**SEMMA on Airplane ticket price prediction**

**Step 1: Sample**

The **Sample** step involves creating a representative subset of the data that can be used for analysis and model training. As a beginner, I would follow these guidelines for sampling:

* **Objective**: The goal is to select a sample that accurately represents the full dataset, ensuring that all significant patterns and characteristics are included without overwhelming the analysis with too much data.
* **Method**: I would typically use random sampling methods. If the dataset is large, I would select a percentage (e.g., 10-20%) of the total data randomly to ensure that the sample is manageable while still retaining diversity in the data.
* **Considerations**:
  + Ensure that the sample captures all key variables present in the full dataset.
  + Check for stratification if necessary, especially if the dataset has categories (like different airlines or stops) to ensure all are represented in the sample.
* **Documentation**: It's essential to document the sampling method, the size of the sample taken, and any characteristics of the data that were particularly important in guiding the sampling decision.

### Step 2: Explore

The **Explore** step in the SEMMA methodology focuses on examining the data to understand its characteristics, identify patterns, and detect anomalies. Here’s how I would approach this step as a beginner:

The primary goal is to gain insights into the dataset, assess the quality of the data, and discover relationships among variables. This is crucial for informing subsequent modeling steps.

1. **Descriptive Statistics**:
   * Calculate basic statistics (mean, median, mode, minimum, maximum, standard deviation) for numerical variables such as **Price** and **Duration**. This helps in understanding the distribution and central tendencies.
   * For categorical variables like **Airline** and **Source**, frequency counts can reveal the most common categories.
2. **Data Visualization**:
   * Use visualizations to identify patterns and distributions. Common visualizations include:
     + **Histograms**: To examine the distribution of numerical variables (e.g., Price).
     + **Box Plots**: To identify outliers and understand the spread of the data.
     + **Bar Charts**: To compare the frequency of categorical variables (e.g., count of flights by Airline).
     + **Heatmaps**: To visualize correlations between numerical variables.
3. **Correlation Analysis**:
   * Analyze correlations between numerical features to see how they relate to each other. For example, examining if there's a correlation between **Duration** and **Price** can provide insights into pricing strategies.
4. **Missing Values**:
   * Identify any missing or null values in the dataset and consider how to handle them. This may involve:
     + Imputing missing values using techniques like mean/mode substitution.
     + Removing rows or columns with excessive missing data.
5. **Outlier Detection**:
   * Identify any outliers in the dataset, especially in key numerical variables like **Price** and **Duration**. Understanding these outliers can impact model performance, and decisions will need to be made on how to address them.
6. **Data Quality Assessment**:
   * Evaluate the overall quality of the data. This includes checking for consistency, validity, and accuracy of entries.

### Step 3: Modify

The **Modify** step in the SEMMA methodology focuses on transforming and preparing the dataset for modeling. This involves selecting, creating, and refining variables to improve the model's performance and ensure that the data is in a suitable format. Here’s how I would approach this step as a beginner:

#### 1. Variable Selection:

* **Identify Relevant Features**: Based on the insights gained during the Explore step, I would identify which features (variables) are most relevant for predicting the target variable (e.g., Price).
* **Remove Irrelevant Features**: Eliminate any features that do not provide useful information for the analysis or could add noise to the model, such as redundant or highly correlated variables.

#### 2. Creating New Variables:

* **Feature Engineering**: Create new variables that may capture additional information or relationships in the data. For example:
  + **Duration Conversion**: Convert the Duration from a string format (e.g., "2h 50m") into a numerical format (total minutes) for easier analysis.
  + **Extracting Date Features**: From Date\_of\_Journey, extract additional features like day of the week, month, or holiday indicators, as these could influence flight prices.
* **Encoding Categorical Variables**: Convert categorical variables into numerical format using techniques like one-hot encoding or label encoding to prepare them for modeling.

#### 3. Handling Missing Values:

* **Imputation**: Decide on methods to handle missing values identified in the Explore step. This may include:
  + Filling missing values with the mean, median, or mode of the respective columns.
  + Dropping rows or columns with excessive missing values if they are deemed insignificant.

#### 4. Normalization and Scaling:

* **Data Scaling**: Normalize or standardize numerical features to bring them into a similar range. This is especially important for algorithms sensitive to the scale of input data, like k-nearest neighbors or gradient descent-based algorithms.
* **Transformations**: Apply transformations (like logarithmic transformation) if the data is skewed, particularly for variables like Price.

#### 5. Outlier Treatment:

* **Identifying Outliers**: Use box plots or z-scores to detect outliers in numerical variables, especially in Price.
* **Handling Outliers**: Decide on strategies to manage outliers, such as capping, removal, or transforming them to reduce their impact on the model.

#### 6. Documentation:

* Document all modifications made to the dataset, including feature selection, creation of new variables, and any changes to existing variables. This will provide clarity for future steps and aid in replicating the process.

### Step 4: Model

The **Model** step in the SEMMA methodology involves selecting and applying various modeling techniques to the modified dataset to predict the target variable. This is a crucial phase where the data is transformed into actionable insights. Here’s how I would approach this step as a beginner:

#### 1. Selecting Modeling Techniques:

* **Model Choice**: Based on the problem type (in this case, regression for predicting Price), I would choose suitable modeling techniques. Common choices for regression include:
  + **Linear Regression**: A fundamental model that establishes a linear relationship between the features and the target variable.
  + **Decision Trees**: A model that splits the data based on feature values to make predictions.
  + **Random Forest**: An ensemble method that combines multiple decision trees to improve accuracy and reduce overfitting.
  + **Gradient Boosting Machines (GBM)**: Another ensemble method that builds models sequentially to optimize predictions.

#### 2. Splitting the Data:

* **Training and Testing Sets**: Divide the modified dataset into training and testing subsets (e.g., 80% training and 20% testing) to evaluate the model's performance. The training set is used to train the models, while the testing set is used to validate their accuracy.

#### 3. Training the Models:

* **Model Fitting**: Fit each selected model to the training data. This involves adjusting the model parameters to learn from the data patterns.
* **Hyperparameter Tuning**: For more complex models (like Random Forest or GBM), tune hyperparameters using techniques like grid search or random search to optimize model performance.

#### 4. Model Evaluation:

* **Performance Metrics**: Use appropriate metrics to evaluate the models. For regression, common metrics include:
  + **Mean Absolute Error (MAE)**: Measures the average magnitude of errors in predictions.
  + **Mean Squared Error (MSE)**: Evaluates the average of the squares of the errors.
  + **R-squared**: Indicates how well the model explains the variability of the target variable.
* **Cross-Validation**: Employ cross-validation techniques (like k-fold cross-validation) to ensure that the model's performance is robust and not just specific to the training set.

#### 5. Comparison of Models:

* **Selecting the Best Model**: Compare the performance metrics of different models to identify which one performs best on the validation data.
* **Interpreting Results**: Analyze the model outputs, including coefficients for linear models or feature importance for tree-based models, to gain insights into which features are most influential in predicting the target variable.

#### 6. Documentation:

* Document the modeling process, including the models chosen, parameters used, evaluation metrics, and the rationale behind selecting the final model. This documentation will be crucial for the final step of SEMMA.

### Step 5: Assess

The **Assess** step in the SEMMA methodology involves evaluating the effectiveness and robustness of the models developed in the previous step. This phase is crucial for understanding how well the chosen model performs and ensuring that it meets the project's objectives. Here’s how I would approach this step as a beginner:

#### 1. Final Model Evaluation:

* **Testing Performance**: Use the testing dataset (the portion that was not used during training) to evaluate the final model's performance. This step helps assess how well the model generalizes to unseen data.
* **Comparison of Metrics**: Recalculate performance metrics (like MAE, MSE, and R-squared) on the testing set to see how they compare to the training results. A significant difference may indicate overfitting or underfitting.

#### 2. Validation of Results:

* **Residual Analysis**: Analyze the residuals (the differences between predicted and actual values) to check for patterns. Ideally, residuals should be randomly distributed; patterns may suggest that the model is missing some key relationships.
* **Model Robustness**: Test the model's robustness by applying it to different subsets of data or through techniques like bootstrapping to estimate the stability of the model's predictions.

#### 3. Interpreting Results:

* **Feature Importance**: If applicable, review the feature importance scores from the model to understand which variables have the most significant impact on the target variable (e.g., Price). This insight can guide further analysis and decision-making.
* **Visualization of Predictions**: Visualize the predicted vs. actual values using scatter plots or line graphs to assess how closely the model's predictions match the actual data.

#### 4. Documentation of Findings:

* Document all findings, including performance metrics, visualizations, insights gained from residual analysis, and any patterns observed. This documentation serves as a valuable reference for understanding the model's performance.

#### 5. Decision Making:

* **Business Implications**: Consider the implications of the model results for business decisions or strategies. For instance, understanding how different factors influence flight prices can inform pricing strategies for airlines.
* **Next Steps**: Based on the assessment, determine if the model is ready for deployment or if further iterations are needed, such as additional feature engineering, trying different modeling techniques, or collecting more data.