# Flight Ticket Price Analysis and Prediction Using Python and Machine Learning

*Analysis of the flight booking dataset obtained from "Ease My Trip" website for the time frame February 11th to March 31st 2022 and ticket price prediction using Machine Learning algorithms like Random Forest, Linear Regression and Polynomial regression.*

**Naad Dantale**

Manipal University Jaipur
1ST Year B.Tech CSE

# ABSTRACT

Air travel has become an essential mode of transportation for people around the world. As a result, the airline industry has become highly competitive, and airlines are always looking for ways to improve their services and reduce costs. In this context, understanding the factors that affect flight ticket prices is crucial for both airlines and customers.

In this project, I have analyzed  the flight booking dataset obtained from the "Ease My Trip" website for the time frame of February 11th to March 31st, 2022, to identify the trends and patterns in flight ticket prices. I have used Python libraries like Pandas, Numpy, Matplotlib, Seaborn and SciKit Learn to perform exploratory data analysis (EDA) to gain insights into the data and prepare it for machine learning algorithms.

Furthermore, I have implemented machine learning algorithms such as Random Forest, Linear Regression, and Polynomial Regression to predict the ticket prices based on various features like the departure and arrival cities, the airlines, the date of travel, and other factors. I have used these algorithms to evaluate the performance of each model and select the best one based on accuracy and other evaluation metrics.

The insights gained from this project can be useful for airlines, travel agencies, and customers to make informed decisions about flight booking, pricing, and demand forecasting. By leveraging the power of data and machine learning, we can unlock new opportunities and drive innovation in the airline industry.

# DATASET

The Dataset was obtained from kaggle which is a popular website for Datasets and Machine Learning. The Dataset contains information about flight booking options from the website Easemytrip for flight travel between India's top 6 metro cities. There are 300261 data points and 11 features in the cleaned dataset.

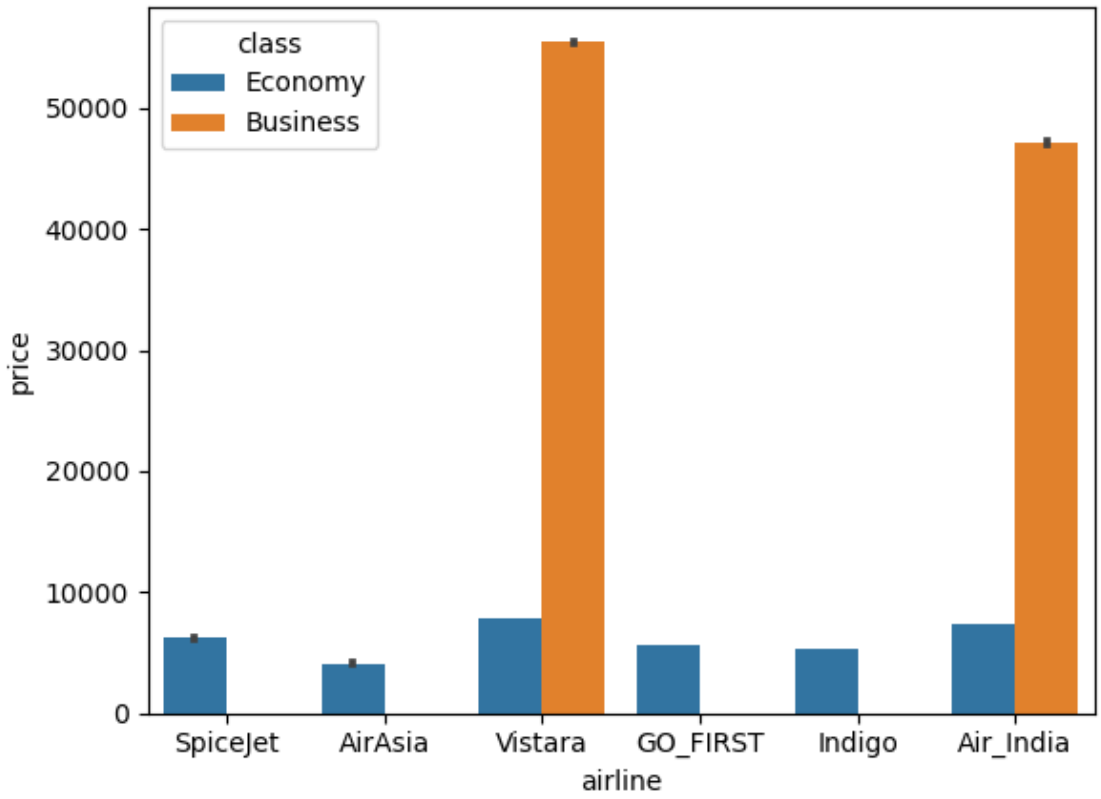
Time Frame - February 11th to March 31st, 2022.

The various features of the cleaned dataset are explained below:

1. Airline: The name of the airline company is stored in the airline column. It is a categorical feature having 6 different airlines.
2. Flight: Flight stores information regarding the plane's flight code. It is a categorical feature.
3. Source City: City from which the flight takes off. It is a categorical feature having 6 unique cities.
4. Departure Time: This is a derived categorical feature obtained by grouping time periods into bins. It stores information about the departure time and has 6 unique time labels.
5. Stops: A categorical feature with 3 distinct values that stores the number of stops between the source and destination cities.
6. Arrival Time: This is a derived categorical feature created by grouping time intervals into bins. It has six distinct time labels and keeps information about the arrival time.
7. Destination City: City where the flight will land. It is a categorical feature having 6 unique cities.
8. Class: A categorical feature that contains information on seat class; it has two distinct values: Business and Economy.
9. Duration: A continuous feature that displays the overall amount of time it takes to travel between cities in hours.
10. Days Left: This is a derived characteristic that is calculated by subtracting the trip date by the booking date.
11. Price: Target variable stores information of the ticket price.

# EXPLORATORY DATA ANALYSIS

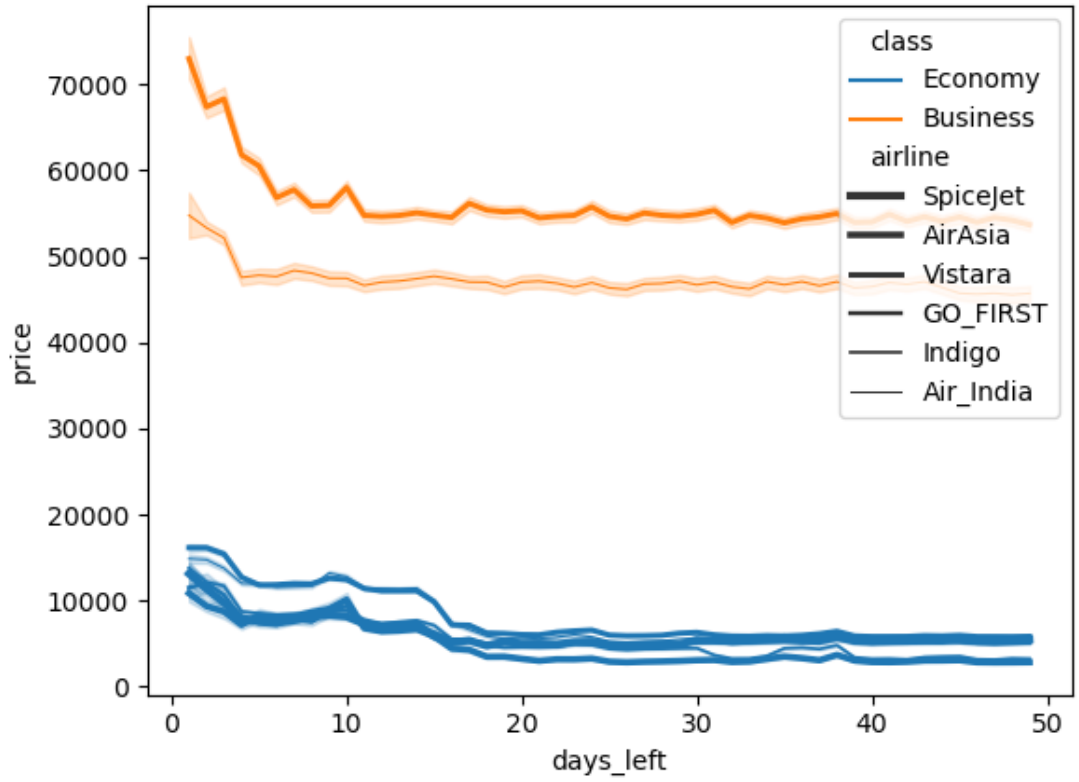
1. **Does price vary with Airlines?**

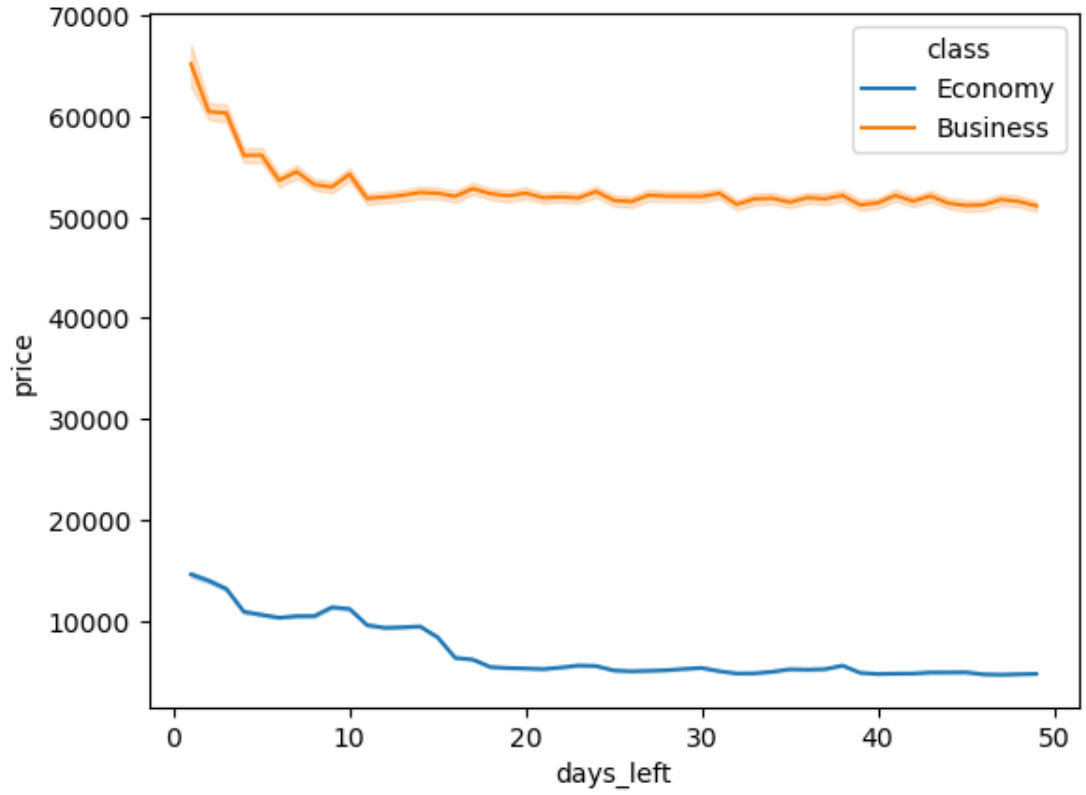



Airline choices are a significant factor in determining ticket prices. Specifically, the data indicates that ticket prices can vary significantly depending on the airline and the cabin class.

In Economy Class, Vistara was found to be the most expensive airline, while Air Asia was the least expensive. However, in Business Class, Vistara was the most expensive airline, while Air India was the least expensive.

2. **How is the price affected when tickets are bought just 1 or 2 days before departure?**



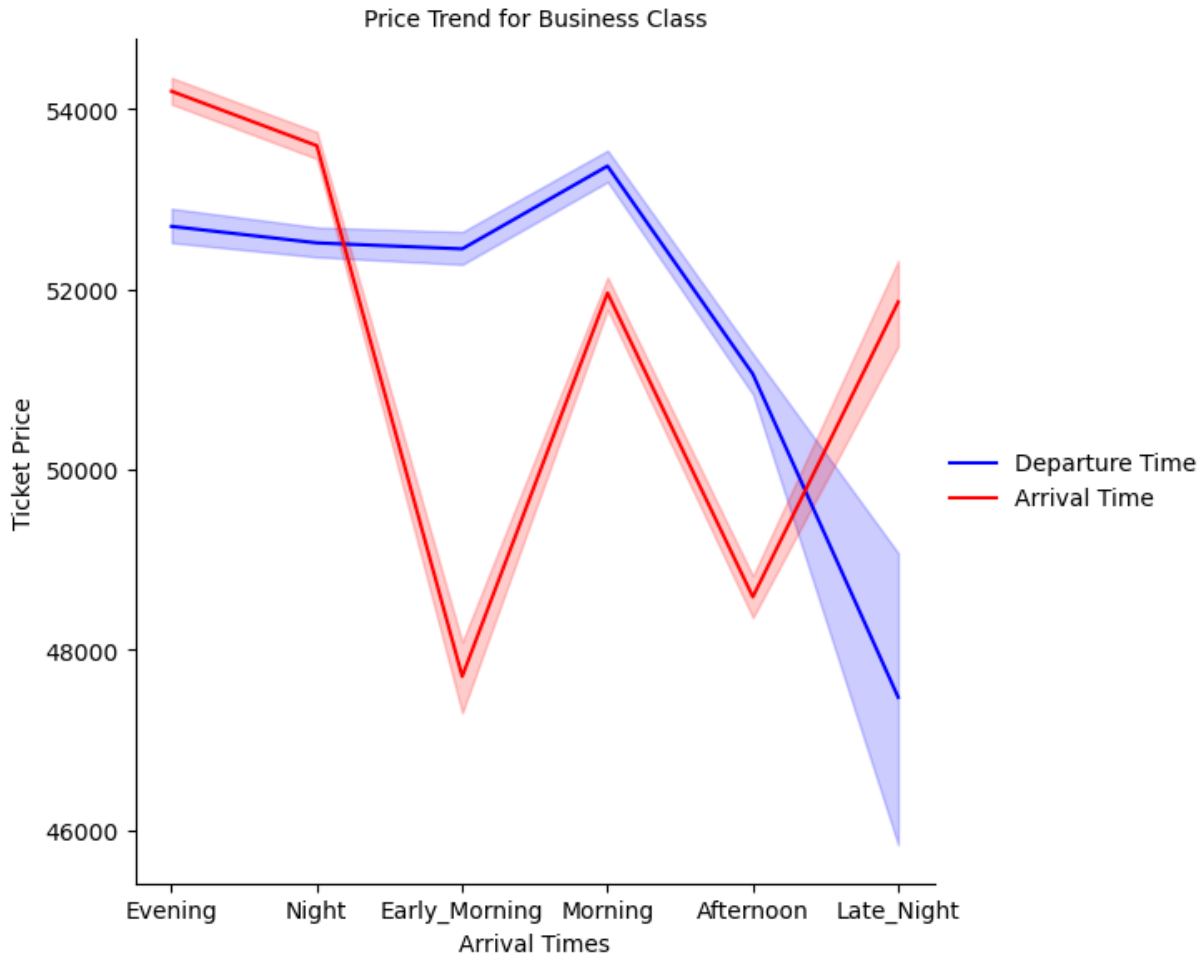



Based on the data analysis, it is evident that the prices of airline tickets rise considerably when purchased just 1-2 days before the scheduled departure, with Business class tickets being subject to even higher price hikes compared to Economy class tickets.

Furthermore, the data reveals that the prices of tickets remain relatively stable for an extended period before experiencing a sudden increase of around 30-35% for Business class tickets on Vistara Airlines, and approximately 15% for Air_India.

To secure the most cost-effective tickets, it is recommended to purchase Economy class tickets at least three weeks prior to the departure date, while for Business class tickets, it is advisable to buy them at least ten days before the scheduled departure.

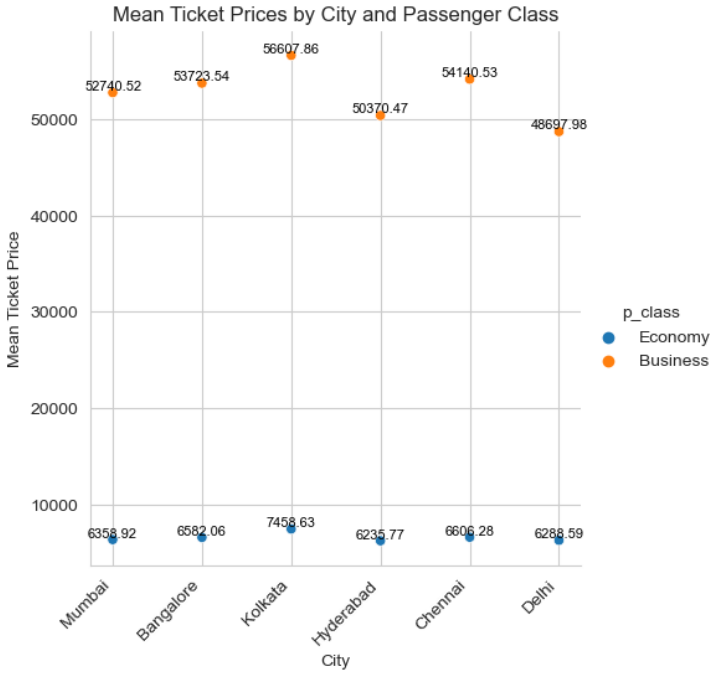
### 3. Does ticket price change based on the departure time and arrival time?

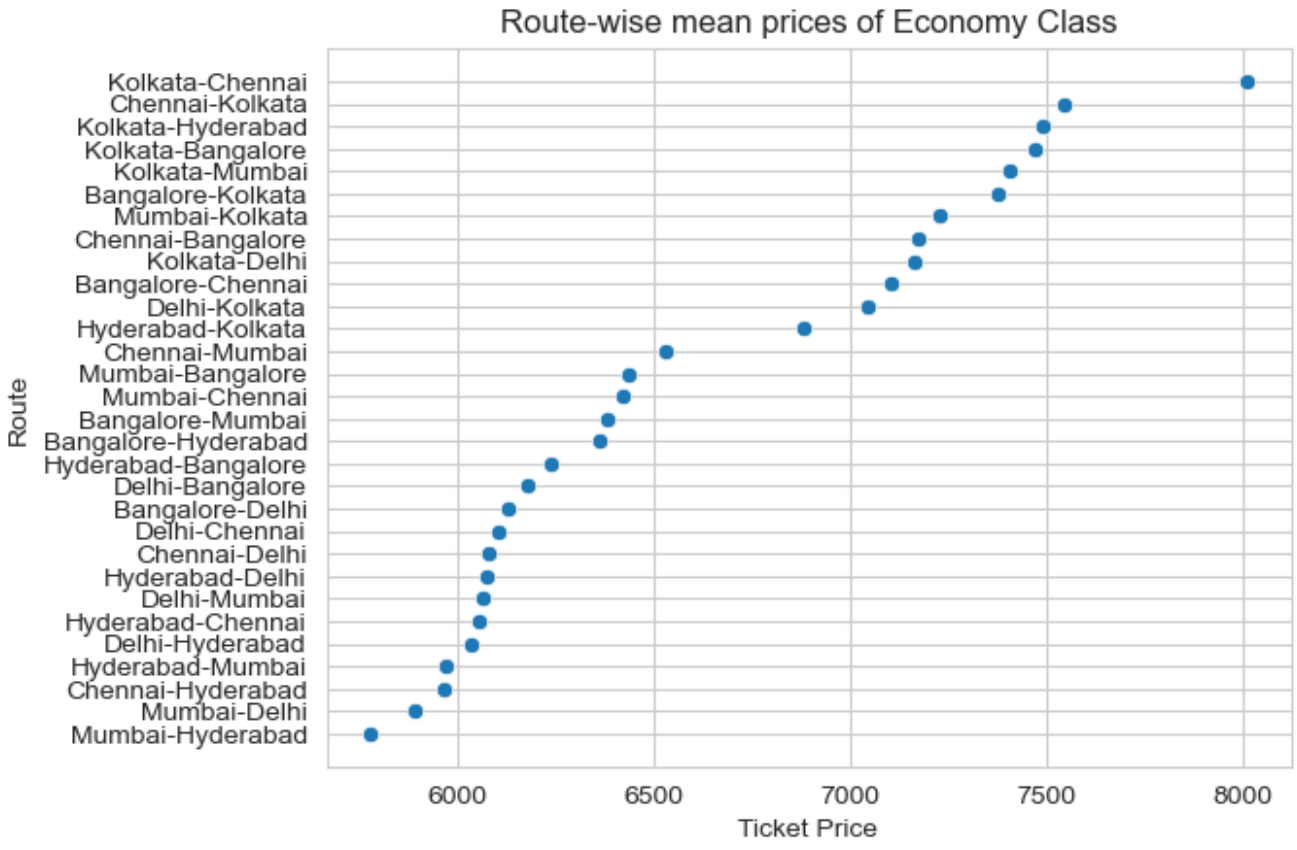


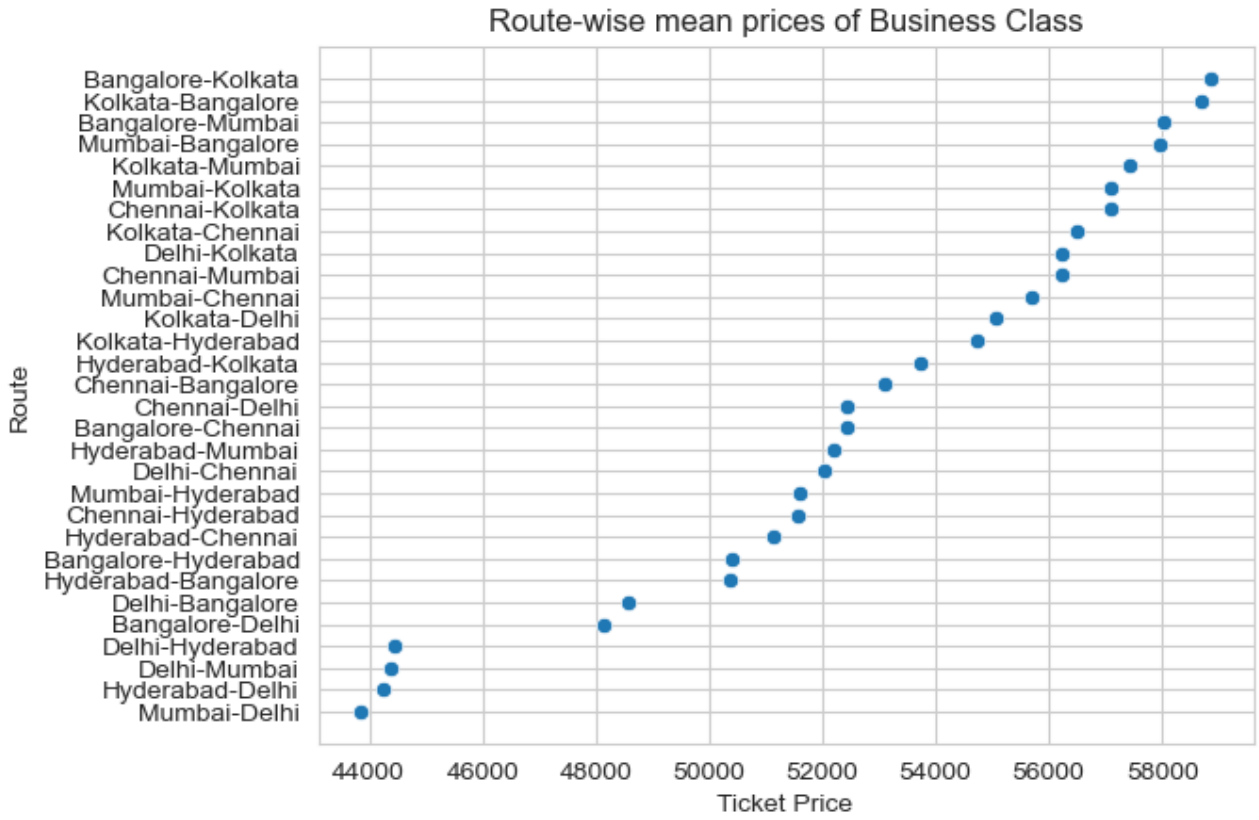

In the first grid, the departure time line plot is shown in blue and the arrival time line plot is shown in red. The x-axis represents the departure or arrival times, and the y-axis represents the ticket price. It is evident that the price tends to be higher for the Economy class when the departure time is in the morning, and it decreases as the day progresses, hitting a low in the evening before increasing again at night. Similarly, the price tends to be higher when the arrival time is in the afternoon or evening, with lower prices in the morning or at night.

In the second grid, the arrival time line plot is shown in blue, and the departure time line plot is shown in red. The x-axis represents the arrival time, and the y-axis represents the ticket price. The trend for the Business class is slightly different than that of the Economy class, with the ticket price tending to be highest in the afternoon and evening, and lower in the morning or at night.

## 4. How does the price change with change in Source and Destination?

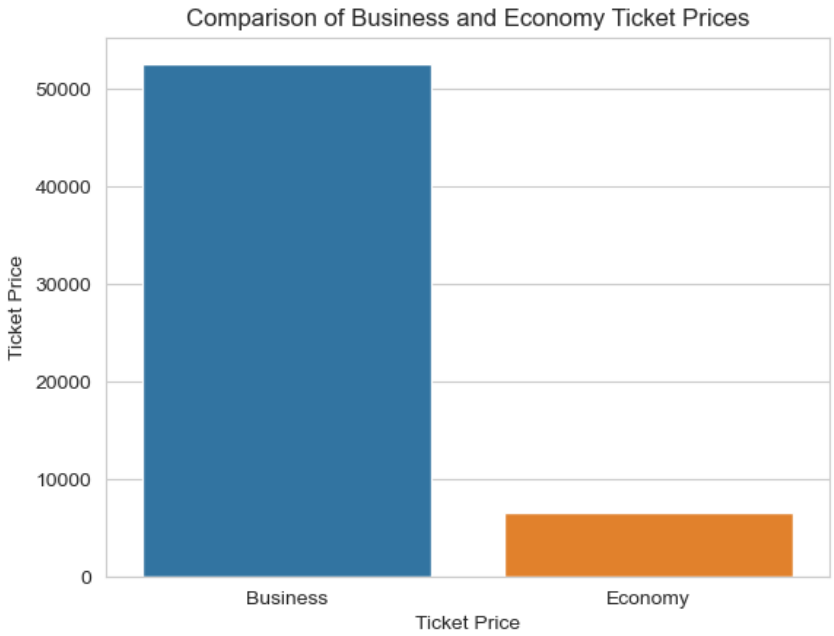



The data shows that flights departing from Delhi have the least expensive tickets, followed by Hyderabad, Mumbai, Bangalore, Chennai, and Kolkata, where tickets are the most expensive for departing flights. Similarly, tickets for flights arriving in Delhi are the least expensive, followed by Hyderabad, Mumbai, Bangalore, Chennai, and Kolkata, where tickets for arriving flights are the most expensive.

Regarding average ticket prices for specific routes, Chennai-Kolkata and Kolkata-Chennai have the most expensive economy tickets on average, while business class tickets for Bangalore-Kolkata and Kolkata-Bangalore are the most expensive. This is interesting because it does not make geographical sense and is likely due to logistical reasons.

For economy class, Mumbai-Hyderabad is the cheapest route, followed by Mumbai-Delhi and Chennai-Hyderabad. For business class, Mumbai-Delhi is the cheapest route, followed by Hyderabad-Delhi and Delhi-Mumbai.

## 5. How does the ticket price vary between Economy and Business class?




It is a well-known fact that Business Class tickets are generally more expensive than Economy Class tickets. This is due to the additional amenities and perks that come with Business Class, such as more legroom, better food, and priority boarding.

However, it is interesting to note that there are certain exceptions to this general rule. For instance, the most expensive Economy Class ticket from Chennai to Kolkata is priced at a whopping 42349Rs, which is higher than some Business Class tickets for other routes.

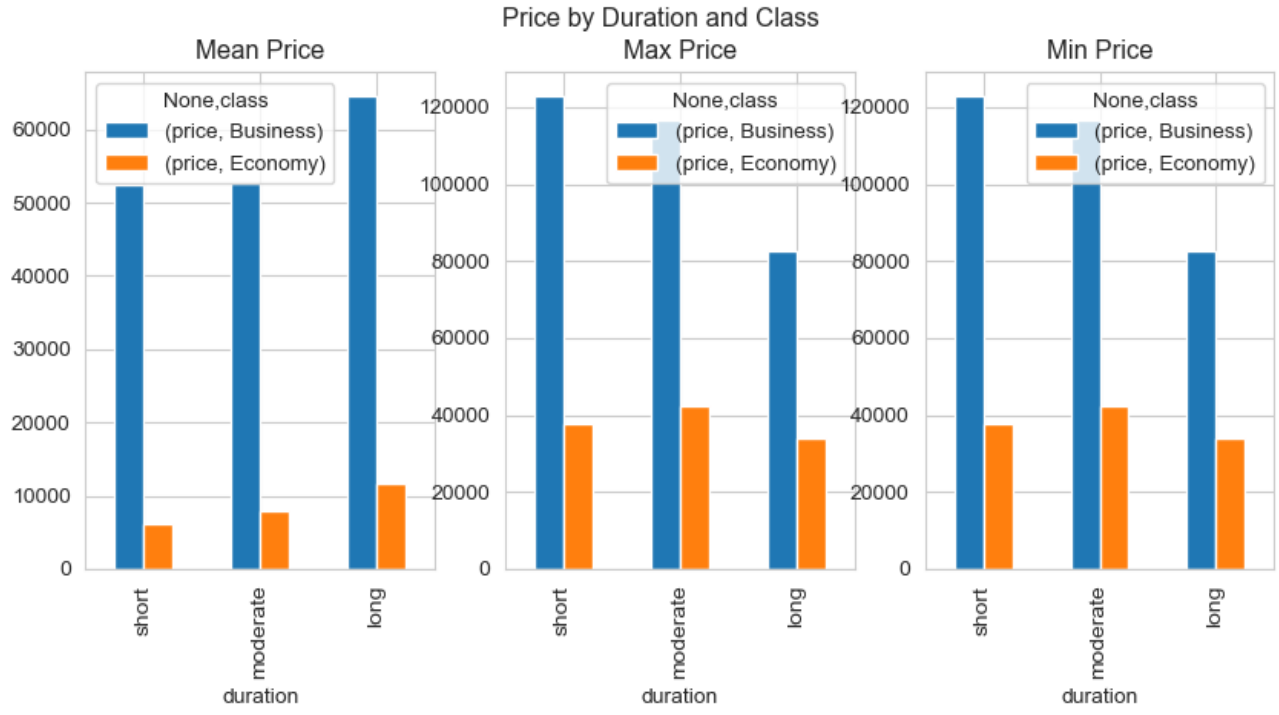
On the other hand, the cheapest Business Class ticket for the route from Bangalore to Hyderabad is only priced at 12000 Rs, which is significantly lower than the average Business Class ticket price. This phenomenon can be attributed to logistical and market factors, such as competition among airlines and demand for certain routes.

Another Fun Fact: Air India appears to offer both the most expensive economy class ticket and the least expensive business class ticket.It's not uncommon for airlines to have some variability in their pricing, and it's possible that Air India's pricing strategy includes offering both high-priced economy tickets and low-priced business tickets. There could be several reasons for this pricing strategy. For example, Air India may be

trying to attract more customers to its business class seats by offering them at a lower price point. At the same time, the airline may be looking to maximize revenue by charging a premium for its most expensive economy tickets. It's also possible that these pricing variations are a result of supply and demand factors, such as the availability of seats on particular flights or the popularity of certain routes. Overall, it's difficult to say for certain why Air India is offering both high-priced economy tickets and low-priced business tickets without more information about their pricing strategy and market factors.

Overall, while the price difference between Business and Economy Class tickets may be a general trend, there can be notable exceptions that are worth exploring and understanding in the context of the airline industry.

## 6. Does Flight Duration Affect the Ticket Prices?




On average, flights with longer durations tend to be more expensive than those with moderate or short durations, regardless of whether it is economy or business class. However, there are some interesting findings when it comes to the maximum ticket

prices. For economy class fliers, short distance flights surprisingly have costlier tickets than longer distance flights, which defies logical expectations. On the other hand, for business class, moderate distance flights tend to have costlier tickets. Overall, these results suggest that flight duration does indeed have an impact on ticket prices, and it is important to consider this factor when planning a trip.

## 7. How does the number of stops affect flight prices?


In analyzing the impact of flight duration on ticket prices, it was found that flights with two or more stops have the highest average prices, followed by one-stop flights and non-stop flights. However, it is interesting to note that one-stop flights have the highest maximum ticket prices, while two or more stops and non-stop flights have lower maximum prices. On the other hand, non-stop flights have the lowest minimum prices, while one-stop and two or more stops flights have higher minimum prices. This trend was observed for both economy and business classes, indicating a uniform pattern. Overall, flight duration plays a significant role in determining ticket prices, with longer flights typically costing more, especially when there are additional stops involved.

## Concluding the EDA:

**If you're someone who is looking to save money on flight tickets, there are several factors to consider before making your purchase. One of the most important factors is the timing of your flights. It's worth noting that the time of day and the day of the week can have a significant impact on the price of a ticket. In general, flights that depart and arrive at late-night times tend to be cheaper than those that depart and arrive during peak hours.****

**For flights that have a 0-stop journey time of 2-3 hours, late-night departure and arrival times are often a viable option. These flights typically depart after 9 or 10 pm and arrive at the destination in the early hours of the morning. While these flights may not be the most convenient in terms of scheduling, they can be a great way to save money on your ticket.**

**Another factor to consider when booking your flight is the number of stops it takes. In some cases, a flight with more stops may be cheaper than one with fewer stops, depending on the route and the airline. It's worth doing some research to find out which airlines offer the best deals for your chosen route and how many stops they typically make.**

**When it comes to finding the best deal on flight tickets, it's important to be flexible and open to different options. By considering factors such as timing, number of stops, and airline options, you can increase your chances of finding a great deal on your next flight.**

# Predictions Based on Data

**This code uses feature engineering techniques and machine learning models to predict flight prices. The steps involved are:**

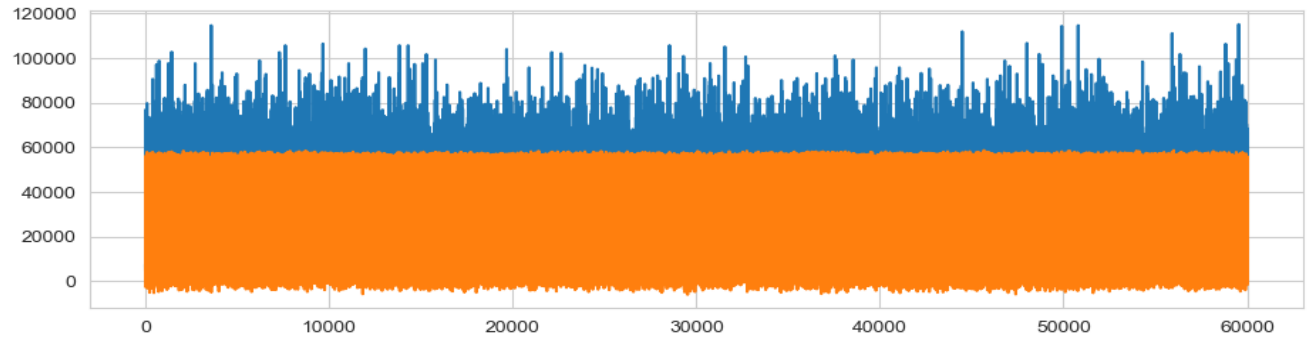
1. **Target encoding all categorical variables (airline, source city, departure time, stops, arrival time, destination city, class).**
2. **Scaling the data between 0 and 1.**
3. **Splitting the data into training and testing sets.**
4. **Training a Random Forest Regressor, a Linear Regression and a Polynomial Regressor on the training set with hyperparameters tuned through experimentation.**
5. **Predicting the flight prices using the trained model on the testing set.**
6. **Calculating accuracy, mean absolute error, and R-squared of the model.**

**The Random Forest Regressor achieved an accuracy of 98.55%, which is a very good score. The mean absolute error is 0.009, which means that the model's average prediction is off by around 0.9% of the actual price. The R-squared value of 0.98 indicates that the model can explain 98% of the variability in the data.**

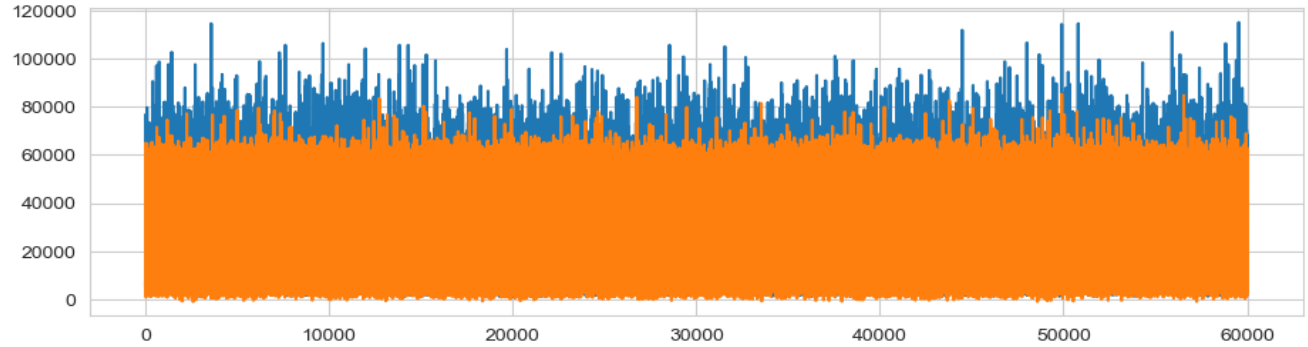
**The Polynomial Regression model performed relatively well, with an R-squared value of 0.943, accuracy score of 95.74% and a mean absolute error of 0.025. This indicates that the model can explain 94.3% of the variability in the data and has an average error of 2.5% in predicting the actual flight ticket prices.**

**The Linear Regression model performed relatively poorly, with an R-squared value of 0.902, accuracy score of 90.21% and a mean absolute error of 0.038. This indicates that the model can explain 90.21% of the variability in the data and has an average error of 3.8% in predicting the actual flight ticket prices.**

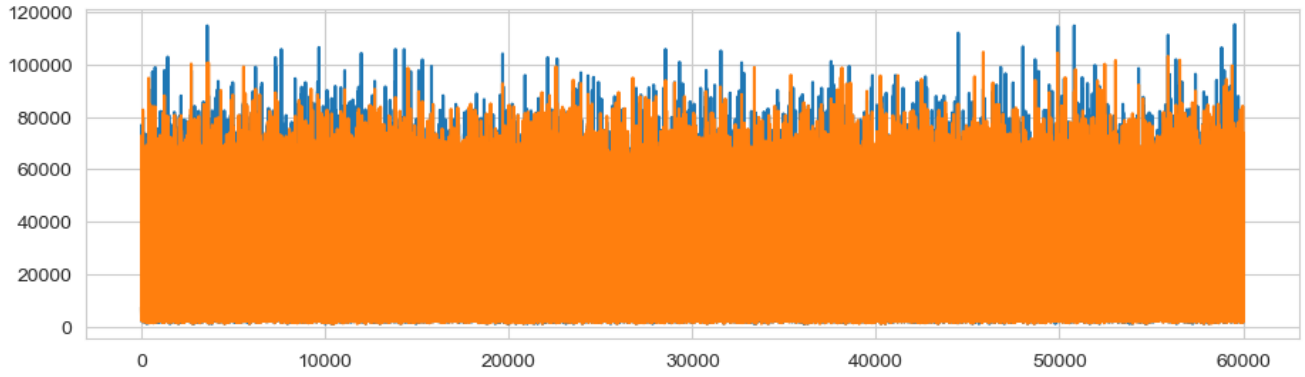
## Here are the plots of Predicted Values Overlayed on True Values for each model




LINEAR REGRESSION




POLYNOMIAL REGRESSION




RANDOM FOREST

After performing linear regression, polynomial regression, and random forest regression on the given dataset, it is evident that the model's fit improves as we move from linear regression to polynomial regression and then to random forest regression.

While linear regression is a simple and straightforward model, it may not capture the complex relationship between the variables. The polynomial regression, on the other hand, is more flexible and can fit non-linear relationships between variables by adding polynomial terms. Therefore, the polynomial regression model has a better fit than the linear regression model.

However, the random forest regression model shows even better performance as compared to both linear and polynomial regression models. This is because the random forest algorithm uses an ensemble of decision trees to make predictions, which allows it to capture complex non-linear relationships between the variables. Additionally, the random forest model reduces the risk of overfitting by aggregating the results of multiple decision trees.

Overall, the results demonstrate that the choice of model has a significant impact on the accuracy and performance of the predictions. The random forest regression model appears to be the best choice for this particular dataset, followed by polynomial regression and linear regression models.

# Experimental Hyperparameter Tuning for better accuracy scores

## 1. Random Forest - 98.5486%

**96.487%: max_depth=10**

**97.224%: max_depth=12**

**97.837%: max_depth=14**

**98.251%: max_depth=16**

**98.460%: max_depth=18**

**98.537%: max_depth=20** ¶

**98.543%: *max_depth=21***

- **98.544%: n_estimators=120**
- **98.5447%: n_estimators=150**
- **98.5476%: n_estimators=200**
- ***98.5486%: n_estimators=300***
- **98.5483%: n_estimators=500**
- **98.5486%: n_estimators=1000**

**98.5457%: max_depth=22**

- **98.5471%: n_estimators=300**

**98.533%: max_depth=23**

**98.525%: max_depth=24**

# CONCLUSION

This project aimed to predict flight ticket prices using machine learning algorithms. The dataset contained information about various flight parameters such as airline, route, source, destination, departure and arrival times, etc.

Three machine learning algorithms - Linear Regression, Polynomial Regression, and Random Forest - were implemented to predict ticket prices. The results showed that the Random Forest algorithm provided the best predictions with the highest accuracy score.

The feature importance analysis revealed that the airline, route, and the number of stops were the most significant factors in determining the ticket price.

The model can be used as a powerful tool to predict flight ticket prices by inputting a few parameters. This can help customers make informed decisions about flight bookings and plan their travel budget accordingly.

The project highlights the importance of using machine learning techniques in the airline industry to predict and optimize ticket prices. It also showcases the potential of data science in solving complex business problems.

Future work could involve incorporating more features such as weather conditions, holidays, and time of the year to improve the accuracy of the model. Overall, the project demonstrates the power of machine learning in providing valuable insights and predictions for the airline industry.

# REFERENCES


1. Dataset - https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction
2. Pandas Documentation - https://pandas.pydata.org/docs/
3. Numpy Documentation - https://numpy.org/doc/stable/index.html
4. Matplotlib Documentation - https://matplotlib.org/stable/index.html
5. Seaborn Documentation - https://seaborn.pydata.org/index.html
6. Scikit-Learn Documentation - https://scikit-learn.org/stable/index.html
7. Jupyter Notebooks - https://jupyter.org/