Flight Price Prediction

Team 6

Abdallah Hussien

David Raafat

François Adham

Karim Mohamed

Submitted To:

Eng. **Omar Samir**

Agenda

1. Business Part
    a. Motivation
    b. Added Value

2. Technical Part
    a. Preprocessing
    b. Visualization
    c. Insights
    d. MapReduce
    e. Results

Business Part

Motivation

As stated by Harvard Business Review (HBR) travelling is one of the most common hobbies around the world as well as being essential for some careers, so we thought about making a model to book the best flight an easier decision.

Added Value

Our model is to predict flight prices knowing some data like destination, airline and class.

This model can be used in many applications like :

1. Airlines can get use of it by predicting competitors' prices and make offers to get a bigger market share.
2. Travellers can use it to know estimated flight prices for a future trip to pick the best airline and save money.

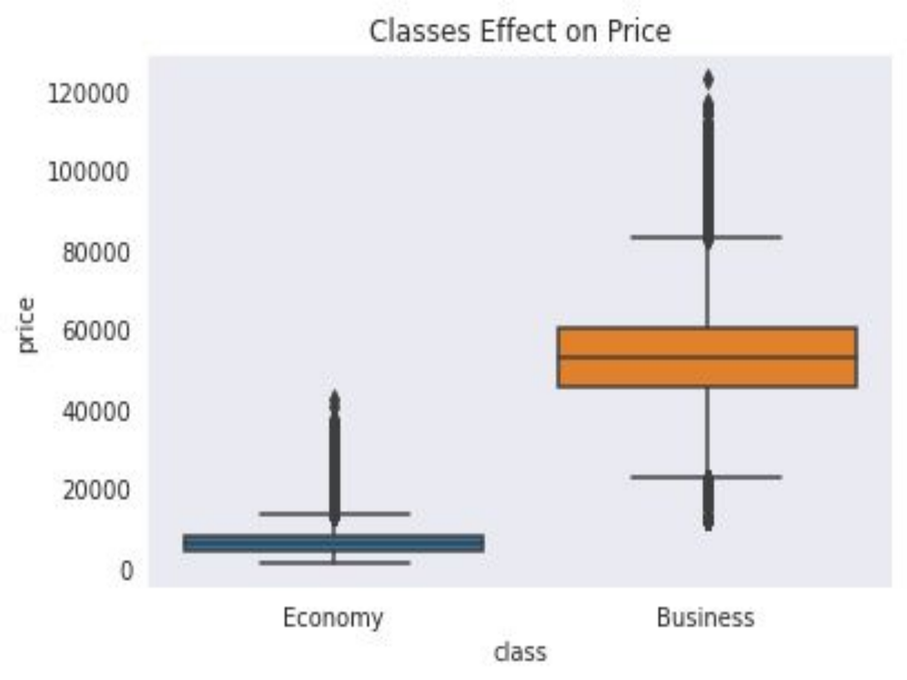
Technical Part

Preprocessing

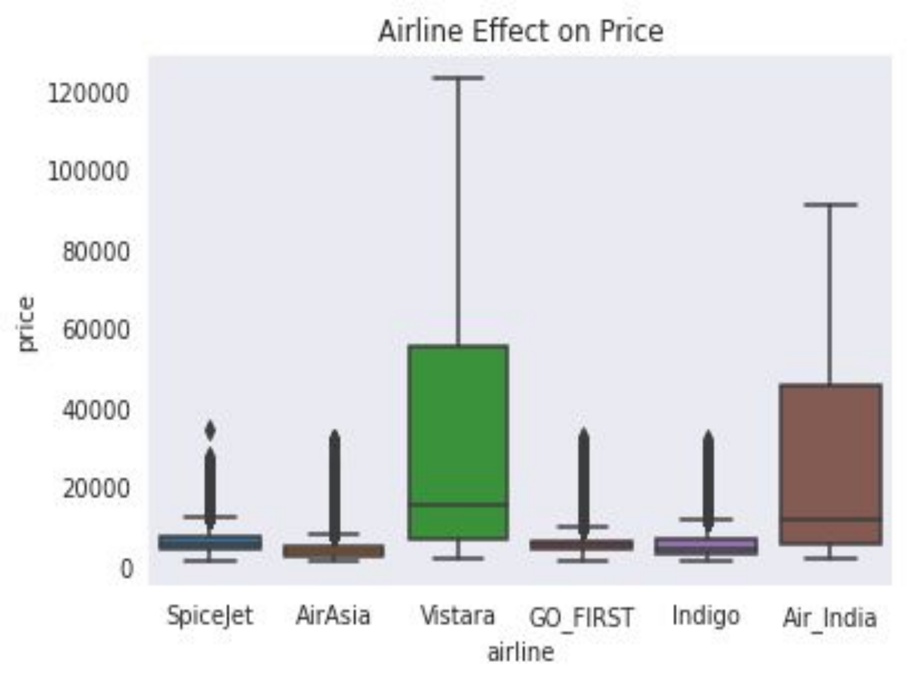
1. Explore the dataset
   a. Check features
   b. Check number of records
2. Check existing NAN values
3. Check how many “Categorical” feature and what are they - if exist.
4. Check how many unique values in these categorical features.
5. Drop useless features like “flight” which is the flight ID.

Visualizations

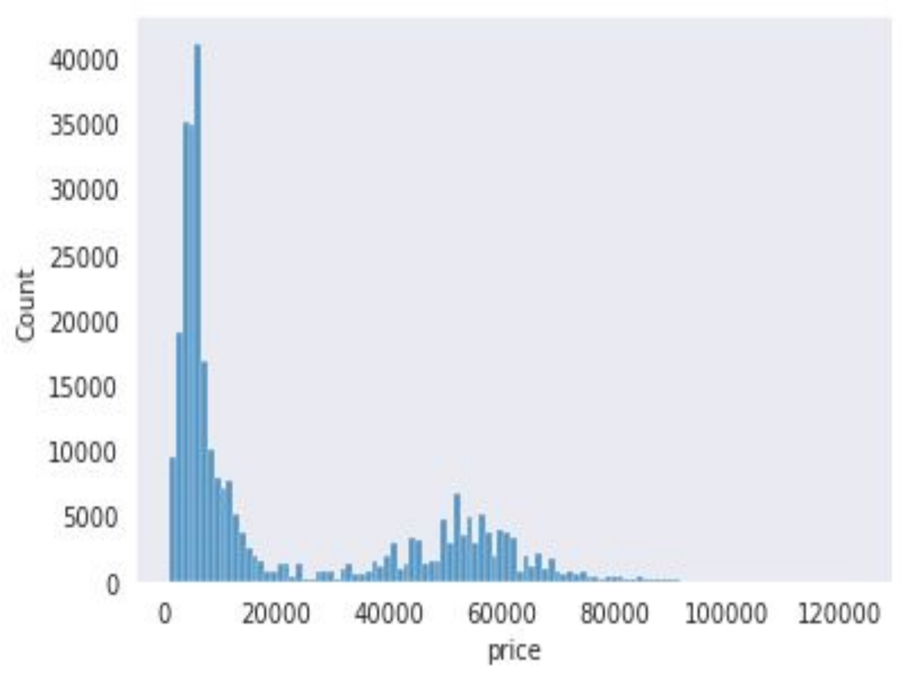
Flight class correlation with price



Airline correlation with price

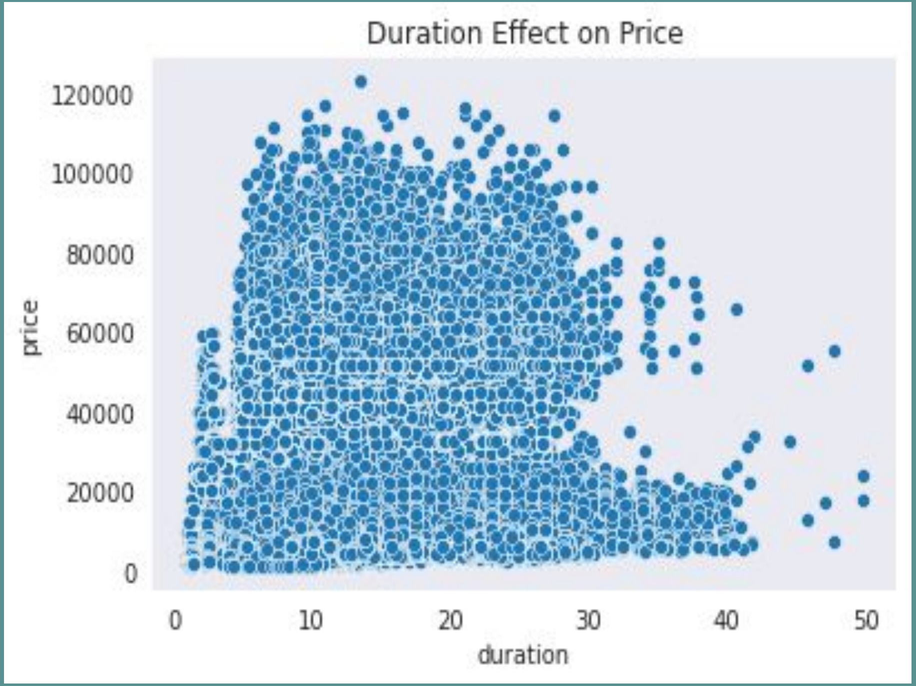


Distribution of flights

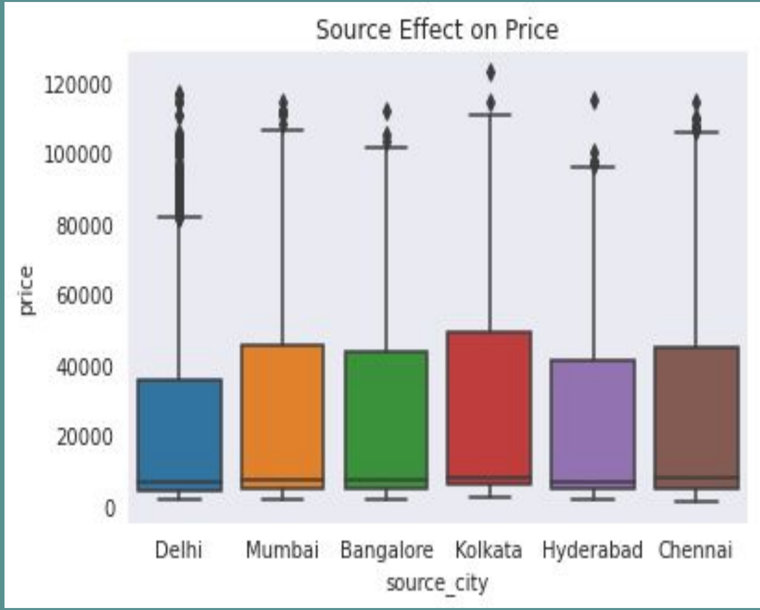


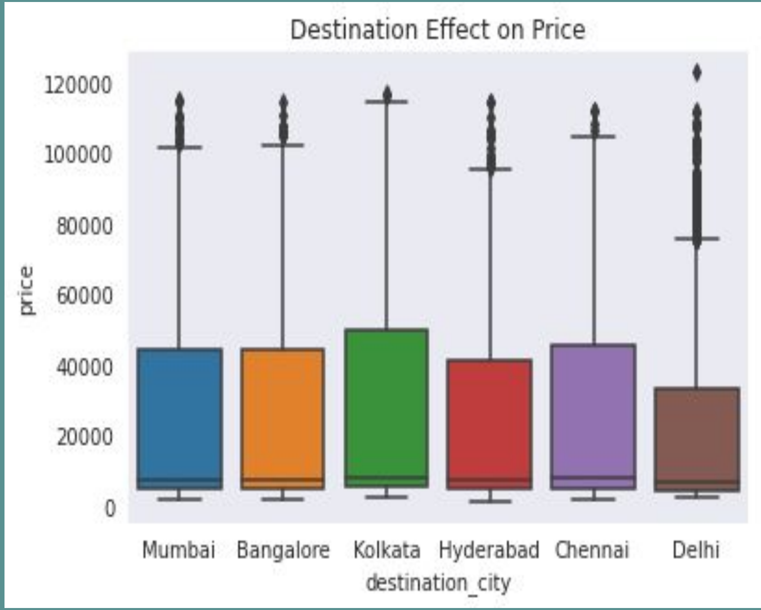
Days left correlation with price

Duration effect on price



Source and Destination effect on price independently





Source_Destination effect on price

Number of stops effect on price

Insights

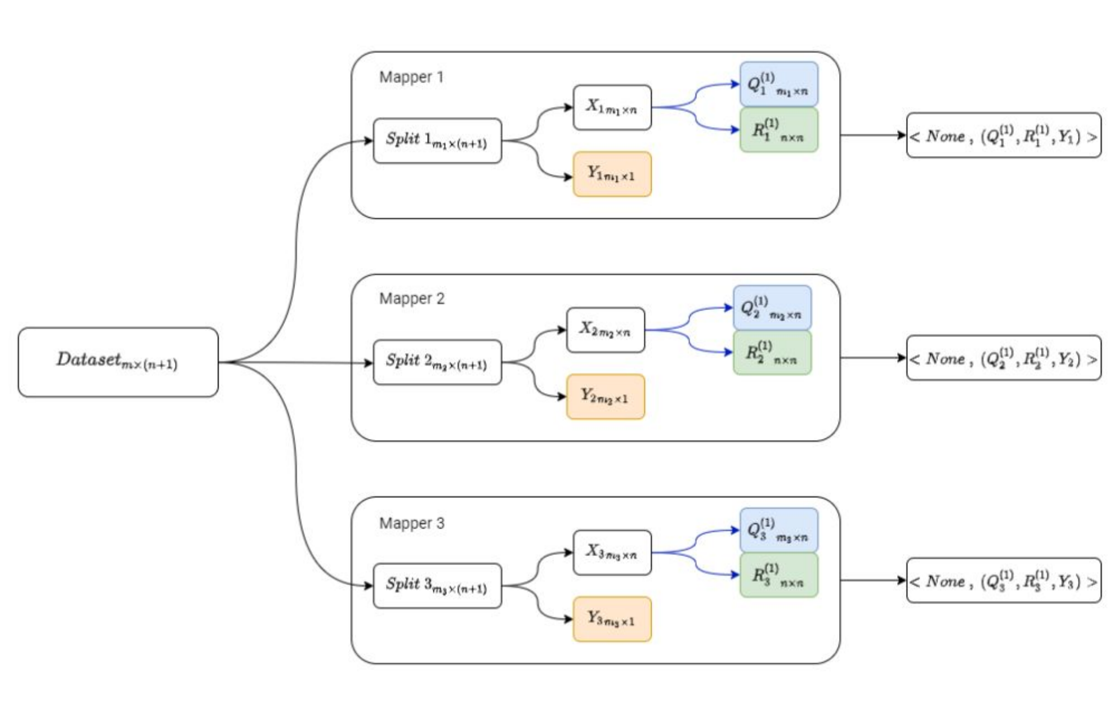
1. Business class price is higher than economy class.
2. Number of cheap flights is higher than number of expensive flights.
3. Less “days_left” means higher price.
4. More “stops” means higher price.
5. Source and destination do not have a significant effect on the price, but pairs of (source, destination) has a higher effect.

MapReduce

Split the dataset to m splits

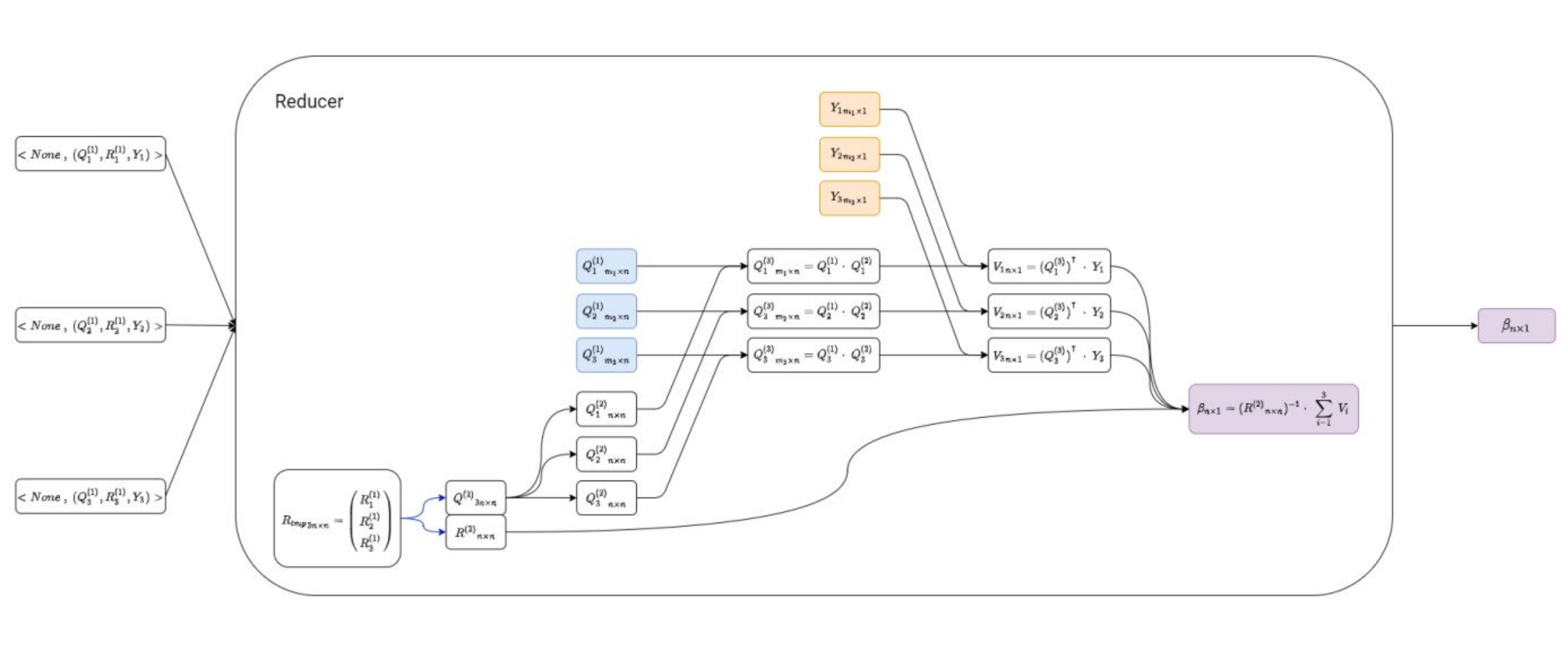
a. Map phase:
   i. Each mapper will decompose the feature vector using QR decomposition i.e. $X_j = Q_j^{(1)} R_j^{(1)}$ where j is the block number.
   ii. Map phase output will be a key-value pair where key will be NONE and the value will be (Q, R)
b. Reduce phase:
   i. Decompose each R matrix using QR decomposition i.e. $R_j^{(1)} = Q_j^{(2)} R_j^{(2)}$
   ii. Compute $Q^{(3)}$ where $Q_j^{(3)} = Q_j^{(2)} Q_j^{(1)}$
   iii. Compute V_j where $V_j = Q_j^{(3)} y_j$ where y_j is a vector of the predicted values
   iv. Compute the weights for each feature $\beta = [\, R^{(2)} \,]^{-1} \Sigma V_j$

Mapper



Mapper 1
Mapper 2
Mapper 3
$Dataset_{m\times(n+1)}$
$Split\ 1_{m_1\times(n+1)}$
$X_{1_{m_1\times n}}$
$Y_{1_{m_1\times 1}}$
$Q_1^{(1)}{}_{m_1\times n}$
$R_1^{(1)}{}_{n\times n}$
$< None\ ,\ (Q_1^{(1)}, R_1^{(1)}, Y_1) >$
$Split\ 2_{m_2\times(n+1)}$
$X_{2_{m_2\times n}}$
$Y_{2_{m_2\times 1}}$
$Q_2^{(1)}{}_{m_2\times n}$
$R_2^{(1)}{}_{n\times n}$
$< None\ ,\ (Q_2^{(1)}, R_2^{(1)}, Y_2) >$
$Split\ 3_{m_3\times(n+1)}$
$X_{3_{m_3\times n}}$
$Y_{3_{m_3\times 1}}$
$Q_3^{(1)}{}_{m_3\times n}$
$R_3^{(1)}{}_{n\times n}$
$< None\ ,\ (Q_3^{(1)}, R_3^{(1)}, Y_3) >$

Reducer



Reducer
$< None\ ,\ (Q_1^{(1)}, R_1^{(1)}, Y_1) >$
$< None\ ,\ (Q_2^{(1)}, R_2^{(1)}, Y_2) >$
$< None\ ,\ (Q_3^{(1)}, R_3^{(1)}, Y_3) >$
$Y_{1_{m_1 \times 1}}$
$Y_{2_{m_2 \times 1}}$
$Y_{3_{m_3 \times 1}}$
$Q_1^{(1)}{}_{m_1 \times n}$
$Q_2^{(1)}{}_{m_2 \times n}$
$Q_3^{(1)}{}_{m_3 \times n}$
$Q_1^{(2)}{}_{n \times n}$
$Q_2^{(2)}{}_{n \times n}$
$Q_3^{(2)}{}_{n \times n}$
$R_{tmp_{3n \times n}} = \begin{pmatrix} R_1^{(1)} \\ R_2^{(1)} \\ R_3^{(1)} \end{pmatrix}$
$Q^{(2)}{}_{3n \times n}$
$R^{(2)}{}_{n \times n}$
$Q_1^{(3)}{}_{m_1 \times n} = Q_1^{(1)} \cdot Q_1^{(2)}$
$Q_2^{(3)}{}_{m_2 \times n} = Q_2^{(1)} \cdot Q_2^{(2)}$
$Q_3^{(3)}{}_{m_3 \times n} = Q_3^{(1)} \cdot Q_3^{(2)}$
$V_{1_{n \times 1}} = (Q_1^{(3)})^\intercal \cdot Y_1$
$V_{2_{n \times 1}} = (Q_2^{(3)})^\intercal \cdot Y_2$
$V_{3_{n \times 1}} = (Q_3^{(3)})^\intercal \cdot Y_3$
$\beta_{n \times 1} = (R^{(2)}{}_{n \times n})^{-1} \cdot \sum_{i=1}^{3} V_i$
$\beta_{n \times 1}$

Results

We used the coefficient of determination equation to compute our model score which is

Coefficient of Determination (R Square)

$$R^2 = \frac{SSR}{SST}$$

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2$$

$$SST = \sum_i (y_i - \bar{y})^2$$

Where,

- SSR is Sum of Squared Regression also known as variation explained by the model
- SST is Total variation in the data also known as sum of squared total
- y_i is the y value for observation i
- y_bar is the mean of y value
- y_bar_hat is predicted value of y for observation i

Results contd.

We tried to remove the duration feature but our model's score dropped a bit so we returned it back.

Dropping source and destination affected a lot our model's score so we returned them back.

Final results:

- Train data: **90.97%**
- Test data: **91.09%**

Thank You