PREDICTING FLIGHT PRICES

Jake Fox

4/20/2023

OUTLINE

- Summary
- Methodology
- Results
- Discussion
- Conclusion
- Appendix

SUMMARY

- Playing the role of an analyst for a new, luxury Indian airline

- Tasked with using market data to help correctly price our flights compared to other airlines.

- The data was downloaded via .csv from Kaggle.

- The user from Kaggle received this data through a website called "Ease My Trip" that helps customers plan their vacations and getaways (~300,000 flight instances).

METHODOLOGY

- Linear regression model that takes 2 predictors to accurately predict flight price based on:
  - Whether a flight is business class or not
  - Total flight time

- Departure time of day, airline, and From-To locations, play a role in pricing and will be illustrated as well.

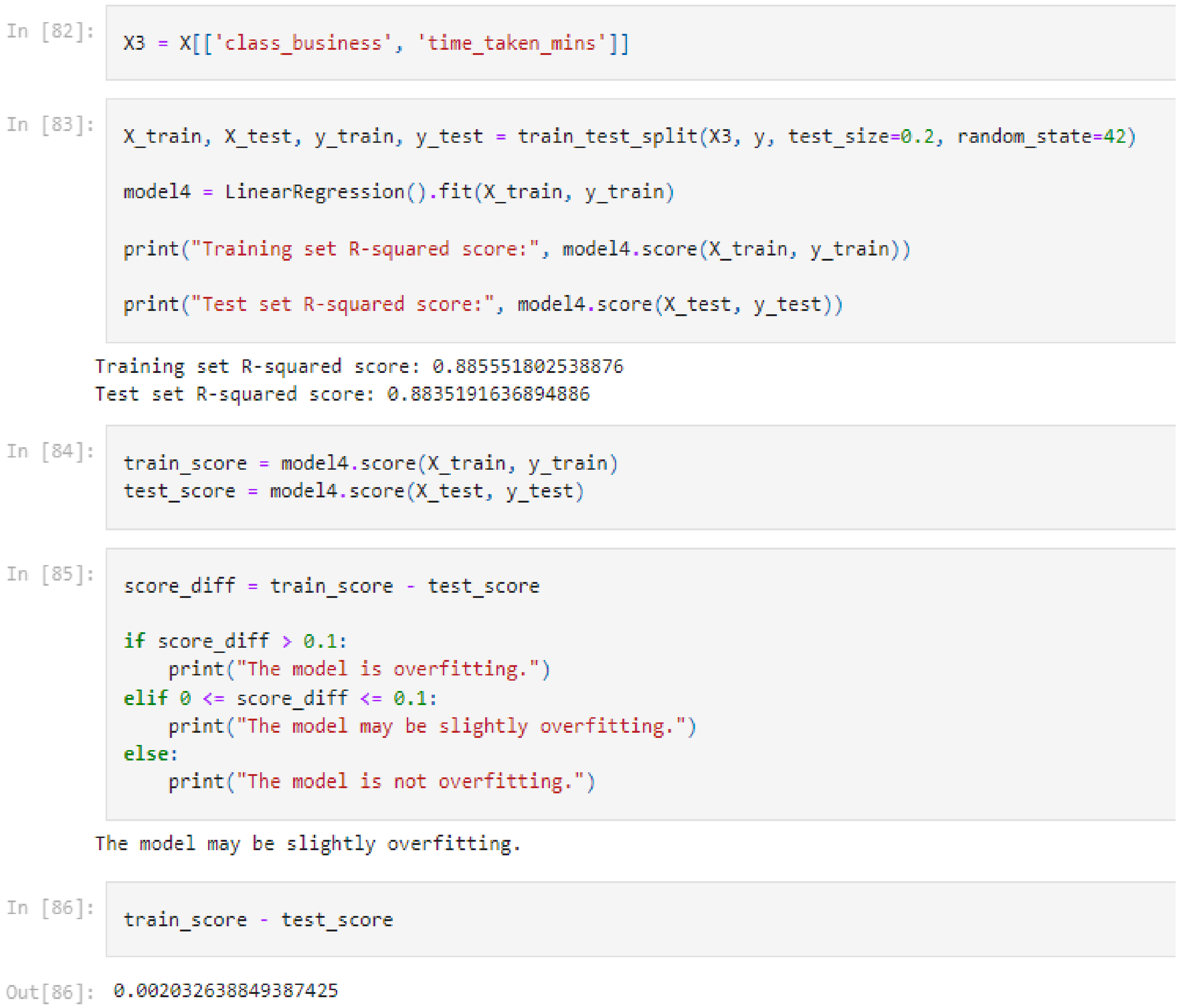
- With this model, as well as supporting data visualizations, we will be able to understand where we stand in the Indian airline market and what kind of airline we'd like to be represented as

RESULTS

Optimal linear regression model: about 88% confident that it will provide an accurate prediction.

Given all other predictors of price, this model was optimal because of its simplicity and high accuracy.

** Slight overfit: Importance of having a generalizable model that performs well on unseen data



```
In [82]:  X3 = X[['class_business', 'time_taken_mins']]

In [83]:  X_train, X_test, y_train, y_test = train_test_split(X3, y, test_size=0.2, random_state=42)

          model4 = LinearRegression().fit(X_train, y_train)

          print("Training set R-squared score:", model4.score(X_train, y_train))

          print("Test set R-squared score:", model4.score(X_test, y_test))

         Training set R-squared score: 0.885551802538876
         Test set R-squared score: 0.8835191636894886

In [84]:  train_score = model4.score(X_train, y_train)
          test_score = model4.score(X_test, y_test)

In [85]:  score_diff = train_score - test_score

          if score_diff > 0.1:
              print("The model is overfitting.")
          elif 0 <= score_diff <= 0.1:
              print("The model may be slightly overfitting.")
          else:
              print("The model is not overfitting.")

         The model may be slightly overfitting.

In [86]:  train_score - test_score

Out[86]: 0.002032638849387425
```

RESULTS (CONT.)

I needed to check how accurately this model was predicting prices, so I created a couple of fake flights to test it.

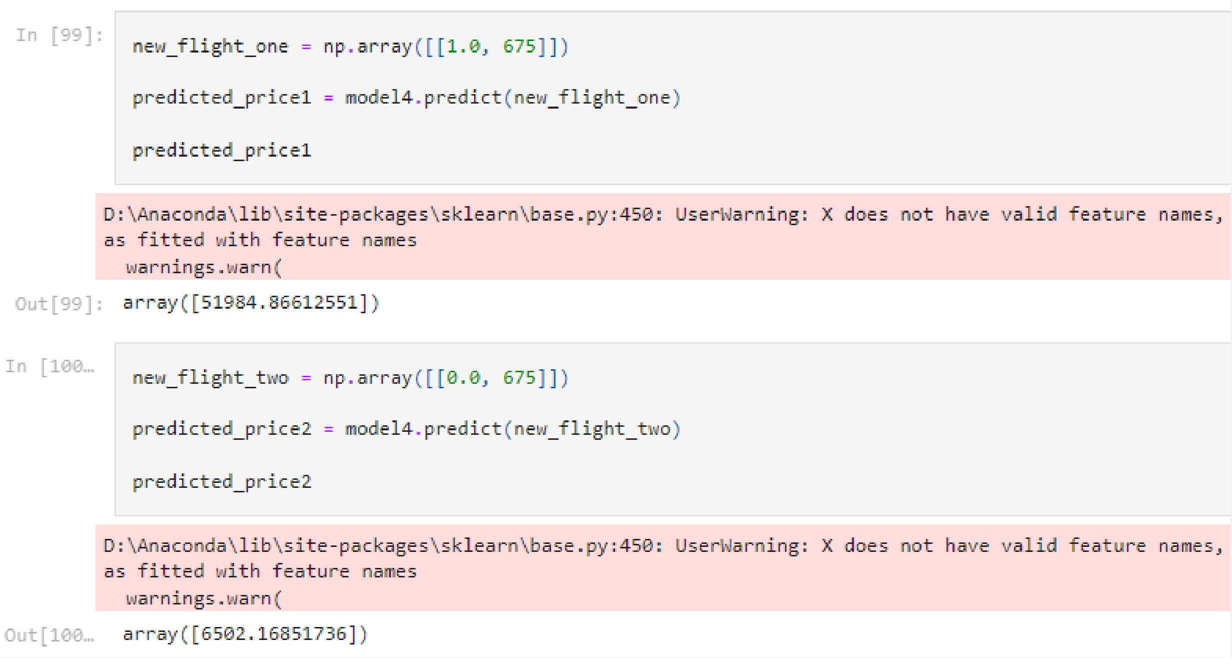
Let's compare predicted prices of business VS. economy flights that are 675 minutes long (11.25 hours) and 1,500 minutes long (25 hours).

Business -> 11.25 hours ~ **$51,985**

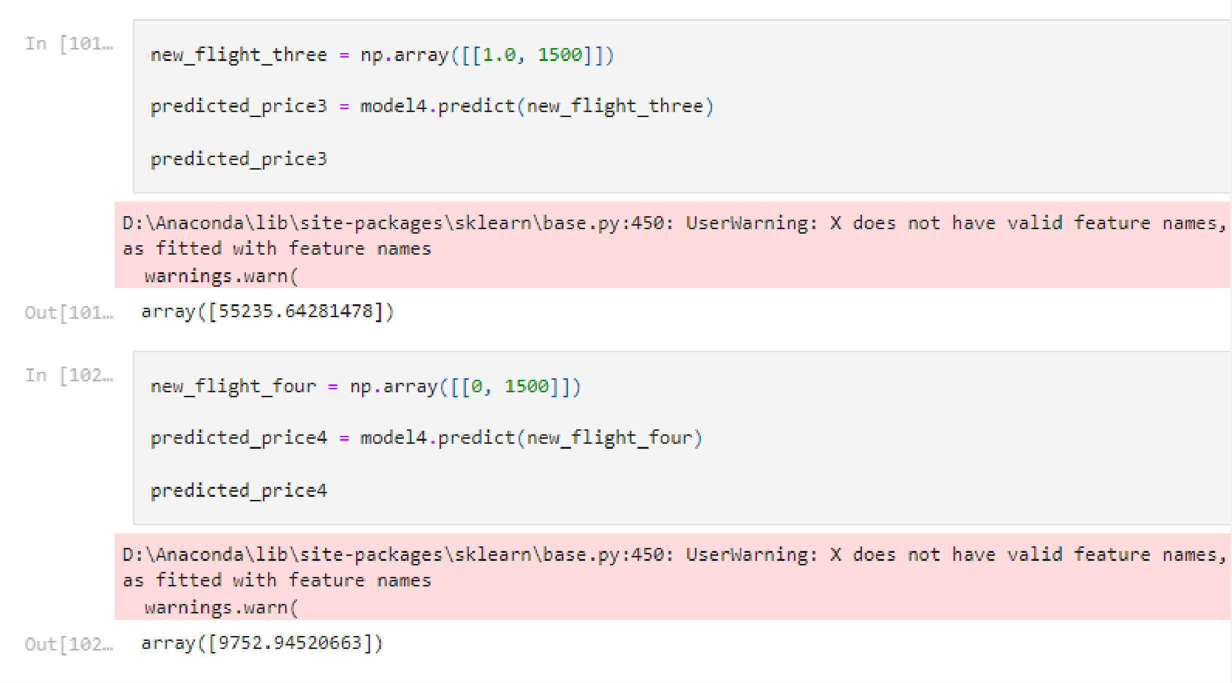
Economy -> 11.25 hours ~ **$6,502**

Business -> 25 hours ~ **$55,236**

Economy -> 25 hours ~ **$9,753**



```
In [99]:  new_flight_one = np.array([[1.0, 675]])

          predicted_price1 = model4.predict(new_flight_one)

          predicted_price1

D:\Anaconda\lib\site-packages\sklearn\base.py:450: UserWarning: X does not have valid feature names,
as fitted with feature names
  warnings.warn(
Out[99]:  array([51984.86612551])

In [100…  new_flight_two = np.array([[0.0, 675]])

          predicted_price2 = model4.predict(new_flight_two)

          predicted_price2

D:\Anaconda\lib\site-packages\sklearn\base.py:450: UserWarning: X does not have valid feature names,
as fitted with feature names
  warnings.warn(
Out[100…  array([6502.16851736])
```



```
In [101…  new_flight_three = np.array([[1.0, 1500]])

          predicted_price3 = model4.predict(new_flight_three)

          predicted_price3

D:\Anaconda\lib\site-packages\sklearn\base.py:450: UserWarning: X does not have valid feature names,
as fitted with feature names
  warnings.warn(
Out[101…  array([55235.64281478])

In [102…  new_flight_four = np.array([[0, 1500]])

          predicted_price4 = model4.predict(new_flight_four)

          predicted_price4

D:\Anaconda\lib\site-packages\sklearn\base.py:450: UserWarning: X does not have valid feature names,
as fitted with feature names
  warnings.warn(
Out[102…  array([9752.94520663])
```

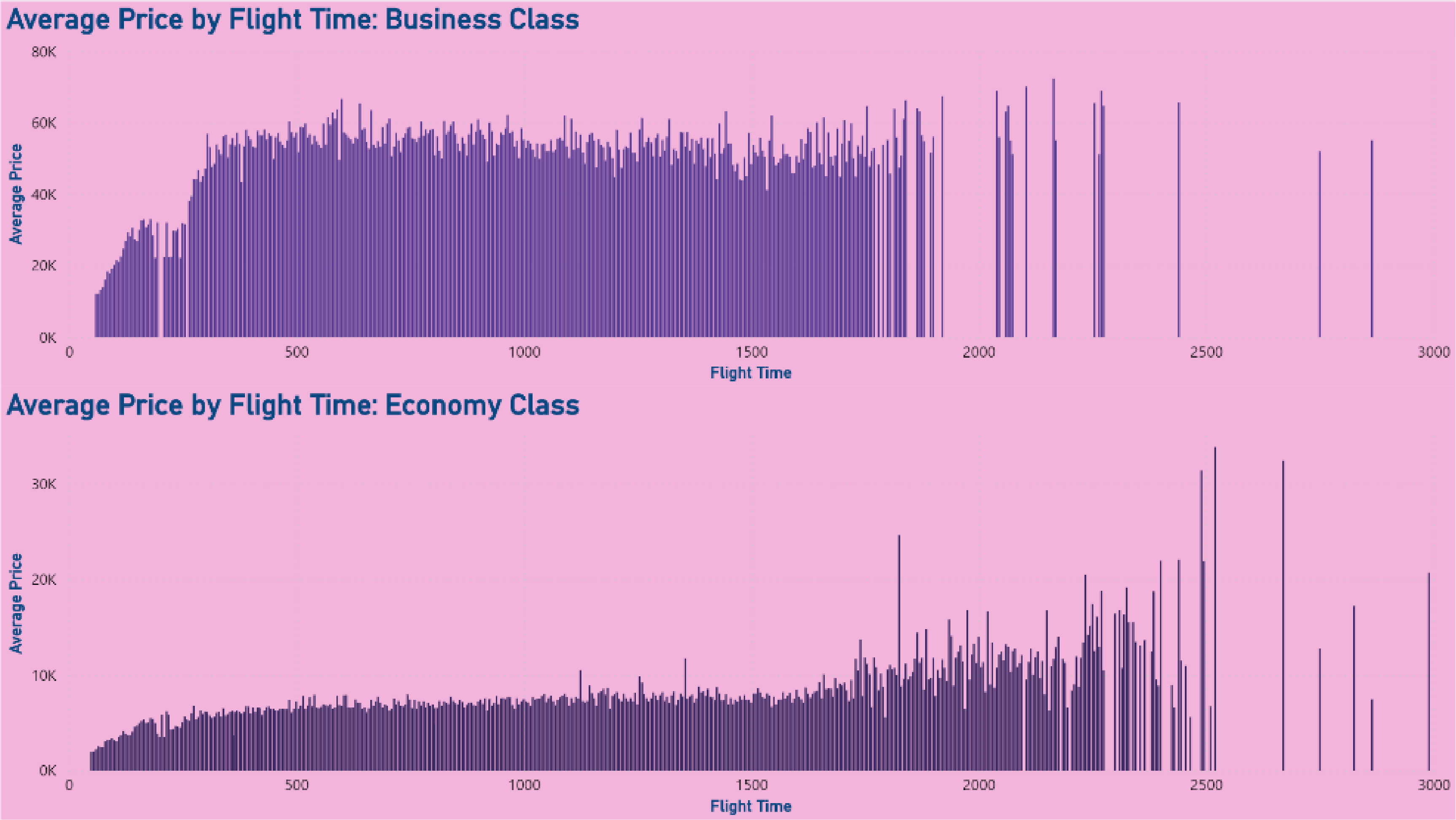
AVERAGE PRICES BY FLIGHT TIME: BUSINESS V. ECONOMY

Using this chart, we can directly see the accuracy of our model given flight time and class:

Business -> (675) 11.25 hours ~ **$51,985**
Economy -> (675) 11.25 hours ~ **$6,502**

Business -> (1500) 25 hours ~ **$55,236**
Economy -> (1500) 25 hours ~ **$9,753**

Compared to total flight time, the class of the ticket is much more prominent in predicting the price of a flight.


Average Price by Flight Time: Business Class

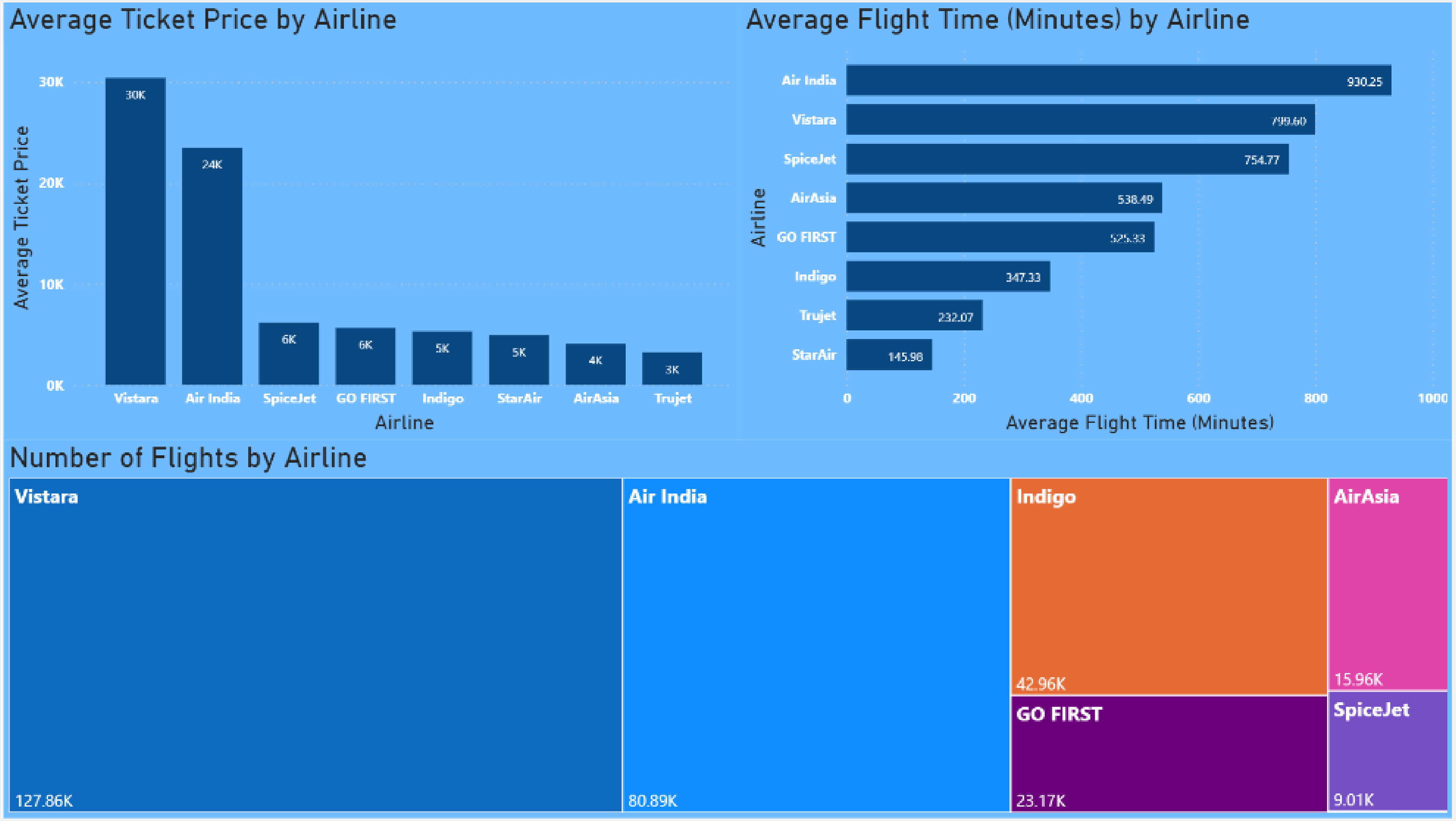
Average Price by Flight Time: Economy Class

RESULTS (CONT.)

Now that the model has been created, we need to start determining our prices based off the type of airline we are.

Given that our goal is to be represented as a high-tier, luxury airline, we must be more specific in comparing higher-tier airlines to our own.

STATS BY AIRLINE



Average Ticket Price by Airline

Average Flight Time (Minutes) by Airline

Number of Flights by Airline

In this dashboard sheet, I notice that airlines Vistara and Air India have the 2 highest average ticket prices, average flight times, and number of total flights.

Because we are trying to be in the luxury airline market, this is a great discovery because it allows us to be pinpoint which airlines we will be competing against the most.

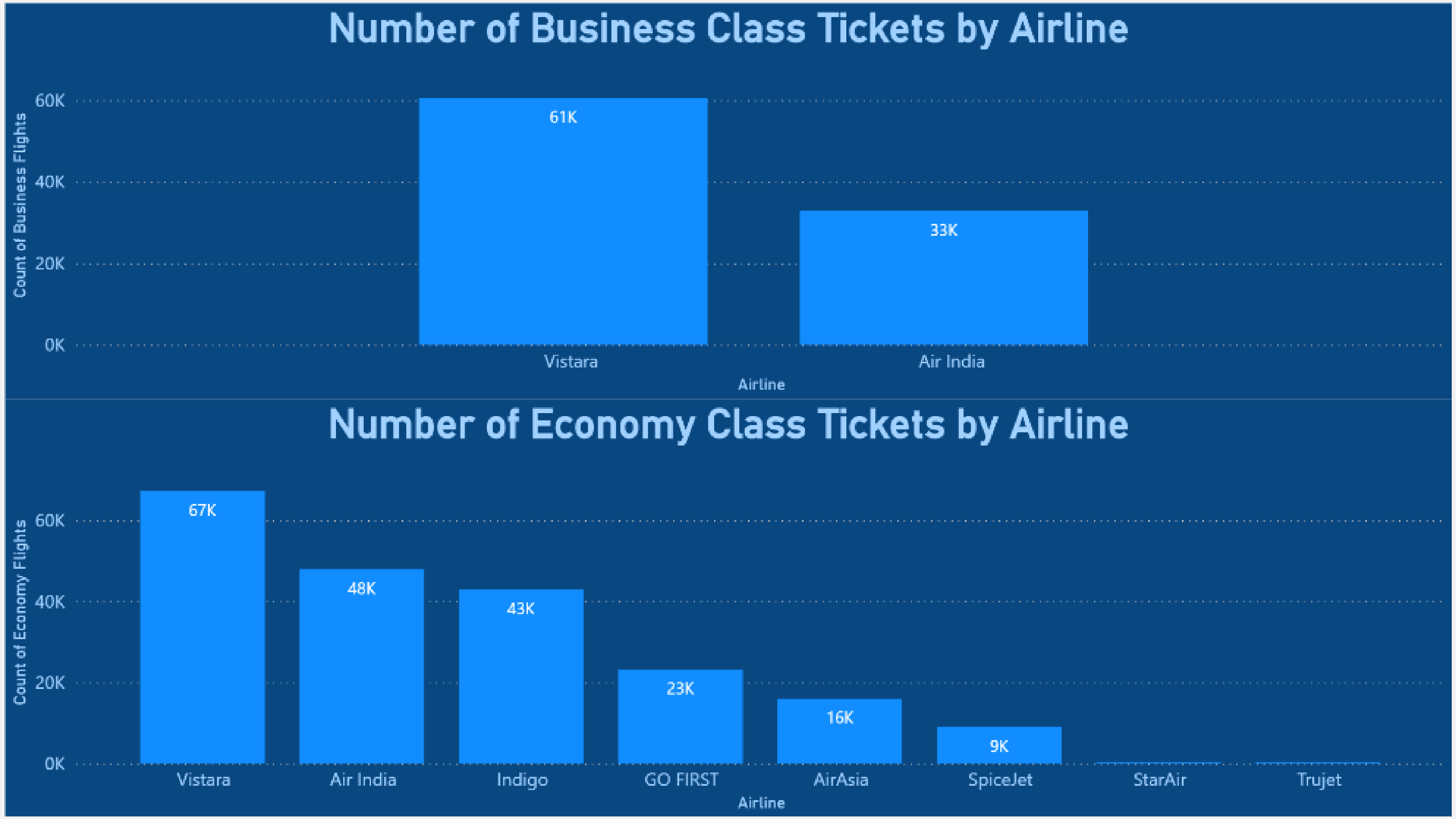
NUMBER OF TICKETS PER FLIGHT CLASS

Here, I noticed that the only two airlines that offer business class flights are Vistara and Air India.

Almost 50% of all Vistara flights and just under 70% of all Air India flights are business class.

This chart explains why they lead the categories in the previous slide.

Thus proving to me that these two airlines are going to be our biggest competitors.



DISCUSSION

Now that we've discovered what type of airline we want to be seen as and we understand our competitors, what are we offering?

In other words, how do we compare to Vistara and Air India?

- What quality of customer service do we provide?
- How is the quality of our food and drink menu?
- Are the plane cabins up to luxury standard compared to other airlines?
- ETC.

With the correct and honest answers to these questions, and more, we will be able to price our tickets comparable to our competitors.

- If we know we can offer more, we can charge more.

- If we know we offer less, we can charge at a lower rate than our competitors.

CONCLUSION

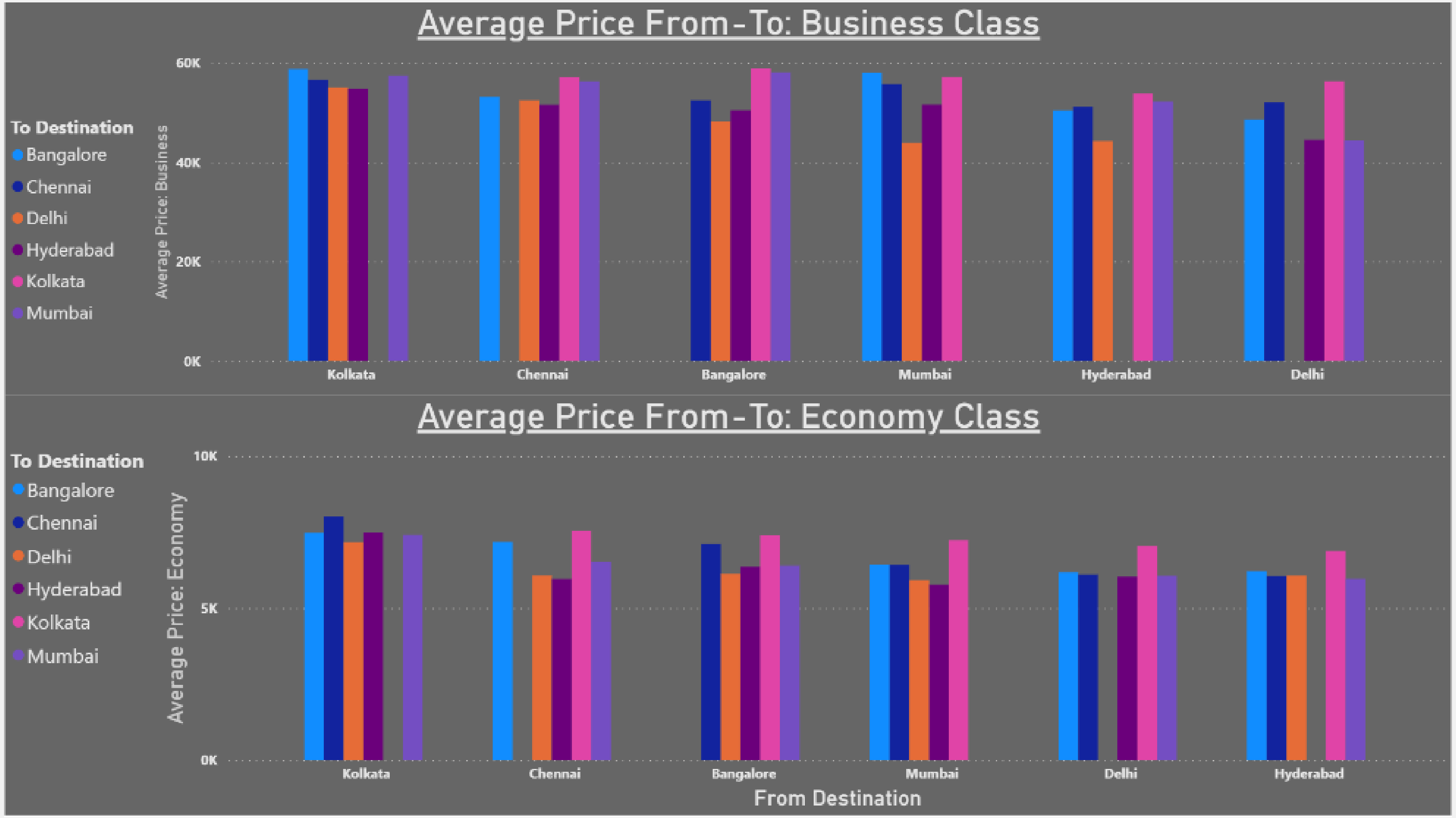
After building the model, using market data, and understanding our competitors, we know exactly what type of luxury airline we'd like to be.

While the model won't give us a direct price for our tickets, it gives us a clear picture as to what the market price for a ticket would be.

The next steps would be to take into account smaller things, like departure time of day and From-To locations, and compare it to our competitors.

From there, we will be able to create our flight prices correctly to fit the market and what we provide as an airline.

APPENDIX



Average Price From-To: Business Class

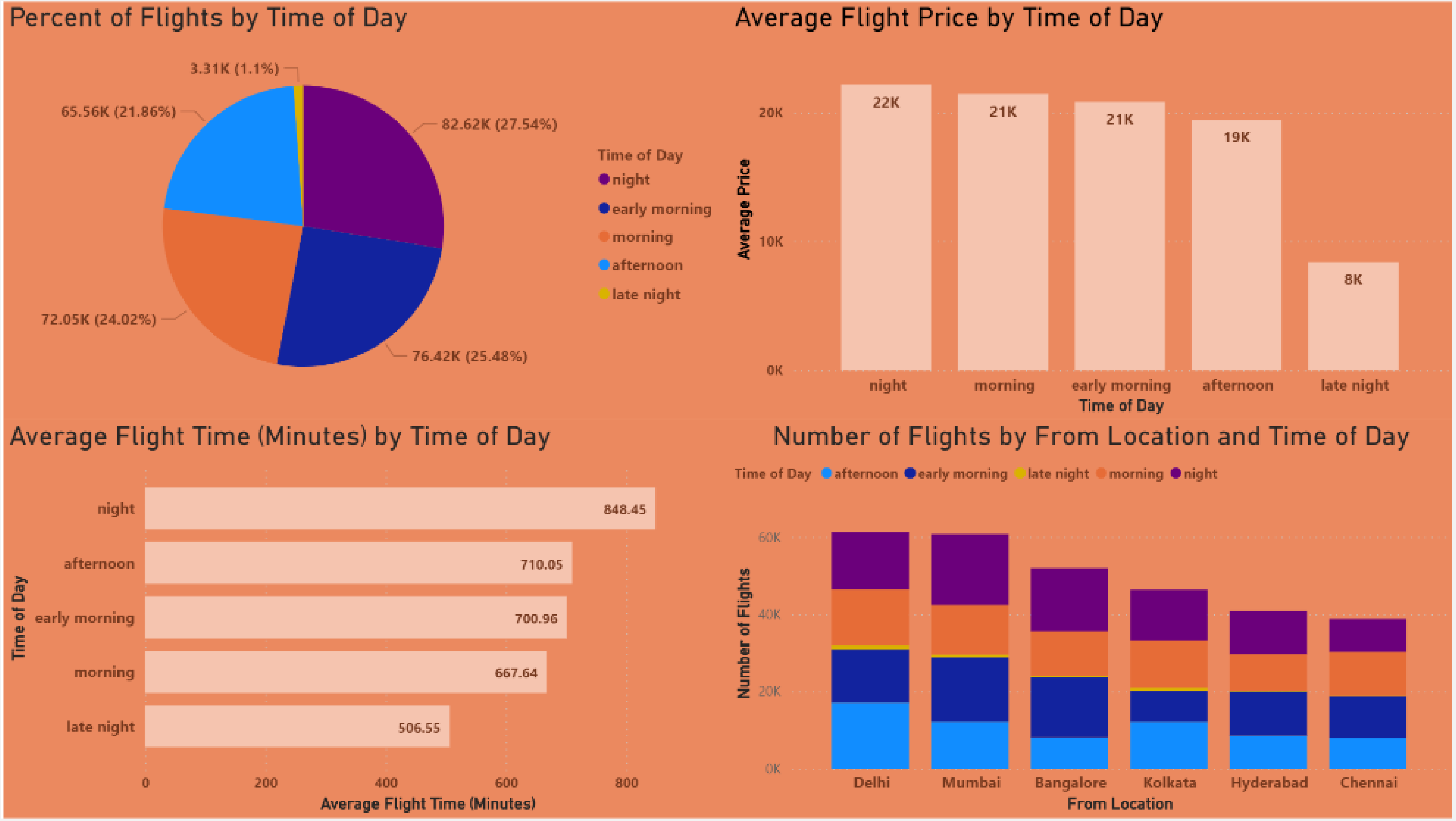
Average Price From-To: Economy Class

Using this chart, we can more accurately choose our prices based off where the flight is going.

For example, we use the colored bars to see that flights to Kolkata cost the most compared to all other destinations.

Vice versa, we can use the bars to see that flights to Delhi and Hyderabad are the two of the lowest costing destinations.

APPENDIX (CONT.)



Using this chart, we can more accurately choose our prices based off the departure time of day.

For example, we can see that night flights (5PM-11:59PM) are the longest, most expensive, and occur the most, on average.

Vice versa, we can see that late night flights (12AM-3:59AM) are the shortest, least expensive, and occur the least, on average.