ANALYSIS OF AIR FAIR'S & INDUSTRY TRENDS

Exploratory Data Analysis Project |

K.Siva Sankar, M.Pavan Kumar Reddy

Contents

Analysis of Air Fair's & Industry Trends

Introduction :

In today's fast-paced world, the aviation industry plays a crucial role in connecting people and facilitating global mobility. As air travel continues to grow in popularity, travelers are constantly seeking the most convenient and cost-effective flight options available. For both individuals and businesses, accessing timely and accurate information about flights, prices, and schedules is essential for making informed travel decisions.

This project aims to investigate the determinants of ticket prices across different classes in the airline industry. By analyzing a comprehensive dataset encompassing various attributes related to flights, passengers, routes, and pricing strategies, we seek to uncover the underlying patterns and drivers shaping pricing differentials among different travel classes.

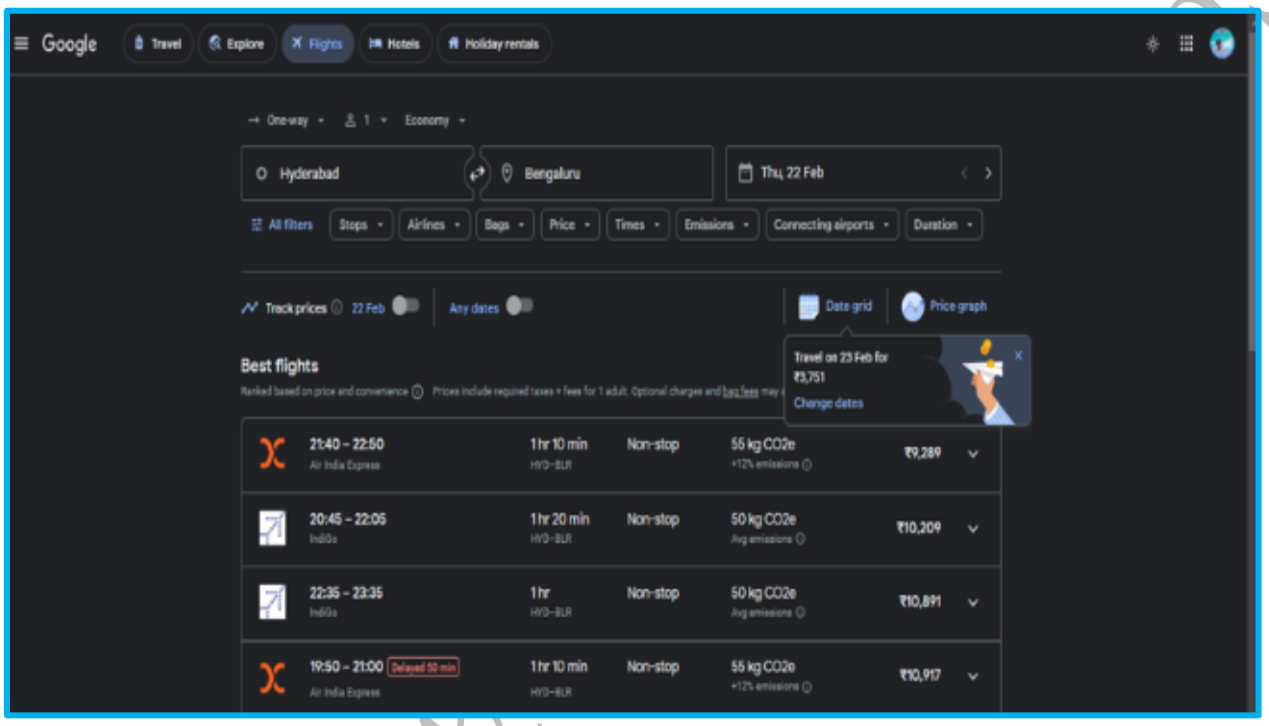
Steps: The steps in our analysis of air fair's & industry trends project is defined as follows:

1. **Binary Classification:** The primary objective of this project is to predict whether the price of a flight will be high or low based on certain features or attributes.
2. **Feature Set:** The features used for classification encompass a variety of flights, departure and arrival airports, flight distance, number of layovers, booking class and price.
3. **Dataset Source:** The dataset used in this project is sourced from google flights. This dataset has been curated to serve as the foundation of our analysis and modeling efforts in the context of airline pricing and flight booking. It provides a representation of various factors influencing flight prices and booking preferences, making our project both practical and applicable in the domain of air travel.
4. **Data Preprocessing and Analysis:** Data preprocessing will involve tasks such as handling missing values, encoding categorical variables, and scaling features. We will also perform data analysis to gain insights into the dataset's characteristics and distributions.

Data Source:

The dataset was collected from Google flights.

Dataset name: Analysis of Air Fair's
Dataset link: Google Flights – Find cheap flight options and track prices



Importing Libraries:

```
In [1]: import numpy as np # linear algebra
        import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
        import regex as re
        from datetime import datetime
        import matplotlib.pyplot as plt
        import seaborn as sns
        import networkx as nx
        import plotly.express as px
```

```
In [2]: # ignore warnings

        import warnings
        warnings.filterwarnings('ignore')
```

Import Dataset:

I will import the dataset with the usual pandas read_csv() function which is used to import CSV (Comma Separated Value) files.

```
In [3]: economy=pd.read_csv(r"C:\Users\pc\batch 261 innomatics\Flights project EDA\economy.csv",index_col=0)
        p_economy=pd.read_csv(r"C:\Users\pc\batch 261 innomatics\Flights project EDA\premiumeconomy.csv",index_col=0)
        business=pd.read_csv(r"C:\Users\pc\batch 261 innomatics\Flights project EDA\business.csv",index_col=0)
        first=pd.read_csv(r"C:\Users\pc\batch 261 innomatics\Flights project EDA\first.csv",index_col=0)

In [4]: data=pd.concat([economy,p_economy, business,first],ignore_index=True)
```

Raw Dataset:

```
In [7]: data
```

Out[7]:

	Name	Cabin	Departure Time	Arraival Time	Departure Place	Arraival Place	Total Travel Time	Stops	Laguage Weight	Prices
0	IndiGo	Economy	4:35 AM	9:45 AM	HYD	DEL	5 hr 10 min	1	140 kg CO2e	₹7,690
1	SpiceJet	Economy	6:05 AM	8:15 AM	HYD	DEL	2 hr 10 min	0	155 kg CO2e	₹9,538
2	Akasa Air	Economy	8:10 AM	10:40 AM	HYD	DEL	2 hr 30 min	0	105 kg CO2e	₹9,745
3	Air India	Economy	6:15 AM	8:35 AM	HYD	DEL	2 hr 20 min	0	104 kg CO2e	₹9,830
4	Vistara	Economy	1:15 PM	3:40 PM	HYD	DEL	2 hr 25 min	0	125 kg CO2e	price unavailable
...	...	...	...	...	...	...	...	...	...	...
3697	Air India, Singapore Airlines	First	10:30 PM	7:40 AM+2	HYD	SYD	27 hr 40 min	2	4,074 kg CO2e	₹814,585
3698	Air India, Singapore Airlines	First	5:05 PM	9:40 PM+1	HYD	SYD	23 hr 5 min	2	5,132 kg CO2e	₹958,892
3699	Air India, Singapore Airlines	First	1:20 PM	9:40 PM+1	HYD	SYD	26 hr 50 min	2	5,162 kg CO2e	₹978,548
3700	Air India, Singapore Airlines	First	2:10 PM	9:40 PM+1	HYD	SYD	26 hr	2	5,408 kg CO2e	₹1,010,720
3701	Air India, Singapore Airlines	First	4:25 PM	9:40 PM+1	HYD	SYD	23 hr 45 min	2	5,332 kg CO2e	₹1,010,720

3702 rows × 10 columns

Data Description:

- Totally there are 10 input features and target class label in the dataset:

Airline: Name of the flights

Cabin: Economy, Premium Economy, Business, First

Departure Time: The time when the flight departs.

Arrival Time: The time when the flight arrivals.

Departure Place: The location from which the flight departs.

Arrival Place: The location from which the flight arrivals.

Total Duration: The duration of the flights.

Stops: The number of stops during the flights.

Emission(in kgs): The emission level associated with the flight in kilograms.

Prices: The price associated with the flight.

What is Data Cleaning?

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.

Steps involved in data cleaning:

- Remove duplicate or irrelevant observations
- Fix structural errors
- Filter unwanted outliers
- Handle missing data

Data Cleaning In Dataset:

- **Handling Missing Values:** Identify missing values using “isna()” or “info()” and handle them using “dropna()” methods.
- **Standardizing Data Formats:** Ensure consistency in data formats, such as date or text fields, using the “to_datetime()” function.
- **Correcting Errors:** Identify and correct data errors, such as typos or incorrect values, using “replace()” function.

- **Leading Whitespace Removal:**
  - In dataset, 'object' type column names and string values often started with leading whitespace.
  - To enhance data consistency, we systematically removed these leading whitespaces.
- **Data Type Casting:**
  - In our dataset, the 'prices' and 'emission(in kgs)' features were originally in string format, represented with the object data type.
  - To facilitate analysis and modeling, these features were transformed into discrete numerical values, resulting in a change of data type to integers (int).
  - This conversion enables us to work with these features in a more suitable format for our classification tasks.
- **Handling Outliers:** Identify and handle outliers using statistical methods like the interquartile range (IQR) or z-score.
- **Data Validation:** Validate data integrity and consistency to ensure it meets expected standards and business rules.

Cleaned Dataset:

```
In [50]: data
```

Out[50]:

	Airline	Cabin	Departure Time	Arrival_Time	Departure Place	Arrival_place	Flight_Durations	Stops	Emission(in kgs)	Prices	Departure_Hour	Arrival_Hour
0	IndiGo	Economy	4:35 AM	9:45 AM	HYD	DEL	310	1	140	7690	4	9
1	SpiceJet	Economy	6:05 AM	8:15 AM	HYD	DEL	130	0	155	9538	6	8
2	Akasa Air	Economy	8:10 AM	10:40 AM	HYD	DEL	150	0	105	9745	8	10
3	Air India	Economy	6:15 AM	8:35 AM	HYD	DEL	140	0	104	9830	6	8
4	IndiGo	Economy	1:50 AM	7:15 AM	HYD	DEL	325	1	140	7690	1	7
...	...	...	...	...	...	...	...	...	...	...	...	...
2508	Air India, Singapore Airlines	First	10:30 PM	7:40 AM	HYD	SYD	1660	2	4074	814585	22	7
2509	Air India, Singapore Airlines	First	5:05 PM	9:40 PM	HYD	SYD	1385	2	5132	958892	17	21
2510	Air India, Singapore Airlines	First	1:20 PM	9:40 PM	HYD	SYD	1610	2	5162	978548	13	21
2511	Air India, Singapore Airlines	First	2:10 PM	9:40 PM	HYD	SYD	1560	2	5408	1010720	14	21
2512	Air India, Singapore Airlines	First	4:25 PM	9:40 PM	HYD	SYD	1425	2	5332	1010720	16	21

2507 rows × 12 columns

Summary of Dataset:

```
In [37]: data.info()

<class 'pandas.core.frame.DataFrame'>
Index: 2507 entries, 0 to 2512
Data columns (total 12 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   Airline           2507 non-null   object
 1   Cabin             2507 non-null   object
 2   Departure Time    2507 non-null   object
 3   Arrival_Time      2507 non-null   object
 4   Departure Place   2507 non-null   object
 5   Arrival_place     2507 non-null   object
 6   Flight_Durations  2507 non-null   int32 
 7   Stops             2507 non-null   int32 
 8   Emission(in kgs)  2507 non-null   int32 
 9   Prices            2507 non-null   int32 
 10  Departure_Hour    2507 non-null   int32 
 11  Arrival_Hour      2507 non-null   int32 
dtypes: int32(6), object(6)
memory usage: 195.9+ KB
```

Statistical Properties of Dataset:

```
In [39]: # statistical properties of dataset
         data.describe().T
```

Out[39]:

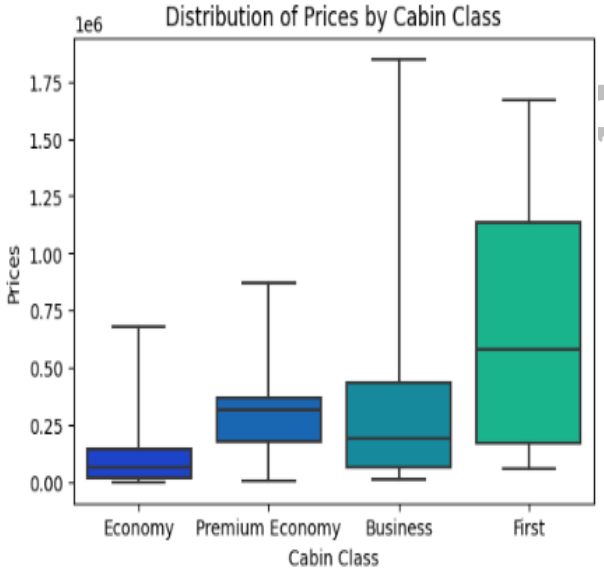
	count	mean	std	min	25%	50%	75%	max
Flight_Durations	2507.0	1476.881532	666.360231	60.0	1070.0	1510.0	1900.0	4034.0
Stops	2507.0	1.853211	1.039903	0.0	1.0	2.0	2.0	6.0
Emission(in kgs)	2507.0	1674.777423	1754.060689	49.0	397.0	959.0	2181.0	7984.0
Prices	2507.0	230834.005983	264984.288396	3208.0	49247.5	133185.0	336247.0	1847785.0
Departure_Hour	2507.0	12.141604	7.484721	0.0	6.0	13.0	20.0	23.0
Arrival_Hour	2507.0	13.794575	5.737654	0.0	10.0	14.0	19.0	23.0

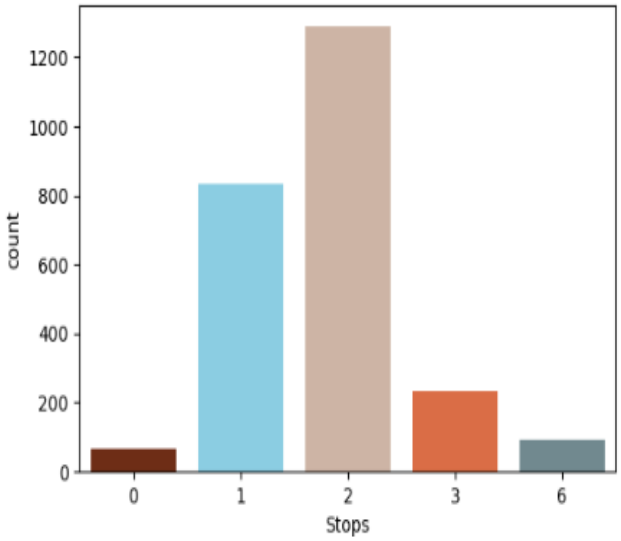
- The above command df.describe() helps us to view the statistical properties of numerical variables. It excludes character variables.
- If we want to view the statistical properties of character variables, we should run the following command -df.describe(include=['object'])
- If we want to view the statistical properties of all the variables, we should run the following command -df.describe(include='all')

Exploratory Data Analysis (EDA):

Price Feature:

- **Class-Based Pricing**: Airlines offer different cabin classes (economy, premium economy, business, first), each with its own price range based on factors like seat comfort, amenities, and services.
- Airlines often employ dynamic pricing algorithms to adjust ticket prices based on factors such as demand, time until departure, competitor pricing, and seat availability.



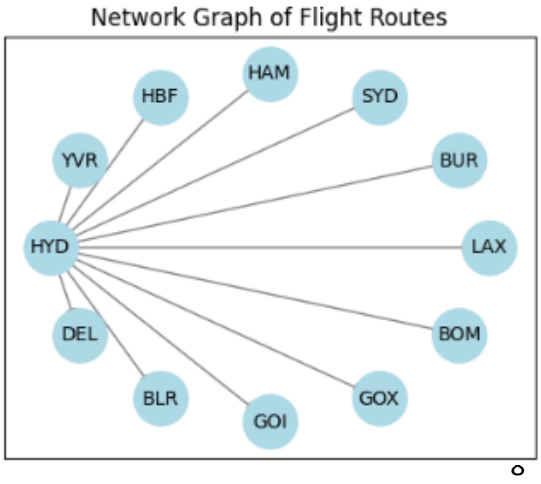


Stops:

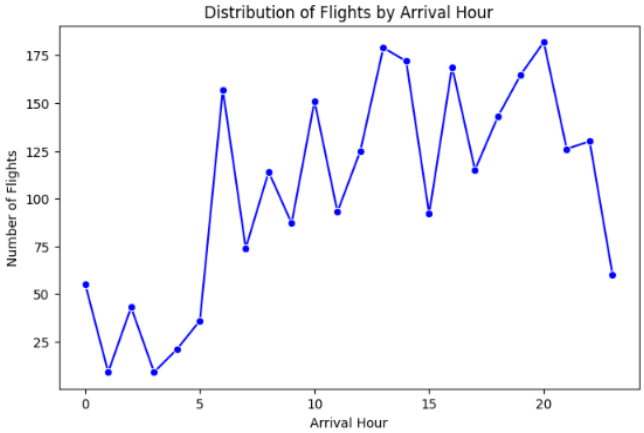
- **Direct Flights:** Flights with no intermediate stops between the origin and destination. These flights offer the quickest and most efficient route. (0: Direct Flights)
- **Connecting Flights:** Involve one or more intermediate stops (layovers) at different airports before reaching the final destination. (1, 2, 3, 6: Connecting Flights)

Departure & Arrival Place:

- Refers to the location where the journey begins.
- Typically an airport or a city.
- Key factors include accessibility, facilities, and transportation options.
- Refers to the destination where the journey ends.



- Important considerations include amenities, proximity to final destination, and transportation connections.

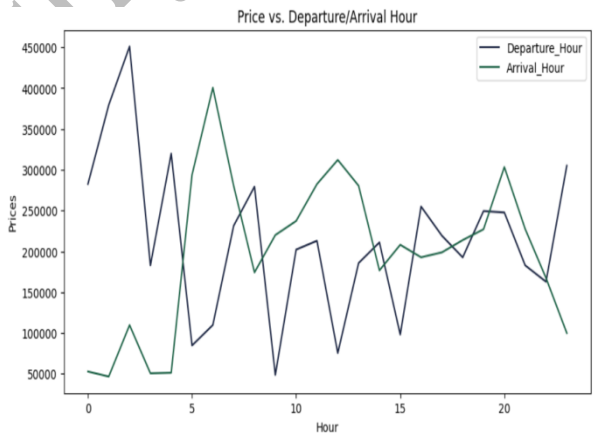


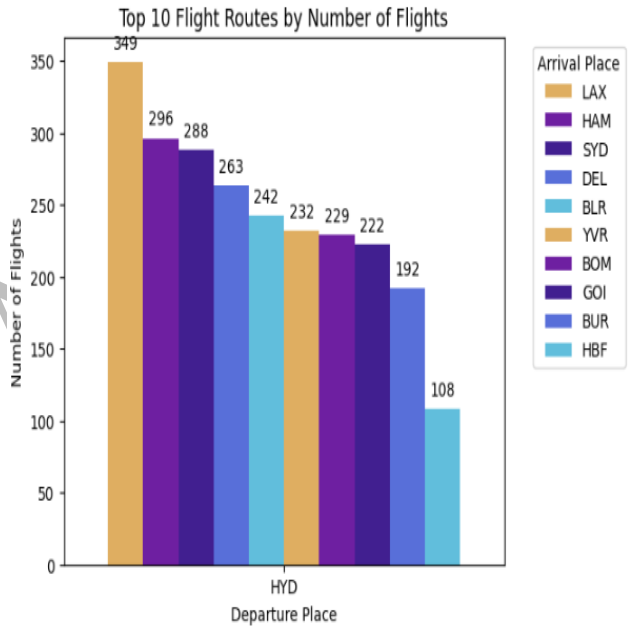
Arrival Time:

- Marks the completion of a journey or the arrival at a destination.
- Critical for coordinating pick-up, accommodation check-in, and scheduling subsequent plans.
- Variability is common due to factors like transportation delays or unforeseen circumstances.

Departure/Arrival Hour v/s Prices:

- Daily fluctuations in flight prices, highlighting higher costs during peak hours and lower prices during off-peak times.
- This dynamic pattern suggests a correlation between pricing and the demand for flights throughout the day.



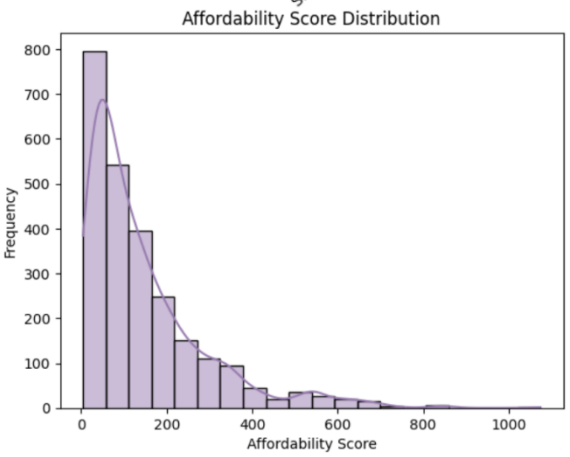


Flights for various places :

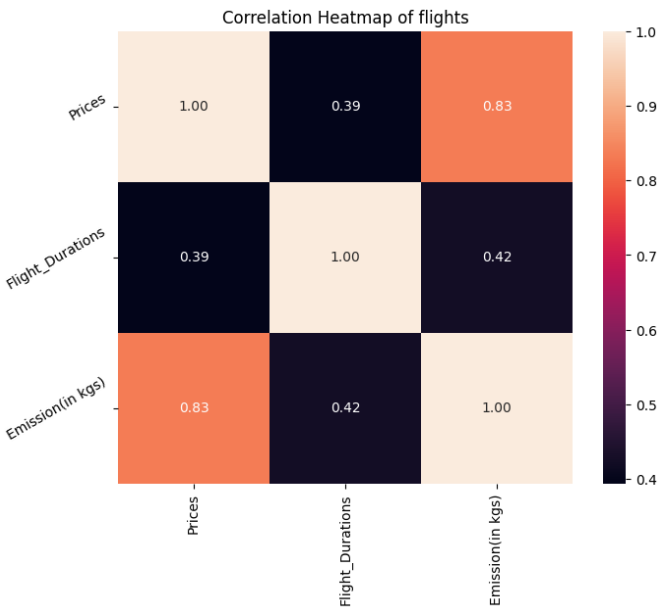
- Flights travel from hyderabad to different locations.
- These flights can connect you to various cities, countries, or continents around the world.
- Understanding the flight routes and options available from your starting point allows you to explore different places and make informed travel decisions.

Affordability Score Distribution Based on Flights:

- Etihad and British Airways:
- Affordability: Excellent (1000) Summary: Most affordable with a perfect score, offering a great balance of quality and price. THAI and Cathay Pacific:
- Affordability: Very Good (Slightly above 800) Summary: Highly affordable with good quality, providing a reasonable balance of service and cost. Air India, United, and Air Canada:
- Affordability: Moderate (Slightly below 800) Summary: Moderately affordable, offering a good mix of quality and fair prices. Singapore Airlines:
- Affordability: Low (Just above 700) Summary: Least affordable, but with very high quality. Suitable for those prioritizing premium service over cost. In choosing an airline, travelers



Correlation Heat Map:



- Prices and emissions show a strong positive correlation (0.83), indicating that as flight prices increase, emissions tend to rise due to factors like fuel costs and operational expenses.
- Flight durations have a moderate positive correlation with both prices (0.39) and emissions (0.42), suggesting longer flights are more expensive and produce higher emissions, reflecting a trade-off between cost and environmental impact.
- Perfect correlation (1.00) for each variable with itself is expected.

Recommendations:

- **Compare Prices**: Utilize comparison websites like Skyscanner or Google Flights to find the best deals.
- **Flexibility in Travel Dates**: Be open to adjusting travel dates for potential savings.
- **Explore Nearby Airports**: Check nearby airports as they may offer cheaper options.
- **Price Drop Notifications**: Enable notifications for price drops on preferred routes.
- **Consider Layovers**: Longer layovers can sometimes result in cheaper fares.
- **Beware of Hidden Fees**: Watch out for hidden charges when booking flights.
- **Travel During Off-Peak Times**: Opt for quieter times of the year for better prices.
- **Use Rewards Points or Miles**: Redeem any accumulated points or miles to reduce flight costs.
- **Read Reviews**: Check reviews to ensure you're selecting a reputable airline.

.

Conclusion:

Analyzing and recommending flight options based on cost and convenience offers travelers the opportunity to make informed decisions that align with their preferences and priorities. By considering factors such as ticket prices, fees, flight durations, and amenities, travelers can optimize their travel experiences to meet their needs while maximizing value. Tools like Google Flights and AirHint provide valuable insights and guidance, empowering travelers to find the most suitable and satisfying flights for their trips. However, it's essential for travelers to recognize the limitations of these tools and to balance their decisions with other considerations such as safety, and environmental impact.