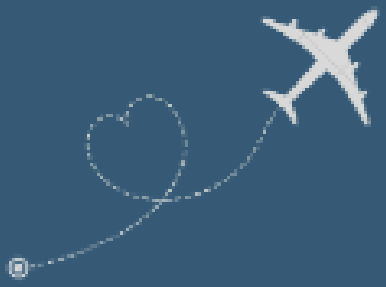
Prepared For :

Big Data Project

AIR FLIGHT FARES

15/05/2023

TEAM3



- Team Members:

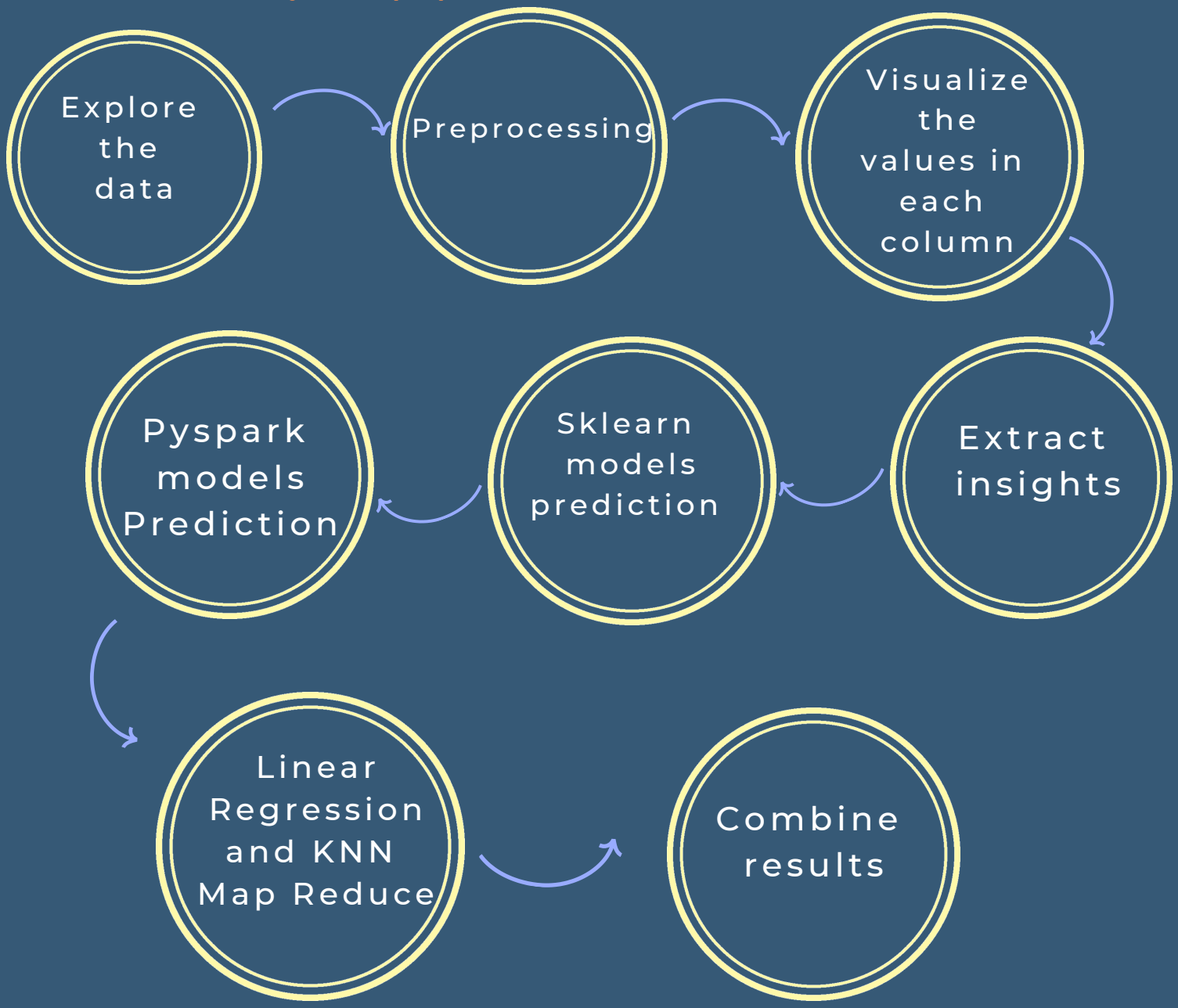
Name	Section	BN
Donia Abdel-fattah	1	28
Raghad Khaled	1	30
Menna Allah Ahmed	2	29
Nada Elsayed	2	32

- Presented To:
- Dr. Lydia Wahid
- Eng. Omar Samir

• Brief Problem Description

Our problem contains information about flights in India and fares for each flight. The goal of the problem is to provide users with information that could help them make informed decisions about when and where to purchase flight tickets. What factor affects the fare of the price? By analyzing patterns in flight fares over time, users can identify the best times to book tickets and potentially save money.

• Project pipeline



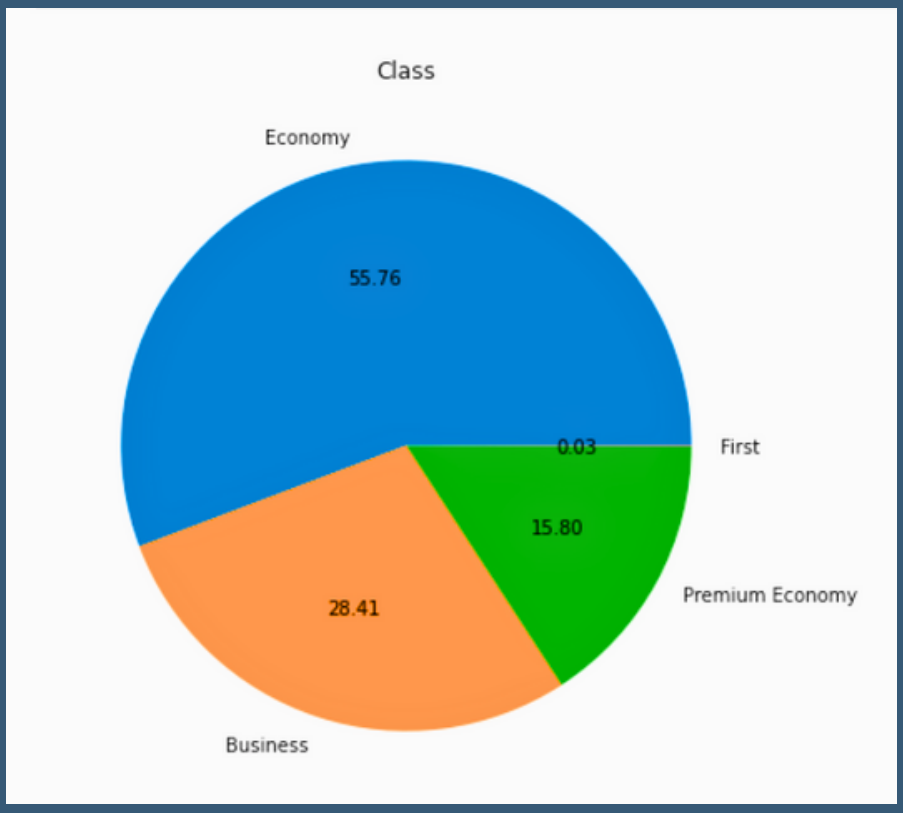
- ## Analysis and solution of the Problem
  - ### Explore the data

1. Get info about each column in the data and the type of data in this column and check the null values and unique values for each column to get the correlation between numerical columns. Get a description for the Fare values (Mean, Std, min, max, ...).

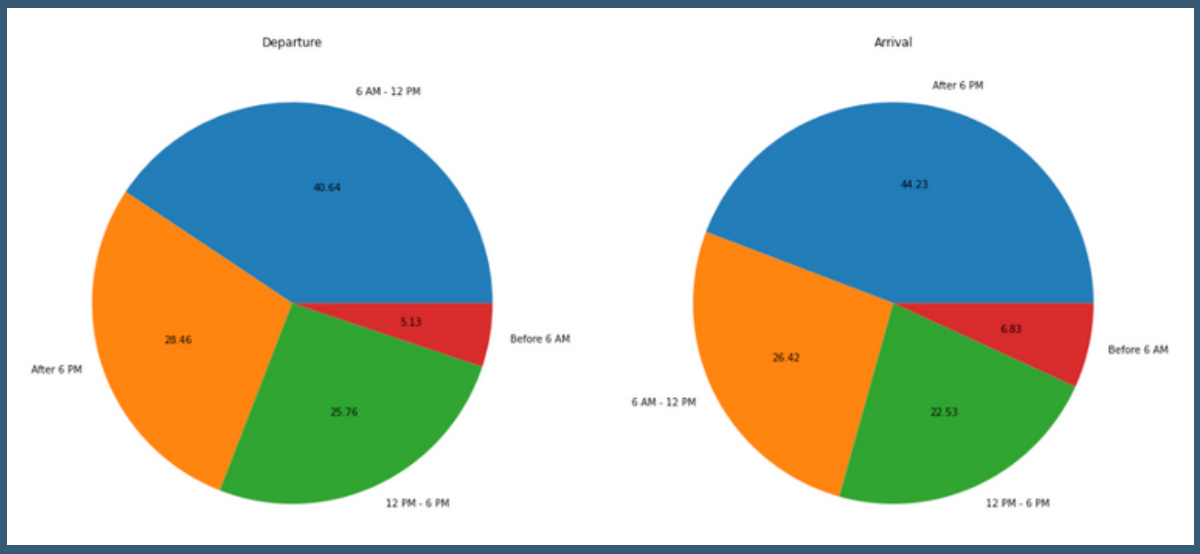
  - ### Preprocessing

1. Remove duplicate rows in the data.
2. Reformate the date “%Y-%m-%d”.
3. Split the “Date_of_journey” column to day and month (no need for a year as all data is in 2023).
4. Extract flight code, Airline, and class columns from the Airline_class column.
5. Extract the Arrival Time, Source, Destination, and Departure Time columns.
6. Convert Arrival Time and Departure Time to categories (Before 6 Am, 12 PM to 6 Pm, ..).
7. Convert day to name (eg. from 16-01-2023 to Monday).
8. Convert Duration Time to decimal (eg. 2h 30m to 2.0833).
9. Convert Months to categorical.
10. Encode categorical columns to numerical ones.
11. Normalize the feature before getting into the map reducer model
12. Remove uncorrelated features before training.

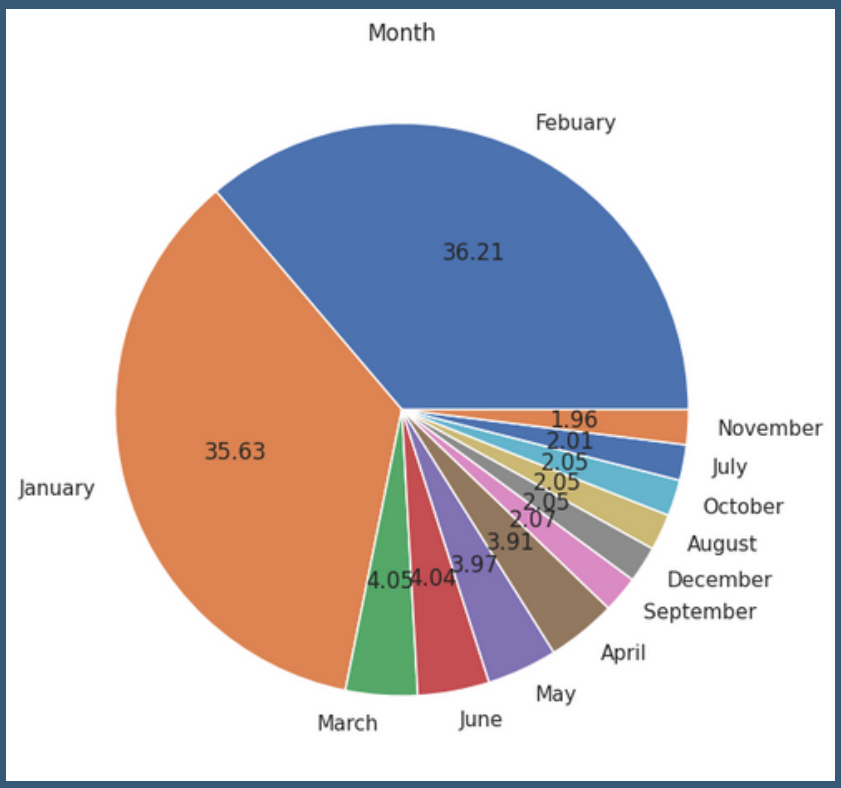
- Data Visualization



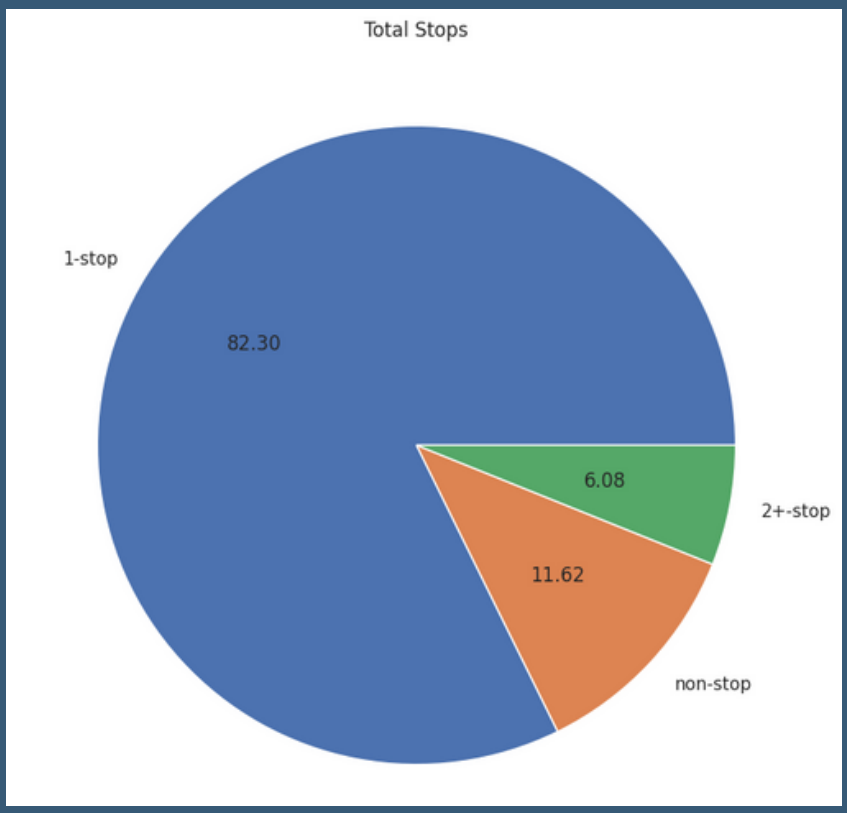
Percentage of each Flight class in data
(Economy class with the highest percentage)



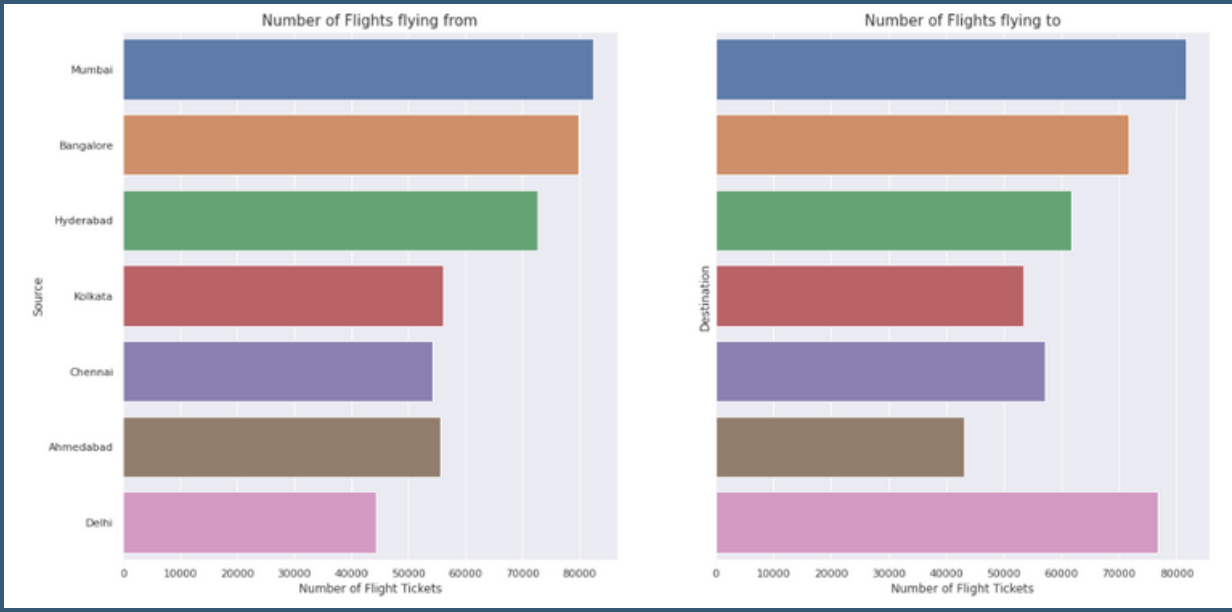
Percentage of each Departure and Arrival Time for the flights
most flights have Departure and Arrival times between 6AM - 12PM



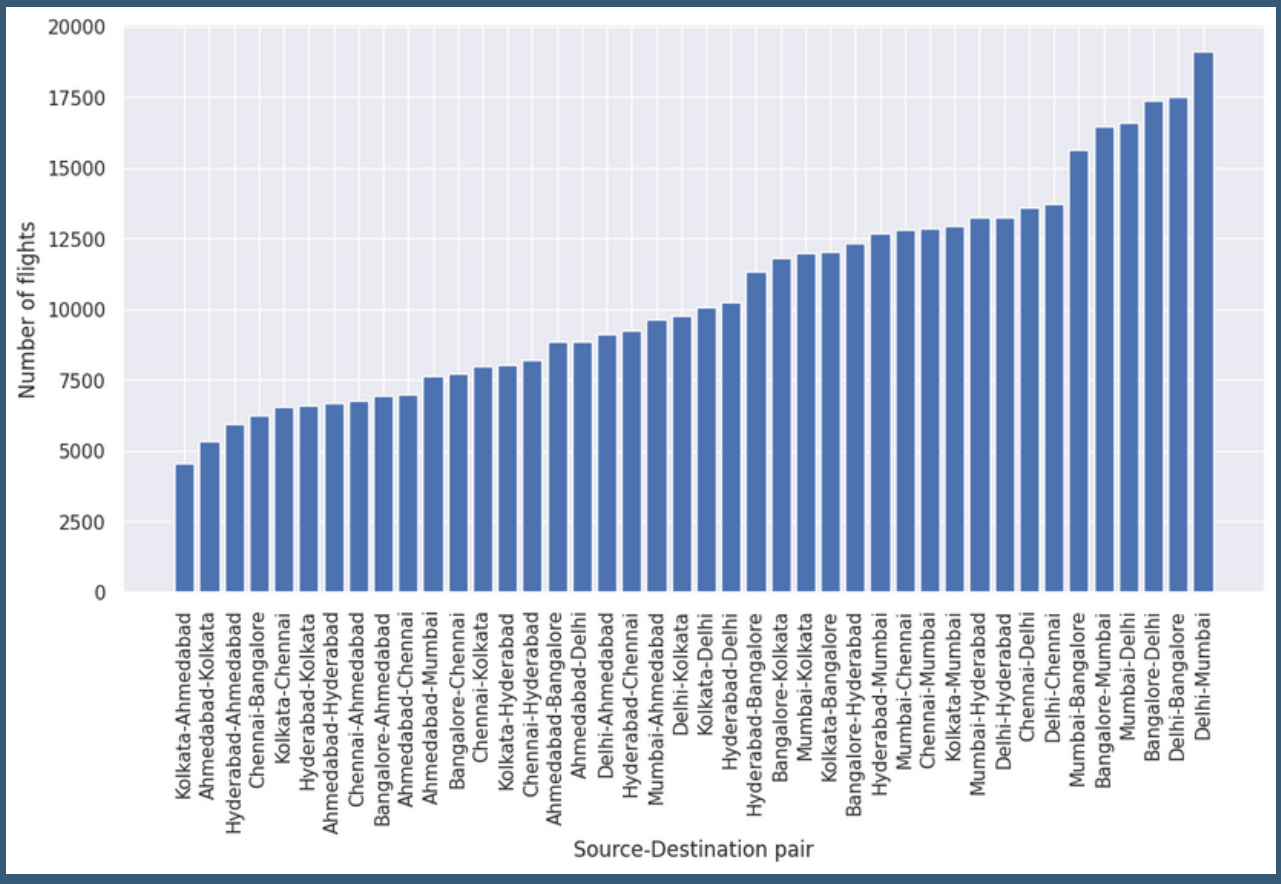
Most flights are in February



Most flights have one stop

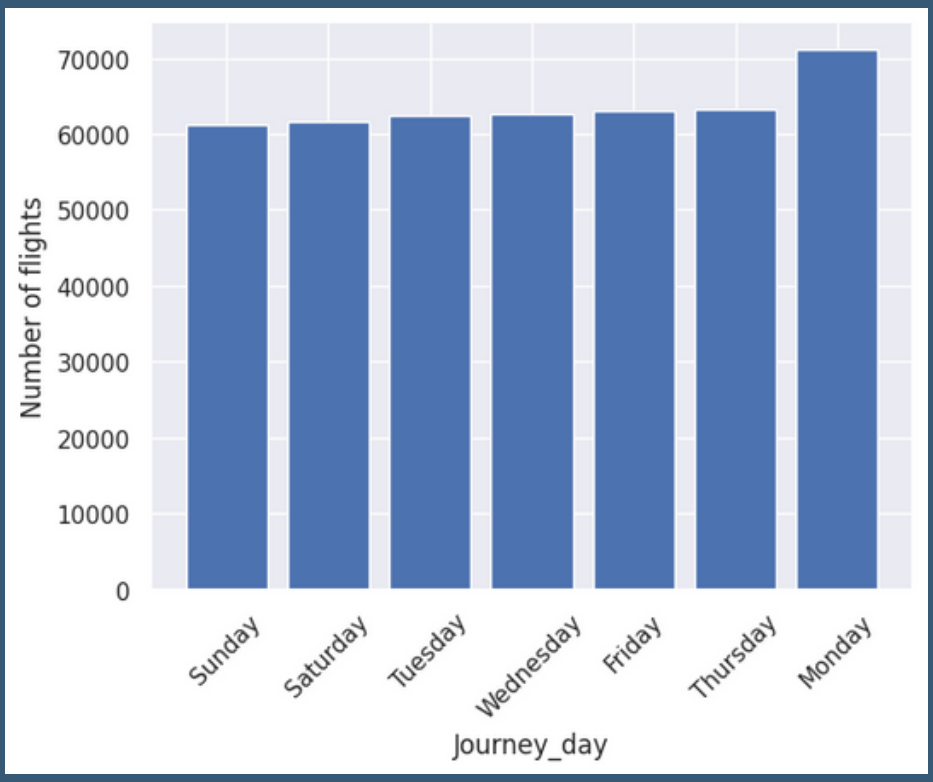


Most flights are booked from Mumbai
and
Most flights booked are to Mumbai and Delhi

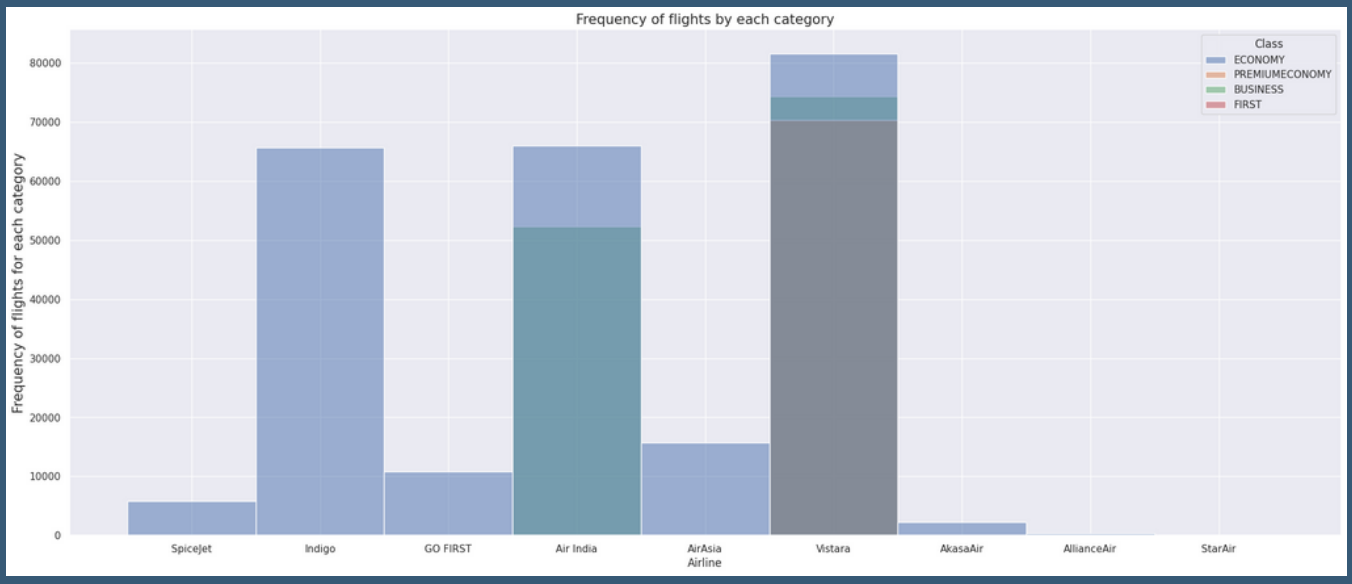


The highest number of flights are between Delhi-Mumbai and Delhi-Bangalore

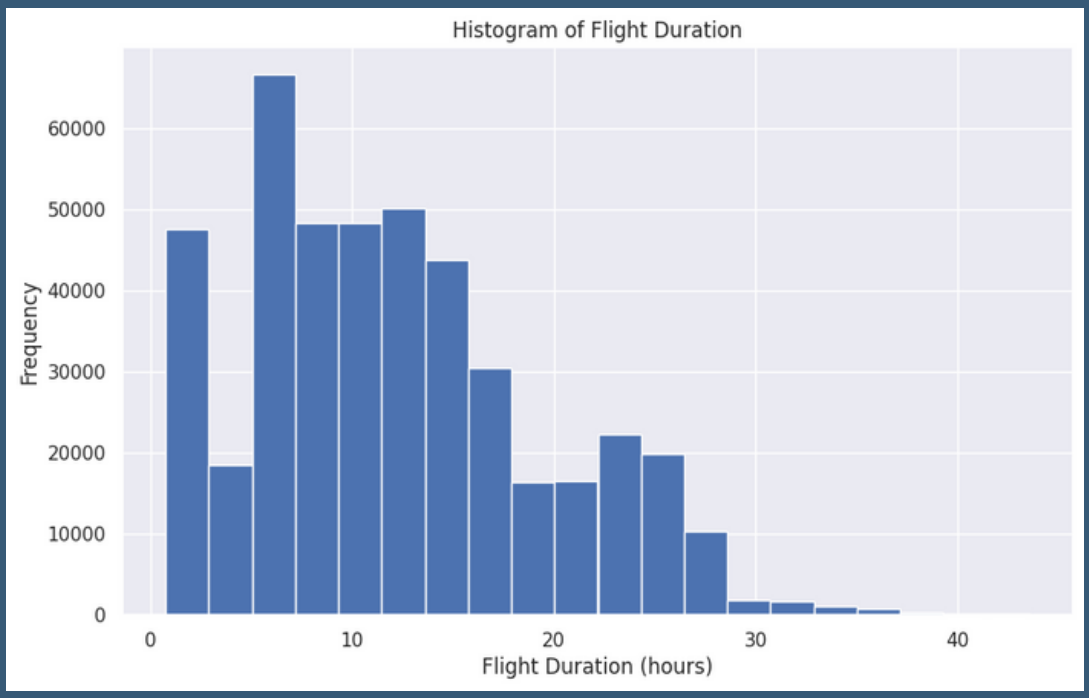
TEAM3



All days has almost the same number of flights but Monday is slightly Higher because it the first day in working.

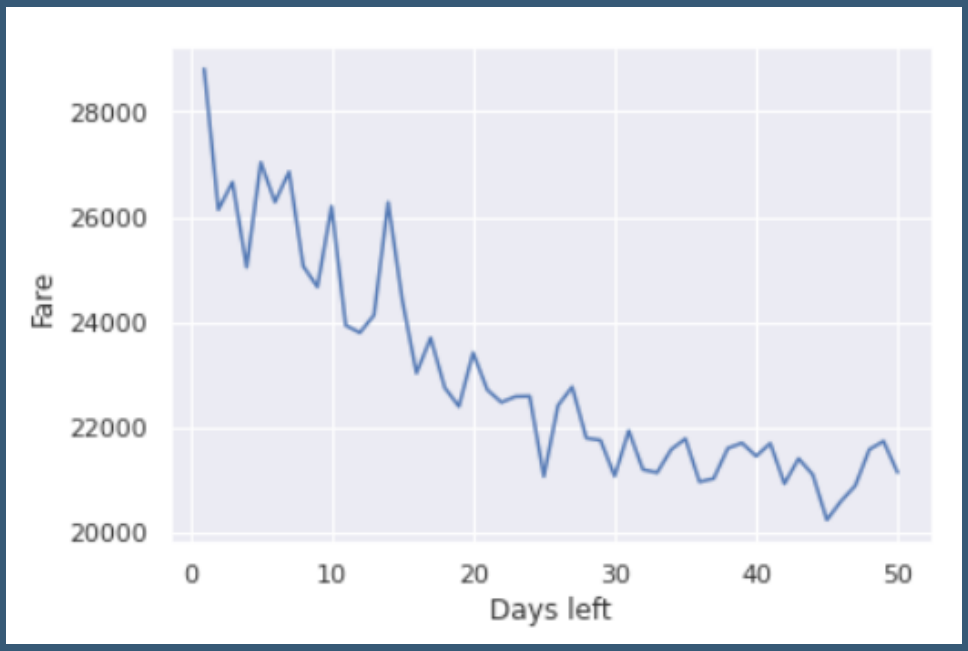


- Vistara is the only airline that has Premium Economy class
- Air India is the only airline that has first class
- Vistara and Air India are the only airlines that have Business class
- all airlines have Economy class

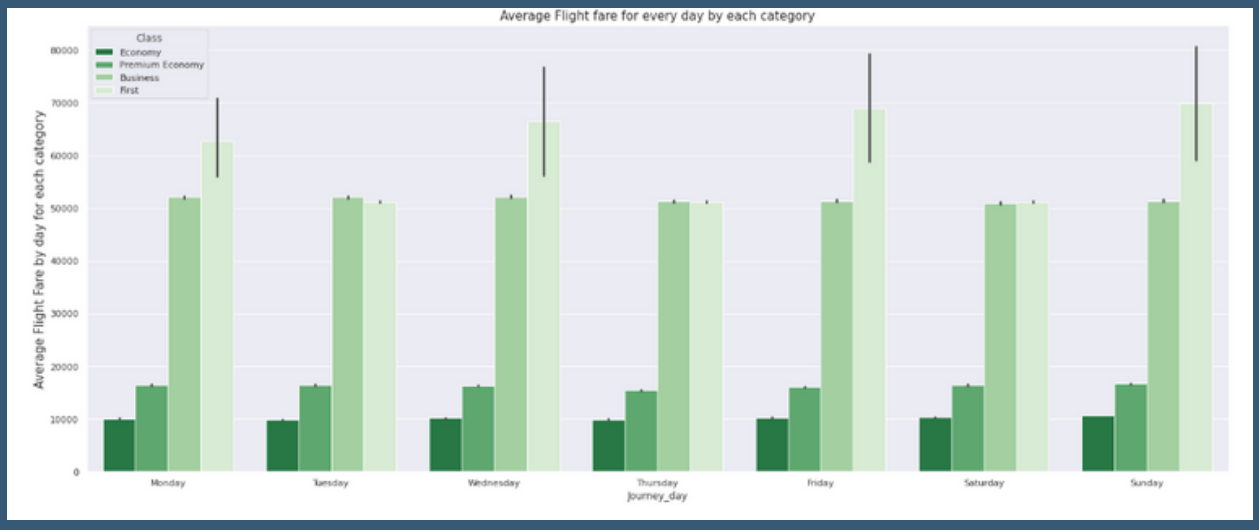


The minimum duration of a flight is 0.75 hours, indicating that there are some very short flights in the dataset. The maximum duration of a flight is 43.58 hours and most of the flights have less than 20 hour

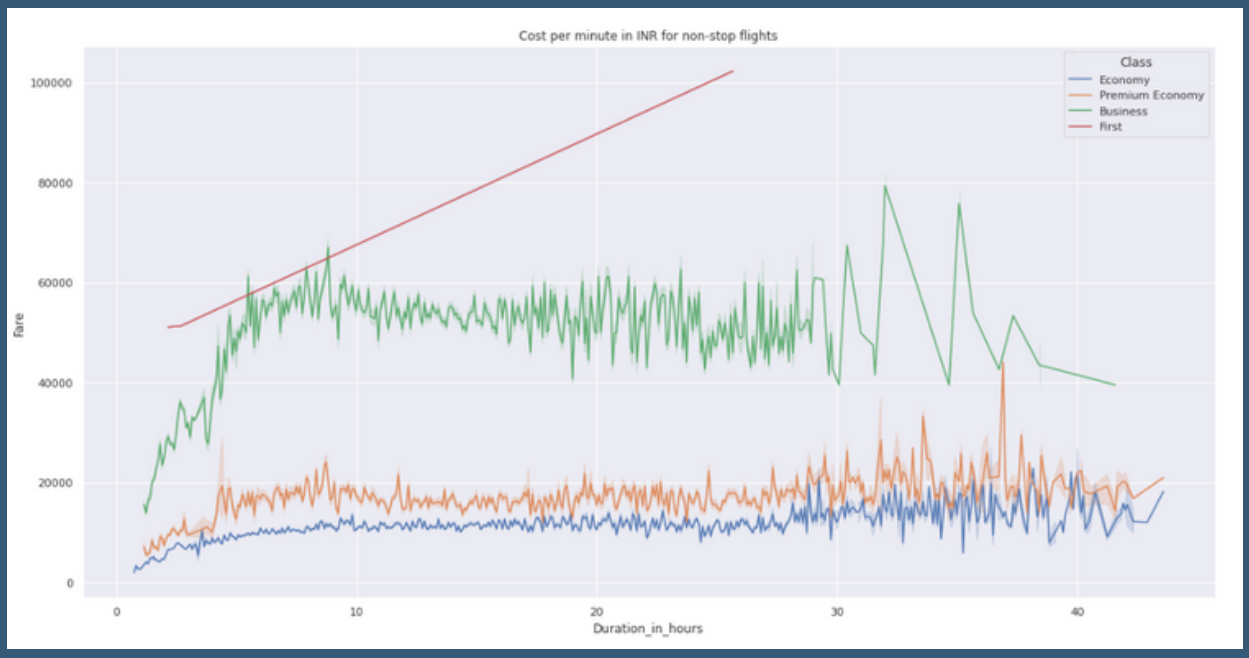
- Extracting insights from data



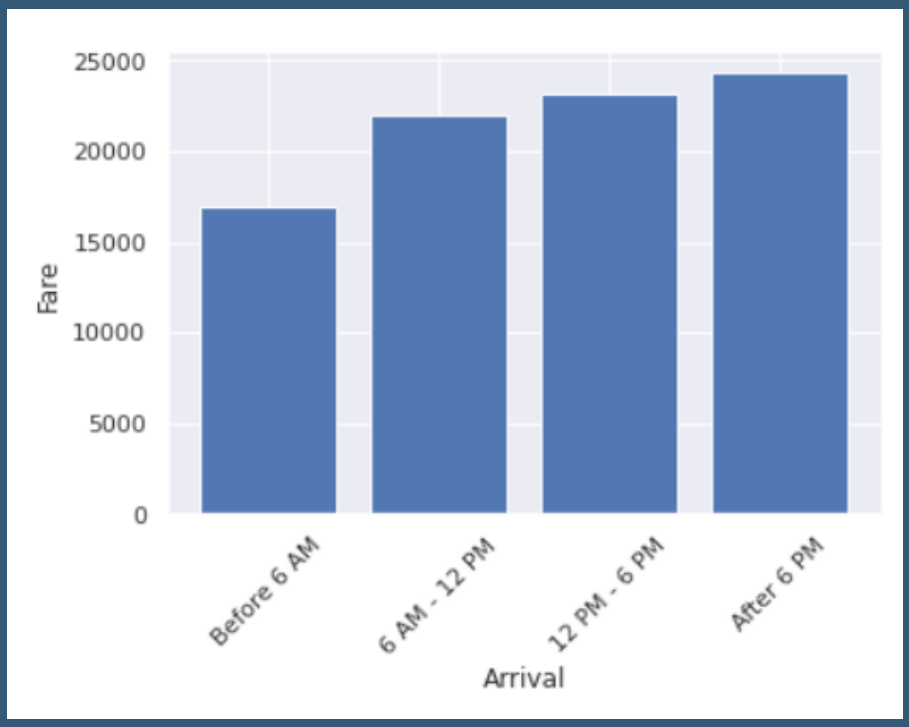
- There is a slight decrease in fares as the number of days left to the journey increases. The fare is highest when there is only one day left for the journey, and it decreases gradually as the days left increase. However, this trend is not linear, and there are some fluctuations in the fare values for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



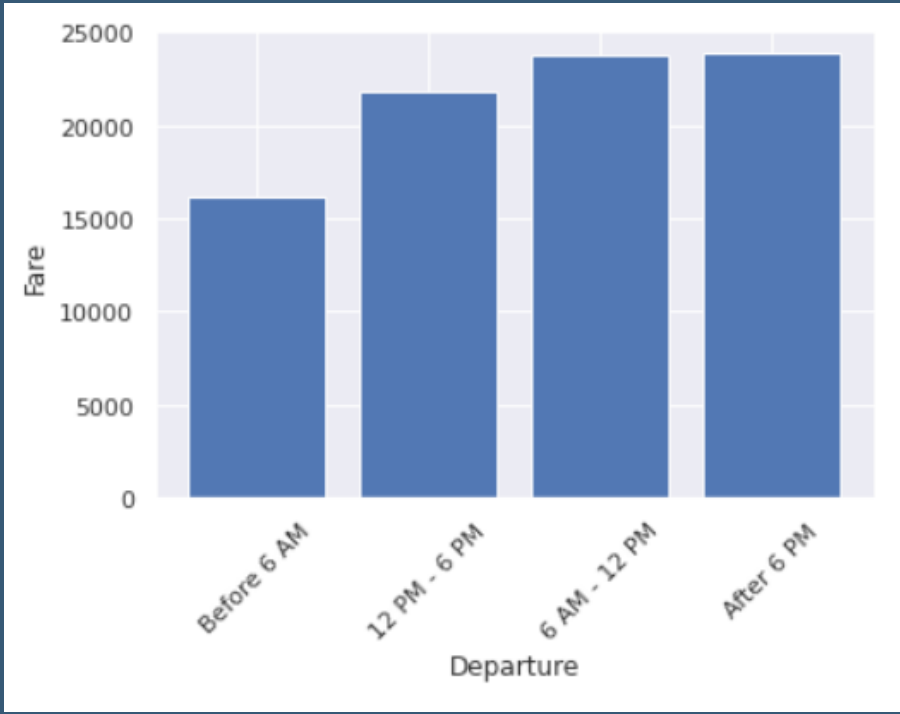
- changing the day not affect average price of the flights except for first category



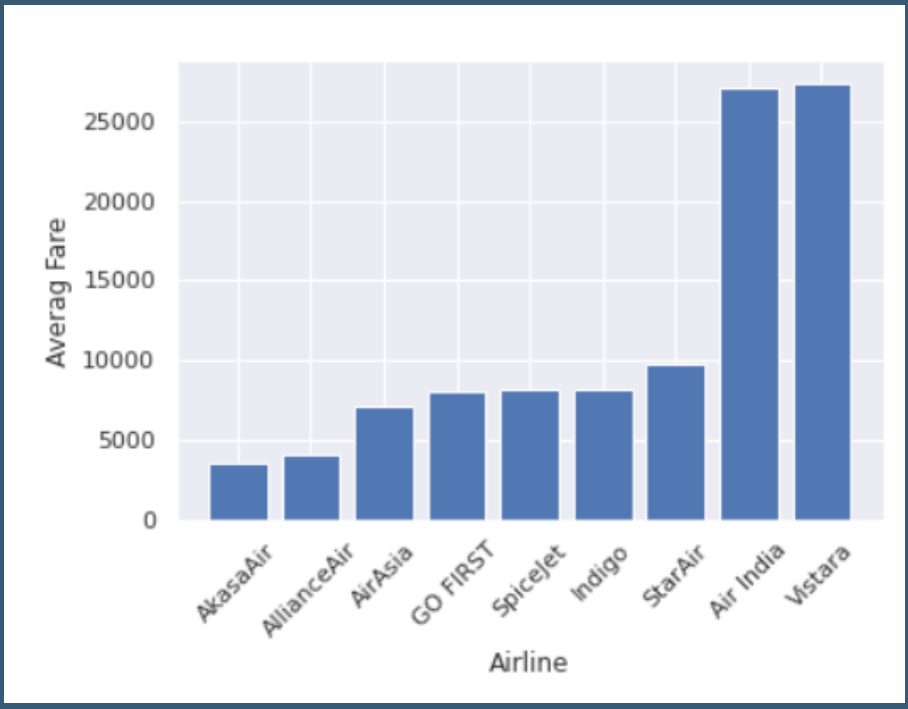
There is a increase in fares as the duration in hour increase for first class. but in other clasess there is high fluctuations in the fare values but not great sign of increase



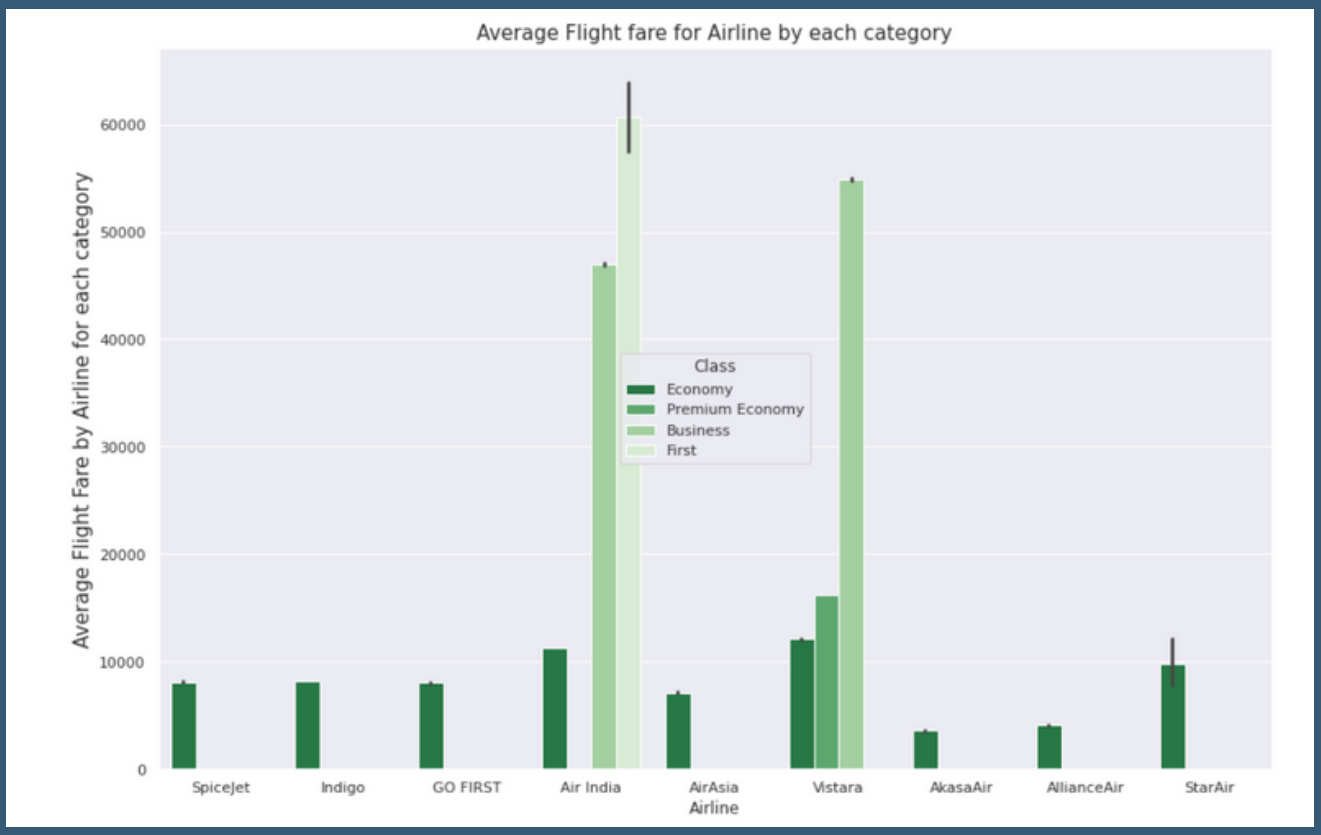
It can be concluded that flight prices vary depending on the arrival time of the flights. The highest fares are observed for flights arriving in the evening (after 6 PM), while the cheapest fares are observed for flights arriving before 6 AM



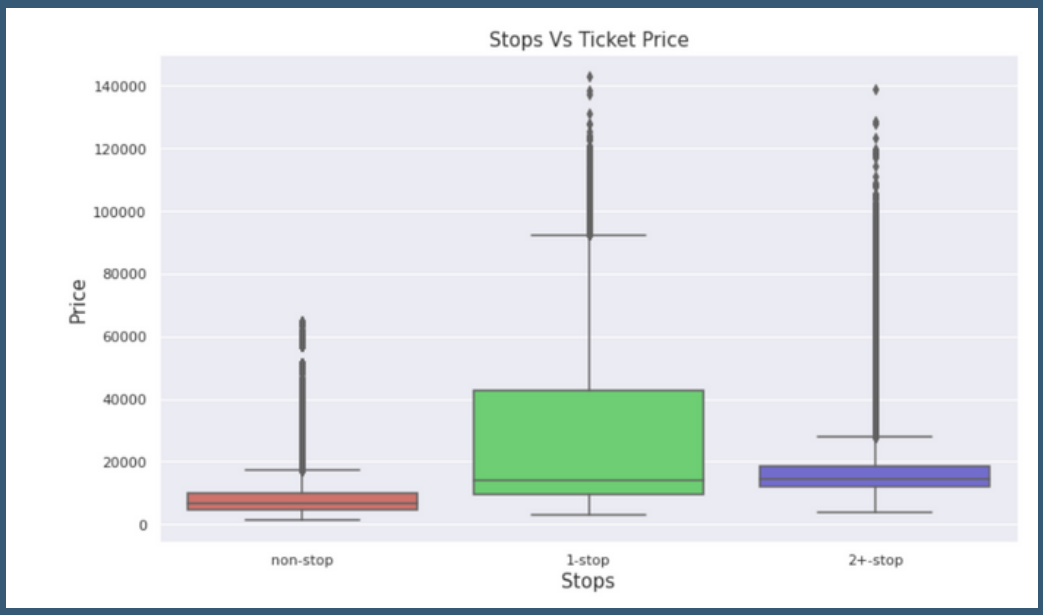
It can be concluded that flight prices vary depending on the Departure time of the flights. The highest fares are observed for flights departure in the evening (after 6 PM), while the cheapest fares are observed for flights departure before 6 AM



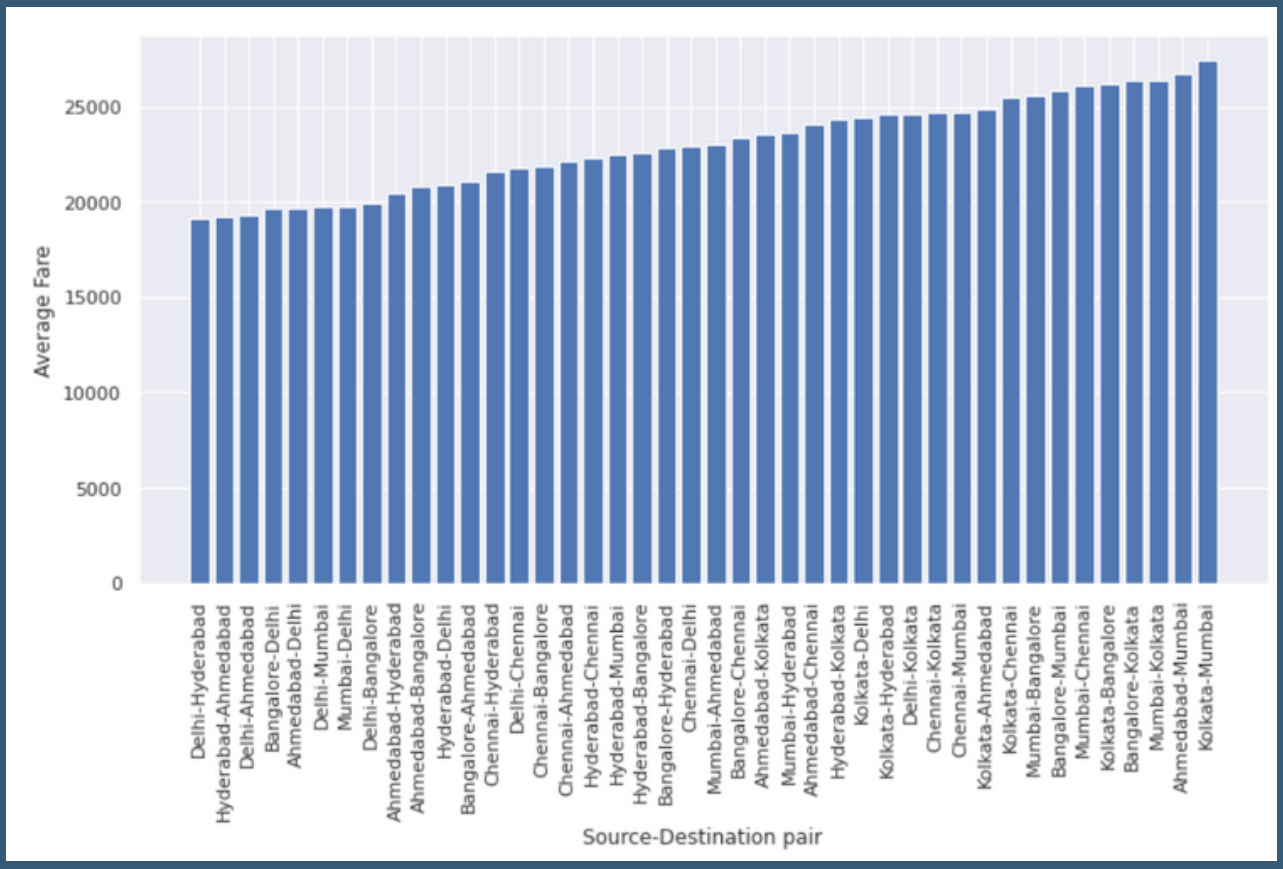
Airline (Air India & Vistara) has the highest average Price while AkasaAir has the cheapest price



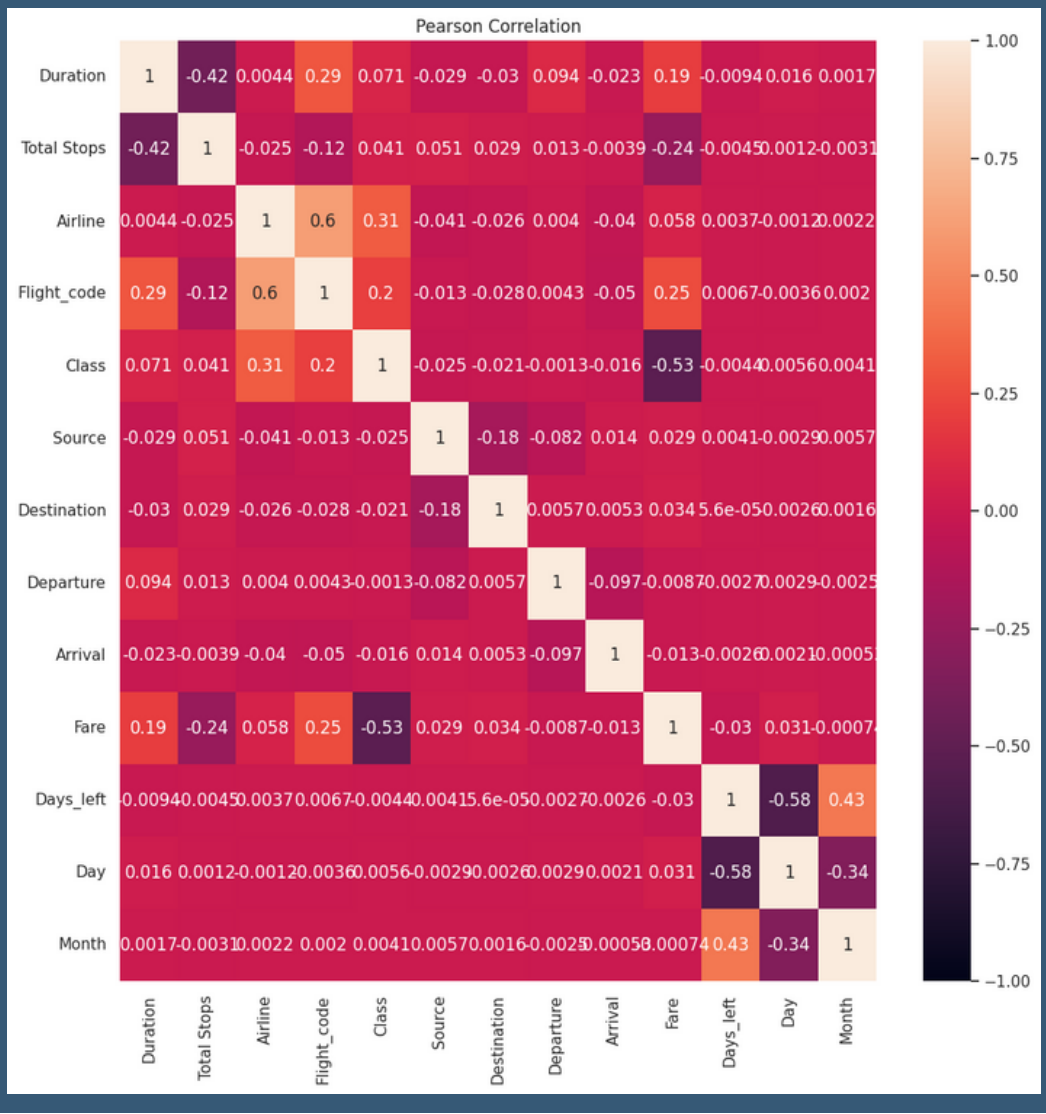
the reason that Air India gets the highest average price because it has first and business classes and Vistara has Pre Economy which is with a high price and for the economy flights they have a slightly higher price and AkasaAir has the lower price



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



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.



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

- Results and Evaluation on train data
    - Sklearn

	Root Mean Square Error	R-squared score
LinearRegression	15096.02463103469	0.4509515602023658
RandomForest Regressor	1480.4535037662426	0.9947195003938359
KNN	373.4125714721	0.999664058

- Results and Evaluation on train data
  - Pyspark

	Root Mean Square Error	R-squared score
DecisionTree Regressor	7176.008040	0.876029
RandomForest Regressor	7842.583577	0.851928
GBTRegressor	6536.481274	0.851928
LinearRegression	15105.226658	0.450701

- 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
KNN	6998.663054	0.8827977

  - 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

- 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.

• Unsuccessful Trials

- We tried to use MRJOB -which runs on Hadoop cluster- python library to implement the mapper-reducer part.

1. First, we Implemented the mapper and reducer on the test data and, we ended up with 32 output files from the reducers with the root-mean error of each data point.
2. Second, we Implemented the mapper and reducer on the train data and split the data into (label: Target, features: x), and define the calculation of the gradient in the reducer. Then we add the final stage for Combining the weights. But, it fails in the final stage and couldn't receive the correct weights.

- We tried to distribute spark with docker; using Bitnami docker image and with a docker-compose file we up master and 2 worker containers. we tried to connect the driver (python code in the local machine) with the master but the IPs configurations were wrong.