VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering
VIT Chennai
Vandalur - Kelambakkam Road, Chennai - 600 127

Final Review Report

Programme: Integrated MTech in Computer Science and Engineering with Specialization in Business Analytics

Course: Marketing Analytics

Slot: D2+TD2

Faculty: Dr. Padmavathy C

Component: J Component

Title: **Flight Price Prediction**

Team Member(s):

Sanjay Kumar (20MIA1008)

Ayush Madurwar (20MIA1009)

Himanshu Mittal (20MIA1035)

Kushagra Joshi (20MIA1054)

Tanmay Tiwari (20MIA1097)

Abhineet Raj (20MIA1146)

EXECUTIVE SUMMARY:

1. The goal of this project is to forecast flight prices using relevant flight information.
2. This project makes use of a four-month-long dataset of flight information and prices.
3. Flight duration, departure time, number of stops, airline, source and destination cities, route details, and seasonality/time of year are identified as significant independent variables in predicting flight prices.
4. The relationship between travel time and flight prices is investigated, with hypotheses about weekday versus weekend prices and peak versus non-peak prices tested.
5. To build predictive models, regression algorithms such as Linear Regression, Random Forest Regression, and Gradient Boosting Regression are used.
6. The data is divided into training and testing sets, and the evaluation metric is RMSE.
7. The project's goal is to create a reliable model for predicting flight prices that can be used by both the airline industry and consumers looking for good deals.
8. The best model is used to forecast flight prices for a second dataset with 2500 observations but no flight prices.
9. The project provides practical applications for the airline industry and consumers looking for the best flight prices.
10. Overall, the project is intended to contribute to the development of more accurate and reliable flight price prediction models.

PROBLEM STATEMENT:

The problem statement for this project is to develop a model that can accurately predict flight prices based on relevant flight details such as flight timings, source and destination details, route details, stops in between, etc. The model should be able to account for various factors that influence flight prices, such as airline, time of day, day of the week, seasonality, and more. The goal is to provide a reliable tool for the airline industry and consumers seeking to find the fairest prices for their flights. Additionally, the project aims to explore the relationship between time of journey and flight prices, testing hypotheses related to weekday versus weekend prices and peak versus non-peak prices. The project uses a dataset of flight information and prices spanning approximately four months, and employs regression algorithms to build predictive models. The models are evaluated using RMSE as the evaluation metric. The ultimate objective is to develop a reliable and accurate model for predicting flight prices that can be used in real-world applications.

OBJECTIVE:

The objective of this project is to predict the price of a flight when a set of relevant flight details are provided. This kind of model can be used to predict the fair price of a flight. The flight details has information related to flight timings, source and destination details, route details, stops in between etc.

The dataset consists of about four months of information about various flights and their respective prices. You are to analyze the data and provide the below.

1. Identify the Independent variables which are significant

2. Establish the relationship between time of journey and flight prices

3. Develop and test the hypothesis: (a) Flight Prices on Weekdays are cheaper than flight prices on weekends.

(b) Flight Prices during peak hours (9 AM till 9 PM) are costlier than flights at other times.

4. Build predictive models to predict flight prices. Split the data into training and test sets. Build supervised models on training data and test it on the test.

Use RMSE as the metric for model evaluation.

5. Use the best model to predict flight prices for which we do not have the flight prices. This is the second dataset attached. (It has 2500 observations for which flight prices have to be predicted).

METHODOLOGY:

Airlines utilize complex algorithms to compute flight prices based on the many factors that exist at the moment. To forecast flight fares, these strategies consider financial, marketing, and societal aspects.

The number of people taking airplanes has increased dramatically in recent years. Prices for airlines are tough to maintain because they alter dynamically due to many variables. As a result, we will attempt to solve this problem using machine learning. This can help airlines forecast what prices they can keep. It can also assist customers in forecasting future flight prices and planning their trip appropriately. So, We have used three methodologies that we are using for flight price prediction:

1. **Multiple Linear Regression**: - Multiple linear regression is a statistical technique that can be used for flight price prediction. It involves identifying a set of independent variables, or predictors, that may impact flight prices, and then building a regression model to estimate the relationship between these predictors and flight prices.

   Here's a general overview of how multiple linear regression could be applied to flight price prediction: Identify the independent variables:

(i) The first step is to identify a set of variables that may impact flight prices. This could include variables such as departure and arrival city, time of year, day of the week, time of day, flight duration, airline, and other factors that may influence flight prices.

(ii)Collect data: Once the variables have been identified, you will need to collect historical data on flight prices and the independent variables. This data can be obtained from a variety of sources, such as airline websites, online travel agencies, or data providers.

(iii)Data preparation: The next step is to prepare the data for analysis. This may involve cleaning and formatting the data, imputing missing values, and transforming variables as needed.

(iv)Build the model: Once the data is prepared, you can build a multiple linear regression model to estimate the relationship between the independent variables and flight prices. The model will produce a set of coefficients that can be used to predict flight prices based on the values of the independent variables.

(v)Evaluate the model: Once the model has been built, you will need to evaluate its performance. This may involve assessing the model's accuracy, identifying any biases or errors, and testing the model on new data to assess its generalizability.

It's important to note that multiple linear regression is just one of many methodologies that can be used for flight price prediction, and its performance will depend on the quality and relevance of the independent variables and the quality of the historical data used for model building.

2. **Decision Tree:-** Decision trees are a type of machine learning algorithm that can be used for flight price prediction. Decision trees are particularly useful when there are multiple independent variables and their interactions need to be considered to make accurate

predictions. Here's a general overview of how decision trees could be applied to flight price prediction:

(i) Identify the independent variables: The first step is to identify a set of independent variables that may impact flight prices. This could include variables such as departure and arrival city, time of year, day of the week, time of day, flight duration, airline, and other factors that may influence flight prices.

(ii) Collect data: Once the variables have been identified, you will need to collect historical data on flight prices and the independent variables. This data can be obtained from a variety of sources, such as airline websites, online travel agencies, or data providers.

(iii) Data preparation: The next step is to prepare the data for analysis. This may involve cleaning and formatting the data, imputing missing values, and transforming variables as needed.

(iv) Build the decision tree: Once the data is prepared, you can build a decision tree to predict flight prices based on the independent variables. The decision tree algorithm will identify the most important variables and create a set of if-then rules to make predictions.

(v) Evaluate the decision tree: Once the decision tree has been built, you will need to evaluate its performance. This may involve assessing the model's accuracy, identifying any biases or errors, and testing the model on new data to assess its generalizability.

One advantage of decision trees is that they are relatively easy to interpret, as the rules used for prediction are displayed in a tree-like structure. This can make it easier to understand how the algorithm is making its predictions and to identify which variables are most important for predicting flight prices. However, it's important to note that decision trees can be prone to overfitting, particularly if the tree is too deep or complex, so it's important to carefully evaluate and validate the model to ensure its performance.

3. **Random Forest Regressor**: - Random Forest Regressor is a machine learning algorithm that can be used for flight price prediction. It is a type of ensemble learning algorithm that combines multiple decision trees to make more accurate predictions. Here's a general overview of how Random Forest Regressor could be applied to flight price prediction:

   (i) Identify the independent variables: The first step is to identify a set of independent variables that may impact flight prices. This could include variables such as departure and arrival city, time of year, day of the week, time of day, flight duration, airline, and other factors that may influence flight prices.

   (ii) Collect data: Once the variables have been identified, you will need to collect historical data on flight prices and the independent variables. This data can be obtained from a variety of sources, such as airline websites, online travel agencies, or data providers.

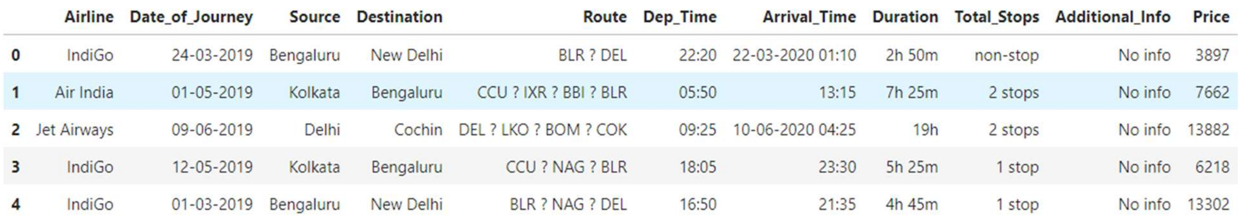
   (iii) Data preparation: The next step is to prepare the data for analysis. This may involve cleaning and formatting the data, imputing missing values, and transforming variables as needed.

   (iv) Build the Random Forest Regressor: Once the data is prepared, you can build a Random Forest Regressor to predict flight prices based on the independent variables. The algorithm will create a set of decision trees and combine their predictions to make a final prediction.

   (v) Evaluate the Random Forest Regressor: Once the Random Forest Regressor has been built, you will need to evaluate its performance. This may involve assessing the model's accuracy, identifying any biases or errors, and testing the model on new data to assess its generalizability.

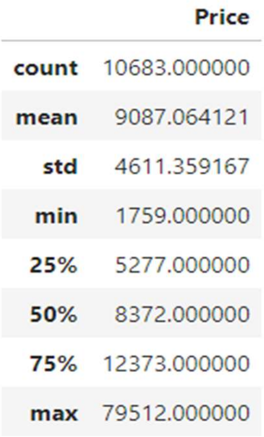
Implementation and findings:

1. We imported the necessary libraries and dataset.

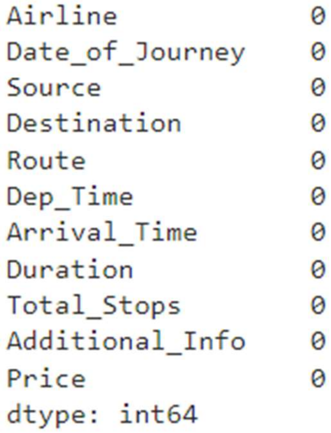


	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price
0	IndiGo	24-03-2019	Bengaluru	New Delhi	BLR ? DEL	22:20	22-03-2020 01:10	2h 50m	non-stop	No info	3897
1	Air India	01-05-2019	Kolkata	Bengaluru	CCU ? IXR ? BBI ? BLR	05:50	13:15	7h 25m	2 stops	No info	7662
2	Jet Airways	09-06-2019	Delhi	Cochin	DEL ? LKO ? BOM ? COK	09:25	10-06-2020 04:25	19h	2 stops	No info	13882
3	IndiGo	12-05-2019	Kolkata	Bengaluru	CCU ? NAG ? BLR	18:05	23:30	5h 25m	1 stop	No info	6218
4	IndiGo	01-03-2019	Bengaluru	New Delhi	BLR ? NAG ? DEL	16:50	21:35	4h 45m	1 stop	No info	13302

2. Further pre-processing of the dataset like number of null values and description and shape was carried in which we found that our data have eleven attributes and our data doesn’t have any null values.

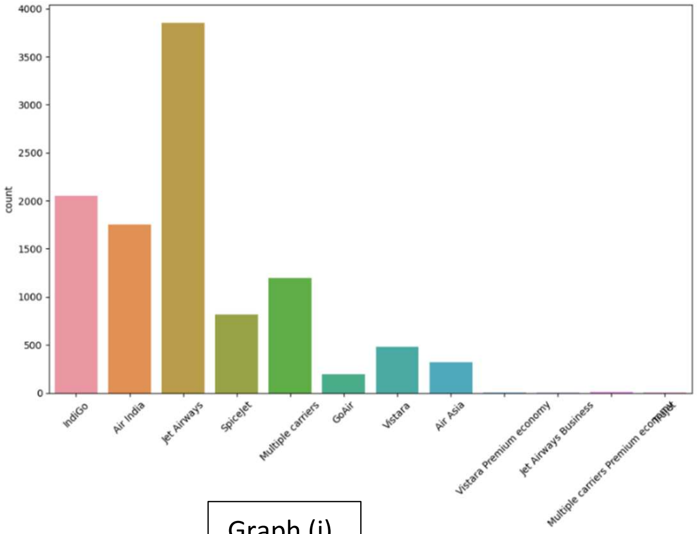


	Price
count	10683.000000
mean	9087.064121
std	4611.359167
min	1759.000000
25%	5277.000000
50%	8372.000000
75%	12373.000000
max	79512.000000

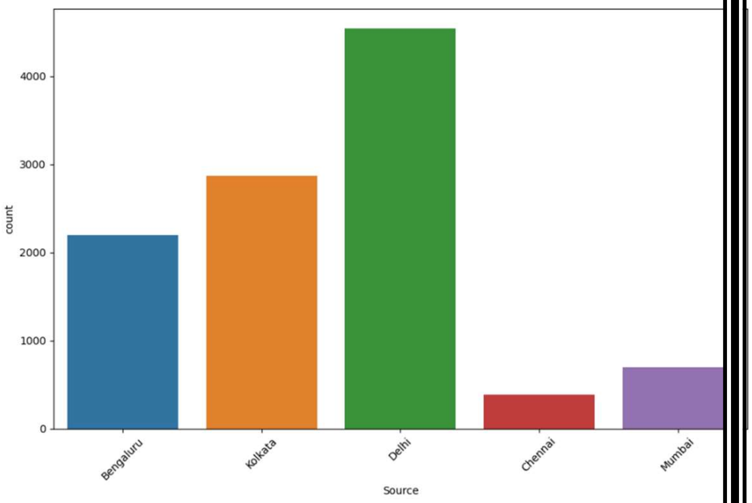


```
Airline            0
Date_of_Journey    0
Source             0
Destination        0
Route              0
Dep_Time           0
Arrival_Time       0
Duration           0
Total_Stops        0
Additional_Info    0
Price              0
dtype: int64
```

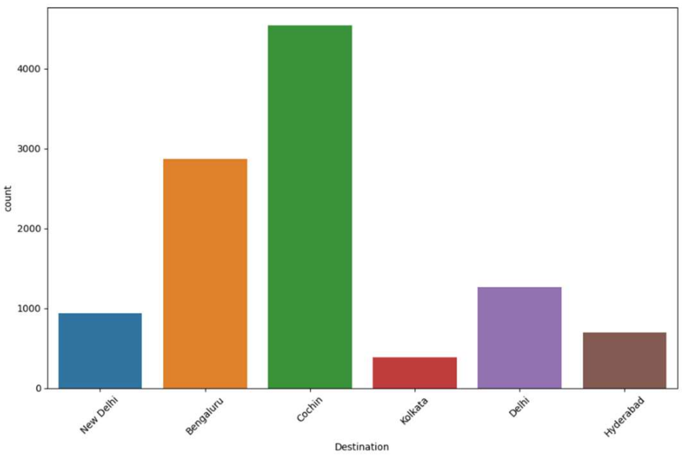
3. Now we have done the explatory data analysis of the dataset and it is described below in the detailed way:



Graph (i)

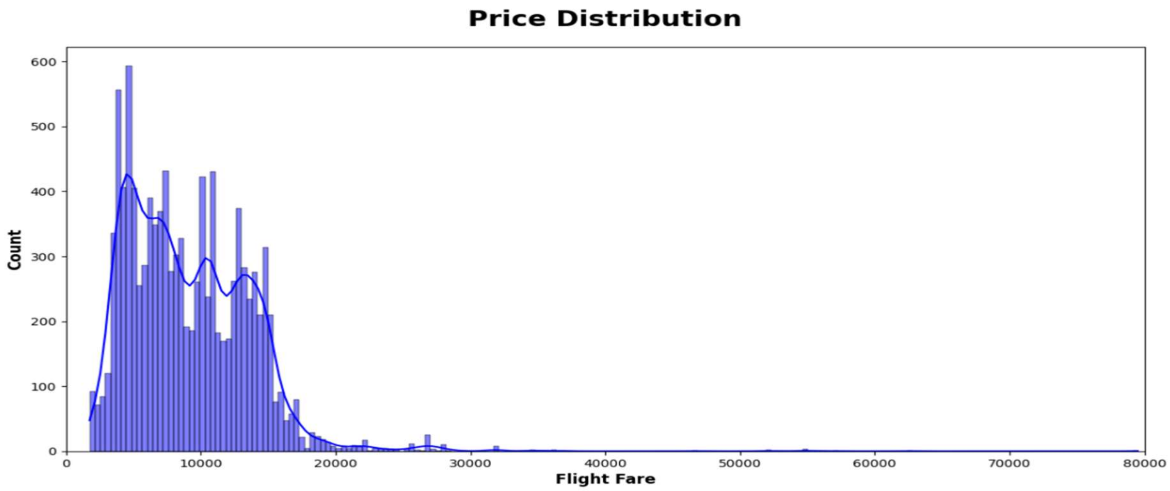


Graph (ii)

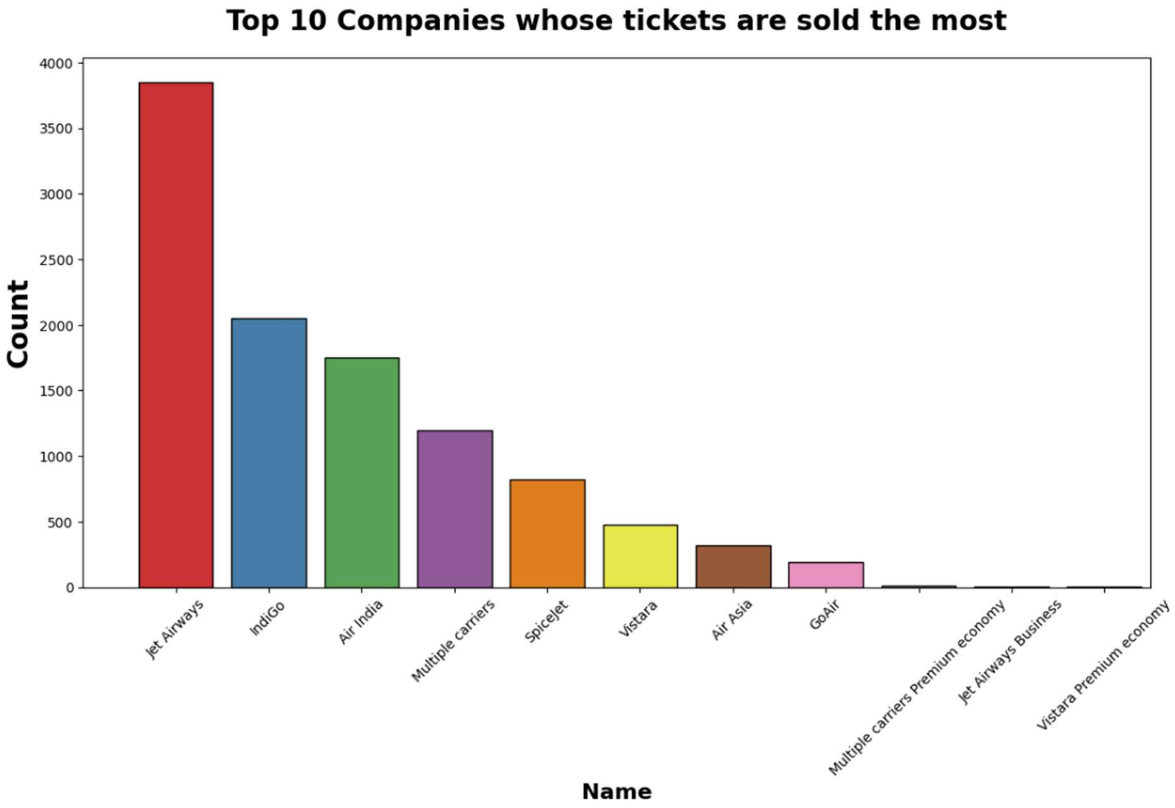


Graph (iii)

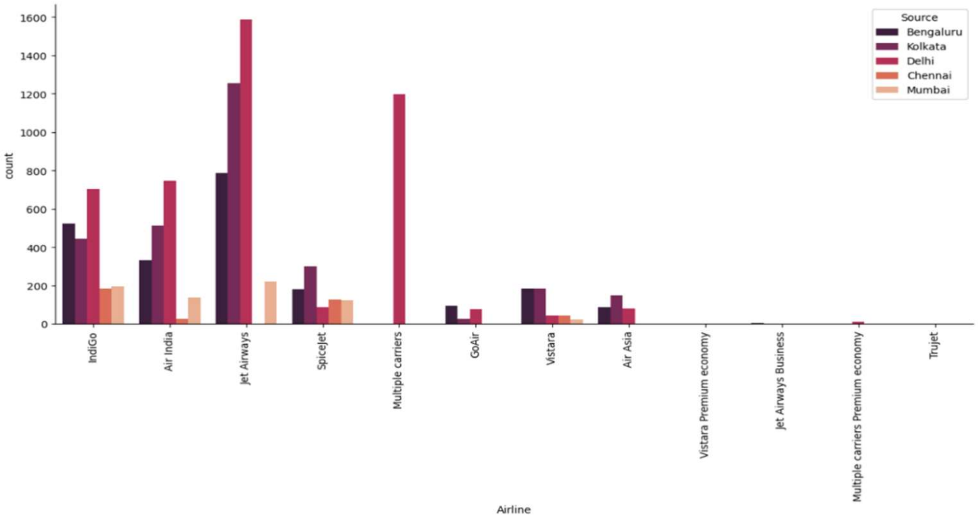
From the graph(i) we can observe that Jet airways have the maximum number of flights and from graph (ii) most of the flights are taking off from Delhi and from graph (iii) we got Cochin is selected as most destination city where flight is going.



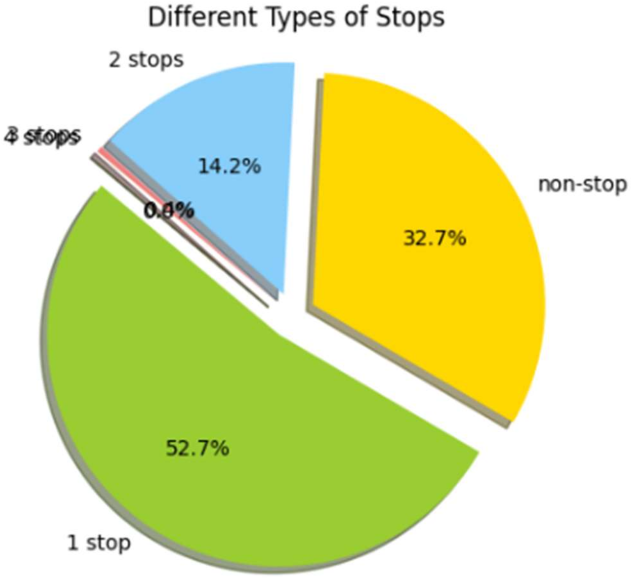
From the above price distribution graph, we can interpret that the most of the flight price ranges between 4000 to 8000 and later with the increase in the price the count of that prices are decreasing exponentially and it also suggest that people mostly travel flight between these price range.



The above graph shows the top 10 Airlines companies whose ticket are mostly sold and that preferred more by the people and top Airlines companies name are Jet Airways which is followed by Indigo and Air India and more.



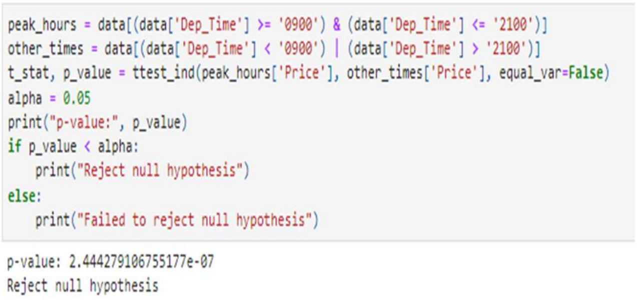
The above graph represents the number of flights tickets sold for each company from each source and we can clearly see that Jet Airways have most of flight from Bangaluru, Kolkata and Delhi and later on Indigo and Air India have highest number of flights.



The above pie graph represents that most of the flight have 1 stop in between the source and destination city and 32.7% are non-stop flights.

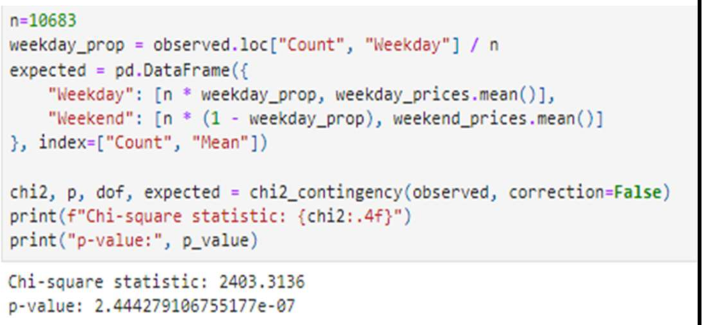
4. Now we have added one column named ‘DAY’ and one hot encoding to the source and destination city to perform the hypothesis testing.

5. As mentioned in the problem statement to do two hypothesis testing we got the p-value which is less than 0.05 which clearly states that there is significant flight prices difference during peak hours and usuals hours also in weekends and weekdays.



```
peak_hours = data[(data['Dep_Time'] >= '0900') & (data['Dep_Time'] <= '2100')]
other_times = data[(data['Dep_Time'] < '0900') | (data['Dep_Time'] > '2100')]
t_stat, p_value = ttest_ind(peak_hours['Price'], other_times['Price'], equal_var=False)
alpha = 0.05
print("p-value:", p_value)
if p_value < alpha:
    print("Reject null hypothesis")
else:
    print("Failed to reject null hypothesis")
```

```
p-value: 2.444279106755177e-07
Reject null hypothesis
```

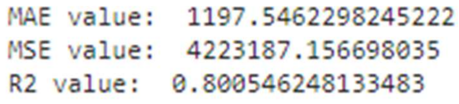


```
n=10683
weekday_prop = observed.loc["Count", "Weekday"] / n
expected = pd.DataFrame({
    "Weekday": [n * weekday_prop, weekday_prices.mean()],
    "Weekend": [n * (1 - weekday_prop), weekend_prices.mean()]
}, index=["Count", "Mean"])

chi2, p, dof, expected = chi2_contingency(observed, correction=False)
print(f"Chi-square statistic: {chi2:.4f}")
print("p-value:", p_value)
```

```
Chi-square statistic: 2403.3136
p-value: 2.444279106755177e-07
```

6. Now for model building we have separated the Date of Journey, Departure time and duration and done one hot encoding for Airline, Source and Destination and we have made a new file named as data_train which we have concat the dataset and dummies of Airline, Source and Destination.
7. We have performed three training model and their accuracy is mentioned below: -
   i. Multiple Linear Regression: We got 57% as accuracy.
   ii. Decision Tree: We got approx. 67% accuracy.
   iii. Random Forest Regressor: We got 90% accuracy.
8. In Random Forest Regressor we got the highest accuracy rate so we took this model and performed certain model evaluation and they are shown below:



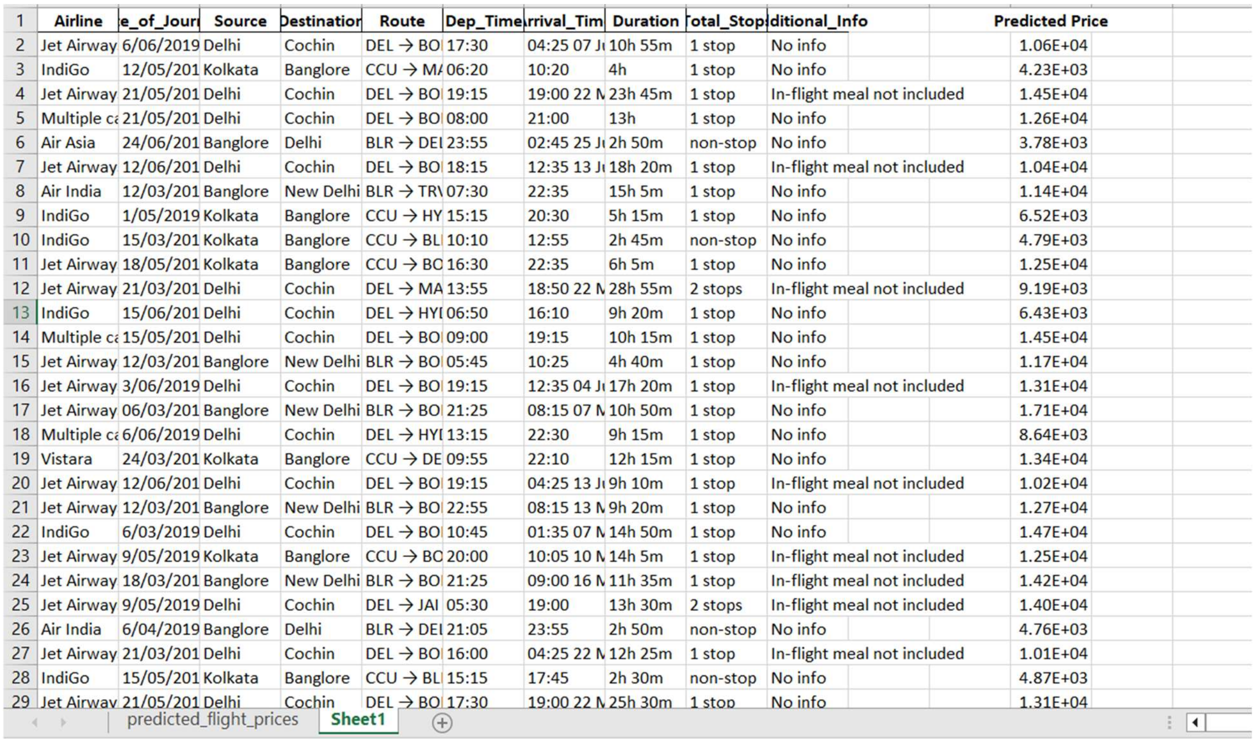
```
MAE value:  1197.5462298245222
MSE value:  4223187.156698035
R2 value:  0.800546248133483
```

The MAE value in this instance is 1197.55. This indicates that the predicted values and actual values diverge on average by 1197.55. The coefficient of determination, or R2 value, quantifies the percentage of variance in the dependent variable (i.e., the variable being predicted) that can be accounted for by the independent variables (i.e., the predictors). In this instance, the R2 value is 0.8005, which indicates that the independent variables can account for 80.05% of the variance in the dependent variable. The regression model may on the whole fit the data reasonably well, according to these results and it is best of all.

9. Now we save this model as a pickle file which help us in testing of our test_data.

10. Now we have loaded our test data and done some pre-processing steps similar to that of training dataset.

11. Finally, we imported our pre-trained pickle file and test the data and save those predicted price in new csv file.



1	Airline	e_of_Jour	Source	Destination	Route	Dep_Time	rrival_Tim	Duration	otal_Stop	ditional_Info	Predicted Price
2	Jet Airway	6/06/2019	Delhi	Cochin	DEL → BO	17:30	04:25 07 J	10h 55m	1 stop	No info	1.06E+04
3	IndiGo	12/05/201	Kolkata	Banglore	CCU → M/	06:20	10:20	4h	1 stop	No info	4.23E+03
4	Jet Airway	21/05/201	Delhi	Cochin	DEL → BO	19:15	19:00 22 N	23h 45m	1 stop	In-flight meal not included	1.45E+04
5	Multiple ca	21/05/201	Delhi	Cochin	DEL → BO	08:00	21:00	13h	1 stop	No info	1.26E+04
6	Air Asia	24/06/201	Banglore	Delhi	BLR → DEI	23:55	02:45 25 J	2h 50m	non-stop	No info	3.78E+03
7	Jet Airway	12/06/201	Delhi	Cochin	DEL → BO	18:15	12:35 13 J	18h 20m	1 stop	In-flight meal not included	1.04E+04
8	Air India	12/03/201	Banglore	New Delhi	BLR → TR\	07:30	22:35	15h 5m	1 stop	No info	1.14E+04
9	IndiGo	1/05/2019	Kolkata	Banglore	CCU → HY	15:15	20:30	5h 15m	1 stop	No info	6.52E+03
10	IndiGo	15/03/201	Kolkata	Banglore	CCU → BL	10:10	12:55	2h 45m	non-stop	No info	4.79E+03
11	Jet Airway	18/05/201	Kolkata	Banglore	CCU → BC	16:30	22:35	6h 5m	1 stop	No info	1.25E+04
12	Jet Airway	21/03/201	Delhi	Cochin	DEL → MA	13:55	18:50 22 N	28h 55m	2 stops	In-flight meal not included	9.19E+03
13	IndiGo	15/06/201	Delhi	Cochin	DEL → HYI	06:50	16:10	9h 20m	1 stop	No info	6.43E+03
14	Multiple ca	15/05/201	Delhi	Cochin	DEL → BO	09:00	19:15	10h 15m	1 stop	No info	1.45E+04
15	Jet Airway	12/03/201	Banglore	New Delhi	BLR → BO	05:45	10:25	4h 40m	1 stop	No info	1.17E+04
16	Jet Airway	3/06/2019	Delhi	Cochin	DEL → BO	19:15	12:35 04 J	17h 20m	1 stop	In-flight meal not included	1.31E+04
17	Jet Airway	06/03/201	Banglore	New Delhi	BLR → BO	21:25	08:15 07 N	10h 50m	1 stop	No info	1.71E+04
18	Multiple ca	6/06/2019	Delhi	Cochin	DEL → HYI	13:15	22:30	9h 15m	1 stop	No info	8.64E+03
19	Vistara	24/03/201	Kolkata	Banglore	CCU → DE	09:55	22:10	12h 15m	1 stop	No info	1.34E+04
20	Jet Airway	12/06/201	Delhi	Cochin	DEL → BO	19:15	04:25 13 J	9h 10m	1 stop	In-flight meal not included	1.02E+04
21	Jet Airway	12/03/201	Banglore	New Delhi	BLR → BO	22:55	08:15 13 N	9h 20m	1 stop	No info	1.27E+04
22	IndiGo	6/03/2019	Delhi	Cochin	DEL → BO	10:45	01:35 07 N	14h 50m	1 stop	No info	1.47E+04
23	Jet Airway	9/05/2019	Kolkata	Banglore	CCU → BC	20:00	10:05 10 N	14h 5m	1 stop	In-flight meal not included	1.25E+04
24	Jet Airway	18/03/201	Banglore	New Delhi	BLR → BO	21:25	09:00 16 N	11h 35m	1 stop	In-flight meal not included	1.42E+04
25	Jet Airway	9/05/2019	Delhi	Cochin	DEL → JAI	05:30	19:00	13h 30m	2 stops	In-flight meal not included	1.40E+04
26	Air India	6/04/2019	Banglore	Delhi	BLR → DEI	21:05	23:55	2h 50m	non-stop	No info	4.76E+03
27	Jet Airway	21/03/201	Delhi	Cochin	DEL → BO	16:00	04:25 22 N	12h 25m	1 stop	In-flight meal not included	1.01E+04
28	IndiGo	15/05/201	Kolkata	Banglore	CCU → BL	15:15	17:45	2h 30m	non-stop	No info	4.87E+03
29	Jet Airway	21/05/201	Delhi	Cochin	DEL → BO	17:30	19:00 22 N	25h 30m	1 stop	No info	1.31E+04

MAJOR FINDINGS:

- There is a significant difference flight prices during weekends and weekdays i.e., during weekends the flight price is more as compared to the weekdays.
- There is a significant difference in flight prices in peak hours (9AM-9PM) and other usual hours i.e., in early morning and late-night flight prices decreases than the usual high fare.
- The most preferred flight Aviation company is Indigo and Jet Airways according to the dataset.
- The flight price gradually decreases if the layover time is increased but in some premium airlines the price increases as layover goes beyond 7hours.
- The most preferred flight in the dataset is having non-stop flights or 1 stop flights.
- The flight prices mainly depend upon the source, destination, number of stops, time and day.
- The flight prices are higher to those areas where number of flights are less as compared to between famous metropolitan cities like Mumbai, Kolkata, Delhi, Bangalore.

Recommendation:

i. **Aviation Company**

- **Analyze supply and demand**: The airline should examine the demand for flights on various routes as well as the supply of available seats. They can modify the flight prices based on this analysis to maximize revenue and occupancy.
- **Use dynamic pricing**: Depending on variables like demand, seasonality, and competition, dynamic pricing adjusts flight prices in real-time. The aviation company can offer competitive prices while ensuring that the flights are profitable by implementing dynamic pricing.
- **Offer discounts and promotions**: To encourage customers to make flight reservations, the aviation company may offer discounts and promotions. Offering discounts for early bookings, group reservations, or loyalty programmes is one way to do this.
- **Data analytics** can be used by the aviation company to gain insights into consumer behaviors and preferences. They can use this information to make more informed decisions about flight costs and enhance the clientele's experience.
- **Offer bundled services**: The aviation company can offer bundled services, such as flights with hotels or car rentals, to increase revenue and provide more value to customers.

- **Invest in technology**: The aviation company can invest in technology to improve the customer experience and reduce costs. This can include implementing self-service kiosks, mobile check-in, and automated baggage handling.

- **Improve customer service**: Good customer service can help to build customer loyalty and increase revenue. The aviation company should focus on providing a positive customer experience through personalized service, efficient processes, and quick issue resolution.

ii. **Consumers**

- Be prepared to make a stopover because direct flights are frequently more expensive than those requiring a stopover. You might be able to find less expensive flights if you're willing to make a layover.

- When travelling off-peak, take advantage of airline sales. Sales are common throughout the year on airline tickets. To save money on your flights, keep an eye out for these sales and take advantage of them.

- Make sure your travel dates are flexible because doing so can help you save money. Peak travel periods like weekends, holidays, and the summer months tend to increase the cost of flights. You might be able to find less expensive flights if you can travel off-peak.

- The earlier you book your flight, the more likely it is that you will receive a lower rate. Booking your flight as far in advance as possible is ideal; however, sometimes doing so can result in lower costs.