**Regression Project :**

1. Exploratory data analysis and Data preprocessing:

* Data types is not to check whether it is categorical or numerical.
* Price is our target variable
* (10683, 11) shape of our dataset
* EDA 5 point summary descriptive analysis
* Data cleaning and Feature Engineering :

Separating date and month from date of journey column and creating new columns

Transforming Departure column with time into 6 categories with interval of 4 hours

Arrival columns data is very inconsistent performed hypotheses to drop as we already have duration column so validated if arrival – departure is equal to duration and checked correlation with arrival if highly correlated then preferred duration column as data is consistent

Converted duration in categories to find inferences.

Compared total stops and route column, as Max total stops are 4 so transformed route column into multiple city columns where each city column is representing each stop.

Handling skewness and outliers - We use log transform method to remove skewness

Handing missing values

Using (matplotlib, seaborn) univariate and bivariate analysis found

1. Encoding the data — One Hot encoding

3. Scaling the data — Standard scaler

4. Fitting the machine learning models

K-Nearest Neighbors (KNN)

* + - Train MSE: 4986880.350799638
    - Test MSE: 4559493.9556346

KNN with Cross-Validation:

* + - Train MSE: 5710663.760674275
    - Test MSE: 6218188.034159126

Linear Regression:

* + - Train MSE: 6586228.212870725
    - Test MSE: 6148430.729531382

Decision Tree:

* + - Train MSE: 799396.3993935853
    - Test MSE: 2888145.4728361536

**Conclusion:**

* From the results, it appears that the Decision Tree model performs the best in terms of MSE on the training set, with a significantly lower MSE compared to the other models. However, it has a higher MSE on the testing set, indicating potential overfitting.
* The KNN model performs relatively well, with lower MSE values compared to the Linear Regression model. However, it shows some overfitting as well, as the training MSE is lower than the testing MSE.
* The Linear Regression model has higher MSE values compared to the other models, suggesting that it may not fit the data as well as the other models.

5. Improving the performance selected model:

* Feature Engineering: Explore additional features or transform existing features to better capture the relationships in the data.
* Feature Selection: Identify and include only the most relevant features for modeling, removing any irrelevant or highly correlated variables.
* Regularization: Apply regularization techniques like L1 or L2 regularization in linear models to prevent overfitting.
* Hyperparameter Tuning: Experiment with different hyperparameters for each model to find the optimal configuration that minimizes the MSE on the testing set.

Inferences :

### Here in the above graph we have plotted the plot for journey in a Date vs several flights and got to see that day 09 of month has the most number of flights.

### Here in the above graph we have plotted the plot for journey in a month vs several flights and got to see that May has the most number of flights.

### Now with the help of cat plot only we are plotting a box plot between the price of the flight and the source place i.e. the place from where passengers will travel to the destination and we can see that Banglore as the source location has the most outliers while Chennai has the least.

### Here with the help of the cat plot we are trying to plot the boxplot between the price of the flight and airline and we can conclude that Jet Airways has the most outliers in terms of price

### Now with the help of cat plot only we are plotting a box plot between the price of the flight and the source place i.e. the place from where passengers will travel to the destination and we can see that Banglore as the source location has the most outliers while Chennai has the least.

### Jet airways are booked more often

### Less than 3 Hours are booked more frequently

### 1 stop flights are booked more often compared to others

### Majority of the flights take a stop in Bombay.

### Majority of the flights have no 2nd stop If there is a second stop, chances are high of the place being Cochin.

Jet airways business class has the highest prices between 50k — 80k

All the high cost flights depart from bangalore, rest of the flights have prices between 3k — 50k

All high cost flights have destination as Delhi, rest of the flights have prices between 3k — 50k

If a flight is of business class, its price would be high

The flights with high prices having 1 stop, have stop in Bombay

Flights with 2 stops, having higher prices, have stop in Cochin.

High price flights are lesser during end of month.

Prices are higher in the month of March