Power BI Dataset Assessment Form

# Your Name : Muhammed Aslam k

# Dataset's Link ( Drive link): https://www.kaggle.com/datasets/rohitgrewal/airlines-flights-data/data

# Project Goal :

To analyze airline flight pricing, duration, stops, and travel patterns using Tableau, with a focus on identifying factors that influence ticket prices across routes, times, and airline carriers.

# . Data Quality

How will you handle missing values in important fields (if any)?

 The dataset has **no missing values** in any columns.

 If any are found during further analysis, I will:

* Use imputation (mean/median) for numerical columns
* Use mode or "Unknown" for categorical columns
* Drop rows if values are critical and can't be inferred

Provide the source of the dataset:

* + Kaggle

# Data Structure

What is the file format of the dataset (e.g., CSV, Excel)?

* CSV (Comma-Separated Values)

How many records (rows) does the dataset contain?

* 300,153 rows

How many columns does the dataset have?

-12 columns

# 5. Data Content

List the columns in the dataset and type of data in them

**List the columns in the dataset and type of data in them**

* index – Integer (row identifier)
* airline – Categorical (text)
* flight – Categorical (text)
* source\_city – Categorical (text)
* departure\_time – Categorical (text)
* stops – Categorical (text)
* arrival\_time – Categorical (text)
* destination\_city – Categorical (text)
* class – Categorical (text)
* duration – Numeric (float)
* days\_left – Numeric (integer)
* price – Numeric (integer)

Do the columns contain consistent data (e.g., no mixed data types within the same column)?

Yes, all columns contain consistent data types.

Are there any calculated fields in the dataset (e.g., profit margin, growth rate)? If so, list them:

Yes:

* days\_left (calculated based on date difference before flight)
* duration (likely calculated flight time in hours)

Does the dataset include any categorical variables (e.g., product category, region)? If yes, list them:

Yes:

* airline, flight, source\_city, departure\_time, stops, arrival\_time, destination\_city, class.

Does the dataset include any time-based data (e.g., dates, timestamps)? If yes, explain how it is formatted:

 Indirectly: departure\_time and arrival\_time are time-based but categorical (e.g., "Morning", "Night")

 days\_left gives a numeric representation of the time until departure

Are there any columns with geographical data (e.g., country, city)? If yes, list them:

Yes:

* source\_city, destination\_city

Are there any numerical variables with large ranges or outliers? If so, how will you handle them?

Yes:

* price ranges from ₹1,105 to ₹1,23,071
* Outliers will be identified using boxplots and handled by:
  + Capping extreme values
  + Using log transformation if needed

Does the dataset include any hierarchical or multi-level categories (e.g., subcategories, regions within countries)?

* No

Does the dataset include any columns that could be used for filtering or grouping data (e.g., product ID, customer segment)?

Yes:

* airline, flight, source\_city, destination\_city, class, departure\_time

Explain whether the dataset allows you to explore trends, comparisons, or correlations:

Yes:

* Trends in ticket pricing vs. days left
* Comparisons between airlines, class types, or routes
* Correlation between flight duration, stops, and price

# 6. Size & Performance

What is the size of the dataset in MB?

* 23.7 mb.