Data Challenge 1

Background:
The goal of this data challenge is to demonstrate your understanding of general machine learning by predicting the price of airline tickets based on various features such as the date of the journey, source, destination, route, departure time, arrival time, duration, total stops, and additional information.

In answering the question:

- You are allowed (and expected to) use external packages and libraries (e.g., scikit-learn, matplotlib, pandas, and others).
- You are allowed to consult external sources (notes, etc.).
- You are allowed to discuss your model's performance and general ideas.

... but you are not allowed to use external data sources, nor are you allowed to discuss specific solutions with others inside or outside the course.

You are required to document your findings and explain all choices you make. You may use a Jupyter notebook to capture the output at every step. Any plots (charts, graphs) you create in terms of documenting or explaining must be constructed using matplotlib or seaborn -OR- must have approval from the instructor. Work hard, and have fun!

Tasks:

- **Task 1:** Before training a regression model, it's essential to analyze, and preprocess the data to ensure that it's suitable for analysis and modeling. This step involves handling missing values, converting categorical variables into numerical representations, and performing feature engineering to extract useful information from the existing features.
- **Task 2:** Develop the best regression model for predicting the price of airline tickets using the provided training dataset.
- **Task 3:** Use the provided test dataset to generate predictions using your model.

Dataset:
You are provided with a dataset containing information about airline tickets. You'll use *train_data* and *train_labels* to train your model, where *train_data* contains features (like Airline, Date_of_Journey, etc.) and *train_labels* contains corresponding target values (Price in this case). You'll use the trained model to predict a set of labels based on *test_data*. However, you will not have access to the real *test_labels*.

Dataset Features:

- Airline: The name of the airline.
- Date_of_Journey: The date of the journey.
- Source: The source of the flight.

- Destination: The destination of the flight.
- Route: The route of the flight.
- Dep_Time: The departure time.
- Arrival_Time: The arrival time.
- Total_Stops: The total number of stops during the flight.
- Additional_Info: Additional information about the flight.
- Normalized_Price: The normalized price of the airline ticket.

Submission:

- Your implementation - Submit a Jupyter notebook (e.g., DC1.ipynb) containing your data exploration, preprocessing, model training, and evaluation steps. Clearly document your process, including feature selection, data preprocessing techniques, model selection, hyperparameter tuning, and evaluation metrics.
- Your predictions for the target on the test set - Submit a CSV file (e.g., airline_price_predictions.csv) containing the predicted prices for the airline tickets in the test dataset.

Grading:

You will receive a Satisfactory grade or above if the ***RMSE*** is ***4.78 or less*** on the test labels. Depending on your model development process, code quality, documentation, quality of feature engineering and performance relative to your peers, you will receive an Exemplary grade.