**Introduction**

Different airline industries in different countries have different rules and regulations due to which airline flight pricing strategies have developed into complex structures with sophisticated rules and mathematical models that drive the price strategies of flight tickets if all the factors and characteristics are taken into consideration. Traditional factors or variables such as distance, duration, and ticket class (economy and business) play a significant role but are no longer the only factors that dominate the pricing strategy. Economic, marketing and sociological factors have been increasingly influential in determining flight ticket rates. The majority of studies on forecasting flight prices have focused on the market. The airport pair between the flight origin and destination is defined as a market. Airlines must be able to forecast airfare trends to alter their strategies and resources for the market along a particular route. However, research that already exists on flight prediction uses conventional statistical models such as linear regression and is based on the assumption that the relationship between dependent variables and independent variables is a linear relationship, which is not true in many cases. The advancement of statistical models and machine learning (ML) makes it possible to let us infer rules and simulate fluctuations in flight costs based on a huge variety of factors, frequently automatically finding hidden links between them. The difficulties in acquiring access to the data in the former case make replicating the results and expanding the study practically impossible. The problem with the latter is that each online booking site's transaction records represent a small percentage of total ticket sales in the entire market because there are several sites and people also book tickets from agents and directly from the airport. For different reasons, data collection from the airport and thousands of agents and sites makes it time-consuming and impossible. This causes the data to be skewed, as most of the data is not normally distributed and thus does not reflect the true nature of the entire market. The purpose of this research paper is to analyze and investigate the relationship between several factors that might be playing an important role in the pricing strategy of airlines and to develop a statistical modeling framework to predict the flight price with the help of publicly available datasets from Kaggle. This research paper will include background information about the dataset, preprocessing, and cleaning of the dataset to make it the best fit to use in our research. Later, we will do some exploratory data analysis and build models and interpret and analyze them based on the knowledge we have gained in the data science course this semester, and finally, we will conclude and end the research paper with a list of references that we considered and used.

**Why did we choose this topic?**

When we are traveling by flight, we are always confused by shifting airplane ticket prices. Why are flight prices so exorbitant one minute and reasonable the next? What is the optimal time to book a flight to obtain the best price? When is the best time to get those great airplane ticket offers and discounts? As a result, we have compiled an interesting list of points to consider when purchasing an airline ticket. As a result, we are inclined to conduct research on predicting flight prices because it is relevant to our daily lives.

**An overview of the dataset**

The dataset contains statistics on flights between India's top six metro cities. The dataset contains 300,153 data points and 11 variables.

The different characteristics of the cleaned dataset are described in the following sections.

1. **Airline:** The airline column contains the name of the airline firm. It has six distinct airlines.
2. **Flight:**   The flight code contains information regarding the plane.
3. **Source City:** The location from which the flight departs. It has six unique cities.
4. **Departure Time**: The derived categorical feature was created by grouping time periods into bins. It stores departure time information and has six unique time labels.
5. **Stops:** A three-valued categorical feature that counts the number of stops made between the origin and destination cities.
6. **Arrival Time:** This is a derived categorical feature created by grouping time periods into bins. It stores arrival time information and has six unique time labels.
7. **Destination City:** The location where the flight will land. It has six unique cities.
8. **Class**: A categorical feature containing seat class information; it has two distinct values: business and economy.
9. **Duration:** the total number of hours required to travel between cities.
10. **Days Left:** Calculated by subtracting the trip date from the booking date.
11. **Price:** Target variable stores ticket price information.

Table

Description automatically generated

**Data Preprocessing**

Data preprocessing is the process of converting raw data into understandable and usable forms. The datasets that we collected from Kaggle were the raw data for us, and the dataset can be characterized by incompleteness, inconsistencies, lack of behavior, and trends while containing errors. Thus, preprocessing is necessary to standardize and resolve all the characteristics mentioned earlier. Preprocessing is essential to handle the missing values and address inconsistencies. During the data preprocessing stage, impossible data combinations such as arrival time and departure time (early morning, morning, night, and late-night) were handled. Missing values, null values, and redundancies were also addressed. Proper headings were given to column names, and we converted all the categorical variables into numeric, which we are going to use in the analysis and modeling. After preprocessing, the dataset is more reliable and relevant.

Table

Description automatically generated with medium confidence

**Normality Test**

The normal distribution, often known as “the Gaussian distribution”, is a probability function that represents the distribution of values for a variable. It is symmetric distribution in which the majority of observations cluster around the center peak, and the probability for values further from the mean tapers down equally in both directions, with fewer outliers at the extremes of the data range.

Chart, histogram

Description automatically generated

Normality tests are used to determine if data is derived from a Gaussian distribution or whether a variable or sample has a normal distribution. We have two methods for determining the normality of data.

1. Shapiro-Wilks Statistical Test.

|  |  |  |
| --- | --- | --- |
| **Variables** | **Statistics** | **P-Value** |
| **Source City** | 0.903 | 0.00 |
| **Departure Time** | 0.888 | 0.00 |
| **Stops** | 0.543 | 0.00 |
| **Arrival Time** | 0.896 | 0.00 |
| **Destination City** | 0.905 | 0.00 |
| **Class** | 0.583 | 0.00 |
| **Duration** | 0.956 | 0.00 |
| **Days Left** | 0.959 | 0.00 |
| **Price** | 0.752 | 0.00 |

**Result:** since all the p-values for all the variables are less than 0.05. Therefore, we will reject the null hypothesis. And thus, we can say that data is not normally distributed.

1. Plotting Q-Q plots.
2. Chart, line chart

   Description automatically generated Chart, line chart

   Description automatically generated Chart, line chart

   Description automatically generated
3. Chart, line chart

   Description automatically generated Chart, line chart

   Description automatically generated Chart, line chart

   Description automatically generatedChart, line chart

   Description automatically generatedChart, line chart

   Description automatically generatedChart, line chart

   Description automatically generated
4. re
5. re

**Result:** We plotted Q-Q plots for all variables to be clearer and to confirm the normality of our data. As we can see, there are horizontal lines on some of the plots, such as the destination city, the source city, class, arrival time, and departure time. These horizontal lines show the categorical nature of the variables. In the price and duration plots, the points follow a strongly nonlinear pattern.

Thus, with both the Shapiro-Wilk test and the Q-Q plot, we can say that the data is not normally distributed.