**MIDS DS200 Project 2**

**Flight Price Analysis**

**Rohan Krishnamurthi, Stephen Rasch, Thomas Young**

**July 2024**

Primary Dataset:

<https://www.kaggle.com/datasets/muhammadbinimran/flight-price-prediction/data>

Github repository: <https://github.com/rohankrishnam223/Project2_Krishnamurthi_Rasch_Young>

**Introduction**

On average, there are over 100,000 commercial flights each day. With so many different options, it can be extremely difficult to book with certainty, ensuring you’ve selected the best flight for your needs at the right price. The goal of this analysis is to uncover meaningful relationships between various flight attributes to better support airline travelers in the Indian market. Our analysis is divided into three sections, addressing the following questions:

1. **Market Analysis**: What does the Indian airline market look like? What are the routes? Which airlines serve those routes?
2. **Price Analysis**: How do different factors influence flight price – airline type, day of the week, number of stops, duration and time of day?
3. **Value Optimization**: For different consumer types (business or leisure), which routes on which airlines offer the best “value”?

Our analysis assumes the following:

1. Due to airline dynamic pricing models, flight prices typically fluctuate based on the booking window. Since the actual booking dates are unavailable, this will not be factored into our analysis.
2. We do not have a full breakdown of every route serviced by the provided airline carriers. Because of this, we'll assume that if no flights are shown for an airline on a specific route, that airline does not operate on that route. IE: if IndiGo does not have data on a flight from BLR to DEL, we assume that IndiGo does not service this route
3. For the purposes of this analysis, we will assume that our data is representative of the full Indian airline market. In reality, we know that this is likely not the case.

**Data**

The Kaggle Dataset contains flight information for the country of India in the year 2019 (March - June). It contains 10,683 rows and 11 columns.

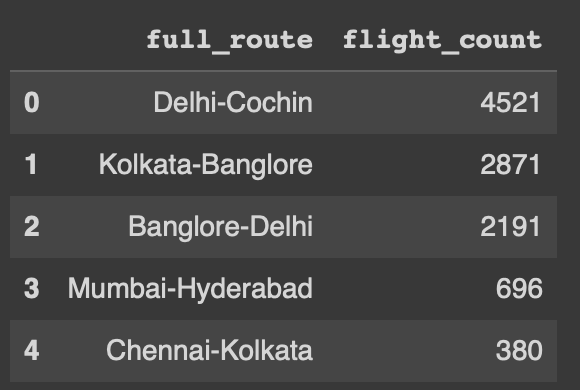
| **Variable Name** | **Variable Type** | **Description** | **Example Value** |
| --- | --- | --- | --- |
| Airline | String | Airline company | Air India |
| Date\_of\_Journey | Date time | Date of flight | 01-05-2019 |
| Source | String | Origin city | Banglore |
| Destination | String | Destination city | New Delhi |
| Route | String | Travel route | BLR ? DEL |
| Dep\_Time | String | Departure time | 22:00 |
| Arrival\_Time | String | Arrival time | 10:00 |
| Duration | String -> int | Flight duration | 2h 50m -> 170 mins |
| Total\_Stops | String -> int | Stops on flight | 2 stops -> 2 |
| Additional\_info | String | Extra information if applicable | No info |
| Price | Int | Flight price | 3897 |

High-Level Observations from EDA

* Duplicate Rows: Two duplicate rows were identified and dropped
* Date Range: We have 3 months worth of data (March 2019 - June 2019)
* Null values: The two duplicate values mentioned above are NULLs, occurring in the total stops and route attributes

**Market Analysis**

**Which routes operate in India and what are their attributes?**

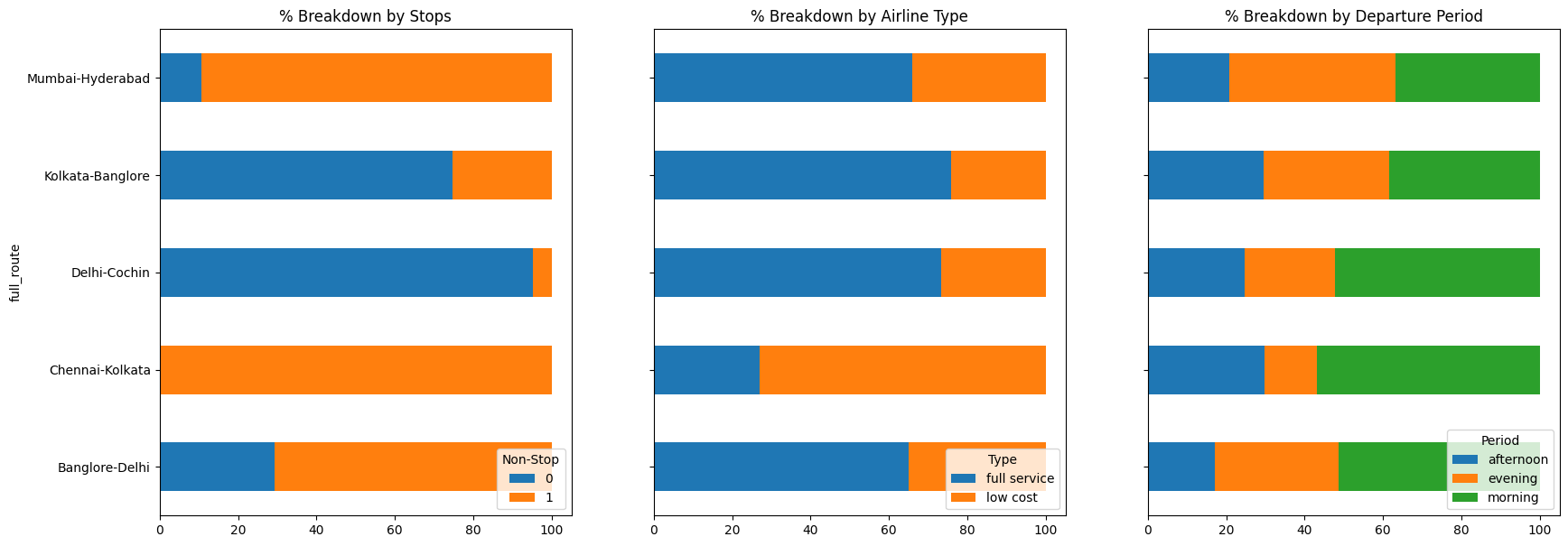


To start our analysis, we created a new variable called ‘full\_route’ which concatenated the source and destination for each flight into a single variable. When doing so, we noticed that Delhi and New Delhi were used interchangeably by different airlines, even though they both mapped to the same airport code, DEL. For consistency, we standardized all instances to ‘Delhi’.

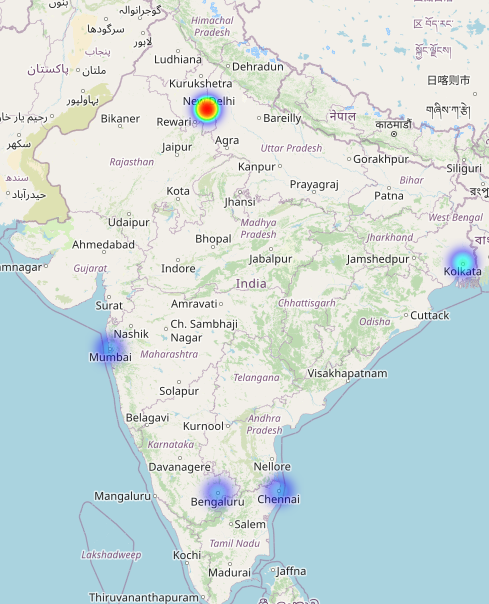
As shown in the table above, there are 5 unique full routes between 8 different cities. Delhi-Cochin is the most frequent route with 4,521 flights.

We identified three different variables that could be used to better explain the different attributes of each route: number of stops, departure period and airline type.

* Number of Stops: We transformed the Total\_Stops feature from its original categorical format into an integer format. Additionally, we introduced a separate binary feature to indicate whether a flight is non-stop or contains stops.
* Departure Period: Within the data, departure time is represented as a time in military format. To make our analysis easier, we developed a categorical variable that classifies departure times into three periods: morning, afternoon, and evening.
* Airline Type: We created features for Airline Type and Airline Primary Hub using information sourced from Wikipedia. We compiled this information into a dictionary, which was then used to map these attributes to the corresponding airlines in our dataset. Categories created are full-service and low-cost airlines.

****

The graphs above show a breakdown of each route attribute as a percentage of total flights. For example, Delhi-Cochin is predominantly a non-stop route that full service airlines serve. Additionally, the majority of these flights depart in the morning time (between 4 am - 11 am). While our data did not include a variable for traveler type, we could hypothesize that this route serves business travelers who are looking for a route with fewer stops and served by a full service airline that leaves in the morning.

This heat map of India displays the distribution of flights from each city in the dataset. The map is for source flights, meaning the values shown below are the flights which have the associated city as the origin of the given flight. New Delhi can clearly be seen as the most common source airport city. 

See the counts of each city as a source below:

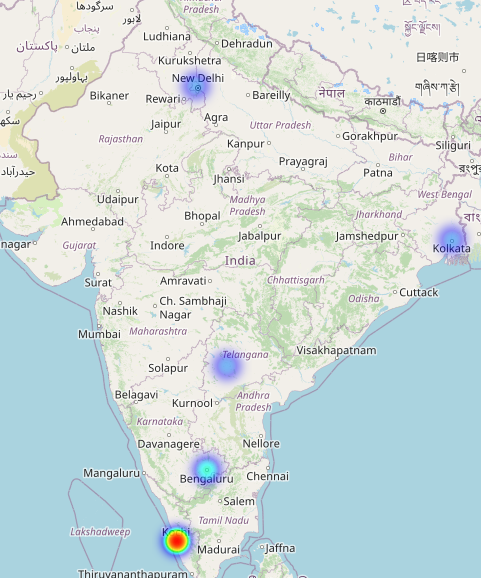
Bangalore (Bengaluru): 2197 flights

Chennai: 381 flights

Kolkata: 2871 flights

Mumbai: 697 flights

New Delhi (Delhi): 4537 flights



The complementary heat map here displays the distribution of flights from the Indian cities as destination cities. The values shown below accompanied by the map display the breakdown of how many flights arrive in each city, as their final destination. Kochi, more commonly known as Cochin, can be seen as the most popular destination city.

Bangalore (Bengaluru): 2871

Cochin (Kochi): 4537

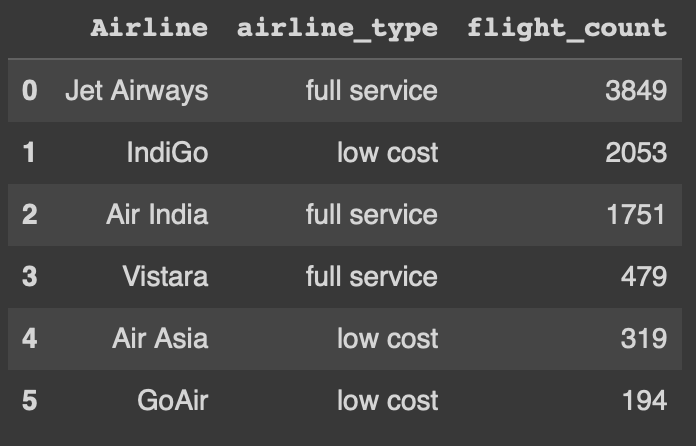
Hyderabad: 697

- *Central city near Telangana*

Kolkata: 381

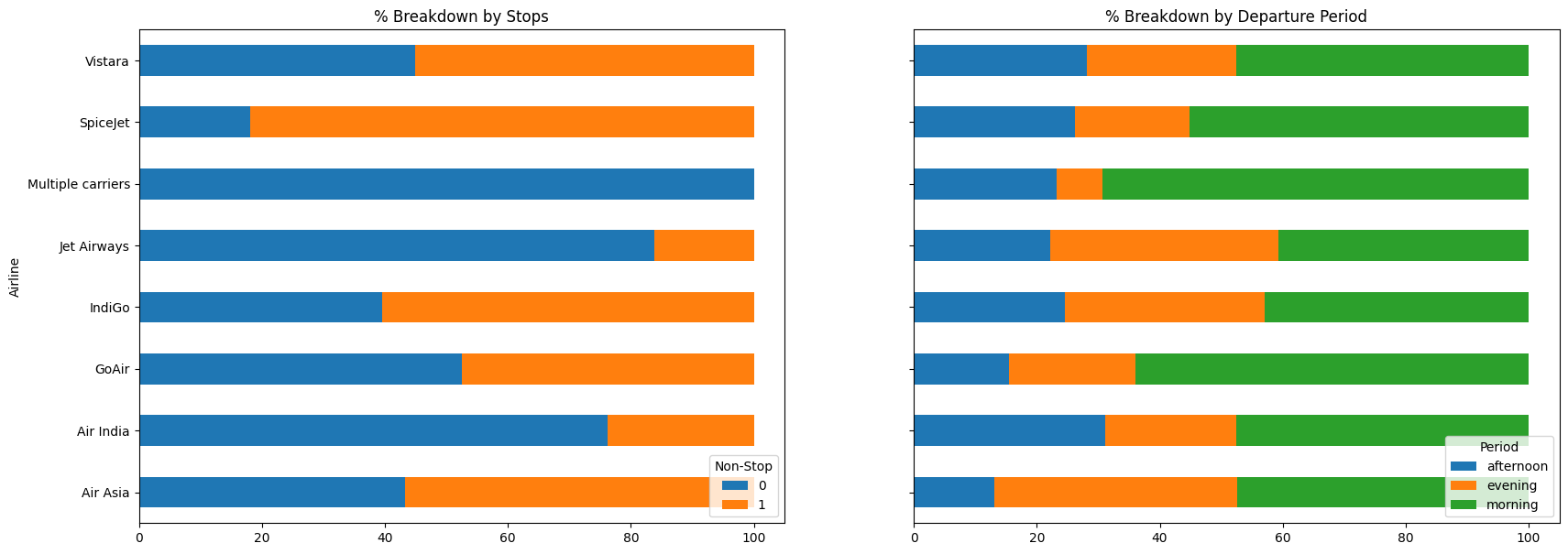
New Delhi (Delhi): 2197

**Which airlines operate in India and what are their attributes?**



In addition to understanding which routes operate in India, we also wanted to figure out which airlines serve these routes. During our EDA, we identified multiple erroneous and or non-significant values for Airline. For example, three rows in the data had Airline values of ‘Jet Airways Business.’ While an attribute for Business ticket would have been helpful in our data, due to the infrequency of these values, we decided to drop all rows with these Airline values. As shown in the table above, Jet Airways, a full service carrier, was the largest airline with 3,849 flights.

Similar to our route analysis, we also wanted to see what made each of these airlines different. For this analysis, we leveraged the non-stop and departure period variables we created during our route analysis.

****

The graph above shows a breakdown of each route attribute as a percentage of total flights. For example, the majority of Jet Airways flights are non-stop flights that leave in the morning and afternoon.

**Price Analysis**

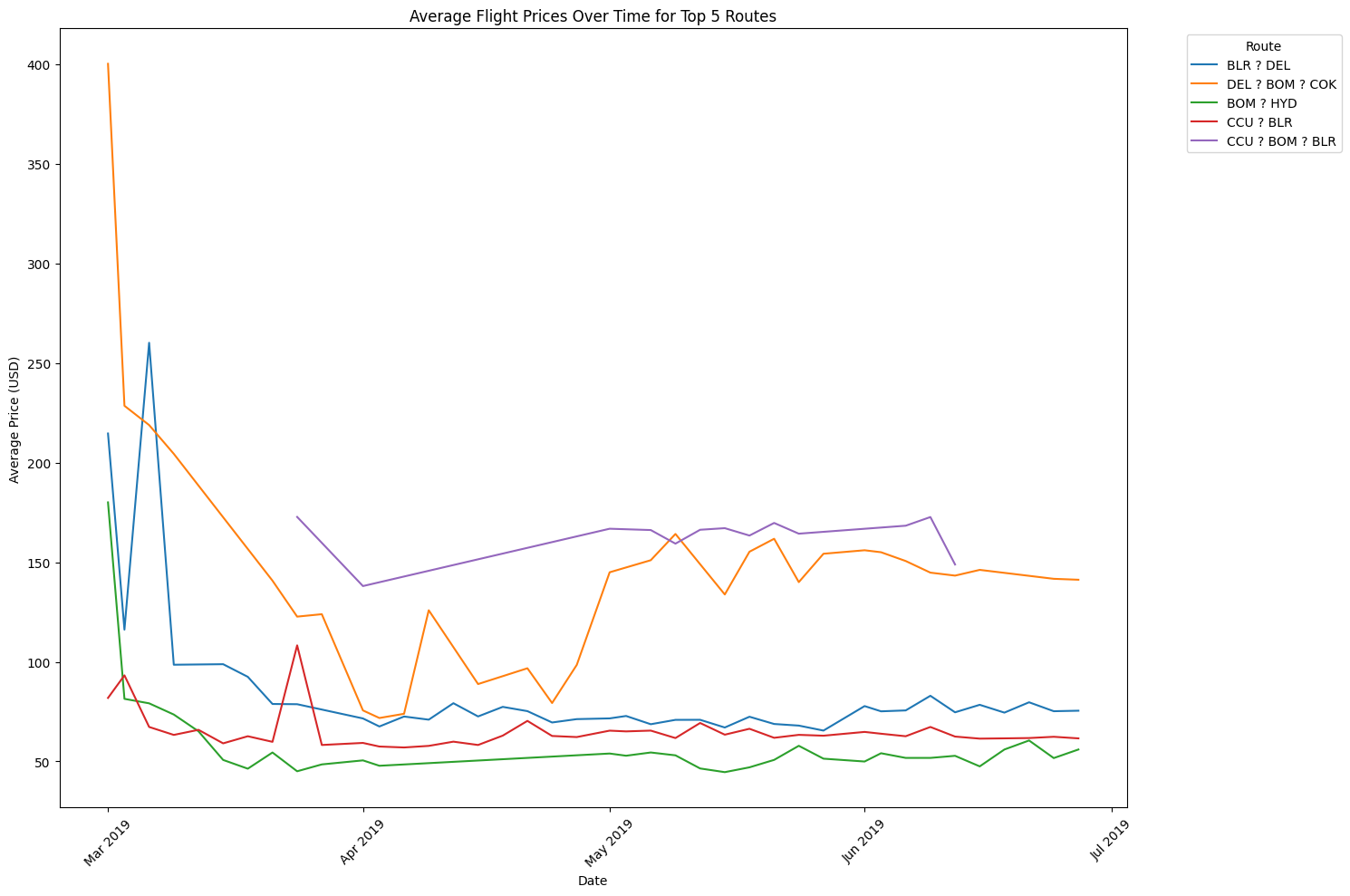
**How do different factors influence flight price: airline type, date, number of stops, duration and time of day?**

To begin our price analysis we ran some data preprocessing explained below:

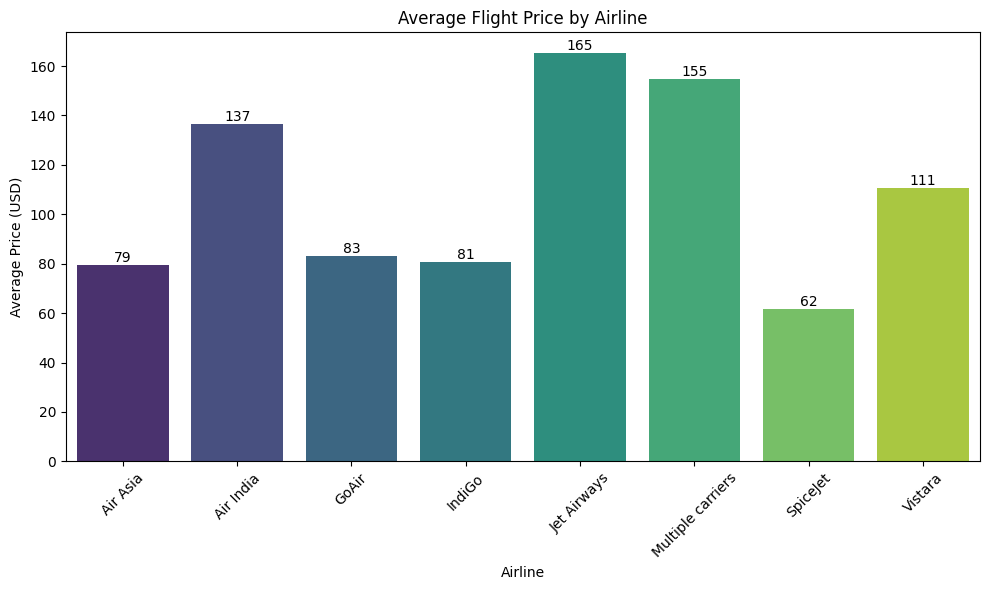
We converted the Price from Indian Rupees (INR) to U.S Dollars (USD) using a conversion rate from 2019 to keep data consistent. The conversion rate at this time was 1 USD = 70.39 INR

Hierarchical Data: We decomposed the Departure Date into a hierarchical structure of month, day, and year. Breaking the departure date into its component parts allows for more granular analysis of how seasonal trends and specific days of the week influence flight prices.

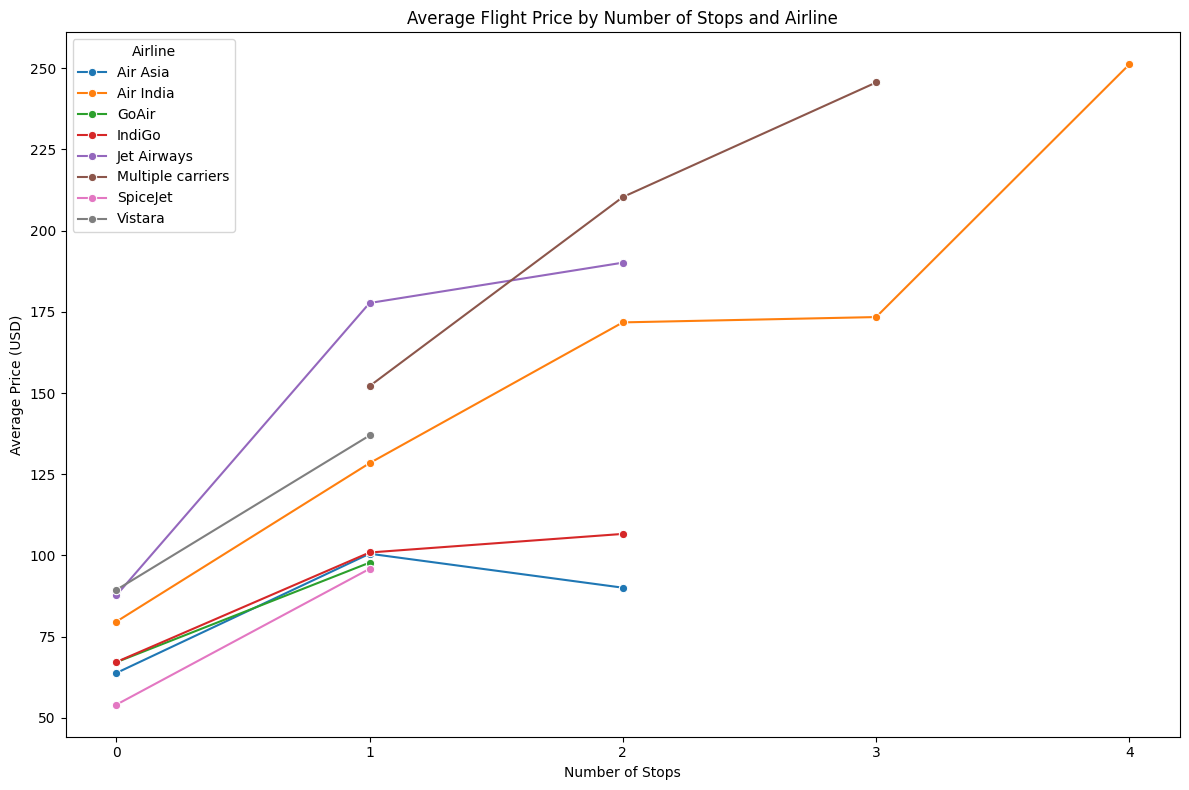
Duration Format: We determined that Flight\_Duration is not in integer format and includes "h" and "m" for hours and minutes. We wrote a function to standardize all values to minutes as an integer variable, making it easier to perform numerical analysis and comparisons.



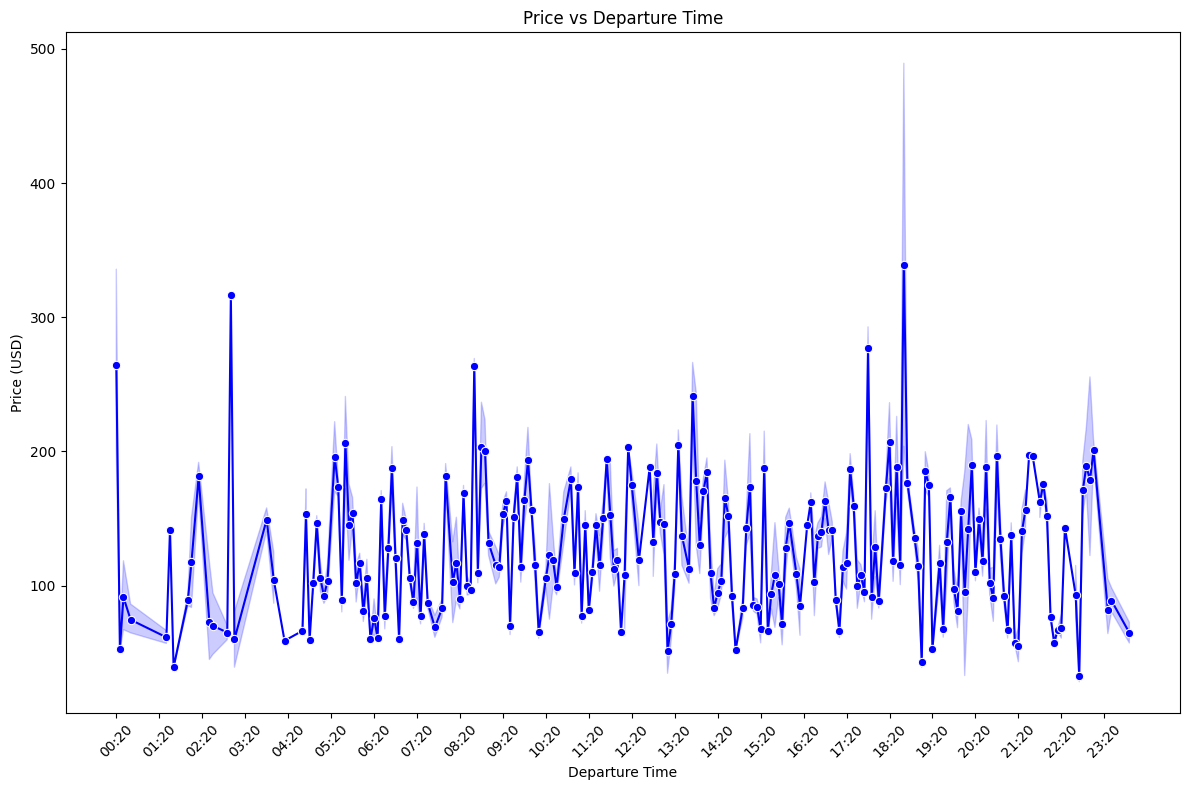
The line graph (left) emphasizes some important metrics when considering the price of flights over time. All of India celebrates Holi in the month of March, meeting our expectation that during holidays, flight prices will increase. There is a big drop off entering April, and prices remain consistent for the remaining months. Over a 3 month period it is difficult to identify patterns and trends to predict future prices, but the month of March is clearly an outlier given the holiday season.



Analyzing the price of flights by airline companies provides insight into which carriers provide the cheapest and most expensive flights. The average cost of all flights is $109.12, meaning any carrier which averages a cost below this can be considered a price-friendly airline. Air Asia, GoAir, IndiGo, and Spicejet fall under this category and are therefore price-friendly carriers. These results match the outside research we performed to create our airline type feature (full service and low-cost carriers).



The line graph here displays the upward trend of flight price as the number of stops on a given flight increases. Air asia (blue line) presents itself as an outlier in this visual, as the average price drops from 1 to 2 stops. We can assume that more stops also indicates a longer flight journey, so this logically makes sense as a longer journey with more stops will typically cost more than a non-stop journey. Therefore, we can say with confidence that number of stops is an influential factor of flight price. Future analysis would include breaking down individual routes by stops and non-stop flights, to understand how price differs on the same route.



There is no discernable pattern shown by the line plot of departure time vs price. There are some times of the day that are clearly more expensive to fly than others, such as the huge spikes around 18:20 and 02:40. There are 2 local maximums around 17:20 - 18:20 which makes sense as this is the end of a typical work day, when fliers may be leaving the office to catch flights home or elsewhere. As expected, unorthodox departure times are the cheapest times to fly, such as the global minimum at 22:40 or the second minimum at 01:40.

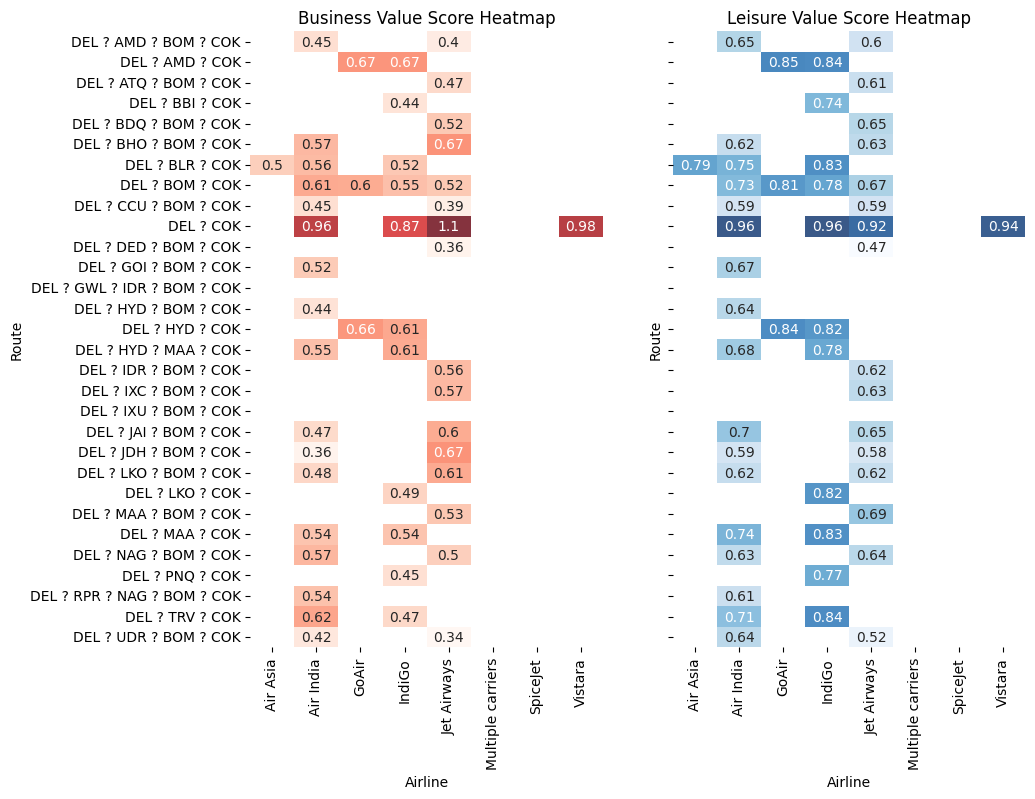
From these analyses we can calculate the cheapest flights possible, based on the influential factors plotted above.

| **Attribute** | **Value** |
| --- | --- |
| Cheapest time to fly | 22:45 |
| Cheapest airline | SpiceJet |
| Cheapest duration | 1h 25 min |
| Cheapest number of stops | Non-stop (0) |
| Cheapest date to fly | 03/21/2019 |

**Value Optimization**

**How can a flight's value be quantified for different consumers?**

As seen in our analysis, things like number of stops, airline type and departure time play a key role in a flight’s price. But what makes a flight valuable to a consumer? It could be the cheapest option, the quickest, or a balance of both. As part of our analysis, we developed a weighted scoring model that combines multiple insights to recommend the best flight (by route and airline) for a given consumer, whether they are a business traveler or a leisure traveler. This model takes into account the following variables:

* Flight Duration, Non-Stop (binary), Airline Type (full service or low cost), Ticket Price, Departure Period (morning, afternoon, evening)

The heatmap to the right shows the results for passengers looking to travel from Delhi to Cochin. For both business and leisure travelers, the non-stop flight emerged as the most valuable option. However, the recommended airline varied based on the type of traveler.

Business Travelers: Jet Airways is considered the best option. As a full-service airline, Jet Airways offers flights on this route in the morning, providing a good flying experience and ensuring business passengers reach their destination promptly.

Leisure Travelers: Air India and IndiGo are tied as the best options. Both are low-cost carriers that offer cheaper non-stop flights on this route. Although these airlines may not provide the same amenities as a full-service carrier, they are suitable for leisure passengers who are budget-conscious.

**Conclusion**

This analysis aims to shed light on the Indian airline market by exploring key attributes, uncovering trends, and more affecting flight pricing and value. The Indian airline market is diverse, with numerous airlines operating a wide range of routes. Our findings indicate that major airlines dominate popular routes, but there is a large presence of regional carriers serving these same routes in a more cost-effective manner. Understanding this breakdown facilitates identifying the best options for each individual flier, as everyone has varying preferences. Our exploration of pricing factors revealed numerous contributors to the price of a given flight. Flights operated by major airline carriers tend to be more expensive, and prices vary over time at all levels: time of day, day of week, month, and holiday vs. non-holiday period. When considering value for different types of travelers, business travelers prioritize direct flights and shorter durations, often opting for premium services despite higher costs. In contrast, leisure travelers may benefit more from considering routes with layovers if it translates to significant cost savings. Tailoring travel decisions based on these preferences can maximize the overall value of a trip. The insights from our analysis are thorough and provide valuable metrics for fliers to maximize value on their journey while not sacrificing a huge cost. However, we must acknowledge the limitations inherent in our data. 3 months of data cannot be considered a complete assessment of the Indian airline market, especially given how flight prices vary with seasons. Future research incorporating a larger dataset and more granular market data would be a great supplement to the research done here. Nevertheless, the analysis here provides a great foundation to make more informed decisions in the complex realm of Indian air travel.