FLIGHT FARE PREDICTION

DATA ANALYSIS USING PYTHON

### INTRODUCTION

An airline is a company that provides air transport services for traveling passengers and freight. Airlines use aircraft to supply these services and may form partnerships or alliances with other airlines for codeshare agreements, in which they both offer and operate the same flight. Generally, airline companies are recognized with an air operating certificate or license issued by a governmental aviation body. Airlines may be scheduled or charter operators.

Airlines assign prices to their services in an attempt to maximize profitability. The pricing of airline tickets has become increasingly complicated over the years and is now largely determined by computerized yield management systems.

The price of an Airline Ticket is affected by a number of factors, such as flight duration, days left for departure, arrival time and departure time etc. Airline organizations may diminish the cost at the time they need to build the market and at the time when the tickets are less accessible. They may maximize the costs. The price may rely upon different factors. Each factor has its own proprietary rules and algorithms to set the price accordingly. Recent advances in Artificial Intelligence (AI) and Machine Learning (ML) makes it possible to infer such rules and model the price variation.

### DATA COLLECTION AND METHODOLOGY

Octo parse scraping tool was used to extract data from the website. Data was collected in two parts: one for economy class tickets and another for business class tickets. A total of 300261 distinct flight booking options was extracted from the site. Data was collected for 50 days, from February 11th to March 31st, 2022.  
Data source was secondary data and was collected from Ease my trip website.

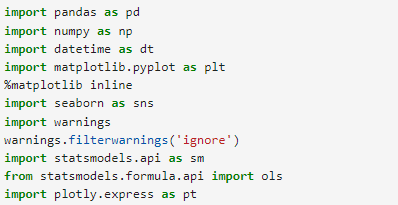
### DATASET

Dataset contains information about flight booking options from the website Easemytrip for flight travel between India's top 6 metro cities. There are 300261 datapoints and 11 features in the cleaned dataset.

### FEATURES

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 created by grouping time periods into bins. It stores information about the departure time and have 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.

### IMPORTING REQUIRED LIBRARIES



Read file and Data Checks to Perform

df = pd.read\_csv("Flight Fare.csv")

* By using ”pd.read\_csv” we can read the file in python.
* After that we can see data types, head and tail of the data set.
* By using “aircraft.duplicated().sum()” we can check is there any duplicated values in the data set.
* By using “aircraft.describe()” we can check the 5 number summary of the dataset.

### Data Visualization

### Flights Count of Different Airlines

### 

### Inference

* Indigo becoming as a most popular Airline

### Ticket Price Of Airlines

### 

### Insights

* From the above lineplot, we can able to understand that ticket price of airline vistara is highest followed by air\_india.
* Then the remaining airlines such as spicejet,airasia,go\_first and indigo have similar price range.

Classes of Different Airlines

### 

### Insights

* As we can see, compared to number of Business class tickets the number of economy class tickets are higher in most of the airlines

### Airlines Vs Price

### 

### Inference

How Does the Ticket Price vary between Economy and Business Class?

* Ticket Price is Maximum for Bussiness Class When compared to Economy Class

### Class Vs Ticket Price

### 

### Inference

How Does the Ticket Price vary between Economy and Business Class?

* Ticket Price is Maximum for Bussiness Class When compared to Economy Class

### Stops Vs Ticket Price

### 

### Inference

How Does the Ticket Price vary with the number of stops of a Flight?

* Flights having one stop has maximum ticket price

### Heatmap of three Columns

### 

### Inference

* Duration and Days Left: Weak or no correlation (near 0, light color)
* Duration and Price: Moderate positive correlation (around 0.2, light red)
* Days Left and Price: Strong negative correlation (around -0.09, dark blue)
* By interpreting these correlations, airlines can enhance their pricing strategies, improve marketing efforts, and optimize operational efficiency. Understanding these relationships allows for more informed decision-making, leading to better business outcomes and customer satisfaction.

### Departure Time Vs Ticket Price & Arrival Time Vs Ticket Price

### 

### Inference

How the Ticket Price change based on the Departure Time and Arrival Time?

1. Departure Time Vs Ticket Price

* Ticket Price is More for the Flights when the Departure Time is at Night
* Ticket Price is almost equal for flights Having Departure time at Early morning, Morning and Evening
* Ticket Price is Low for the Flights Having Departure Time at Late-night

1. Arrival Time Vs Ticket Price

* Ticket Price is More for the Flights when the Arrival Time is at Evening
* Ticket Price is almost equal for flights Having Arrival time is at Morning and Night
* Ticket Price is Low for the Flights Having Arrival Time at Late-night as same as Departure Time

### Source City Vs Ticket Price &

### Destination City Vs Ticket Price

### 

### Inference

How the price changes with change in Source city and Destination city?

1. Source City Vs Ticket Price

* Ticket Price is More for the Flights whose Source City is Kolkata
* Ticket Price is almost equal for flights Having Source Cities as Mumbai and Chennai, Hyderabad and Bangalore
* Ticket Price is Low for the Flights Having Source City as Delhi

1. Destination City Vs Ticket Price

* Ticket Price is More for the Flights whose Destination City is Kolkata and Chennai
* Ticket Price is almost equal for flights Having Destination Cities as Mumbai and Bangalore
* Ticket Price is Low for the Flights Having Destination City as Delhi

### Statistical Inference of Chi-Square Test

Airline & Class Columns

* The p-value of 0.0 indicates very strong evidence against the null hypothesis, implying that there is a significant association between the airline and the class of passengers. This means that the distribution of passengers in Business and Economy classes is not independent of the airline.
* The high Chi-square statistic (60492.80) suggests a large deviation between the observed and expected frequencies, confirming the strong association between the airline type and passenger class.
* The Chi-square test indicates a significant relationship between the airline and the class of passengers, highlighting distinct market strategies among different airlines.

Source City & Destination City

* p-value: The p-value is 0.0 (or extremely close to 0, below the machine precision limit). This indicates strong evidence against the null hypothesis, suggesting that the variables are not independent.
* There is a statistically significant association between the two categorical variables under consideration. The large discrepancy between the observed and expected frequencies, as reflected by the high Chi-squared statistic, supports this conclusion. Therefore, we can infer that the variation observed in the data is not due to random chance, and there is a meaningful relationship between the variables.
* The chi-square test indicates a significant association between source and destination cities, revealing important patterns in flight frequencies. These insights can help airlines optimize their operations, allocate resources more efficiently, and tailor their strategies to meet market demand. Understanding these patterns enables data-driven decision-making, enhancing both operational efficiency and customer satisfaction.

### Conclusion

Flight fare prediction involves analysing various factors that influence ticket prices, such as:

**Route and Distance:** Fares vary based on the distance between departure and destination cities.

**Airline and Class:** Different airlines and travel classes (economy, business) offer varying price points.

**Time and Season:** Prices fluctuate with demand, time of booking, and seasonal variations.

**Additional Factors:** Factors like stopovers, day of the week, and advance booking time also impact fares.

By leveraging Python's data analysis, businesses can make informed decisions regarding pricing strategies and customer offerings, enhancing operational efficiency and customer satisfaction in the airline industry.