AIR FLIGHT FARES

Big Data Project

2023

Our Team

Donia Abdel-fattah	1	28
Raghad Khaled	1	30
Menna Allah Ahmed	2	29
Nada Elsayed	2	32

Presented To:

- Dr. Lydia Wahid
- Eng. Omar Samir

PROJECT PIPELINE

Check data Type

Check the null values

check unique values for each columns

correlation between numerical columns

Get description for the Fare values (Mean, Std, min, max, ...)

Explore the data

Preprocessing

Remove duplicate rows in data

Reformate the date “%Y-%m-%d”

Split “Date_of_journey” column to day and month

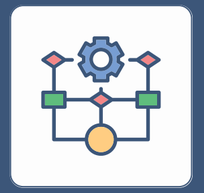
Extract flight code, Airline and class columns

Extract Arrival Time, Source, Destination and Depurature Time

Convert Arrival Time and Departure Time to categories (Before 6 Am , 12 PM to 6 Pm , ..)

Convert day to name (eg. from 16-01-2023 to Monday)

Convert Duration Time to decimal (eg. 2h 30m to 2.0833)



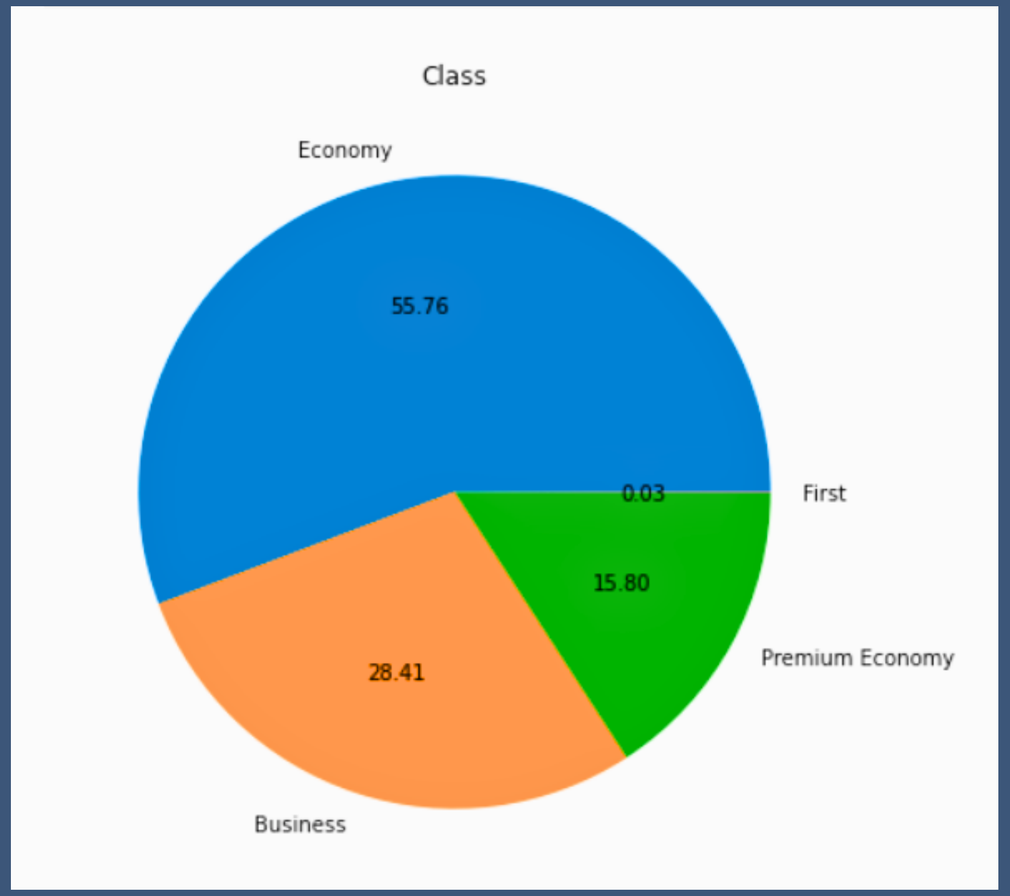
Convert Months to categorical.

Encode categorical columns to numerical.

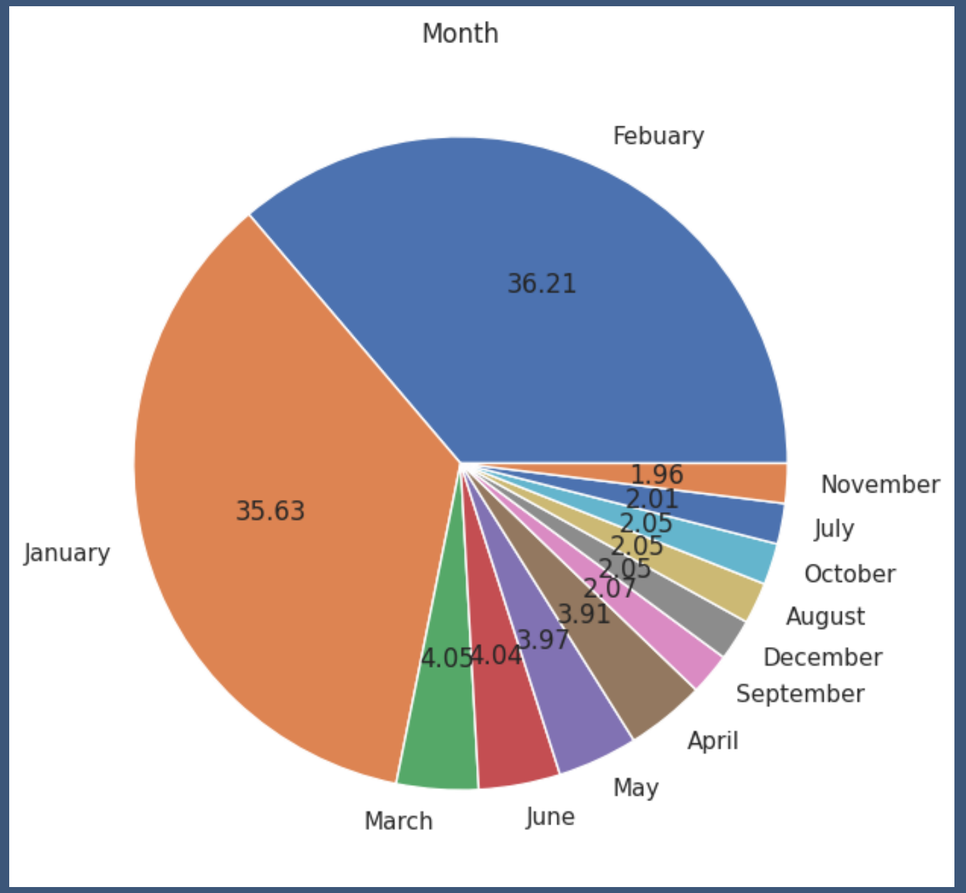
Normalize the feature before get in map reducer model

Remove uncorrelated features before training.

Data Visualization

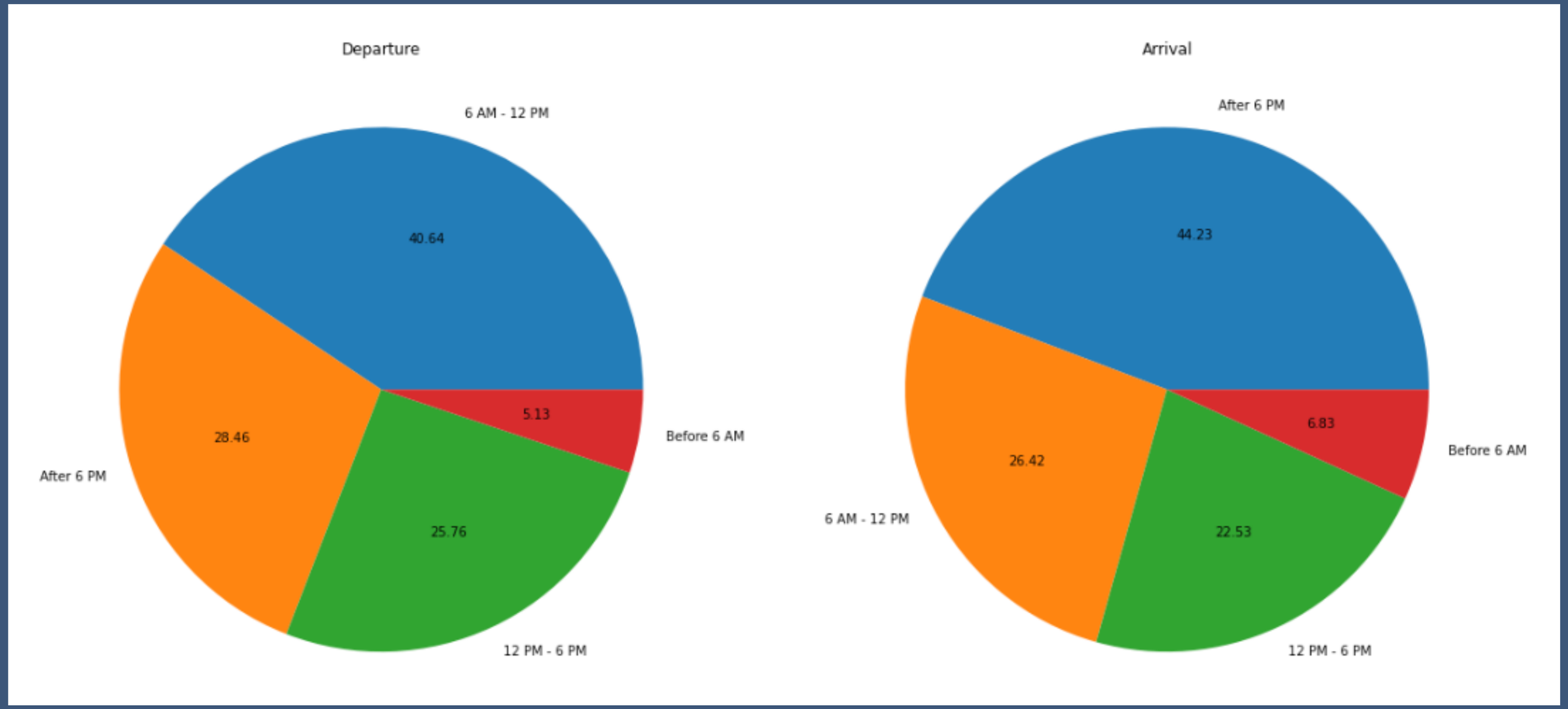


Percentage of each Flight class in data
(Economy class with the highest percentage)



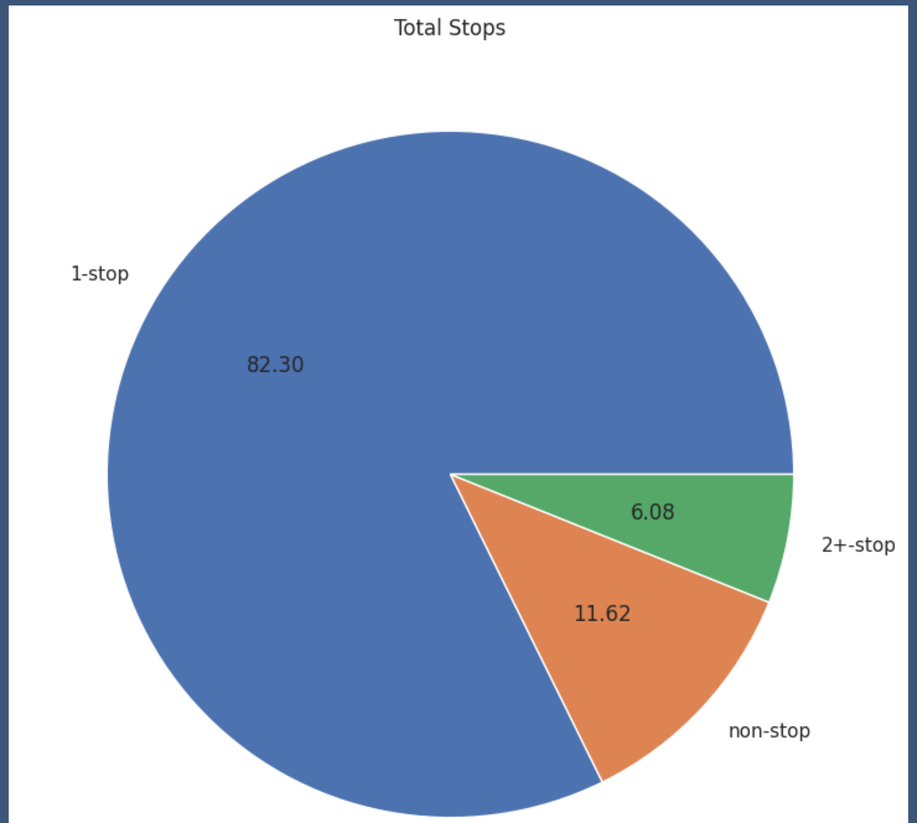
Most flights are in february

Data Visualization

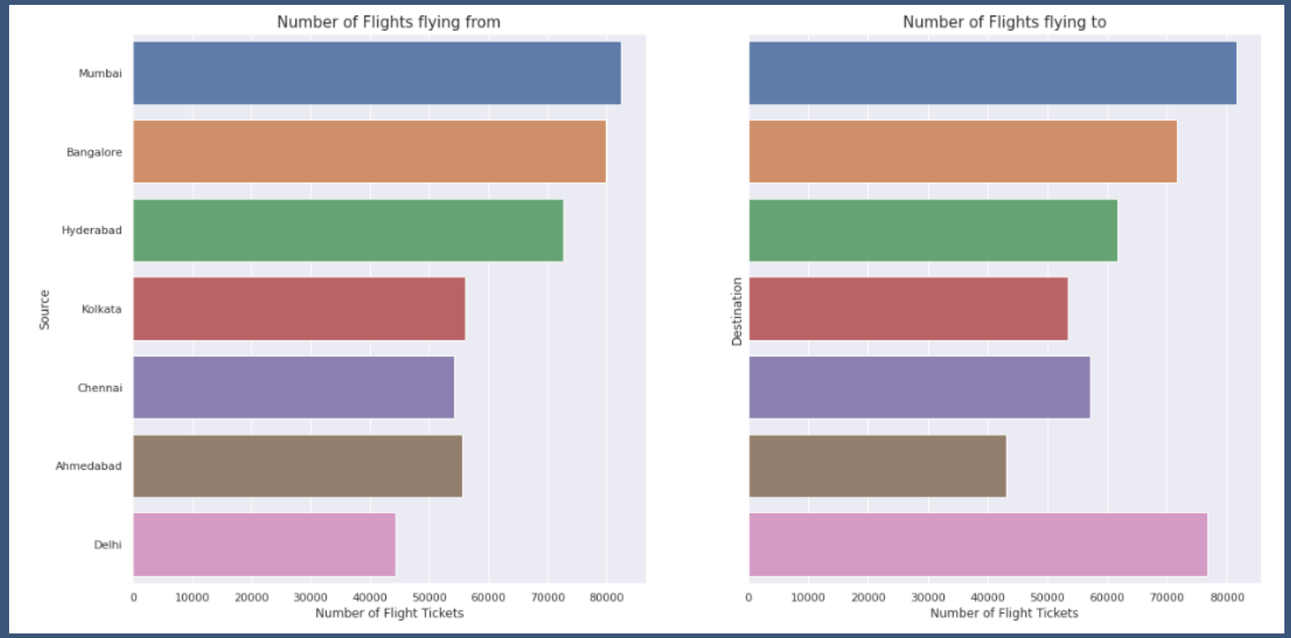


Percentage of each Departure and Arrival Time for the flights
most flights have Departure and Arrival time between 6AM - 12PM

Data Visualization

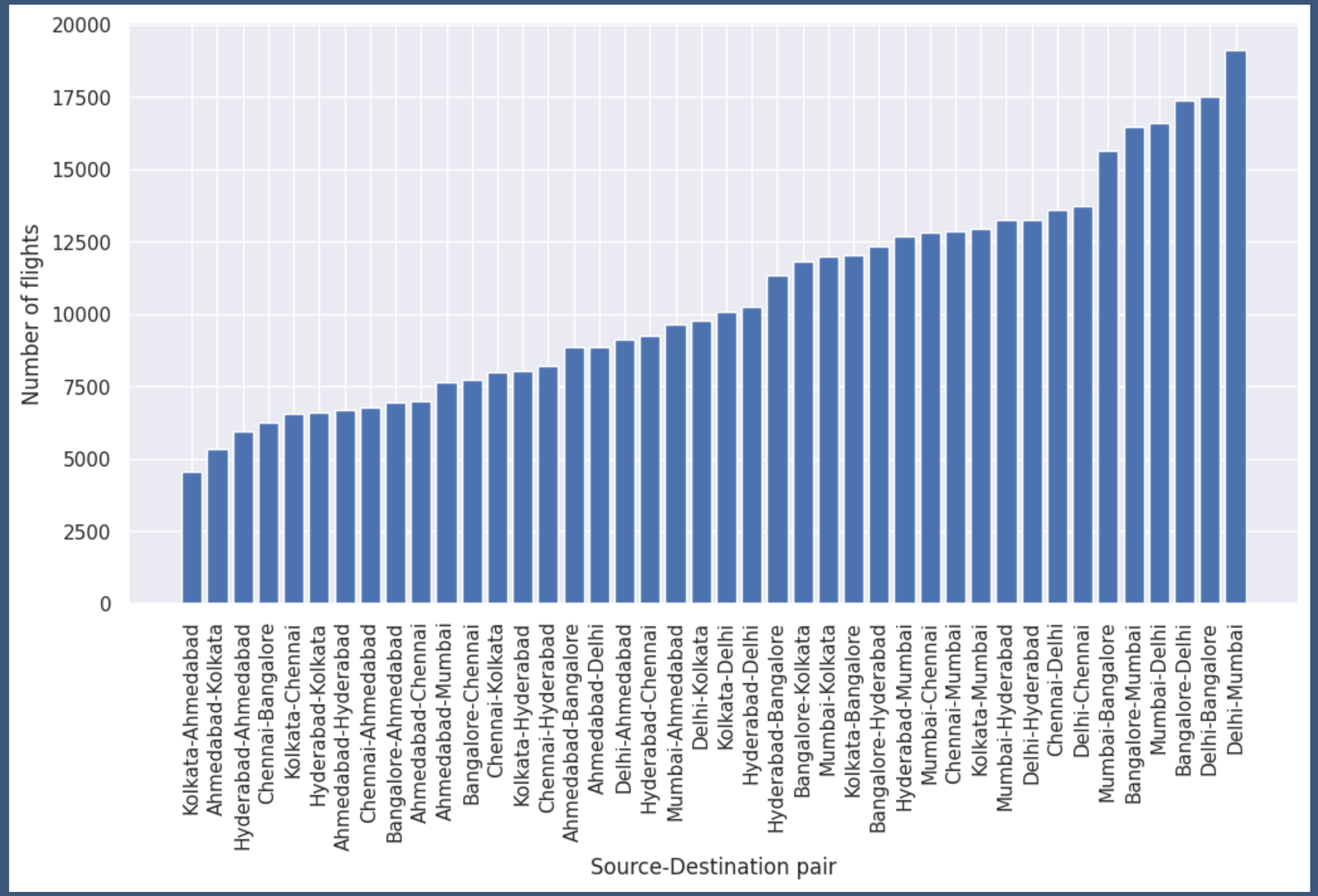


Most flights have one stop



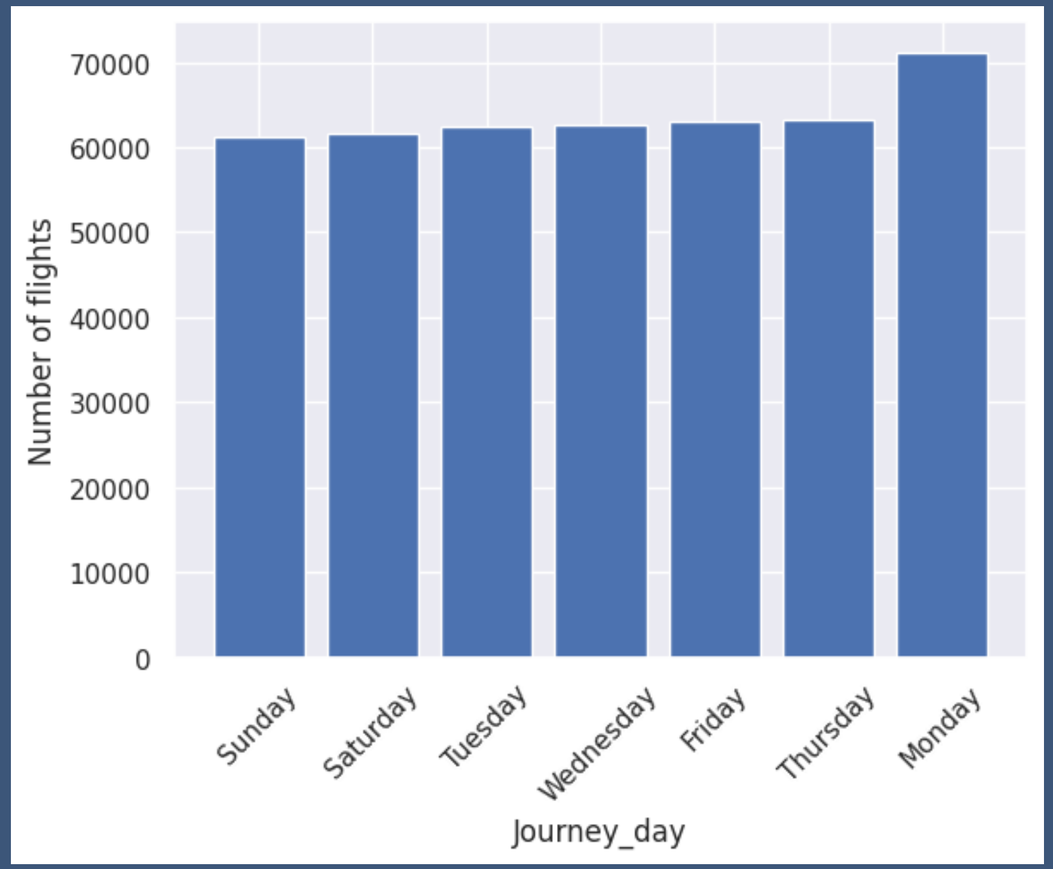
Most flights are booked from Mumbai
and
Most flights booked are to Mumbai and Delhi

Data Visualization

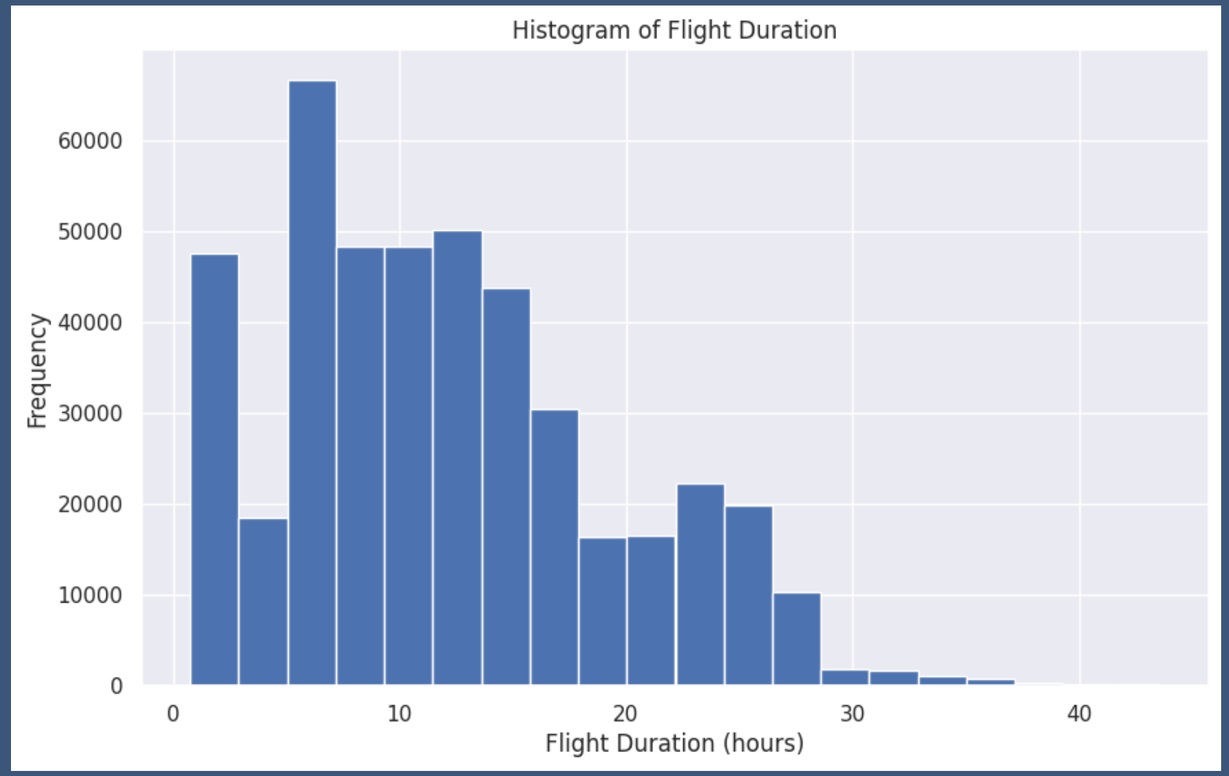


The highest number of flights are between
Delhi-Mumbai and Delhi-Bangalore

Data Visualization

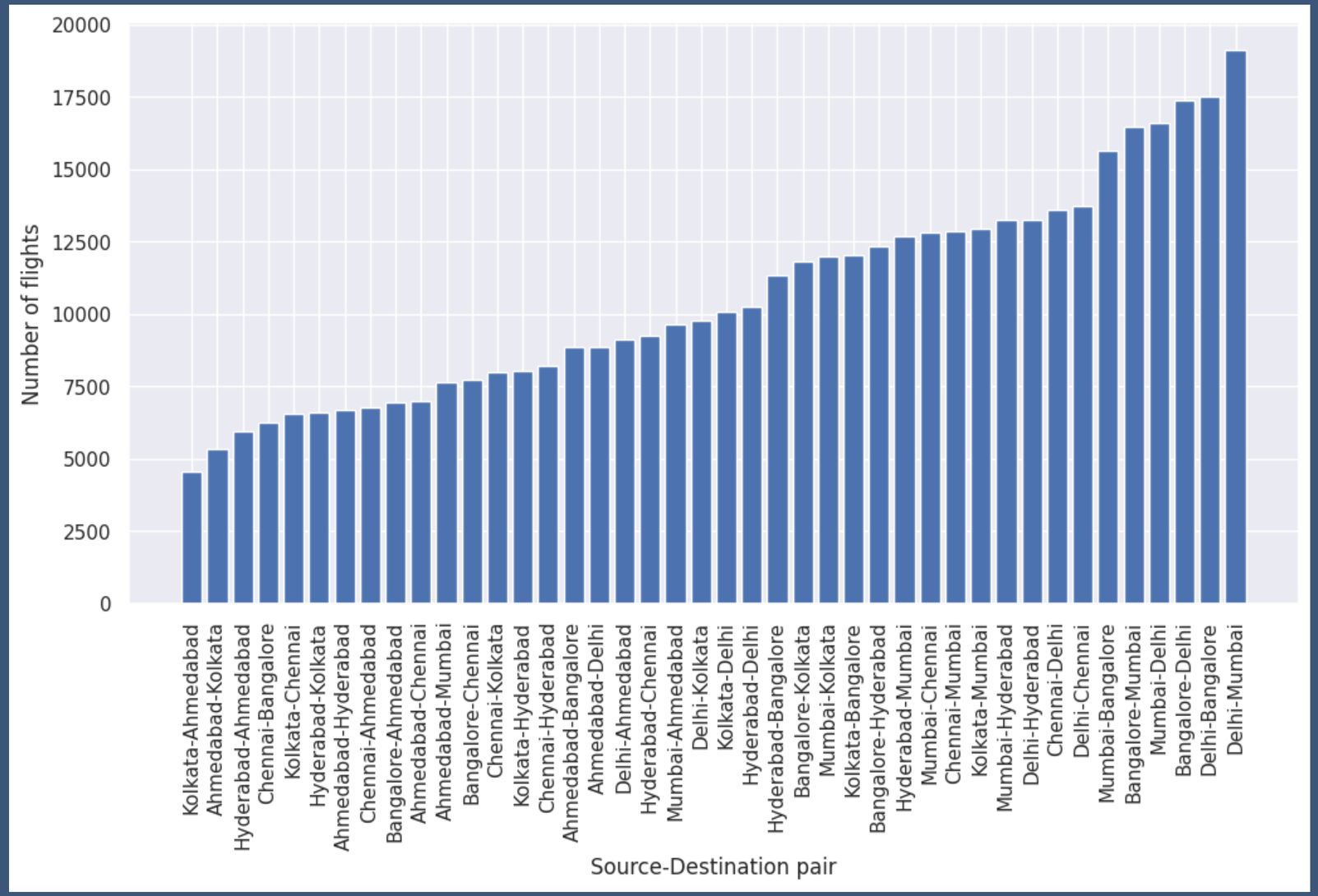


All days has almost the same number of flights



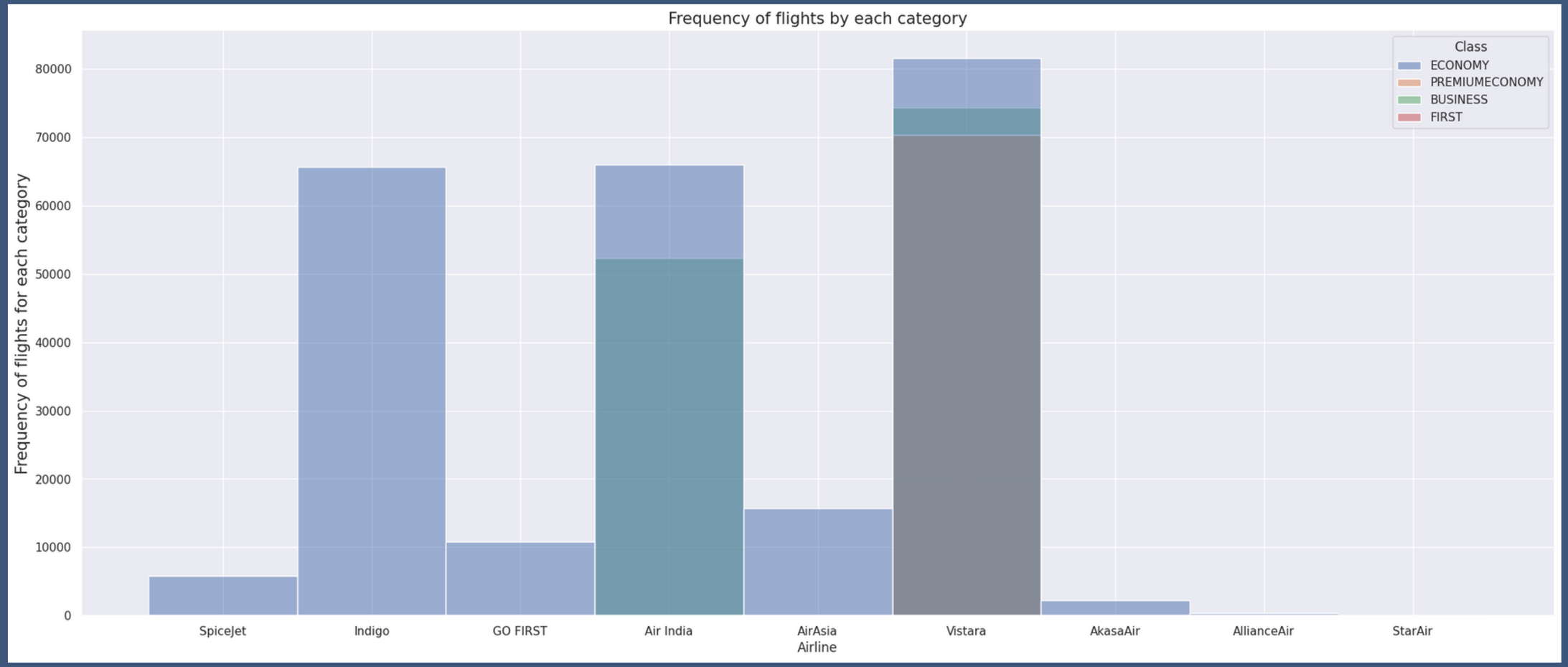
The minimum duration of a flight is 0.75 hours, indicating that there are some very short flights in the dataset. The maximum duration of a flight is 43.58 hours and most of the flights has less than 20 hour

Data Visualization



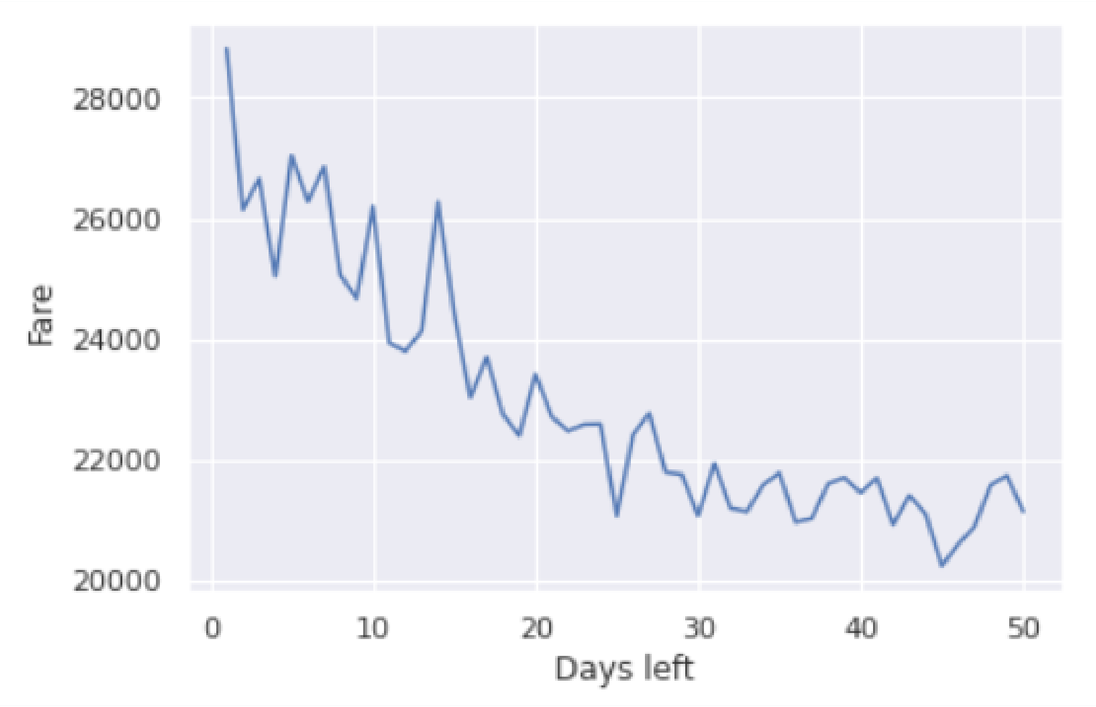
The highest number of flights are between
Delhi-Mumbai and Delhi-Bangalore

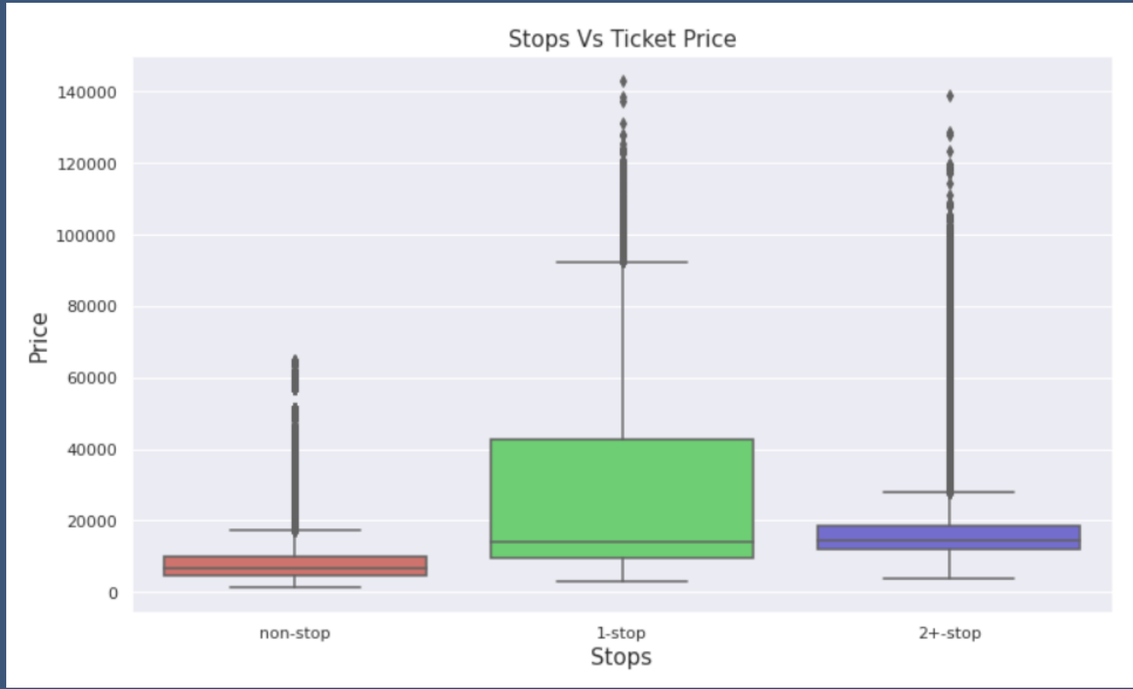
Data Visualization



- Vistara is the only airline that has Premium Economy class
- Air India is the only airline that has first class
- Vistara and Air India are the only airline that has Business class
- all airlines has Economy class

Extracting insights from data

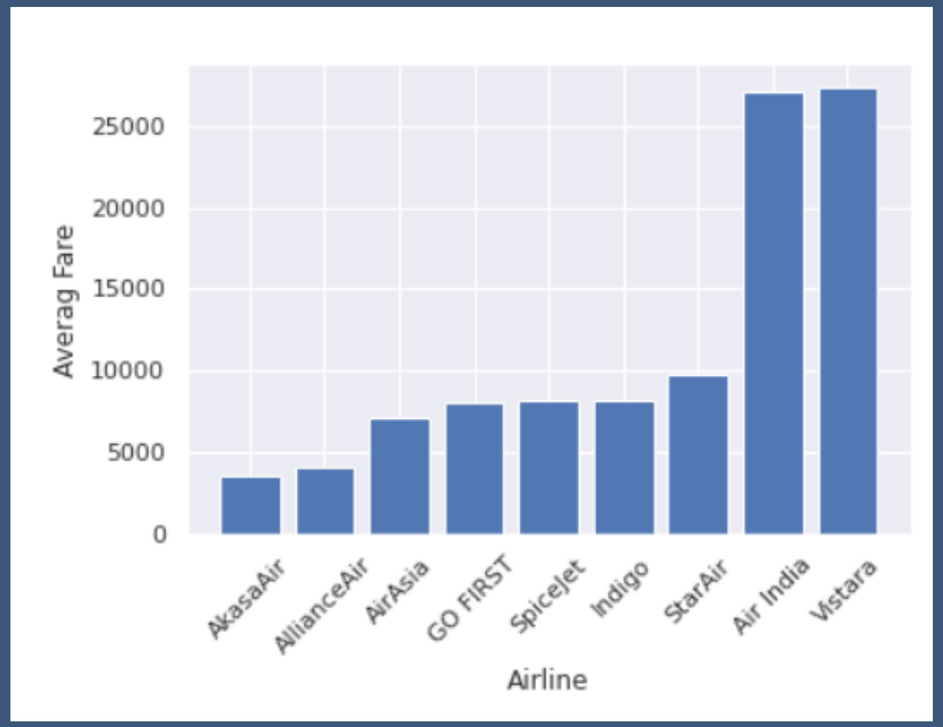


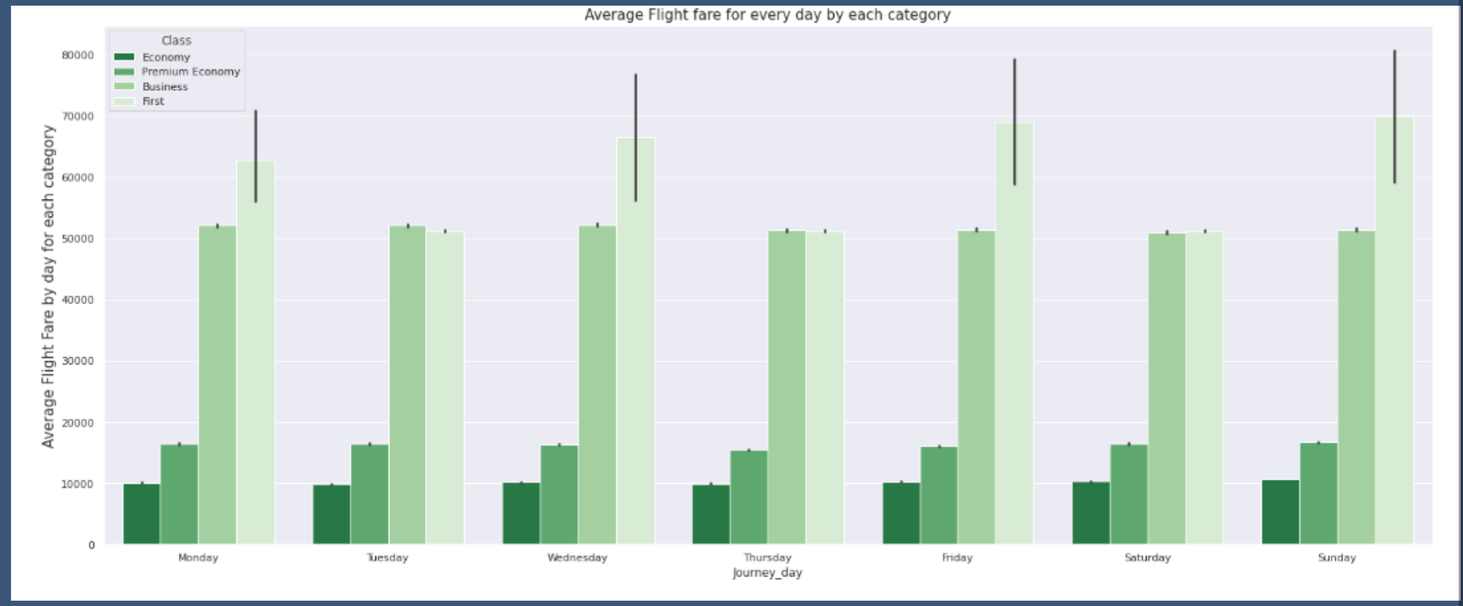


There is a slight decrease in fares as the number of days left to the journey increases. The fare is highest when there is only one day left for the journey, and it decreases gradually as the days left increase. However, this trend is not linear, and there are some fluctuations in the fare values for certain days left.

Flights with 1-stop has the highest mean price and variance and non-stop has the cheapest price but lower variance this result not make sense as non-stop flight expected to have more price than 1-stop and 2+-stop but may other factors affect as the distance

Extracting insights from data

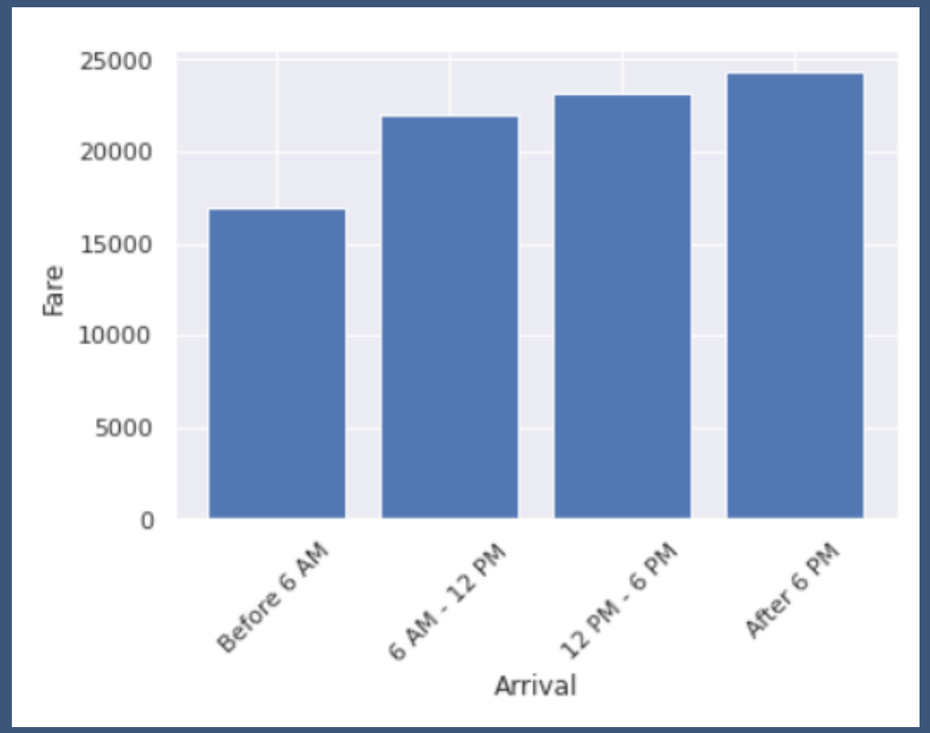


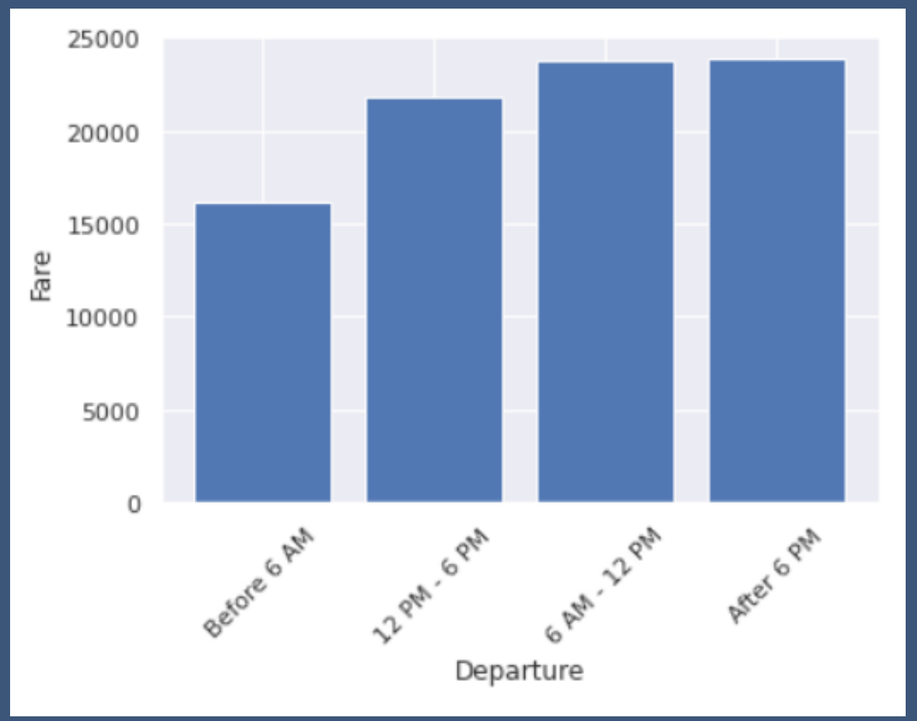


Airline (Air India & Vistara) has the highest average Price while AkasaAir has the cheapest price

Reason for Air India get the highiest average price because it has first and business classes and Vistara has Pre Economy which are with high price and for economy flight they have slightly higher price and AkasaAir has the lower price

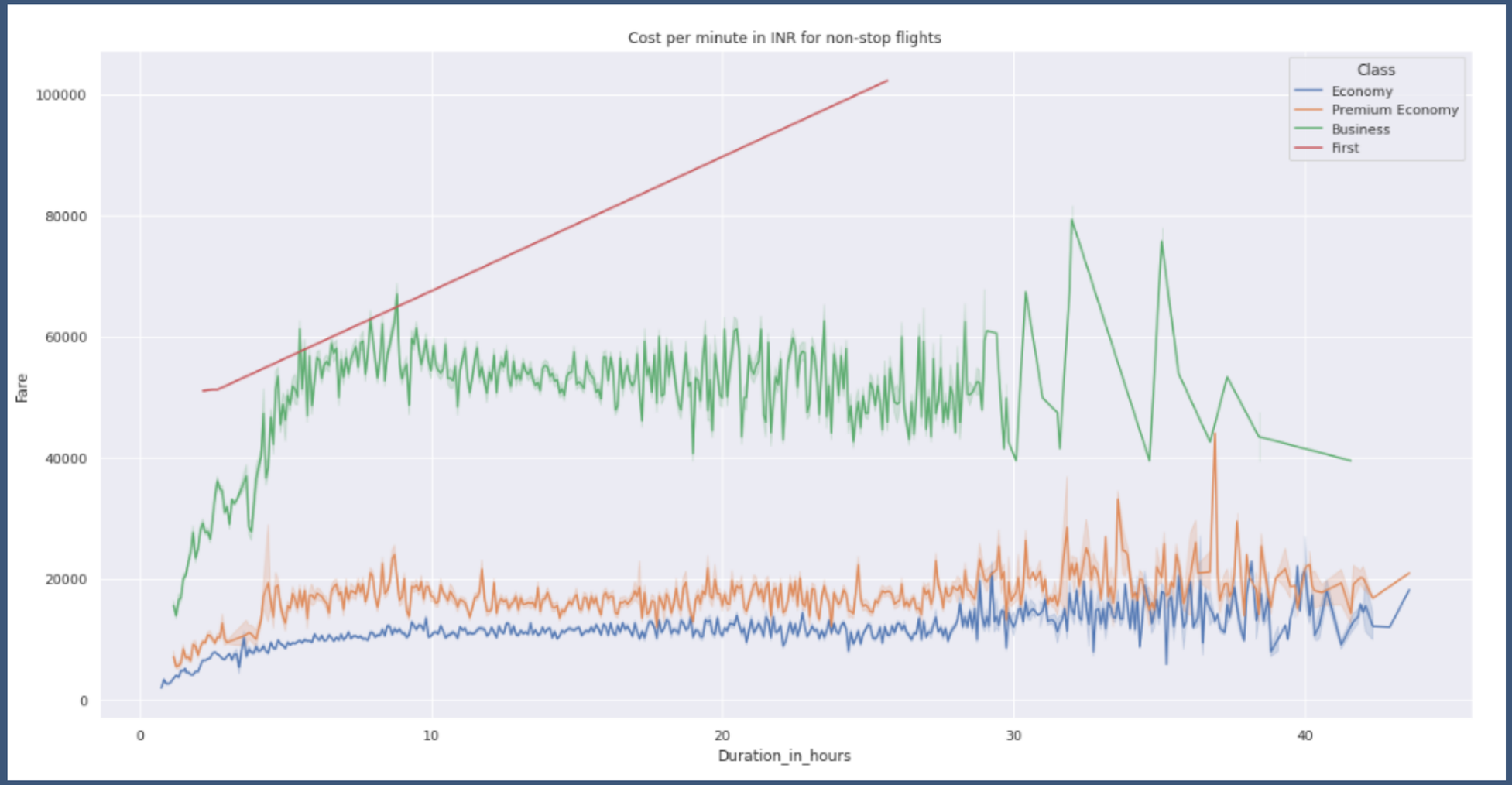
Extracting insights from data





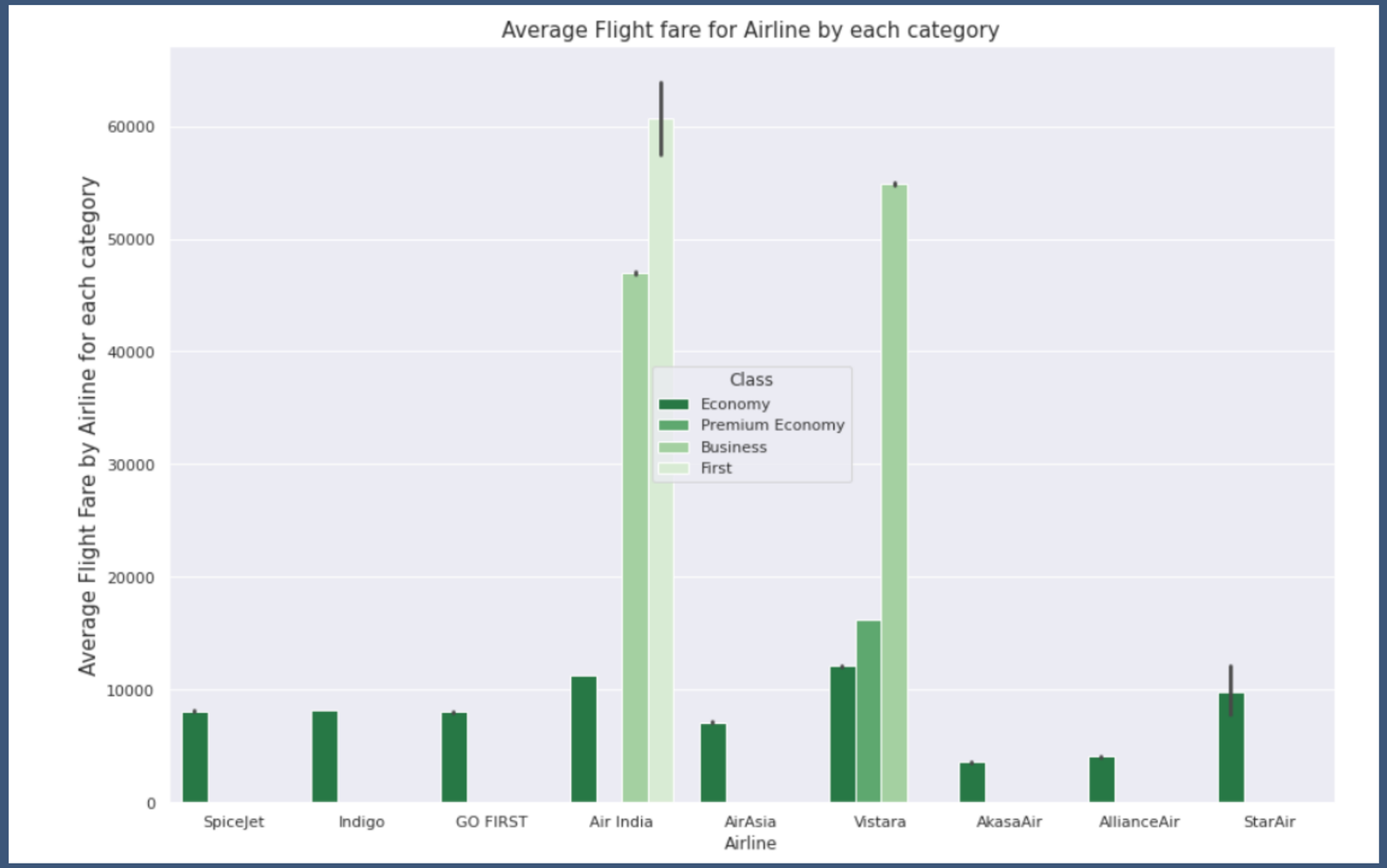
It can be concluded that flight prices vary depending on the Departure/Arrival time of the flights. The highest fares are observed for flights departure/Arrival in the evening (after 6 PM),while the cheapest fares are observed for flights departure/Arrival before 6 AM

Extracting insights from data



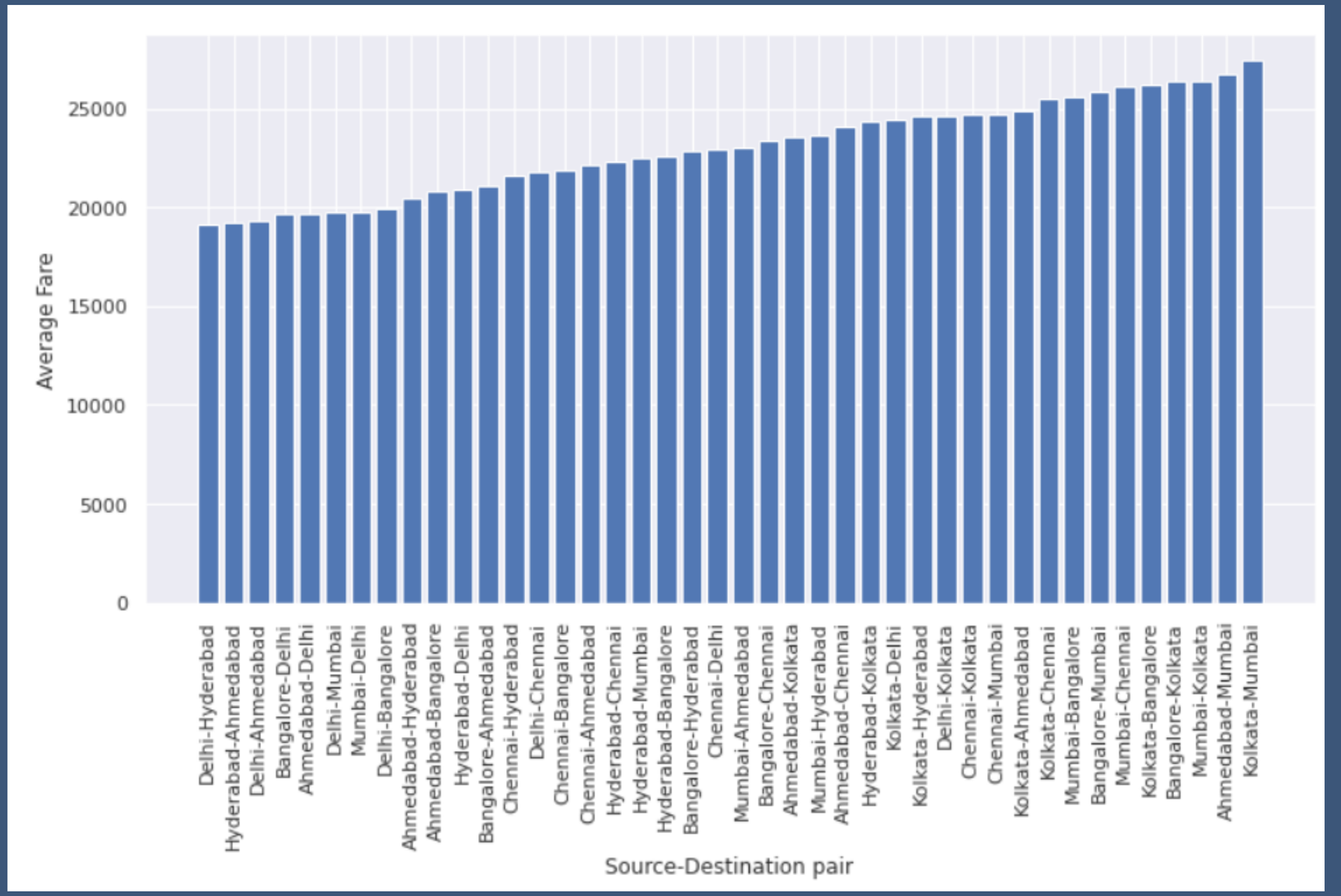
There is a increase in fares as the duration in hour increase for first class. but in other clasess there is high fluctuations in the fare values but not great sign of increase

Extracting insights from data



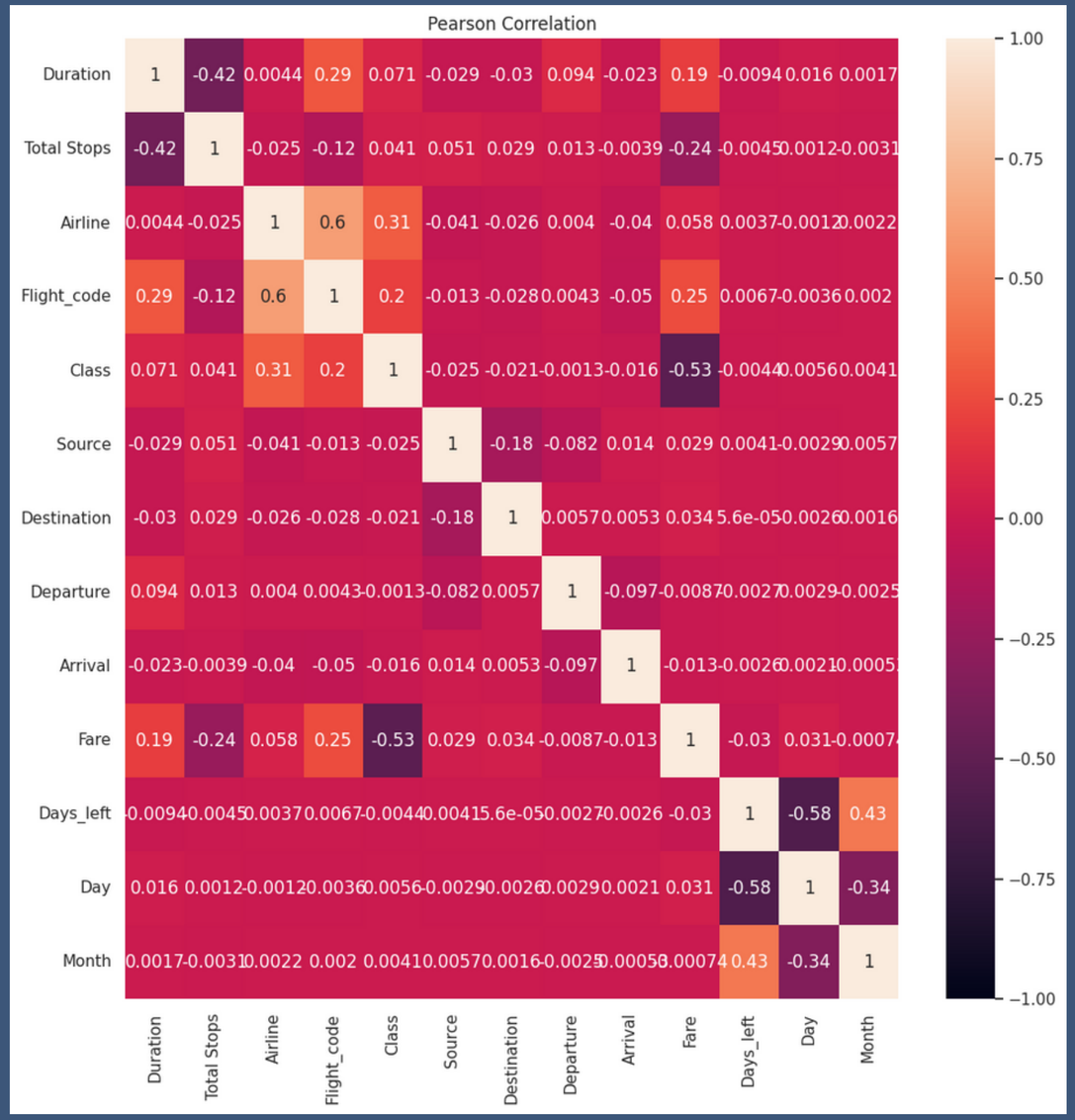
Reason for Air India get the highiest average price
because it has first and business classes and Vistara has Pre Economy which are with high price and for economy flight they have slightly higher price and AkasaAir has the lower price

Extracting insights from data



The route with the highest average fare is Kolkata-Mumbai with an average fare of Rs26997.85. The route with the lowest average fare is Hyderabad-Ahmedabad with an average fare of Rs19001.85.

Extracting insights from data



Correlation map

Model/Classifier Training

- Sklearn Models
  - LinearRegression
  - RandomForestRegressor
- PySpark Models
  - DecisionTreeRegressor
  - RandomForestRegressor
  - GBTRegressor
  - LinearRegression
- Map reducer pyspark
  - LinearRegression
  - Knn (k=1)

Results and Evaluation on test data

- Sklearn

	Root Mean Square Error	R-squared score
LinearRegression	15132.276083465254	0.4520834173630537
RandomForest Regressor	4179.116795947831	0.9582097518936157

Results and Evaluation on test data

- Pyspark

	Root Mean Square Error	R-squared score
DecisionTree Regressor	7212.022478	0.875324
RandomForest Regressor	7872.615217	0.851438
GBTRegressor	6579.925293	0.851928
LinearRegression	15110.782143	0.452677

Results and Evaluation on test data

- Map Reduce

	Root Mean Square Error
LinearRegression	25231.5056
KNN	176406.579

Enhancement And Future Work

To improve prediction accuracy :

- **Collect more data with additional features like current prices of aviation fuel and the distance between the source and destination in terms of longitude and latitude as distance affects the flight fare.**

- **Furthermore, it may be advantageous to incorporate data on flight cancellations, delays, and other elements that can affect flight availability and prices.**

- **Provide details on aspects related to the quality of the flight experience, such as legroom. Including this kind of data could give travelers a more complete understanding of the flight market.**

شُكراً