Beyond Price: Understanding Flight Value

1. Motivation (Why this project exists)

When booking flights, we usually compare tickets based on **price alone**.
However, two flights with similar prices can feel very different in reality; depending on travel time, number of stops, and overall inconvenience.

This project aims to answer a simple but practical question:

> **Which flight actually gives better value for money once time and inconvenience are considered?**

Inspired by the idea of *"feels like" temperature* in weather apps, we introduce a **"feels like cost"** for flights.

2. Dataset Overview

The dataset contains domestic Indian flight data from 2019, including:

- Airline
- Route (Source → Destination)
- Ticket Price
- Duration
- Number of Stops
- Travel Date

While price is the observed cost, **duration and stops represent hidden costs** paid in time, effort, and comfort.

3. Tech Stack

- **Programming Language:** Python
- **Data Manipulation:** Pandas, NumPy
- **Machine Learning:** Scikit-learn
    - Linear Regression (interpretable modeling)
    - KMeans (unsupervised clustering)
- **Visualization:** Matplotlib
- **Environment:** Kaggle Notebook
- **Documentation:** Google Docs

Tools were chosen for **interpretability, reproducibility, and clarity**, rather than model complexity.

4. Translating Travel Experience into Numbers

To model inconvenience quantitatively:

- **Duration** (originally text like 2h 50m) is converted into total minutes.
- **Stops** (non-stop, 1 stop, 2 stops) are converted into integers.

- Duration is normalized so it can be fairly combined with stops.

This allows us to work with **experience as data**, not just text.

5. Value-for-Money (VFM) Score

We define a *perceived cost* of a flight as:

$$\textbf{Adjusted Cost} = \textbf{Price} \times (1 + \alpha \times \textbf{Stops} + \beta \times \textbf{Duration})$$

Where:

- **Price** is the actual ticket cost
- **α (alpha)** represents how painful extra stops are
- **β (beta)** represents how painful longer travel time is

The **Value-for-Money (VFM) score** is simply the inverse of this adjusted cost:

$$\textbf{Higher VFM} = \textbf{better deal}$$

This means:

- A slightly more expensive non-stop flight can rank higher than a cheaper but exhausting one.
- Value is separated from raw price.

6. Where Do α and β Come From? (Key Insight explained in depth)

(Grounding Inconvenience Using Linear Regression)

To avoid choosing penalty weights arbitrarily, we estimate how the market already prices inconvenience using a simple linear regression model.

Regression Model

We model ticket price as:

$$\textbf{Price} = b_0 + b_1 \cdot \textbf{Duration}_{\textbf{minutes}} + b2 \cdot \textbf{Stops} + \varepsilon$$

Where:

- **Price** is the ticket cost
- **Duration** is total travel time
- **Stops** is the number of layovers
- b_1 and b_2 are coefficients learned from data

The model finds values of b_1 and b_2 that minimize overall **mean squared error**, i.e., it chooses coefficients that best explain observed prices across thousands of flights.

Intuition Behind the Coefficients

The coefficients have a direct and interpretable meaning:

- $\mathbf{b_1}$: average price change for one additional minute of travel, *holding stops constant*
- $\mathbf{b_2}$: average price change for one additional stop, *holding duration constant*

This allows us to isolate trade-offs:

- Time mainly reflects distance and operating cost
- Stops reflect inconvenience and uncertainty

In the data, the penalty for stops is significantly larger than for duration, indicating that travelers dislike journey interruptions more than longer travel time.

From Regression to α and β

Regression coefficients are expressed in **rupees**, while the Value-for-Money score requires **relative penalties**.

We convert rupee trade-offs into percentages of a typical ticket price:

$$\boldsymbol{\alpha} = Price\ discount\ per\ stop\ /\ Average\ Ticket\ Price$$

$$\boldsymbol{\beta} = Price\ impact\ of\ an\ average\ duration\ flight\ /\ Average\ Ticket\ Price$$

These values represent how much of the ticket price is implicitly "paid" for inconvenience.

Why This Works

- No arbitrary weights are introduced
- Assumptions are explicit and interpretable
- Results align with real market behavior

Most importantly:

> **Linear regression reveals how the market compensates travelers for inconvenience.**
> **We translate that compensation into human-understandable penalties.**

7. Airline Strategy Profiling

Instead of analyzing individual flights, we aggregate data at the **airline level** to understand strategy.

For each airline, we compute:

- Average price
- Average duration
- Average stops
- Average VFM score
- Price volatility

This reveals distinct behaviors:

- Airlines that compete on **value**
- Airlines that sell **convenience at a premium**
- Airlines with **volatile pricing**

We further use clustering to group airlines with similar strategies.

8. Persona-Based Value

Not all travelers value time the same way.

We simulate:

- **Budget travelers** (low penalties for time/stops)
- **Business travelers** (high penalties for time/stops)

The same flight and the same airline can rank very differently depending on the persona.

This reinforces an important idea:

> **There is no universally "best" airline, only best choices for different people.**

9. Key Takeaway

This project is not about prediction.
It is about **decision-making**.

By combining:

- human intuition
- simple mathematics
- transparent assumptions
- and real market behavior

we move from *"Which flight is cheapest?"*
to

> **"Which flight is actually worth it?"**

10. References & Resources

This project builds on well-established concepts in statistics, machine learning, and data analysis. The following resources and datasets were referenced during development and validation.

Dataset

- Indian Domestic Flights Dataset (2019)

Documentation & Learning Resources

- **Scikit-learn – Linear Regression**
  Used to understand model assumptions, coefficient interpretation, and limitations.
- **Scikit-learn – Clustering (KMeans)**
  Referenced for airline strategy clustering and feature scaling considerations.
- **Pandas Documentation**
  Used extensively for data cleaning, aggregation, and feature engineering.
- **Statistical Interpretation of Regression Coefficients**
  Referenced to correctly interpret regression coefficients as trade-offs rather than causal effects.

11. My Learnings

This project focused as much on clear thinking as on writing code.

- **Price is not value**: Real-world decisions include hidden costs like time, inconvenience, and uncertainty that a single number cannot capture.
- **Simple models can be powerful**: Linear regression, used for interpretation rather than prediction, revealed meaningful trade-offs in how the market prices inconvenience.
- **Coefficients are trade-offs, not truths**: Model outputs describe average market behavior, not universal rules or causal effects.
- **Value depends on the user**: Different traveler personas lead to different "best" choices, there is no one-size-fits-all answer.
- **Interpretability matters in decision support**: For many real problems, explainable models are more useful than complex predictive ones.

Final Note

> **"Good data science is not about complexity, it is about clarity."**

By grounding assumptions in data and focusing on real-world relevance, this project emphasizes better decision-making, not just better metrics.

Author:

Vanshaj Verma

(Applied Data Science/Decision Modeling)