Why Do Flight Prices Fluctuate So Wildly?

An Exploratory Data Analysis of the Factors Driving Airfare.

A case study in data cleaning, feature engineering, and insight generation.

The ‘Dynamic Pricing’ Mystery Creates Confusion for Travelers

Flight ticket prices fluctuate heavily, even for the same route. Without structured analysis, both travelers and booking platforms lack clarity on what truly drives these changes. Key questions remain unanswered:

* Why is one airline consistently more expensive than another?
* Do non-stop flights always cost more?
* How much do departure times and travel dates really impact the final fare?

Our Goal: Replace Guesswork with Data-Driven Facts

This project's objective was to analyze a dataset of over 10,000 flight records to uncover the trends and features that influence airfare variability.

THE DATA

A raw Excel file (`flight_price.xlsx`) containing 10,683 flight records with 11 initial features.

THE METHOD

Exploratory Data Analysis (EDA) using Python, Pandas, and NumPy.

THE OUTCOME

A cleaned, feature-rich dataset ready for machine learning and a set of clear, actionable insights for travelers and analysts.

The Analysis Began with Raw, Unstructured Data

The initial dataset contained multiple challenges that made direct analysis impossible. Key issues included:

- Dates and times stored as text strings ('objects'), not numerical values.
- Flight duration was in a mixed, human-readable format (e.g., '2h 50m').
- Number of stops was also a text category (e.g., 'non-stop', '1 stop').

```
In [2]: df.head()
```

	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price
0	Airline	24/03/2019	Taohi	Mahalta	To-Vietanam Airrport	10:00	13:00	2h 50m	non-stop	NuN	380
1	Airline	24/03/2019	Tatrih	Taishi	To-Vesanam Airtport	08:00	12:00	2h 50m	non-stop	NuN	300
2	Airline	24/03/2019	Einpeta	Mahaika	To-Vietanam Airrport	03:00	14:00	2h 50m	non-stop	NuN	300
3	Airpaan	24/03/2019	Taohi	Barcina	To-Vssanam Ainport	03:00	15:00	2h 70m	non-stop	NuN	250
4	Airline	27/03/2019	Tsohi	Mahalta	To-Vsaanam Airrport	07:00	23:00	2h 50m	non-stop	NuN	100

```
In [4]: df.info()
```

```
class 'summary lists for DataFrame'>
Data columns (nols 11 columns):
 #   Name              Non-Null Count  Dtype
---  ----              --------------  -----
 0   Airline           245 non-null    object
 1   Date_of_Journey   245 non-null    object
 2   Duration          245 non-null    object
 3   Total_Stops       245 non-null    object
 4   Additional_Info   245 non-null    object
 5   Price             245 non-null    object
dtypes: object, (0)
```

Feature Engineering I: Deconstructing Dates and Times

The text-based date and time columns were parsed to extract numerical features, enabling analysis of time-based trends. Date_of_Journey ('24/03/2019') was split into numerical Day (24) and Month (3). Dep_Time ('22:20') was split into Departure_hour (22) and Departure_min (20).

```
df['Date'] = df['Date_of_Journey'].str.split('/').str[0]
df['Month'] = df['Date_of_Journey'].str.split('/').str[1]
```

Date_of_Journey
'24/03/2019'

Before & After

Date	Month
24	3

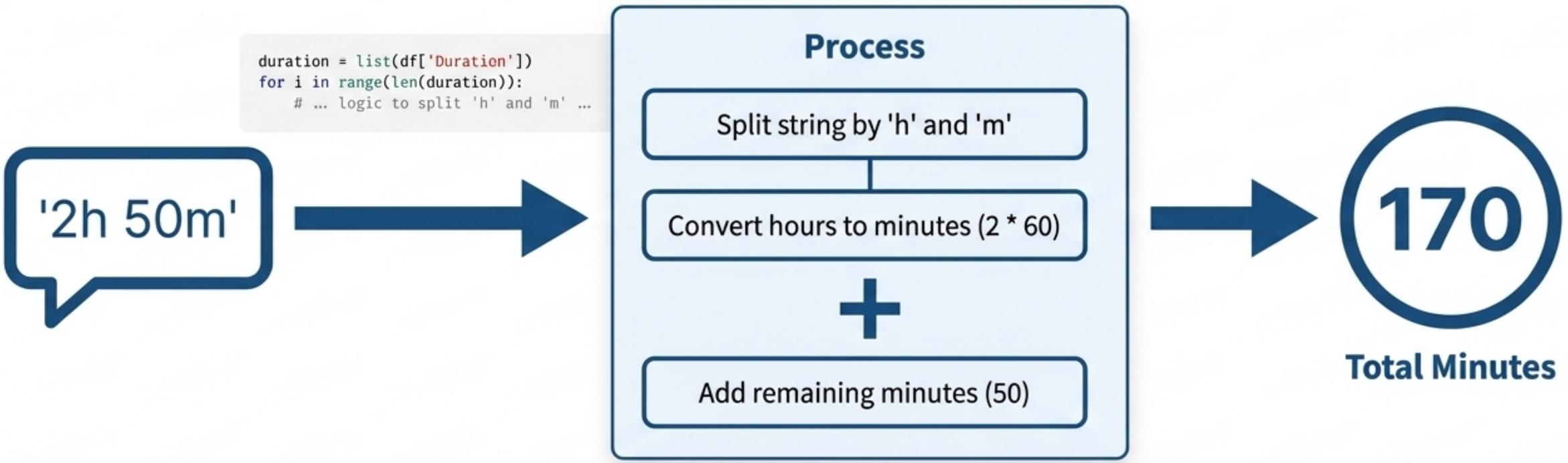
Why This Matters

This transformation unlocks the ability to analyze seasonal spikes (e.g., holidays) and identify patterns related to time of day (e.g., pricier evening flights driven by business travel).

Feature Engineering II: Converting Duration into a Usable Metric

The `Duration` column, with values like '2h 50m', was converted into a single numerical feature: total minutes. This allows for direct correlation analysis with price.

```
duration = list(df['Duration'])
for i in range(len(duration)):
    # ... logic to split 'h' and 'm' ...
```



Why This Matters

A numeric duration is essential for modeling and for quantifying the relationship between flight length and cost.

Final Touches: Standardizing Categorical Data

The remaining text-based features were converted into numerical formats to complete the data preparation.

Ordinal Encoding (`Total_Stops`)

The `Total_Stops` column was mapped from text to integers.

Before & After

Total_Stops
'non-stop'
'1 stop'
'2 stops'

```
df['Total_Stops'].map(
  {'non-stop':0,
  '1 stop':1, '2 stops':2})
```

Total_Stops
0
1
2

One-Hot Encoding

Categorical features like `Airline` were transformed using One-Hot Encoding, creating a binary representation suitable for machine learning.

Airline_IndiGo	Airline_Jet Airways	Airline_SpiceJet	Airline_Air India	Airline_GoAir
1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
...				

Result: A fully cleaned, numerical dataset ready for analysis and prediction.

The Analysis Revealed Four Primary Drivers of Airfare

Our Exploratory Data Analysis (EDA) confirmed that pricing is not random.
Four factors consistently emerged as the strongest predictors of cost.

Number of Stops

The strongest predictor

Airline Brand

Budget vs. Premium

Departure Time

Time of day and season

Total Duration

Flight length

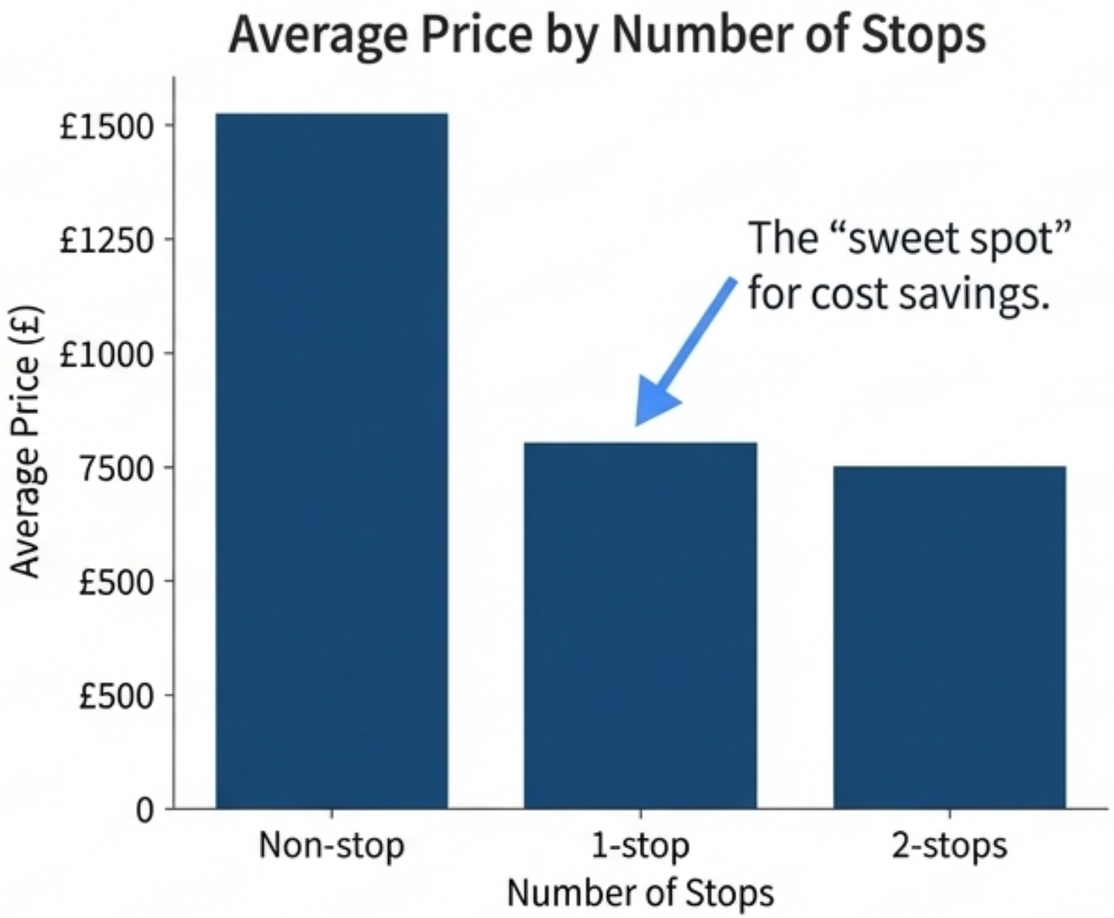
Insight 1: Non-Stop Flights Are the Most Expensive Category

A Premium for Convenience

Contrary to the belief that fewer stops mean lower operational costs, the data shows that travelers travelers pay a significant premium for the convenience of a direct flight.

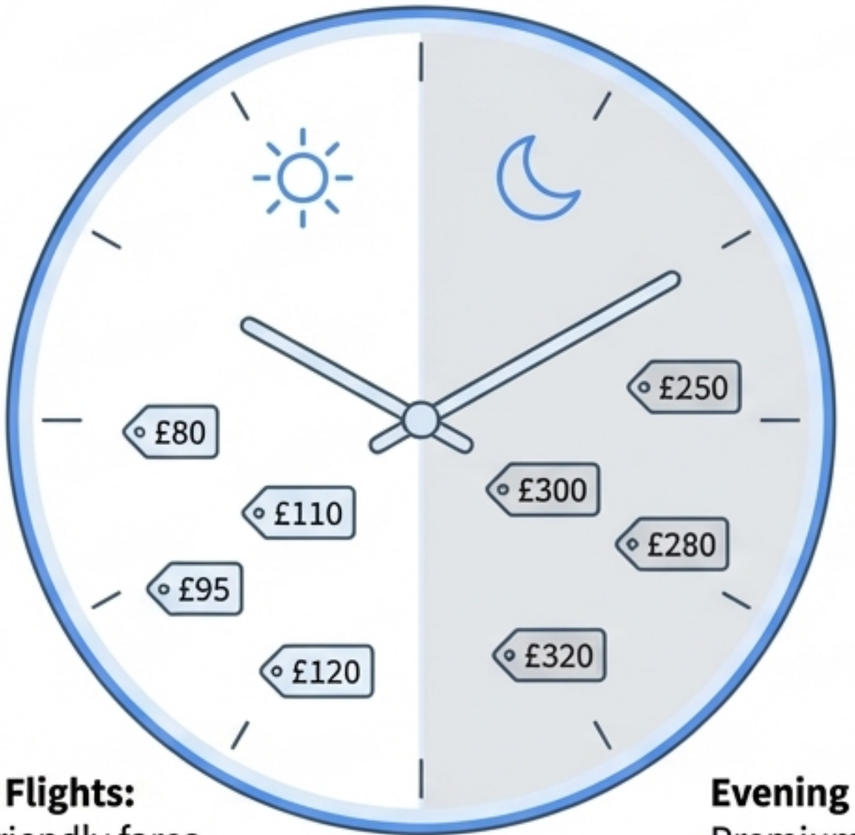
The Data Says

Flights with one or two stops are, on average, significantly more affordable. A 1-stop flight often represents the “sweet spot,” balancing travel time with cost savings.



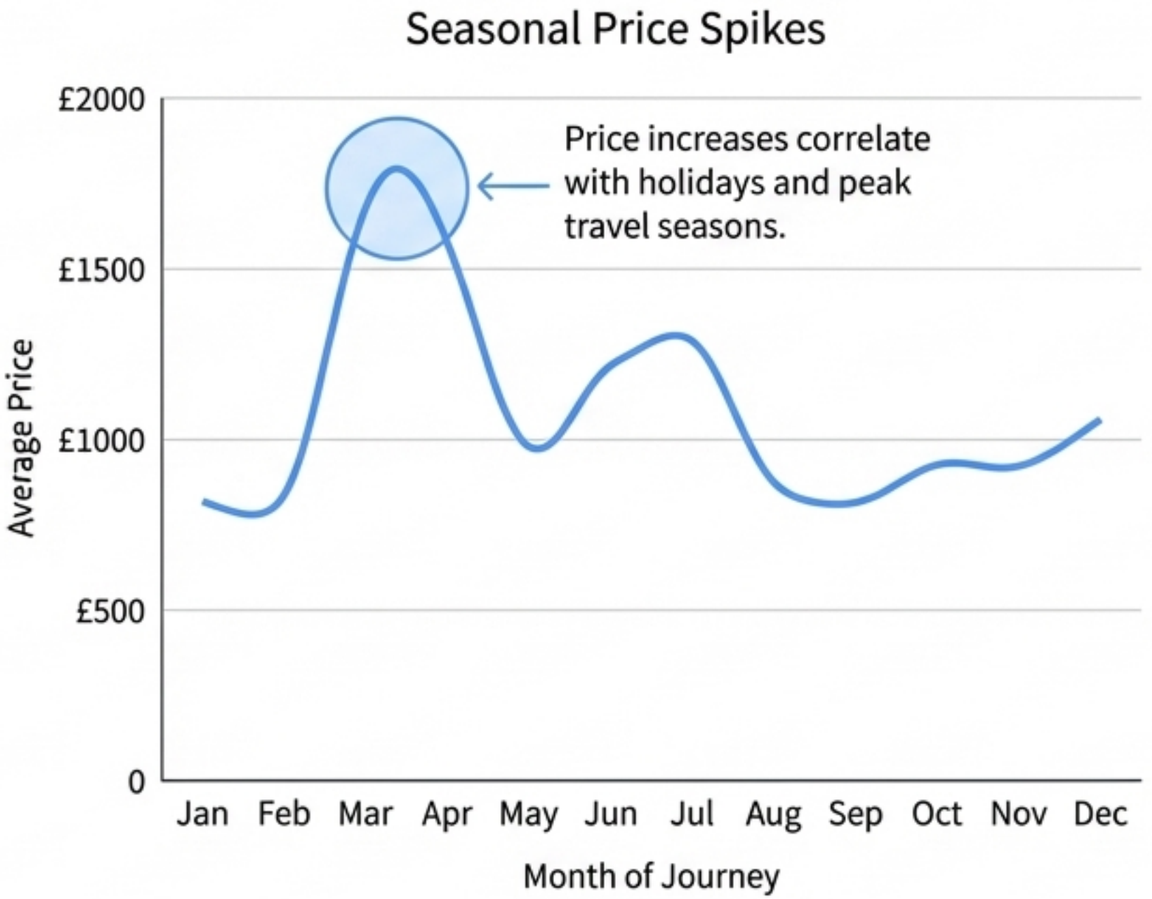
Insight 2: Departure Time and Seasonality Heavily Influence Cost

When you fly is a massive price predictor. Our analysis of the engineered time features revealed clear patterns.



Morning Flights:
Budget-friendly fares

Evening Flights:
Premium prices

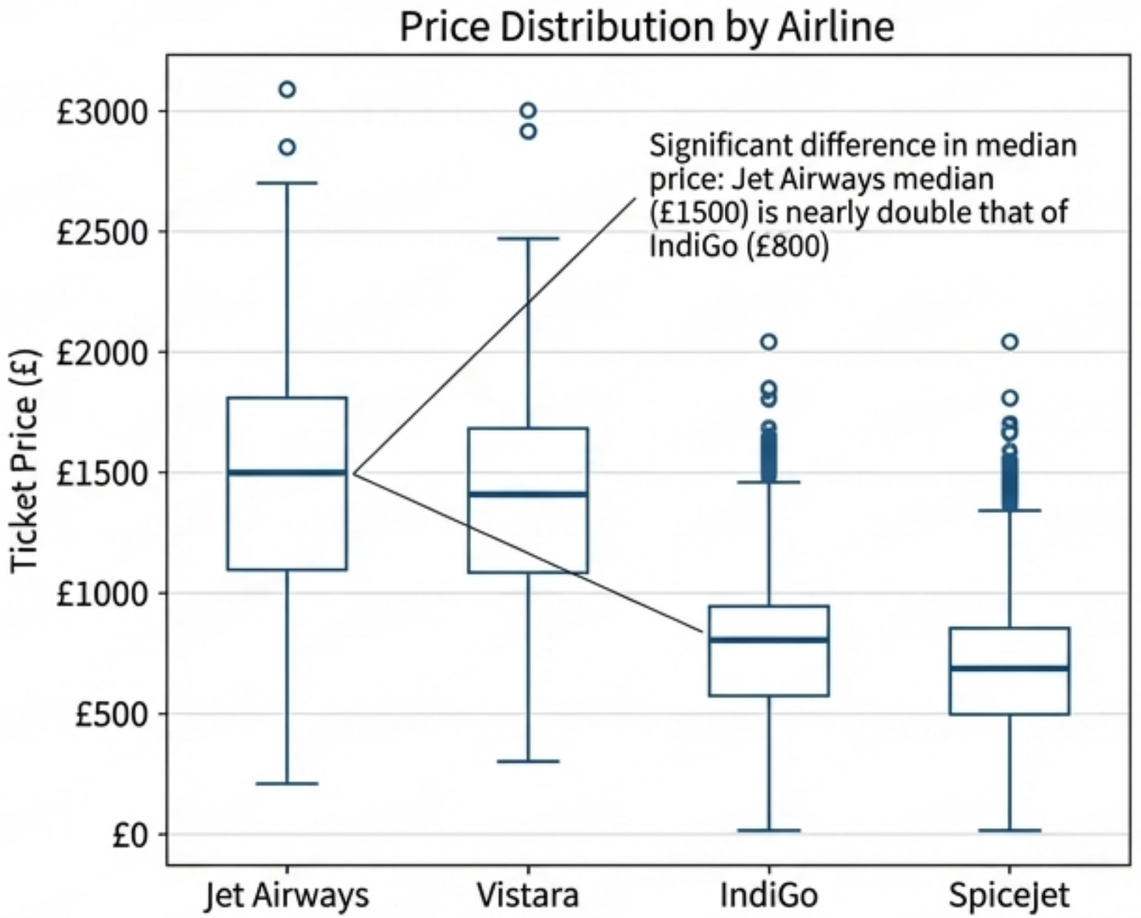


Insight 3: Airline Brand Carries a Clear Price Premium

Paying for the Name

The choice of airline is a significant factor in ticket price. Pricing is uneven across carriers, even for similar routes and durations.

- Premium Airlines: (e.g., Jet Airways, Vistara) consistently charge more, reflecting their brand positioning, service quality, and amenities.
- Budget Airlines: (e.g., IndiGo, SpiceJet) offer lower base fares but exhibit higher price volatility due to aggressive dynamic pricing models.



A Note on Outliers: Investigating Extreme Prices

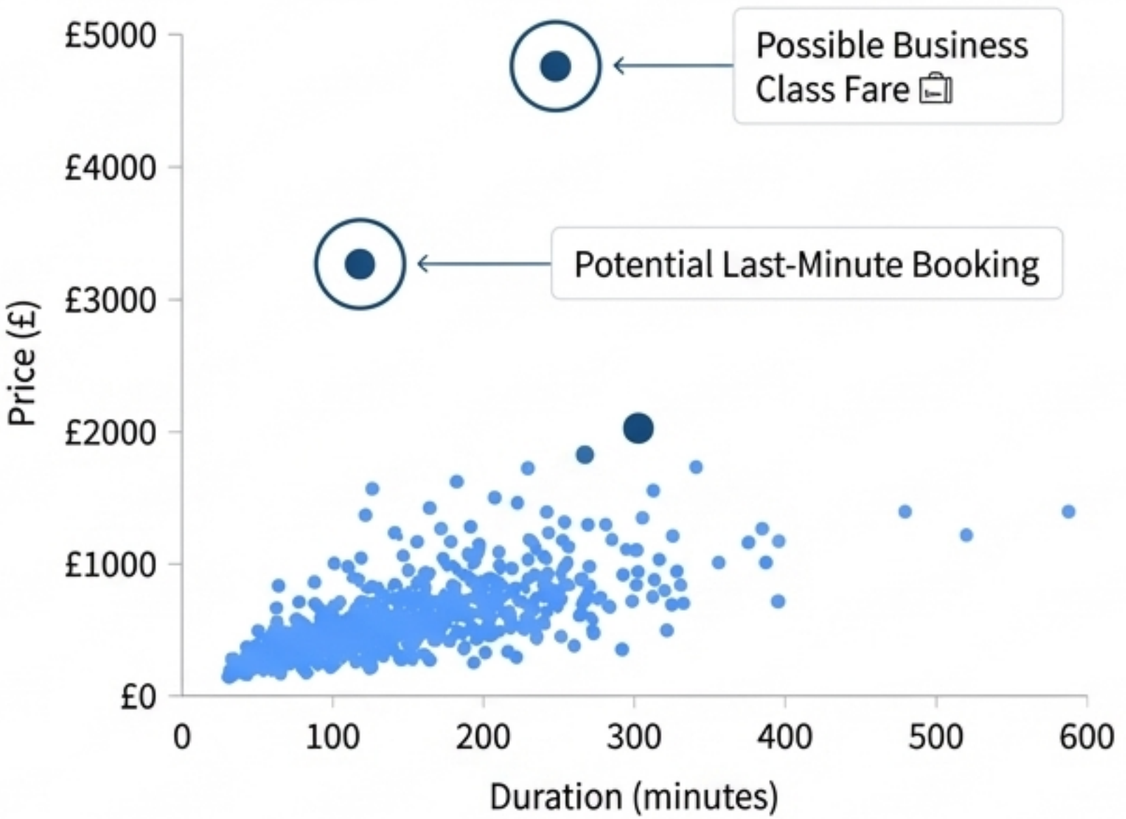
Beyond the Average

The analysis identified several flights costing over 5 times the median price. These were not errors, but real-world pricing phenomena.

- Last-minute bookings for high-demand routes.
- Holiday or festival travel spikes.
- Business class or premium economy fares.

Why this matters: Acknowledging outliers is crucial for building robust predictive models and understanding the full spectrum of real-world pricing strategies.

Price vs. Duration



From Data to Decisions: Actionable Insights for Stakeholders

The findings from this analysis provide clear, practical value for multiple groups.

FOR TRAVELERS

- To save money, consider 1-stop flights and book morning departures.
- Be aware of brand premiums and seasonal price hikes.

FOR AIRLINE PLATFORMS

- Understand competitor pricing patterns on key routes.
- Refine pricing models by weighting the most impactful features.

FOR DATA ANALYSTS

- The feature engineering process provides a roadmap for handling similar datasets.

The Result: A Clean Dataset Ready for Price Prediction

This Exploratory Data Analysis successfully transformed a messy raw dataset into an insight-rich, feature-engineered asset. The groundwork is now complete.

Machine Learning

This cleaned dataset is perfectly structured for training machine learning models to forecast future flight prices with high accuracy.

View the full analysis and Python code on GitHub.

github.com/your-repo-link

#DataScience #Python #EDA #Analytics #TravelTech

Technical Project Summary

Key Parameters

Dataset Size: 10,683 records

Initial Features: 11

Final Features: 15+ (after engineering)

Key Challenge: Parsing and converting string-based date, time, and duration data.

Technology & Libraries

Python

Pandas, NumPy

Jupyter Notebook

Excel (`.xlsx`)