2021 underthe supervision of Mr. Deepak Gupta, Department of Computer Science and Engineering, Jaypee University of Information Technology, Waknaghat.

Pranav Verma
(181333)

The above statement made is correct to the best of my knowledge.

ACKNOWLEDGEMENT

To begin, I would want to offer my heartfelt gratitude and appreciation to Almighty God for His wonderful grace, which has enabled us to successfully finish the project work. I am really grateful and like to express my heartfelt gratitude to my project supervisor, Department of CSE, Jaypee University of Information Technology, Wakhnaghat. My supervisor's extensive knowledge and deep interest in the topic of steganography enabled me to complete this assignment. Her unending patience, intellectual direction, consistent encouragement,persistent and vigorous supervision, constructive criticism, helpful counsel, reading many poor versions and correcting them at all stages, and reading and correcting them at all stages enabled us to accomplish this project. I'd like to convey my heartfelt appreciation to my project supervisior Department of CSE, for his kind assistance in completing my assignment. I would also want to express my gratitude to everyone who has directly or indirectly assisted me in making this project a success. This one-of-a-kind In this case, I'd want to thank the different staff members, both teaching and non-teaching, who have provided convenient assistance and assisted my project. Finally, I must express my gratitude for my parents'unwavering support and patience.

Pranav Verma
(181333)

About the Project

Optimal timing for airline ticket purchasing from the consumer’s perspective is challenging principally because buyers have insufficient information for reasoning about future price movements. In this project we simulate various models for computing expected future pricesand classifying whether this is the best time to buy the ticket.

Why this Project?

Anyone who has booked a flight ticket knows how unexpectedly the prices vary. Airlines useusing sophisticated quasi-academic tactics which they call **"revenue management"** or **"yield management"**. The cheapest available ticket on a given flight getsmore and less expensive over time. This usually happens as an attempt to maximize revenuebased on -

1. Time of purchase patterns (making sure last-minute purchases are expensive)
2. Keeping the flight as full as they want it (raising prices on a flight which is filling up in order to reduce sales and hold back inventory for those expensive last-minute expensive purchases)

FLIGHT TRENDS

Do airfares change frequently? Do they move in small increments or in large jumps? Do theytend to go up or down over time?

BEST TIME TO BUY

What is the best time to buy so that the consumer can save the most by taking the least risk?

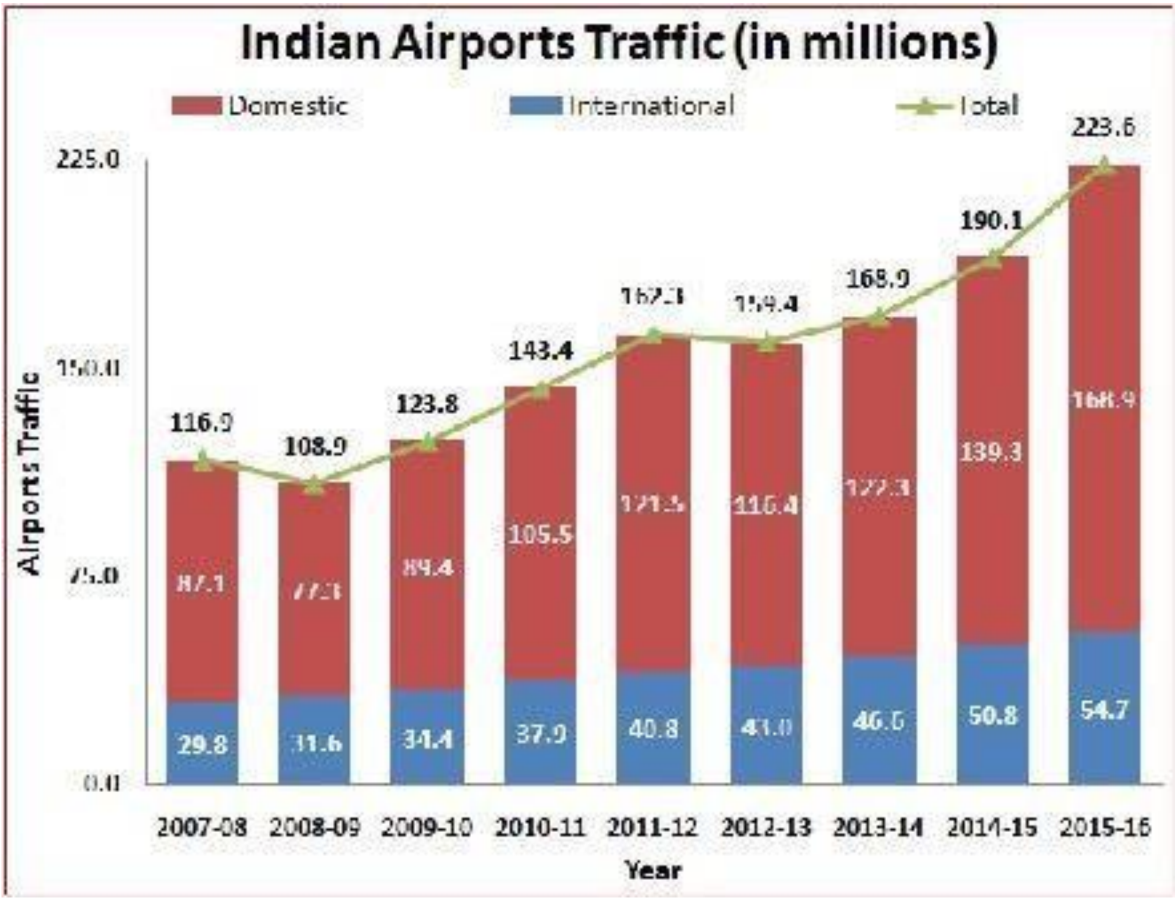
VERIFYING MYTHS

Does price increase as we get near to departure date? Is Indigo cheaper than Jet Airways? Aremorning flights expensive?

Problem Validation & Market Research

According to a report, India’s civil aviation industry is on a high-growth trajectory. India aims to become the **third-largest aviation market by 2020** and the largest by 2030. Indian domestic air traffic is expected to cross **100 million passengers by FY2017**, compared to 81million passengers in 2015, as per Centre for Asia Pacific Aviation (CAPA).

According to Google Trends, the search term - **"Cheap Air Tickets"** is most searched in India. Moreover, as the middle-class of India is exposed to air travel, consumers hunting forcheap prices increases.



Technical Aspects

The project is basically **machine learning & statistic intensive**. We used Python & R for theimplementation of the models & automation.

- Automated Script to Collect Historical Data

For any prediction/classification problem, we need historical data to work with. In this project, past flight prices for each route collected on a daily basis is needed.

Manually collecting data daily is not efficient and thus a python script was run on a remote server which collected prices daily at specific time.

- Cleaning & Preparing Data

After we have the data, we need to clean & prepare the data according to the model's requirements. In any machine learning problem, this is the step that is the most important and the most time consuming. We used various statistical techniques & logics and implemented them using built-in R packages.

- Analyzing & Building Models

Data preparation is followed by analyzing the data, uncovering hidden trends and then applying various predictive & classification models on the training set.

- Merging Models & Accuracy Calculation

Having built various models, we now have to test the models on our testing set and come up with the most suitable metric to calculate the accuracy. Moreover, many a times, merging models and predicting a cumulative target variable proves to be more accurate.

Data Collection

Data Collection is one of the most important aspect of this project. There are various sources of airfare data on the Web, which we could use to train our models. A multitude of consumertravel sites supply fare information for multiple routes, times, and airlines.

We tried various sources ranging from many APIs to scraping consumer travel websites. Inthis section we have discussed in detail on these various sources and the importance of parameters that are collected.

1. Data Sources
2. What and When we collected?
3. Automating the Collection

Data Sources

Historical air flight prices are not readily available on the internet. Therefore the only optionthat we have is to use some resources and collect data over a period of time. There are many **APIs** made available by companies like Amadeus, Sky Scanner. However, the numberof flights they returned for a domestic route in India are limited.

Thus we had to explore APIs by Indian companies. **GoIbibo & Expedia** provide ready to useAPIs which returns a set of variables of a certain route. However, when GoIbibo API was used to extract such flight prices, the returned set of parameters were in a complex form & the raw data had to be cleaned several times.

Therefore, we decided to build a web spider that extracts the required values from a websiteand stores it as a CSV file. We decided to scrape Makemytrip website using a manual spidermade in Python.

Web Scraper (Python2.7)

urllib2 library was used to access the Makemytrip website and load the required page a jsonobject. This object was parsed using inbuilt python functions and a csv database was obtained. dateutil.rrule library was used to obtain a set of dates between two dates on whichthe above functions can be applied to get the data for a range of dates. The whole script is open-sourced on Github.

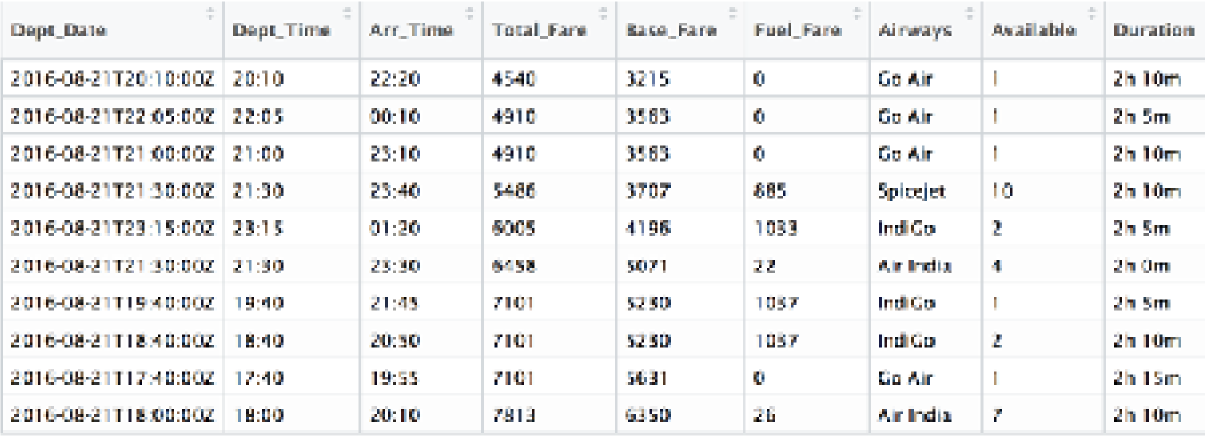
What and When we collected?

What?

The basic structure of the script successfully extracts information from the Makemytrip website and outputs a csv data file. Now an important aspect is to decide the parameters thatmight be needed for the flight prediction algorithm.

Makemytrip returns numerous variables for each flight returned. However not all are requiredand thus we selected the following -

1. Origin City
2. Destination City
3. Departure Date
4. Departure Time
5. Arrival Time
6. Total Fare
7. Airway Carrier
8. Duration
9. Class Type - Economy/Business
10. Flight Number
11. Hopping - Boolean
12. Taken Date - date on which this data was collected

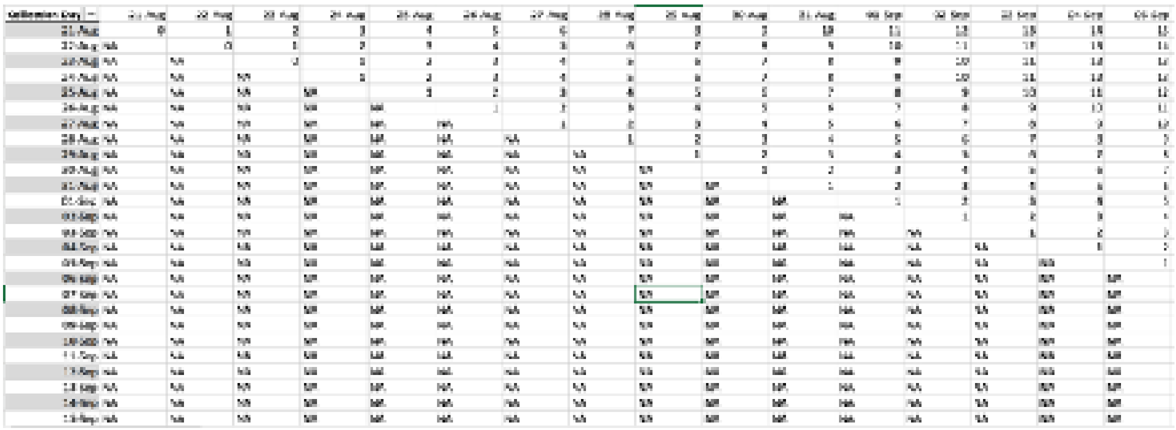


Dept_Date	Dept_Time	Arr_Time	Total_Fare	Base_Fare	Fuel_Fare	Airways	Available	Duration
2016-08-21T20:10:00Z	20:10	22:20	4540	3215	0	Go Air	1	2h 10m
2016-08-21T22:05:00Z	22:05	00:10	4910	3583	0	Go Air	1	2h 5m
2016-08-21T21:00:00Z	21:00	23:10	4910	3583	0	Go Air	1	2h 10m
2016-08-21T21:30:00Z	21:30	23:40	5486	3707	685	Spicejet	10	2h 10m
2016-08-21T23:15:00Z	23:15	01:20	6005	4196	1033	IndiGo	2	2h 5m
2016-08-21T21:30:00Z	21:30	23:30	6458	5071	22	Air India	4	2h 0m
2016-08-21T19:40:00Z	19:40	21:45	7101	5250	1057	IndiGo	1	2h 5m
2016-08-21T18:40:00Z	18:40	20:50	7101	5250	1057	IndiGo	2	2h 10m
2016-08-21T17:40:00Z	17:40	19:55	7101	5631	0	Go Air	1	2h 15m
2016-08-21T18:00:00Z	18:00	20:10	7813	6350	26	Air India	7	2h 10m

When?

We collected prices for flights having departure date from 21st August 2016 to 21st November. This implies that we collected data with max 92 days to departure. Our model isrestricted to predict a maximum of 45 days to departure and thus we collected data for maximum of 92 days. This made sure that we had enough data for 45 days of departure.

As you can observe the cascading effect of dates. Due to this, we have only one day's data offlights having a 92 days to departure. Therefore to restrict our model to 45 days to departure, we had to collevt data till 92 days to departure to get enough valid data.



Collection Day	21-Aug	22-Aug	23-Aug	24-Aug	25-Aug	26-Aug	27-Aug	28-Aug	29-Aug	30-Aug	31-Aug	01-Sep	02-Sep	03-Sep	04-Sep	05-Sep
21-Aug	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
22-Aug	NA	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
23-Aug	NA	NA	0	1	2	3	4	5	6	7	8	9	10	11	12	13
24-Aug	NA	NA	NA	1	2	3	4	5	6	7	8	9	10	11	12	13
25-Aug	NA	NA	NA	NA	1	2	3	4	5	6	7	8	9	10	11	12
26-Aug	NA	NA	NA	NA	NA	1	2	3	4	5	6	7	8	9	10	11
27-Aug	NA	NA	NA	NA	NA	NA	1	2	3	4	5	6	7	8	9	10
28-Aug	NA	NA	NA	NA	NA	NA	NA	1	2	3	4	5	6	7	8	9
29-Aug	NA	NA	NA	NA	NA	NA	NA	NA	1	2	3	4	5	6	7	8
30-Aug	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	2	3	4	5	6	7
31-Aug	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	2	3	4	5	6
01-Sep	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	2	3	4	5
02-Sep	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	2	3	4
03-Sep	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	2	3
04-Sep	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	2
05-Sep	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1
06-Sep	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
07-Sep	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
08-Sep	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
09-Sep	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
10-Sep	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
11-Sep	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
12-Sep	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
13-Sep	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
14-Sep	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
15-Sep	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Automating the Collection

The above script had to be run daily to get the required data. Manually running the file wassenseless because there are many ways to automate the script. There are multiple cloud servers like Microsoft Azure or AWS EC2 which can be used to host the python script.

We used Microsoft Azure initially and the process of setting up a virtual server is quite easy. However, the free trial period was limited and thus we had to shift to our institute's local server. Linux system provides a beautiful option called Cronjobs. It is an in-built scheduling option that is very easy to use. You can read a detailed blog on automating python script here.

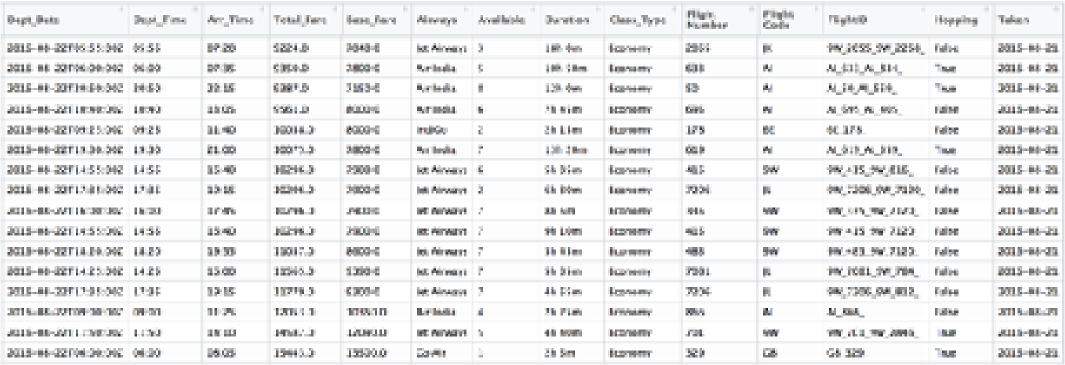
DATA PREPARATION

The kind of data that we collected from the python script was very raw and needed a lot of work. For instance, the price was a character type and not an integer. Moreover, for any model to work efficiently, certain variables need to be introduced by combining or changingthe existing variables.

This section focuses on various techniques we used to clean and prepare the data.

1. Generic Exploration & Cleaning
2. Data Specific Methods
3. Trend Analysis for Predicting Number of Days to wait

Generic Exploration & Cleaning



Dept_Date	Dept_Time	Arr_Time	Total_Fare	Base_Fare	Airways	Available	Duration	Class_Type	Flight Number	Flight Code	FlightID	Hopping	Taken
2015-06-22T05:55:00Z	05:55	07:20	5224.0	3040.0	Jet Airways	3	18h 4m	Economy	2055	9W	9W_2055_9W_2254_	False	2015-06-21
2015-06-22T06:00:00Z	06:00	07:35	5350.0	3800.0	Air India	5	18h 10m	Economy	610	AI	AI_611_AI_614_	True	2015-06-21
2015-06-22T19:55:00Z	19:55	20:15	6287.0	7153.0	Air India	8	13h 4m	Economy	59	AI	AI_14_AI_519_	True	2015-06-21
2015-06-22T18:50:00Z	18:50	15:05	5561.0	8000.0	Air India	6	7h 6m	Economy	695	AI	AI_595_AI_995_	False	2015-06-21
2015-06-22T09:25:00Z	09:25	11:40	10016.0	6000.0	IndiGo	2	2h 11m	Economy	175	6E	6E 175	False	2015-06-21
2015-06-22T13:30:00Z	13:30	21:00	10073.0	3800.0	Air India	7	13h 28m	Economy	619	AI	AI_513_AI_319_	True	2015-06-21
2015-06-22T14:55:00Z	14:55	15:40	10296.0	7000.0	Jet Airways	6	5h 31m	Economy	415	9W	9W_415_9W_815_	False	2015-06-21
2015-06-22T17:01:00Z	17:01	10:18	10266.0	7000.0	Jet Airways	2	6h 00m	Economy	7026	9W	9W_7026_9W_7120_	False	2015-06-21
2015-06-22T16:30:00Z	16:30	17:45	10796.0	7400.0	Jet Airways	7	8h 6m	Economy	416	9W	9W_415_9W_7121_	False	2015-06-21
2015-06-22T14:55:00Z	14:55	15:40	10296.0	7000.0	Jet Airways	7	9h 10m	Economy	415	9W	9W 415 9W 7123	False	2015-06-21
2015-06-22T14:20:00Z	14:20	19:35	11017.0	8000.0	Jet Airways	7	5h 15m	Economy	485	9W	9W_483_9W_7123_	False	2015-06-21
2015-06-22T14:25:00Z	14:25	15:00	11565.0	5300.0	Jet Airways	7	5h 31m	Economy	7081	9W	9W_7081_9W_784_	False	2015-06-21
2015-06-22T17:35:00Z	17:35	10:15	11779.0	6300.0	Jet Airways	7	4h 21m	Economy	7026	9W	9W_7026_9W_812_	False	2015-06-21
2015-06-22T09:30:00Z	09:30	11:25	12811.0	10510.0	Air India	4	7h 21m	Economy	866	AI	AI_866_	False	2015-06-21
2015-06-22T11:50:00Z	11:50	18:10	14567.0	12000.0	Jet Airways	5	4h 60m	Economy	731	9W	9W_711_9W_2095_	True	2015-06-21
2015-06-22T06:30:00Z	06:30	09:05	15445.0	13500.0	GoAir	1	2h 5m	Economy	320	G8	G8 329	True	2015-06-21

The collected data for each route looks like the one above. Because of the large number of flights in the busy routes like Delhi Bombay, the data collected over time is over a million points and hence efficiently handling such big data for faster computation is the first aim. InR the **‘fread’ function in ‘data.table’** package was used.

```
Copybomdel <- fread("bomdel.csv")
```

A few basic cleaning and feature engineering looking at the data. A lot of data preparation needs to be done according to the model and strategy we use, but here are the basic cleaningwe did initially to understand the data bette

- Duplicates

  There were not many, but a few repetitions in the data collected.

- Days to departure

  Our objective is to optimize this parameter. This the difference is the departure dateand the day of booking the ticket. We consider this parameter to be within 45 days.

- Day of departure

  Intuitively we can say that flights scheduled during weekends will have a higher price compared to the flights on Wednesday or Thursday. Since including this in any of themodels we use can be beneficial. We can also try to include the month or if it is a holiday time for better accuracy.

- Duration

  Converting the duration of the flight into numeric values, so that the model can interpret it properly. Also, it will be fair enough to omit flights with a very longduration.

- Time of departure

  Similar to day of departure, the time also seem to play an important factor. Hence wedivided all the flights into three categories: Morning (6am to noon), Evening (noon to9pm) and Night (9pm to 6am)

- Hoppings

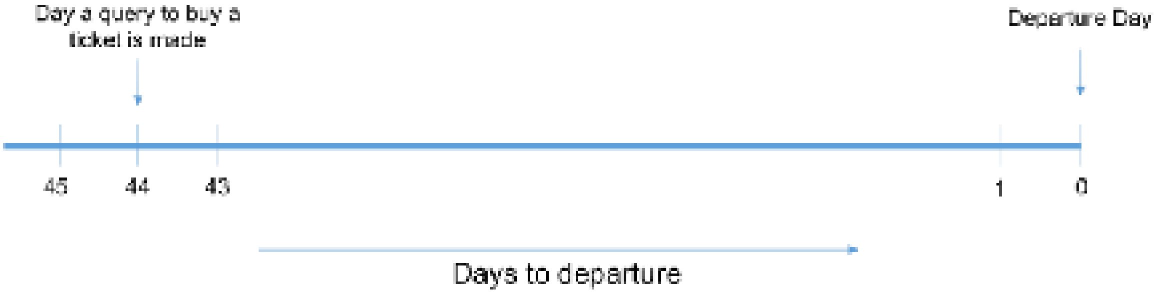
  The data we collected did not give very authentic information about the number ofhops a journey takes. Hence, we calculated the hops using the flight ids.

- Outliers

  We are focusing on minimizing the flight prices, hence we considered only theeconomy class with the following conditions:

a) The minimum value of total fare for all days for a particular flight id is less than themean fare of all the flights b) The duration of the journey is less than 3 times the meanduration.

Data Specific Methods



Suppose a user makes a query to buy a flight ticket 44 days in advance, then our system should be able to tell the user whether he should wait for the prices to decrease or he shouldbuy the tickets immediately. For this we have two options:

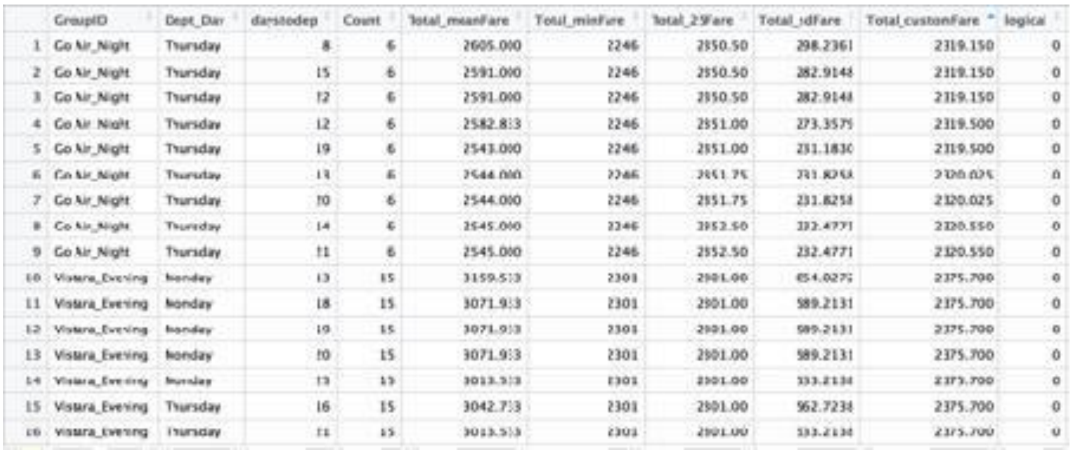
1. **Predict the flight prices for all the days between 44 and 1** and check on which daythe price is minimum.
2. Classify the data we already have into, **"Buy" or "Wait"**. This then becomes a classification problem and we would need to predict only a binary number. However,this does not give a good insight on the number of days to wait.

For the above example, if we choose the first method we would need to make a total of 44 predictions (i.e. run a machine learning algorithm 44 times) for a single query. This also cascades the error per prediction decreasing the accuracy. Hence, the **second method seemsto be a better way to predict**, wait or buy which is a simple binary classification problem. But, in this method, we would need to predict the days to wait using the historic trends.

For this we again have two options:

1. We do the **predictions for each flight id**. The problem with this is that, if there is a change in flight id by the airline (which happens frequently) or there is an introduction or a new flight for a specific route then our analysis would fail.
2. We **group the flight ids according to the airline and the time of departure** and do the analysis on each group. For this we need to combine the prices of the airlines lying in that group such that the basic trend in captured.

Moving ahead with the second option, we created the group according to the airlines and the departure time-slot created earlier (Morning, Evening, Night) and calculated the combined flight prices for each group, day of departure and depart day. Since these three are the most influencing factors which determine the flight prices. Also, we calculated the average number of flights that operated in a particular group, since competition could also play a role in determining the fare.



	GroupID	Dept_Day	daystodep	Count	Total_meanFare	Total_minFare	Total_25Fare	Total_sdFare	Total_customFare	logical
1	Go Air_Night	Thursday	8	6	2605.000	2246	2550.50	298.2361	2319.150	0
2	Go Air_Night	Thursday	15	6	2591.000	2246	2550.50	282.9144	2319.150	0
3	Go Air_Night	Thursday	12	6	2591.000	2246	2550.50	282.9144	2319.150	0
4	Go Air_Night	Thursday	12	6	2582.833	2246	2551.00	273.3575	2319.500	0
5	Go Air_Night	Thursday	19	6	2543.000	2246	2551.00	231.1830	2319.500	0
6	Go Air_Night	Thursday	13	6	2544.000	2246	2551.75	231.8258	2320.025	0
7	Go Air_Night	Thursday	10	6	2544.000	2246	2551.75	231.8258	2320.025	0
8	Go Air_Night	Thursday	14	6	2545.000	2246	2552.50	232.4771	2320.550	0
9	Go Air_Night	Thursday	11	6	2545.000	2246	2552.50	232.4771	2320.550	0
10	Vistara_Evening	Monday	13	15	3159.533	2301	2901.00	654.0275	2375.700	0
11	Vistara_Evening	Monday	18	15	3071.933	2301	2901.00	589.2131	2375.700	0
12	Vistara_Evening	Monday	19	15	3071.933	2301	2901.00	589.2131	2375.700	0
13	Vistara_Evening	Monday	10	15	3071.933	2301	2901.00	589.2131	2375.700	0
14	Vistara_Evening	Monday	13	15	3013.533	2301	2901.00	533.2134	2375.700	0
15	Vistara_Evening	Thursday	16	15	3042.733	2301	2901.00	562.7234	2375.700	0
16	Vistara_Evening	Thursday	11	15	3013.533	2301	2901.00	533.2134	2375.700	0

Combining fare for the flights in one group:

1. Mean fare: This is the average of the fare of all the flights in a particular group corresponding to departure day and days to departure. Because of high standard deviation, taking the mean is not a very good option.

2. Minimum fare: This does not give a very good insight of the trend, as a minimum value could occur because of some offer by an airline.

3. First Quartile: This is a good measure as we are focusing on minimizing the fare and we do not want to consider the flights with high fares.

4. Custom Fare: This is the fare giving more weightage to recent price trend.

Total_customFare = w*(First Quartile for entire time period) + (1-w)*(First quartile of last x days)

(We have considered: w = 0.7 and x = & days)

Calculating whether to buy or wait for the this data:

Logical = 1 if for any d < D the Total_customFare is less than the current Total_customFare

(Here, d is the days to departure and D is the days to departure for the current row.

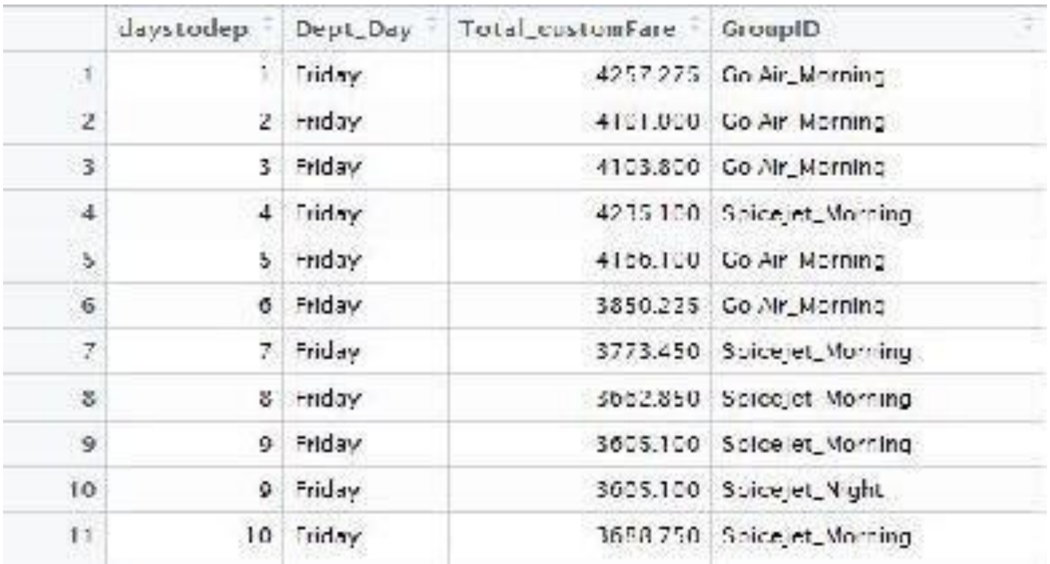
Trend Analysis:

After creating the train file, we shift to create another dataset which is used to predict number of days to wait. For this, we used trend analysis on the original dataset

Determining the minimum CustomFare for a particular pair of Departure Day and Days to Departure

We input the train dataset that has been created and find the minimum of the CustomFare corresponding to each combination of Departure Date and Days to Departure. Now with the obtained minimum CustomFare corresponding to each pair, we do a merge with our initial dataset and find out the Airline corresponding to which the minimum CustomFare is being obtained.

The count on the number of times a particular Airline appears corresponding to the minimumCustom Fare is the probability with which the Airline would be likely to offer a lower price inthe future. This probability of each Airline for having a minimum Fare in the future is exported to the test dataset and merged with the same while the dataset of minimum Fares is retained for the preparation of bins to analyse the time to wait before the prices reduce



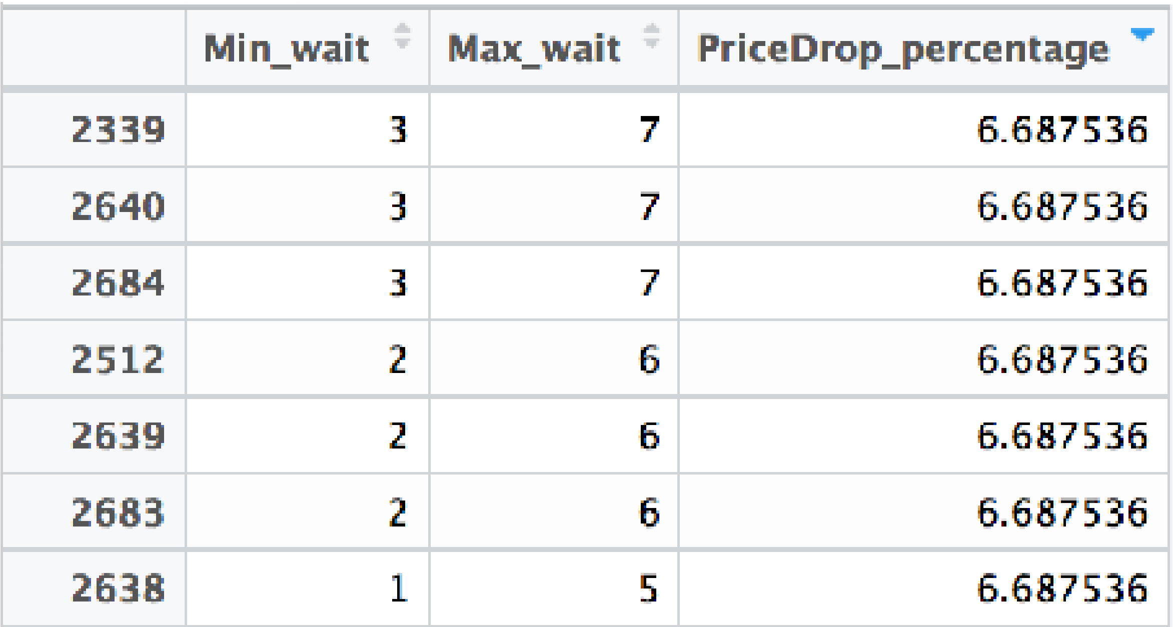
	daystodep	Dept_Day	Total_customFare	GroupID
1	1	Friday	4257.275	Go Air_Morning
2	2	Friday	4101.020	Go Air_Morning
3	3	Friday	4103.800	Go Air_Morning
4	4	Friday	4215.100	Spicejet_Morning
5	5	Friday	4156.100	Go Air_Morning
6	6	Friday	3850.225	Go Air_Morning
7	7	Friday	3773.450	Spicejet_Morning
8	8	Friday	3662.850	Spicejet_Morning
9	9	Friday	3605.100	Spicejet_Morning
10	9	Friday	3605.100	Spicejet_Night
11	10	Friday	3658.750	Spicejet_Morning

Creation of Bins

We next wanted to determine the trend of “lowest” airline prices over the data we were training upon. So the entire sequence of 45 days to departure was divided into bins of 5 days. In intervals of 5, the first bin would represent days 1-5, the second represents 6-10 and so on.

Corresponding to each bin, we required a value of the fare that would be optimal for consideration in suggesting a value for the days to wait to the user. Among all the points thatlie in a bin, the 25th percentile was determined as the value that would be the possible lowestFare corresponding to the bin which indicates days to departure.

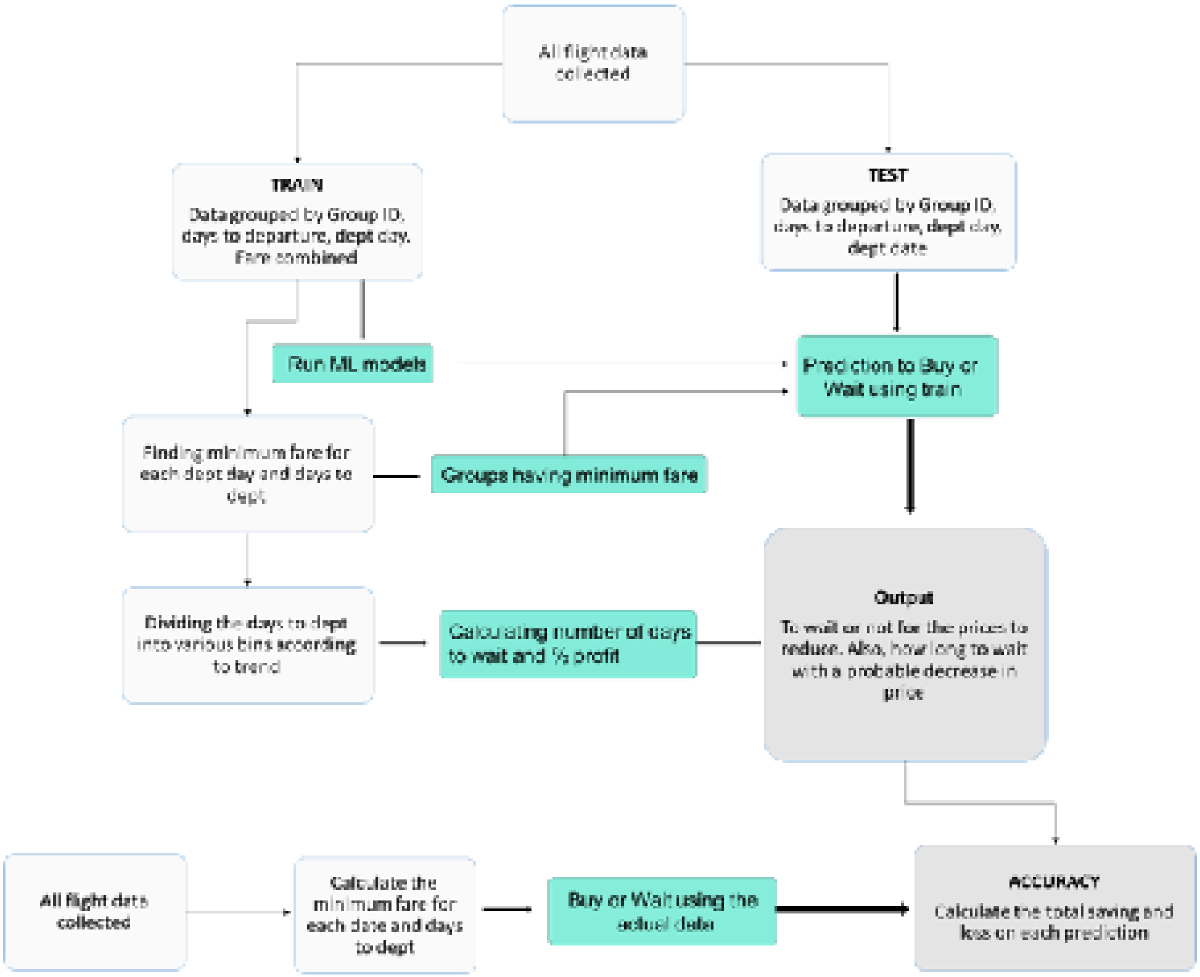
Comparing the present price on the day the query was made with the prices of each of the bin, a suggestion is made corresponding to the maximum percentage of savings that can be done by waiting for that time period.The approximate time to wait for the prices to decrease and the corresponding savings that could be made is returned to the user.



	Min_wait	Max_wait	PriceDrop_percentage
2339	3	7	6.687536
2640	3	7	6.687536
2684	3	7	6.687536
2512	2	6	6.687536
2639	2	6	6.687536
2683	2	6	6.687536
2638	1	5	6.687536

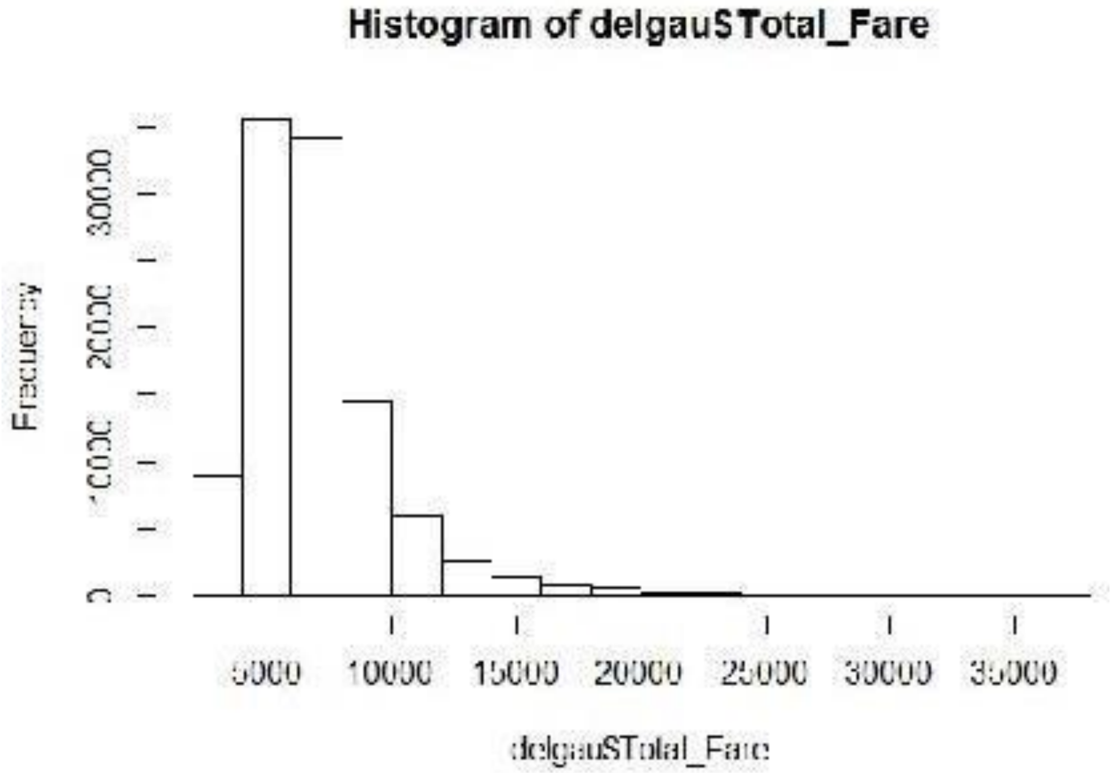
OUR MODEL

This section provides a birdeye view on the whole model we used. The key components ofthe model are the **training data** and the **testing data**. The way we built these data and the various aspects of the model that uses these datasets is very tricky to understand. This flowchart will provide some basic understanding.



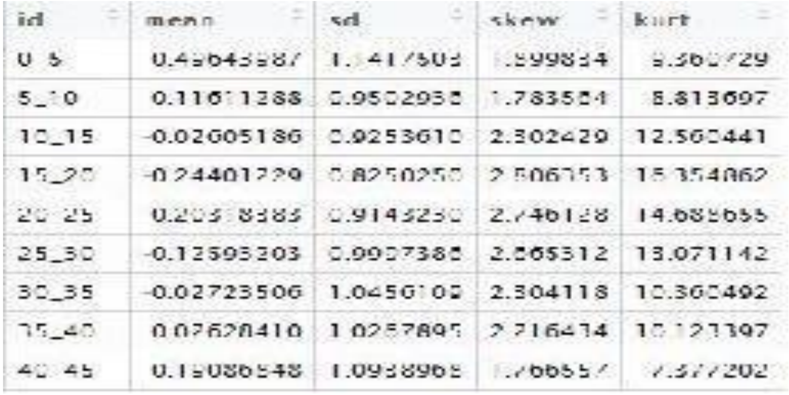
All flight data collected
TRAIN
Datagrouped by Group ID, days to departure, dept day. Fare combined
TEST
Data grouped by Group ID, days to departure, dept day, dept date
Run ML models
Prediction to Buy or Wait using train
Finding minimum fare for each dept day and days to dept
Groups having minimum fare
Dividing the days to dept into various bins according to trend
Calculating number of days to wait and % profit
Output
To wait or not for the prices to reduce. Also, how long to wait with a probable decrease in price
All flight data collected
Calculate the minimum fare for each date and days to dept
Buy or Wait using the actual data
ACCURACY
Calculate the total saving and loss on each prediction

Statistical Analysis



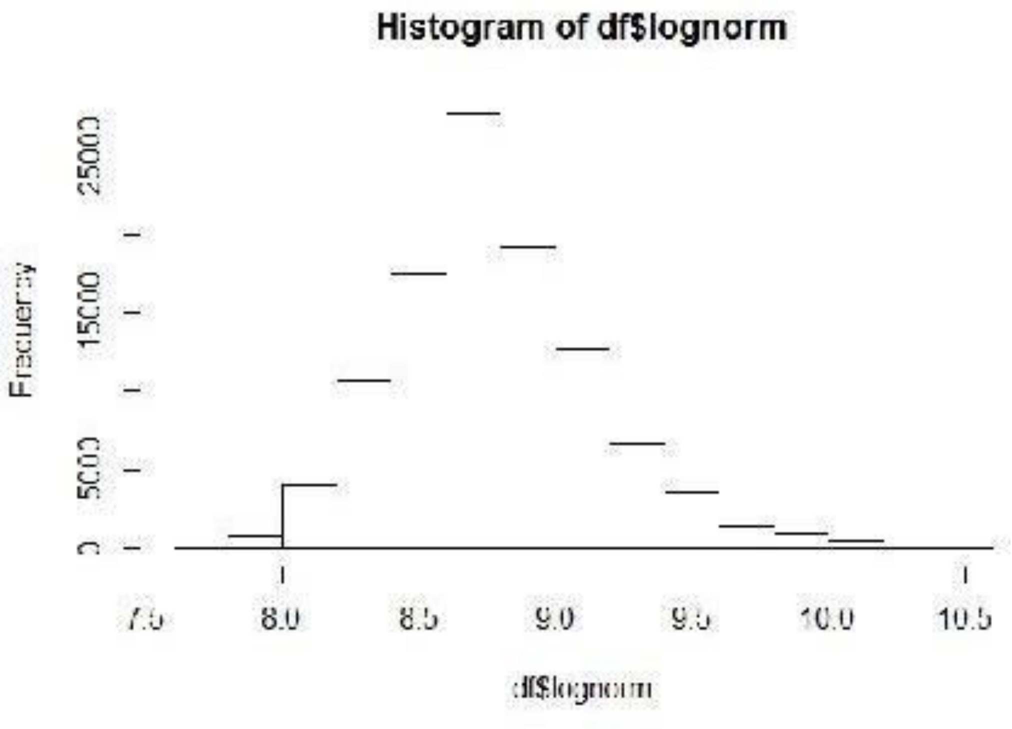
Flight prices after cleaning had the following distribution –

With the moments of Kurtosis was 10.5 and skewness was 3.4 when the entire data was considered. For each interval, all the moments are shown:

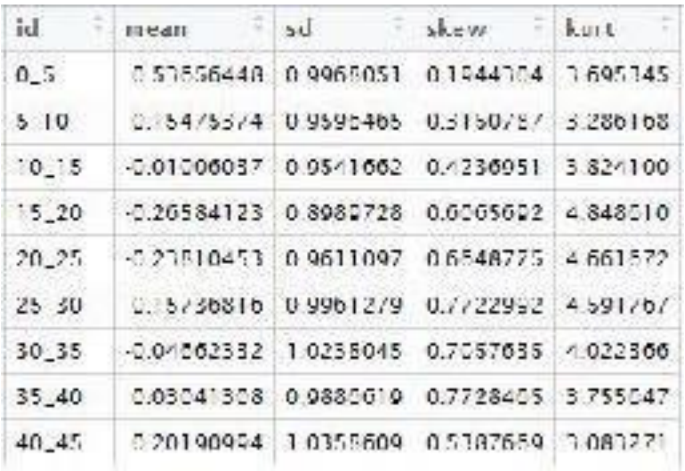


id	mean	sd	skew	kurt
0_5	0.49643987	1.1417503	1.899834	9.360729
5_10	0.11611288	0.9502958	1.783584	8.813697
10_15	-0.02605186	0.9253610	2.302429	12.560441
15_20	-0.24401229	0.8250250	2.506153	15.354062
20_25	0.20318383	0.9143230	2.746128	14.688655
25_30	-0.12593303	0.9907388	2.565312	13.071142
30_35	-0.02723506	1.0456102	2.304118	10.360492
35_40	0.07620410	1.0257895	2.216414	10.123397
40_45	0.15086548	1.0938968	1.766557	7.377202

On observation, it was observed that the data of flight prices followed a lognormal distribution, and upon transformation the distribution came out as follows

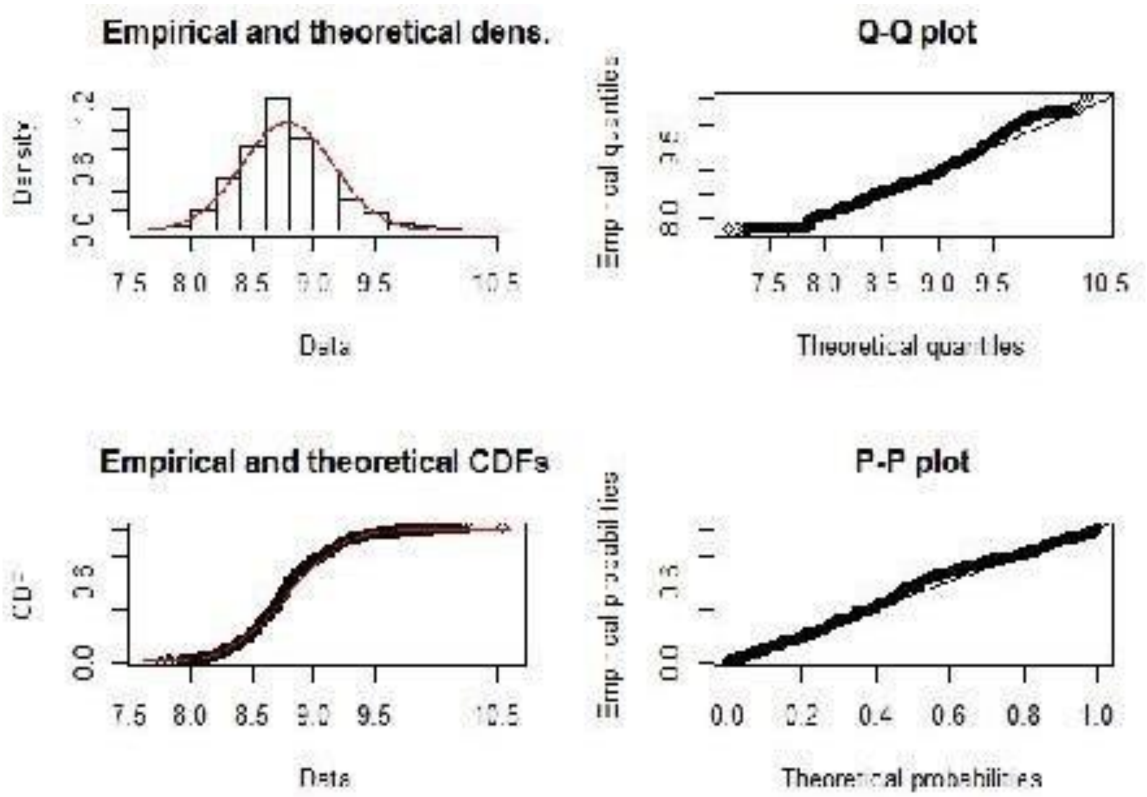


This distribution now appeared to be very much like the normal distribution, so we decided to further examine the data. We calculated the moments as done above to find out that after transformation. The skewness was 0.4 and the kurtosis was 3.8 for the entire data with the moments for each interval as follows -



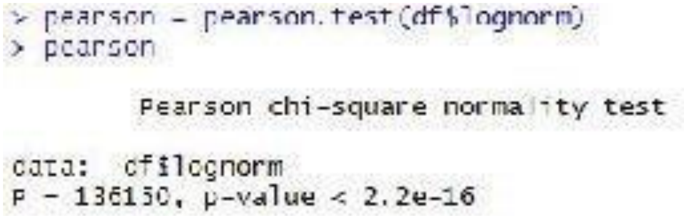
id	mean	sd	skew	kurt
0_5	0.57556448	0.9968051	0.1944704	3.695745
5_10	0.16475574	0.9595465	0.3150787	3.286168
10_15	-0.01006037	0.9541662	0.4236951	3.824100
15_20	-0.26584123	0.8982728	0.6065622	4.848010
20_25	-0.27510453	0.9611097	0.6548775	4.661572
25_30	0.15736816	0.9961279	0.7722952	4.591767
30_35	-0.04562332	1.0238045	0.7057635	4.022366
35_40	0.03041308	0.9885019	0.7728405	3.755047
40_45	0.20190994	1.0355609	0.5187659	3.081271

Further examining the data, we found the QQ plots and the PP plots for the same consideringthe transformed distribution to follow normal distribution -

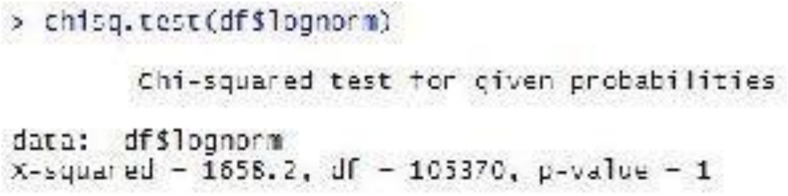


From the analysis we can see that the data almost closely resembles the normal distribution.

Further we decided to use the hypothesis testing and find out the chi-square and p-value. Weused the Pearson normality test for the same and the results are as below -



```
> pearson = pearson.test(df$lognorm)
> pearson

        Pearson chi-square normality test

data:  df$lognorm
P = 136130, p-value < 2.2e-16
```

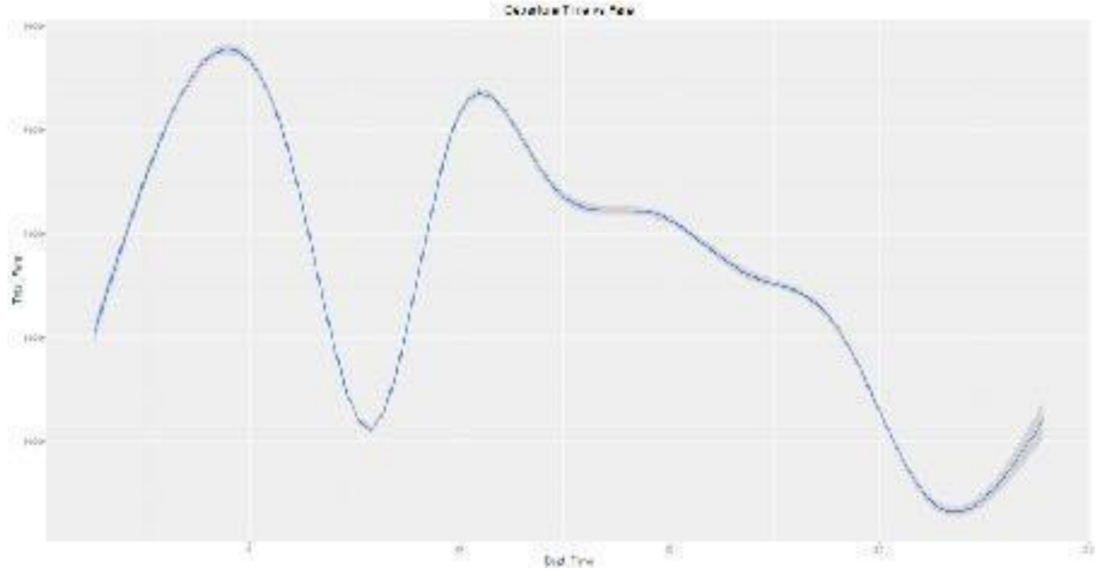


```
> chisq.test(df$lognorm)

        Chi-squared test for given probabilities

data:  df$lognorm
X-squared = 1658.2, df = 105370, p-value = 1
```

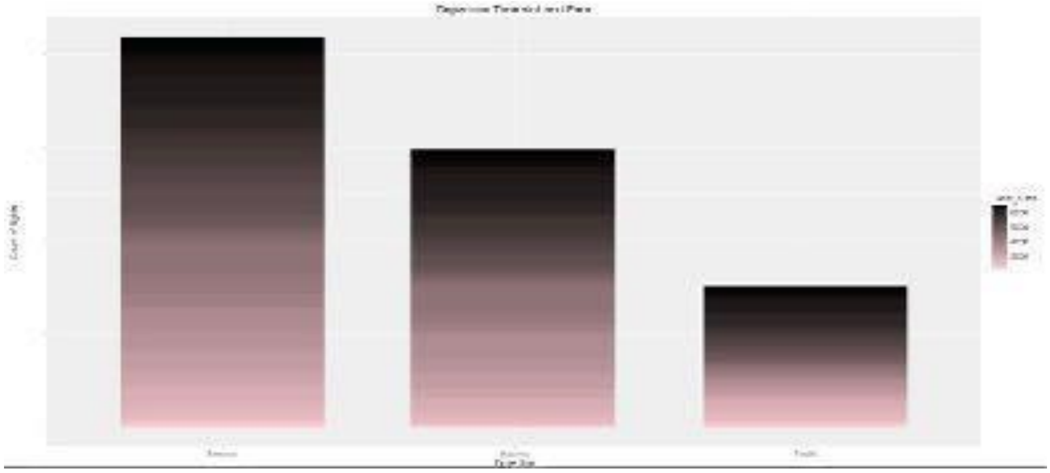
The p-value signifies that our transformed data closely follows normal distribution

Analysis and Graphs

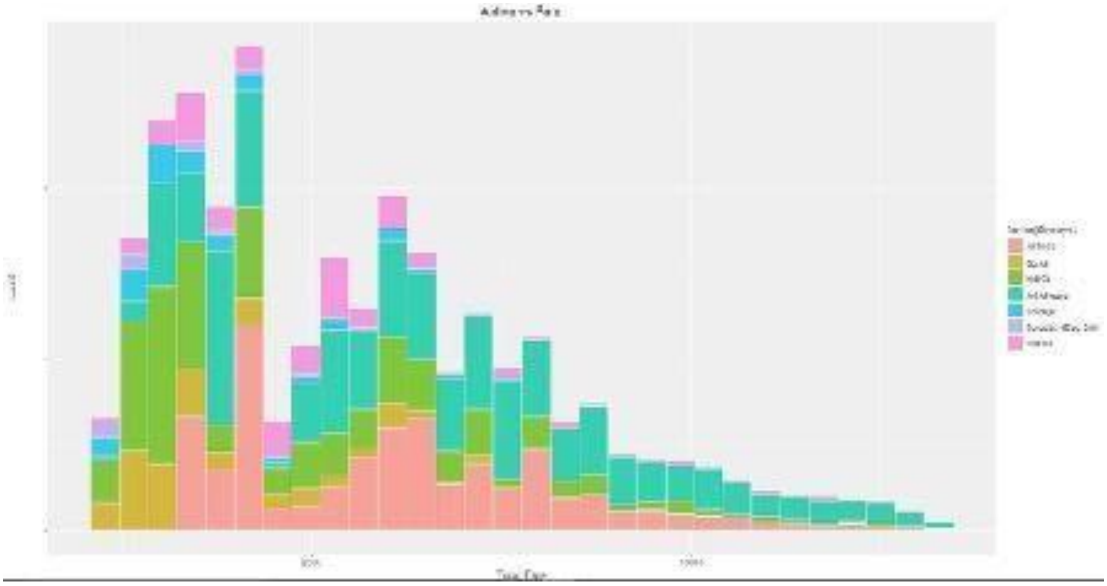
Mumbai-Delhi is one of the busiest route in India and therefore we selected this route tostudy the dynamics of high velocity.



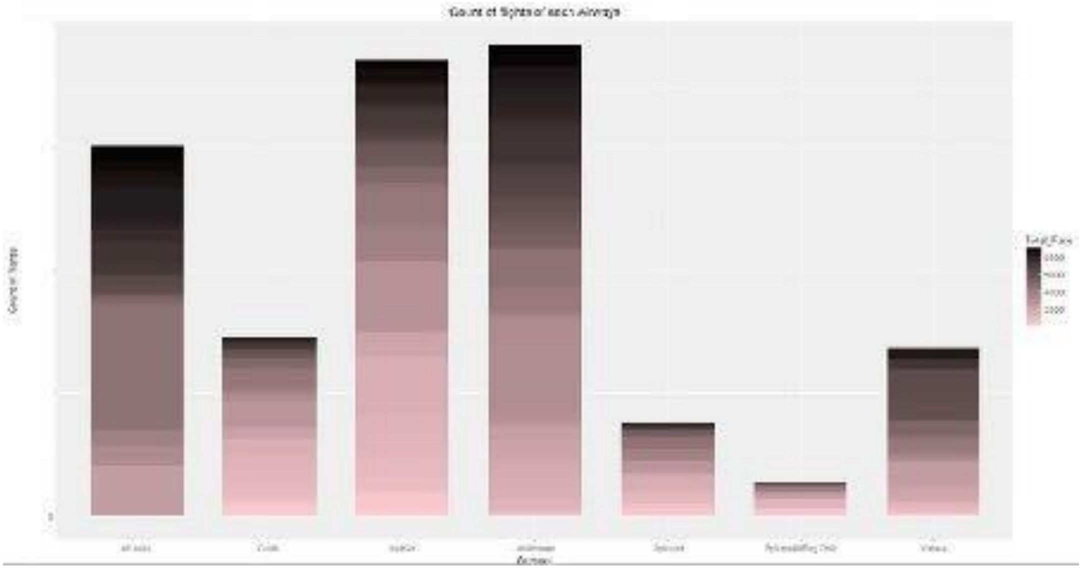
Departure Time vs Fare: Based on the trends, the timeslots can be decided and further expanded.



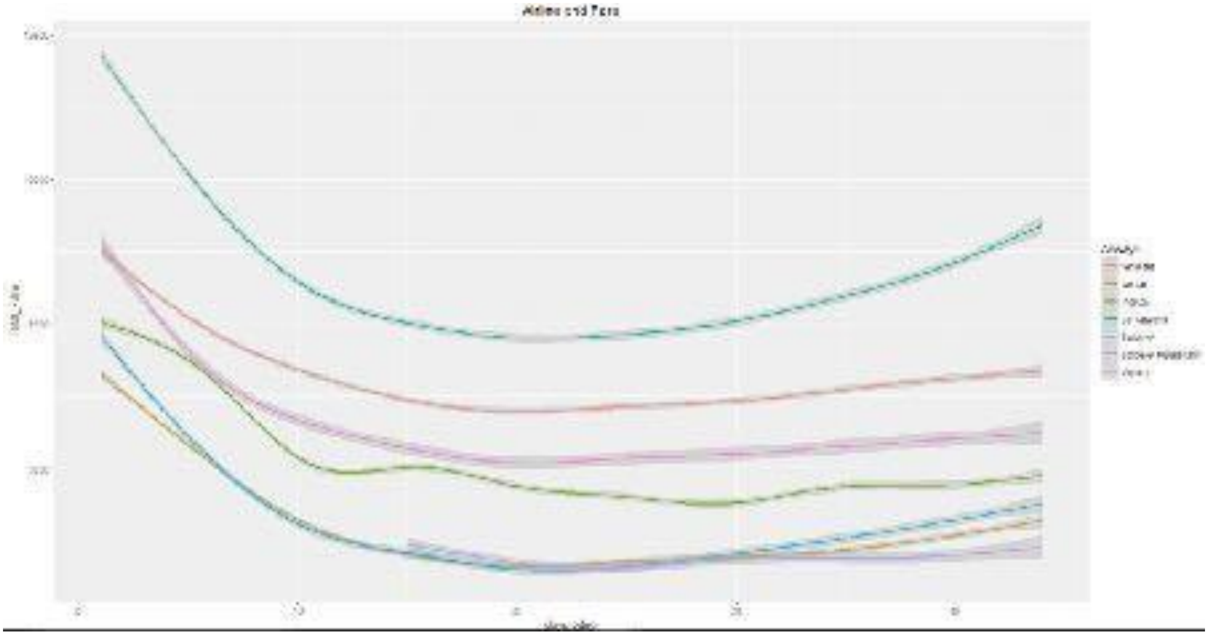
Departure Time slot histogram: Based on division of timeslot, also gives a good insight ofairline fare in each timeslot.



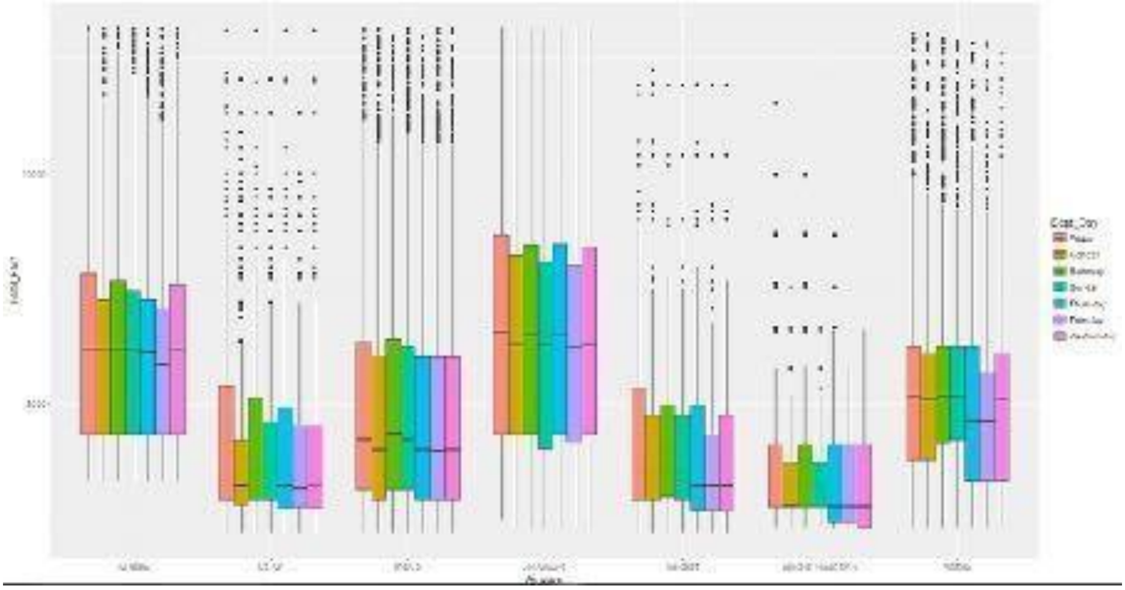
Fare vs Airline: Some are economical while others are well distributed or lie more on thehigher end. Since we are focusing on only the minimum fare, we can ignore few airlines.



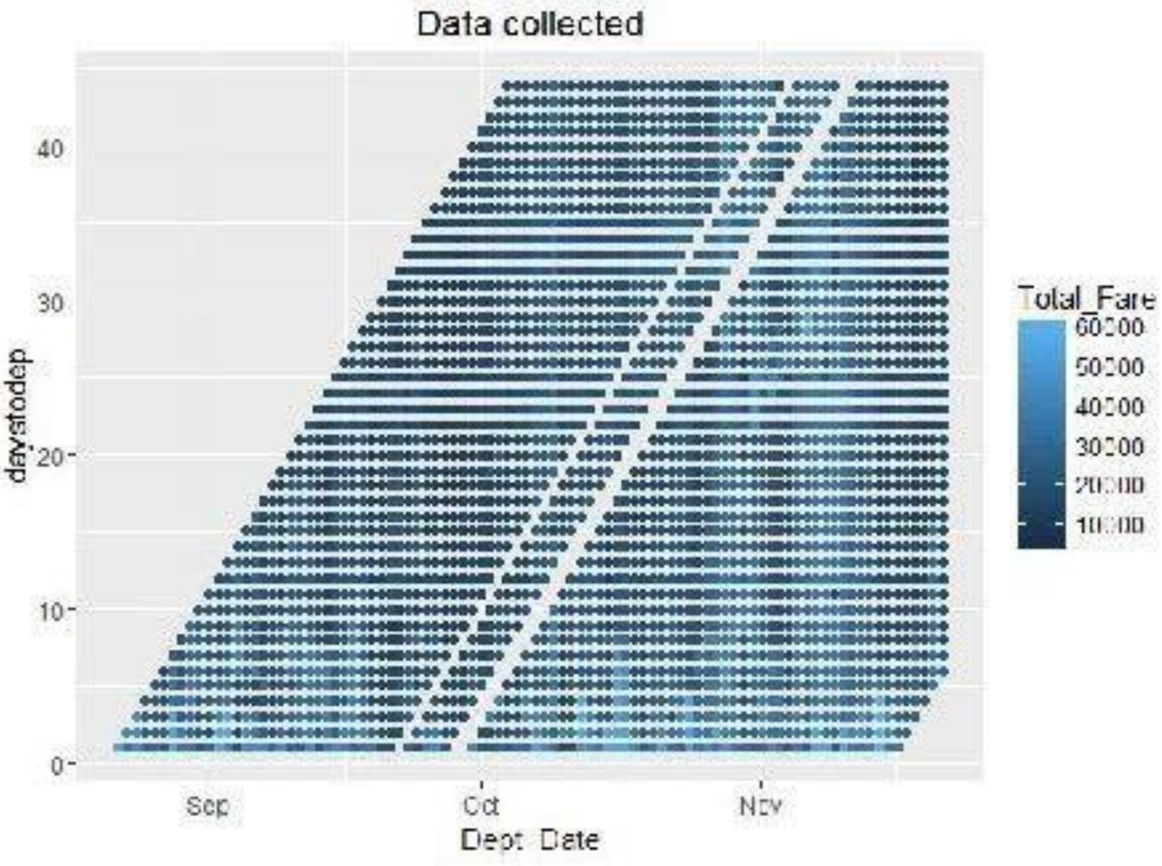
Count of airline: Number of flights operating per airline in different ranges of fare.



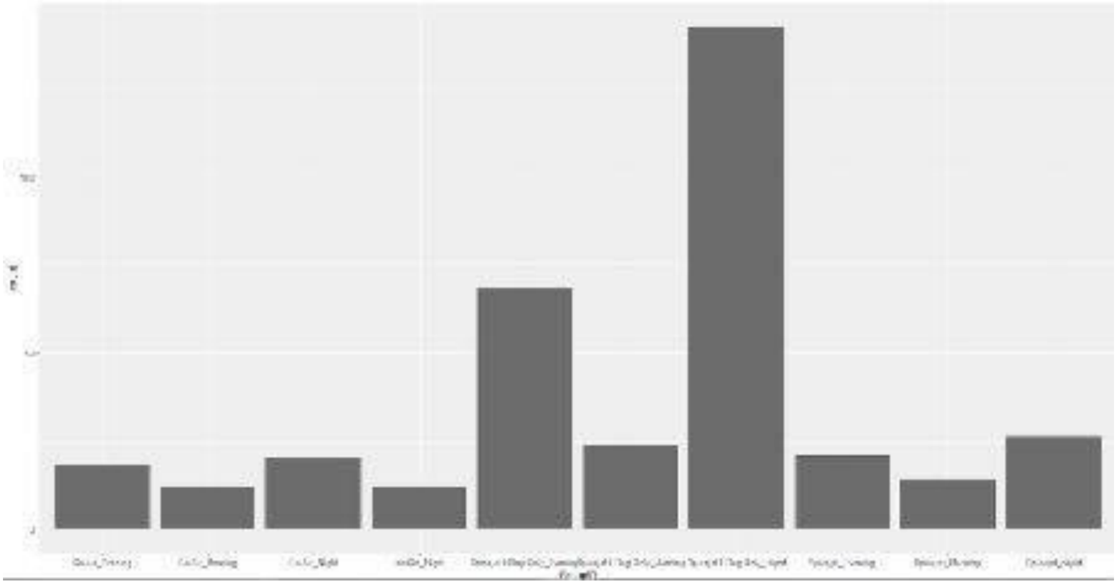
Days to departure vs Fare: Trend of each airline as days to departure varies.



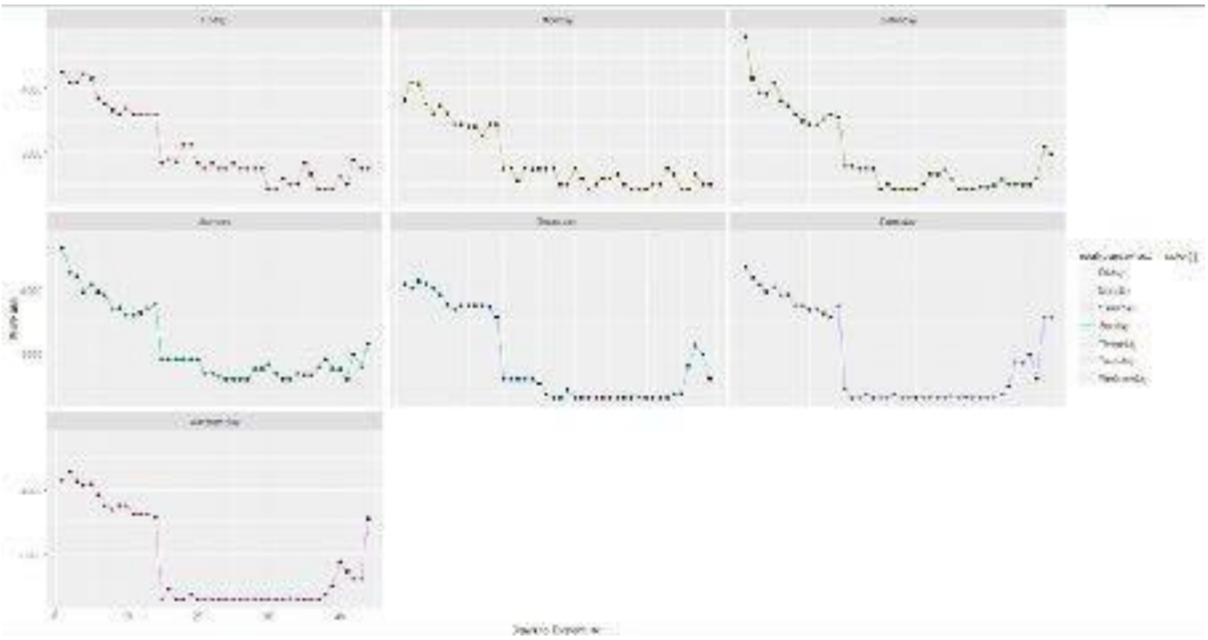
Variation of Fare: Boxplot for fare according to airways and departure day.



Data collected: Departure date vs days to departure.

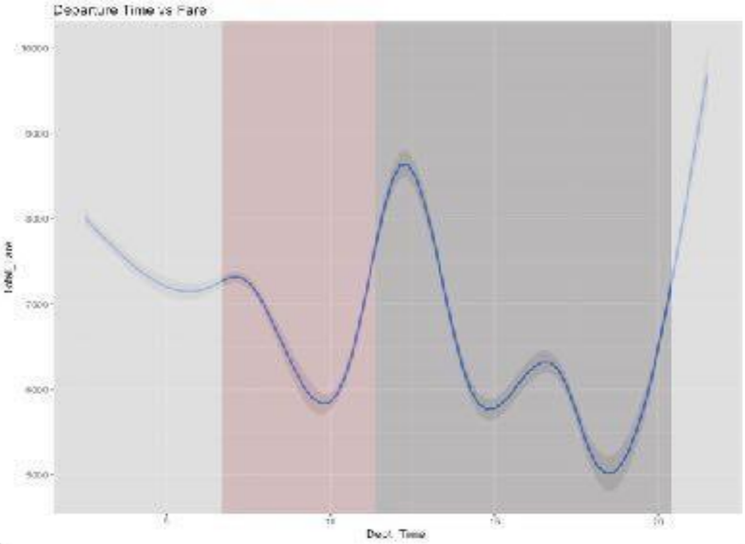


Probability of airline having minimum fare

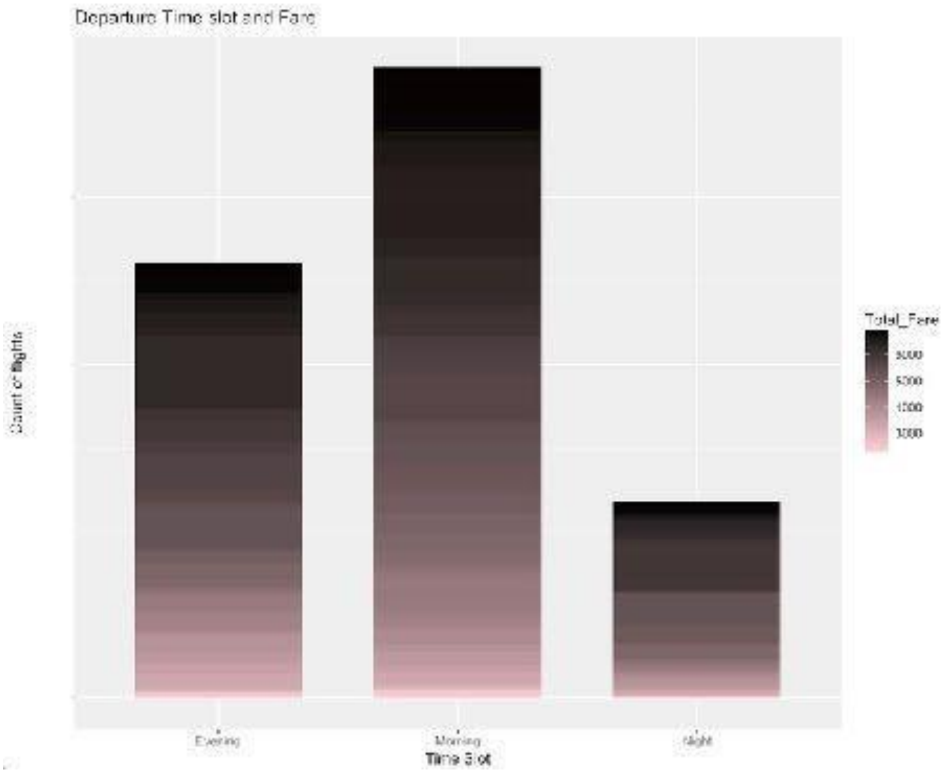


Trend for minimum custom fare as days to departure varies.

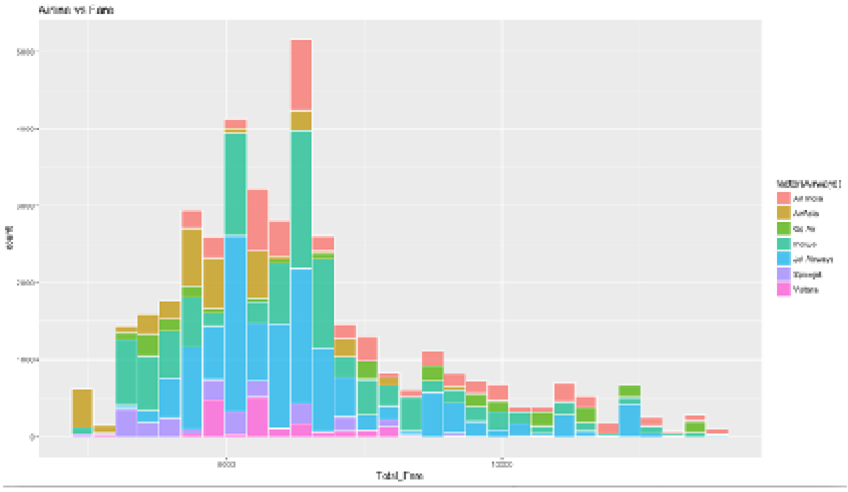
Delhi-Gawahati is a less populated route when compared to Mumbai-Delhi. Our main reason of selecting this route was to understand the dynamics of a low populated domestic Indian route.



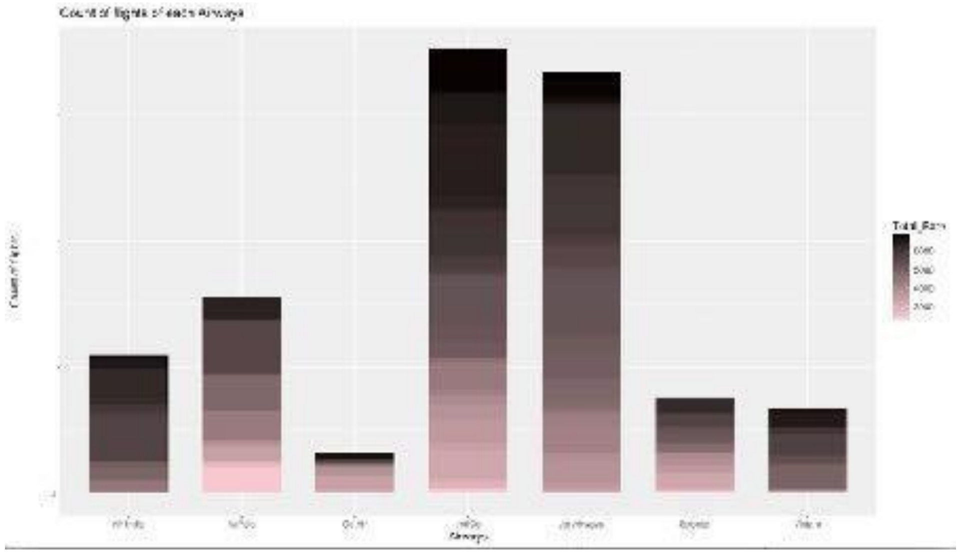
Departure Time vs Fare: Based on the trends, the timeslots can be decided and further expanded.



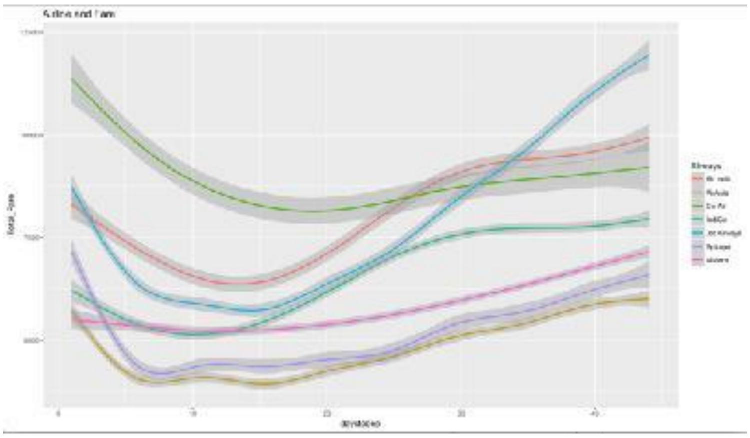
Departure Time slot histogram: Based on division of timeslot, also gives a good insight ofairline fare in each timeslot.



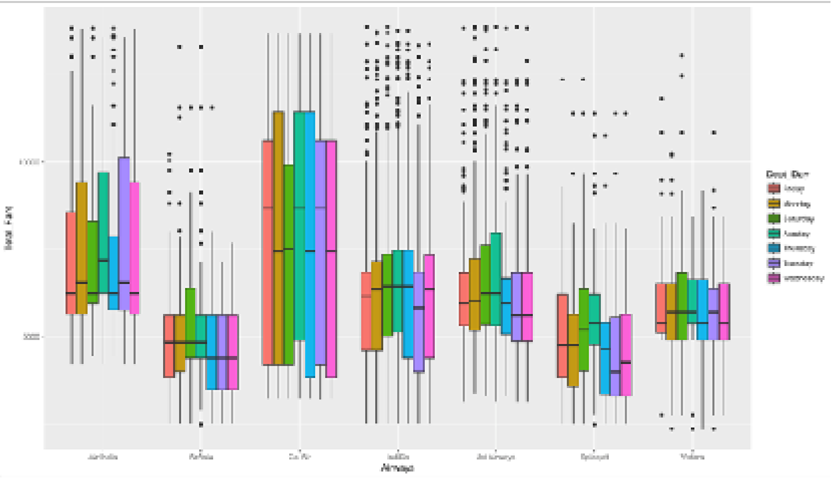
Fare vs Airline: Some are economical while others are well distributed or lie more on thehigher end. Since we are focusing on only the minimum fare, we can ignore few airlines.



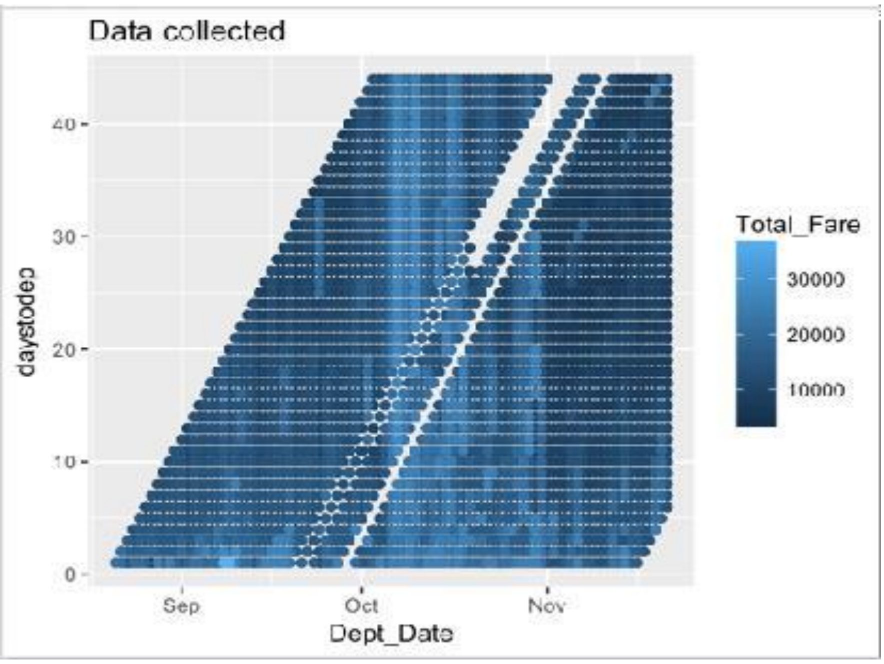
Count of airline: Number of flights operating per airline in different ranges of fare.



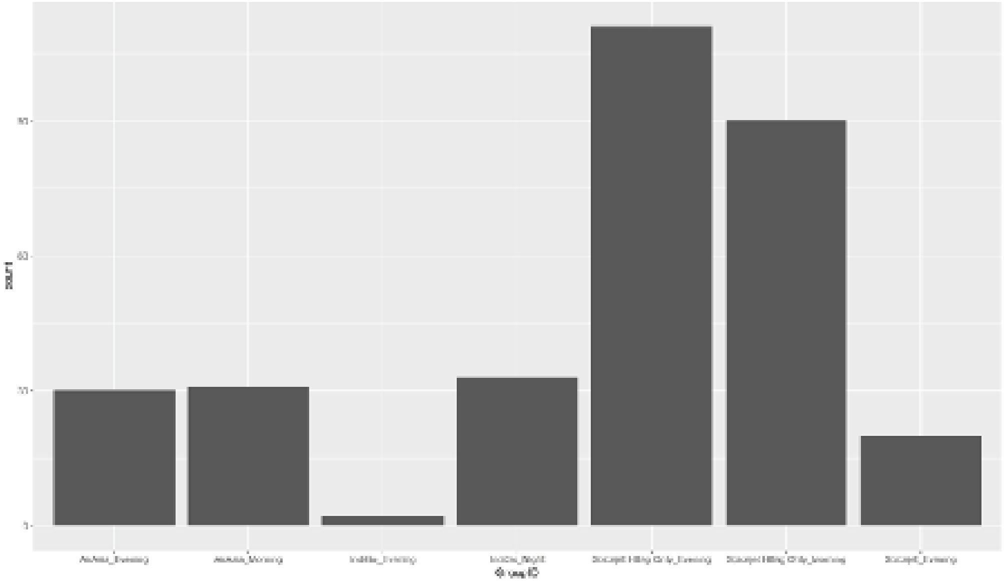
Days to departure vs Fare: Trend of each airline as days to departure varies.



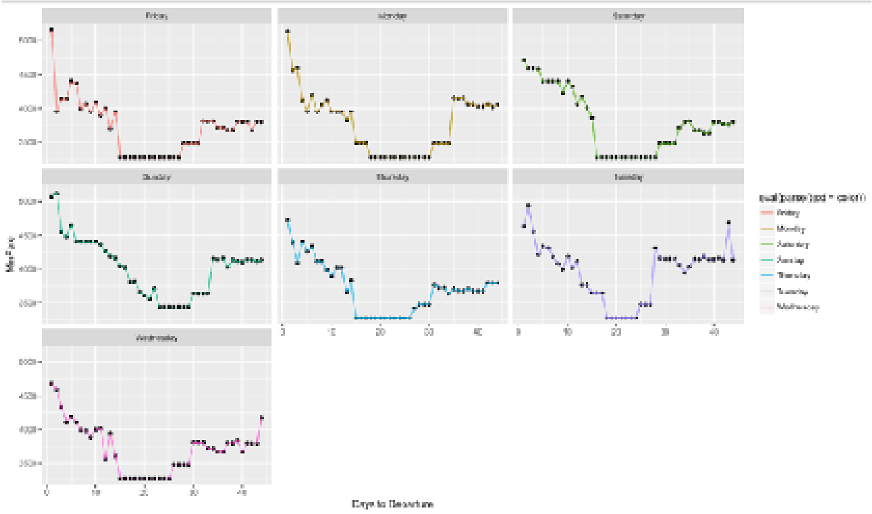
Variation of Fare: Boxplot for fare according to airways and departure day.



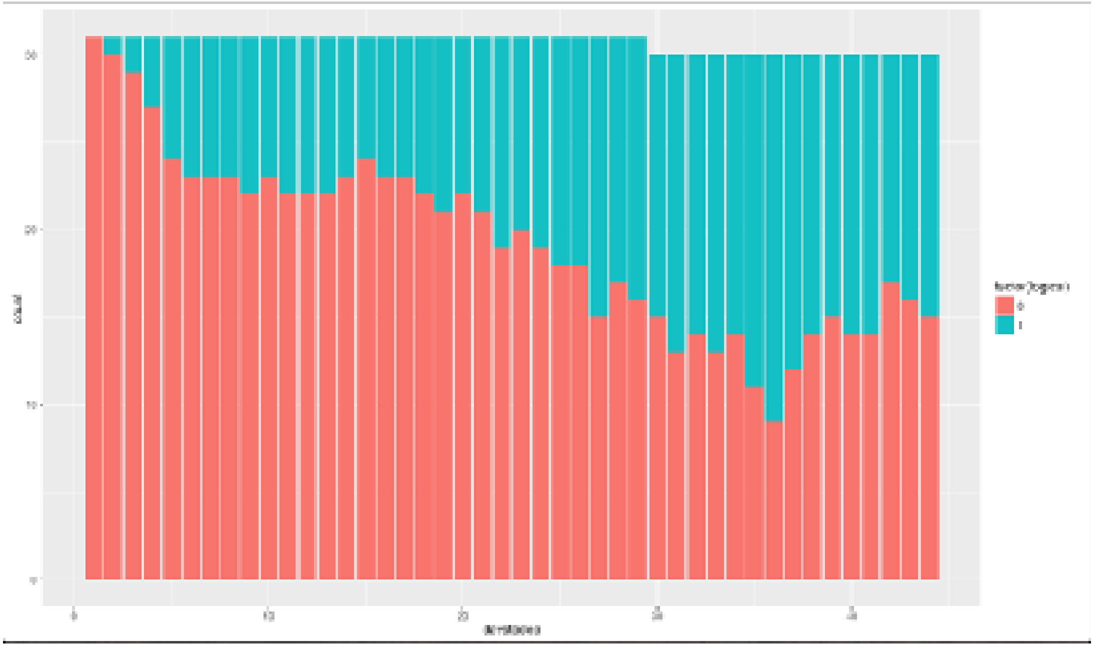
Data collected: Departure date vs days to departure.



Probability of airline having minimum fare



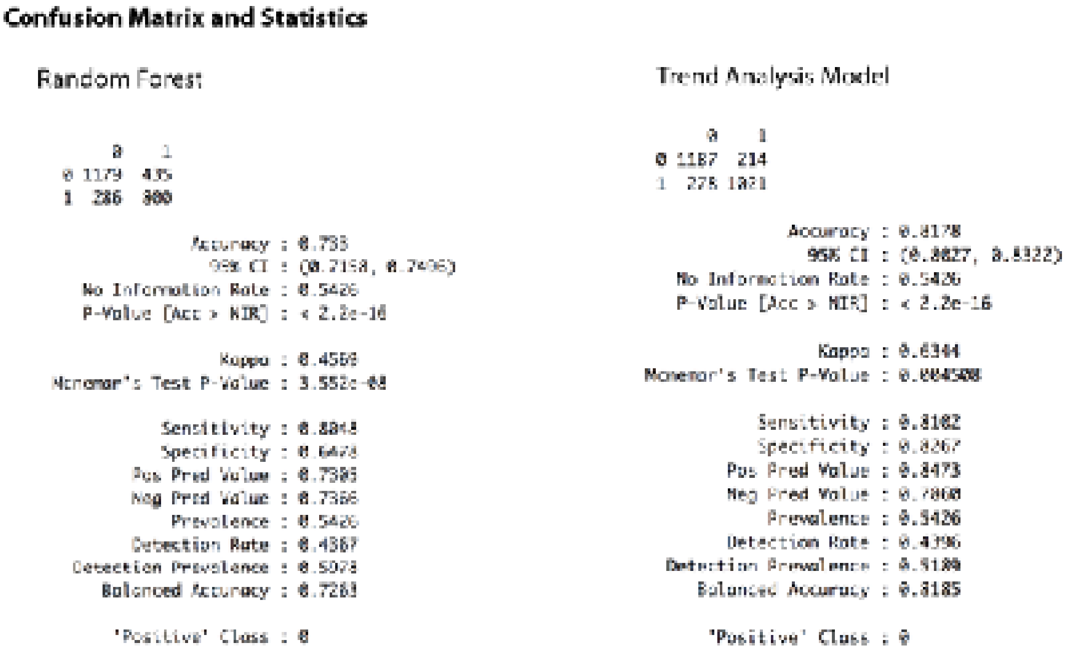
Trend for minimum custom fare as days to departure varies



Number of times it is better to wait for the price to decrease (1 = Wait)

RESULTS & CONCLUSIONS

Results



Confusion Matrix and Statistics

Random Forest

```
     0    1
0 1179  435
1  286  800

               Accuracy : 0.733
                 95% CI : (0.7158, 0.7496)
    No Information Rate : 0.5426
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4568
 Mcnemar's Test P-Value : 3.552e-08

            Sensitivity : 0.8048
            Specificity : 0.6478
         Pos Pred Value : 0.7305
         Neg Pred Value : 0.7366
             Prevalence : 0.5426
         Detection Rate : 0.4367
   Detection Prevalence : 0.5978
      Balanced Accuracy : 0.7263

       'Positive' Class : 0
```

Trend Analysis Model

```
     0    1
0 1187  214
1  278 1021

               Accuracy : 0.8178
                 95% CI : (0.8027, 0.8322)
    No Information Rate : 0.5426
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6344
 Mcnemar's Test P-Value : 0.004508

            Sensitivity : 0.8102
            Specificity : 0.8267
         Pos Pred Value : 0.8473
         Neg Pred Value : 0.7860
             Prevalence : 0.5426
         Detection Rate : 0.4396
   Detection Prevalence : 0.5189
      Balanced Accuracy : 0.8185

       'Positive' Class : 0
```

On 2500 unique test cases, we **saved total of 7 lakh rupees and lost 2 lakh rupees.**

Conclusions

From the data collected and through exploratory data analysis, we can determine the following:

- The trend of flight prices vary over various months and across the holiday
- There are two groups of airlines: the economical group and the luxurious group. Spicejet, AirAsia, IndiGo, Go Air are in the economical class, whereas Jet Airways and Air India in the other. Vistara has a more spread out trend.
- The airfare varies depending on the time of departure, making timeslot used in analysisan important parameter.
- The airfare increases during a holiday season. In our time period, during Diwali the fareremained high for all the values of days to departure. We haven't considered holidayseason as a parameter now, since we are looking at data for a few months.
- Airfare varies according to the day of the week of travel. It is higher for weekends Monday and slightly lower for the other days.
- and
- There are a few times when an offer is run by an airline because of which the prices
- drop suddenly. These are difficult to incorporate in our mathematical models, andhence lead to error.
- Along the Mumbai-Delhi route, we find that the price of flights increases or remains constant as the days to departure decreases. This is because of the high frequency ofthe flights, high demand and also could be due to heavy competition.
- Only about 8-10% of the times, a person should wait according to the data collectedacross the Mumbai-Delhi route, compared to 30-40% in Delhi- Guwahati route.

REFERENCES

1. Amit Kumar Jaiswal, Ivan Panshin, Dimitrij Shulkin, Nagender Aneja, Samuel Abramov. 2019. Semi-Supervised Learning for Cancer Detection of Lymph Node Metastases. arXiv preprint *arXiv:1906.09587*

2. Rane, C., Mehrotra, R., Bhattacharyya, S. *et al.* A novel attention fusion network-based framework to ensemble the predictions of CNNs for lymph node metastasis detection. *J Supercomput* 77, 4201–4220 (2021). https://doi.org/10.1007/s11227-020-03432-6

3. Jaehoon Choi, Minki Jeong, Taekyung Kim, Changick Kim. 2019. Pseudo-LabelingCurriculum for Unsupervised Domain Adaptation. arXiv preprint arXiv:1908.00262

4. Chalbatani GM, Dana H, Memari F, Gharagozlou E, Ashjaei S, Kheirandish P, Marmari V, Mahmoudzadeh H, Mozayani F, Maleki AR et al (2019) Biological function and molecular mechanism of piRNA in cancer. Practical Lab Med 13:e00113

5. Bejnordi BE, Veta M, Van Diest PJ, Van Ginneken B, Karssemeijer N, Litjens G, Van Der Laak JA, Hermsen M, Manson QF, Balkenhol M et al (2017) Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 318(22):2199–2210

6. Nci-dictionaries. URL https://www.cancer.gov/publications/dictionaries/cancer-terms/def/ metastasis.

7. Li C, Wang X, Liu W, Latecki LJ, Wang B, Huang J (2019) Weakly supervised mitosis detection inbreast histopathology images using concentric loss. Med Image Anal 53:165–178

8. Kassani SH, Kassani PH, Wesolowski MJ, Schneider KA, Deters R (2019)

Classification of histopathological

9. biopsy images using ensemble of deep learning networks. arXiv preprint arXiv:1909.11870

10. Litjens G, Kooi T, Bejnordi BE, Setio AA, Ciompi F, Ghafoorian M, Van Der Laak JA, Van Ginneken B, Sánchez CI (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88

11. Liu Y, Gadepalli K, Norouzi M, Dahl GE, Kohlberger T, Boyko A, Venugopalan S, TimofeevA, Nelson PQ, Corrado GS, et al. (2017)