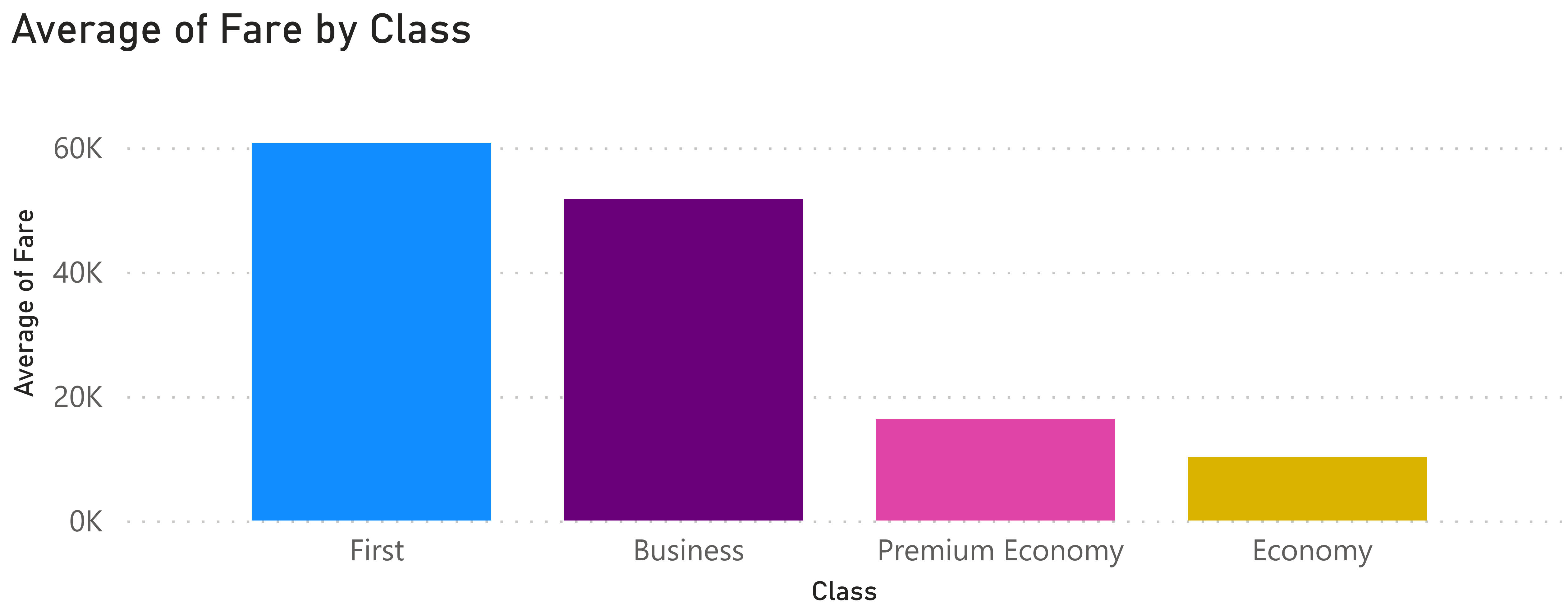
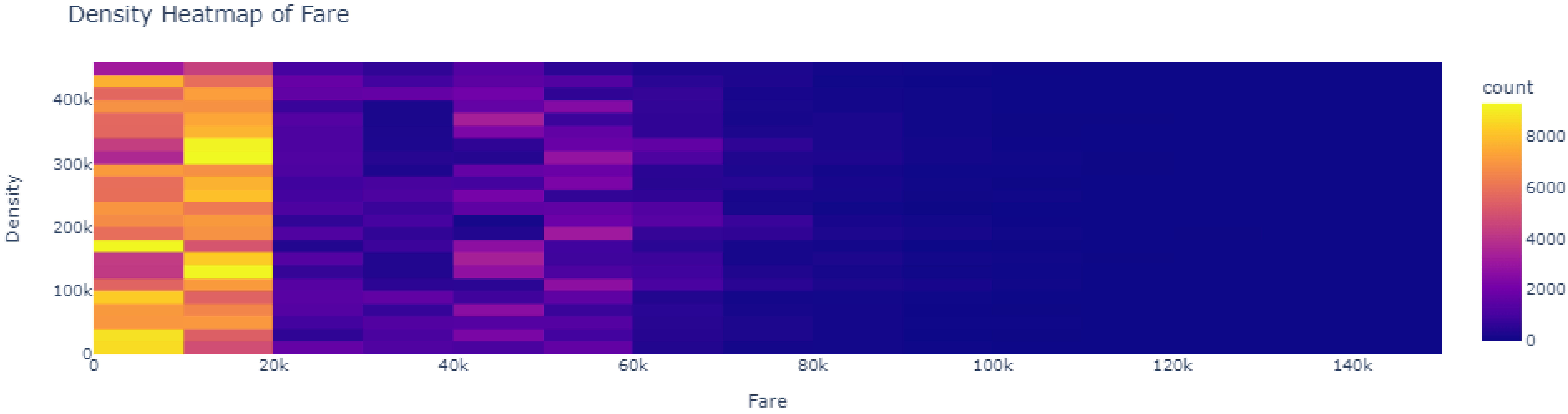
The Airfare ML: Predicting Flight Fares dataset is a collection of flight prices and related features for various routes between different cities. The dataset is available on Kaggle. The purpose of this dataset is to provide a useful resource for building machine learning models that can predict the price of flights between different cities.

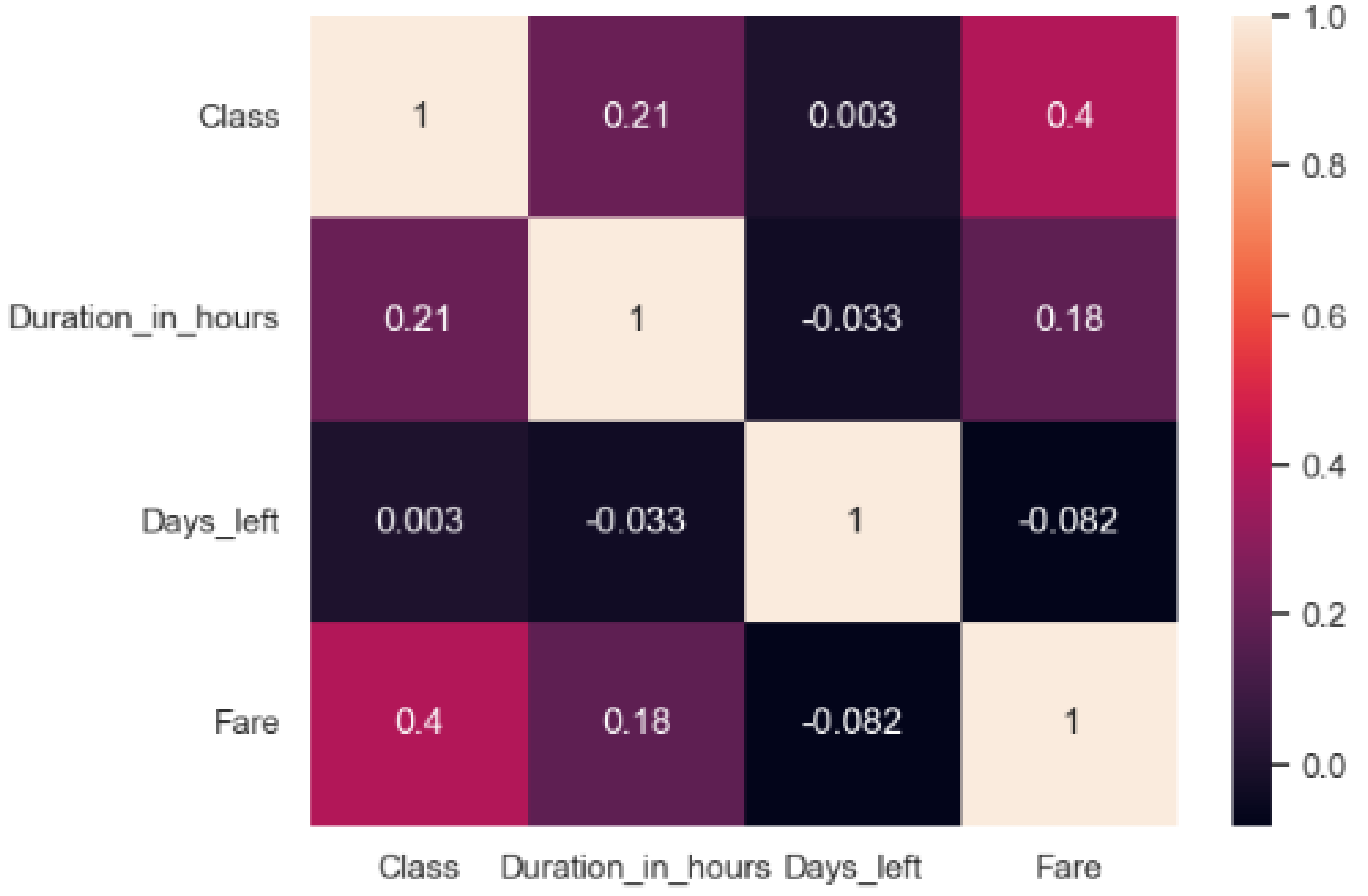
The dataset contains over 40,000 records, each representing a unique flight route. The features provided include the airline, source and destination airports, duration of the flight, distance between the airports, number of stops, and various other attributes. The target variable is the price of the flight, which is provided in Indian Rupees (INR).

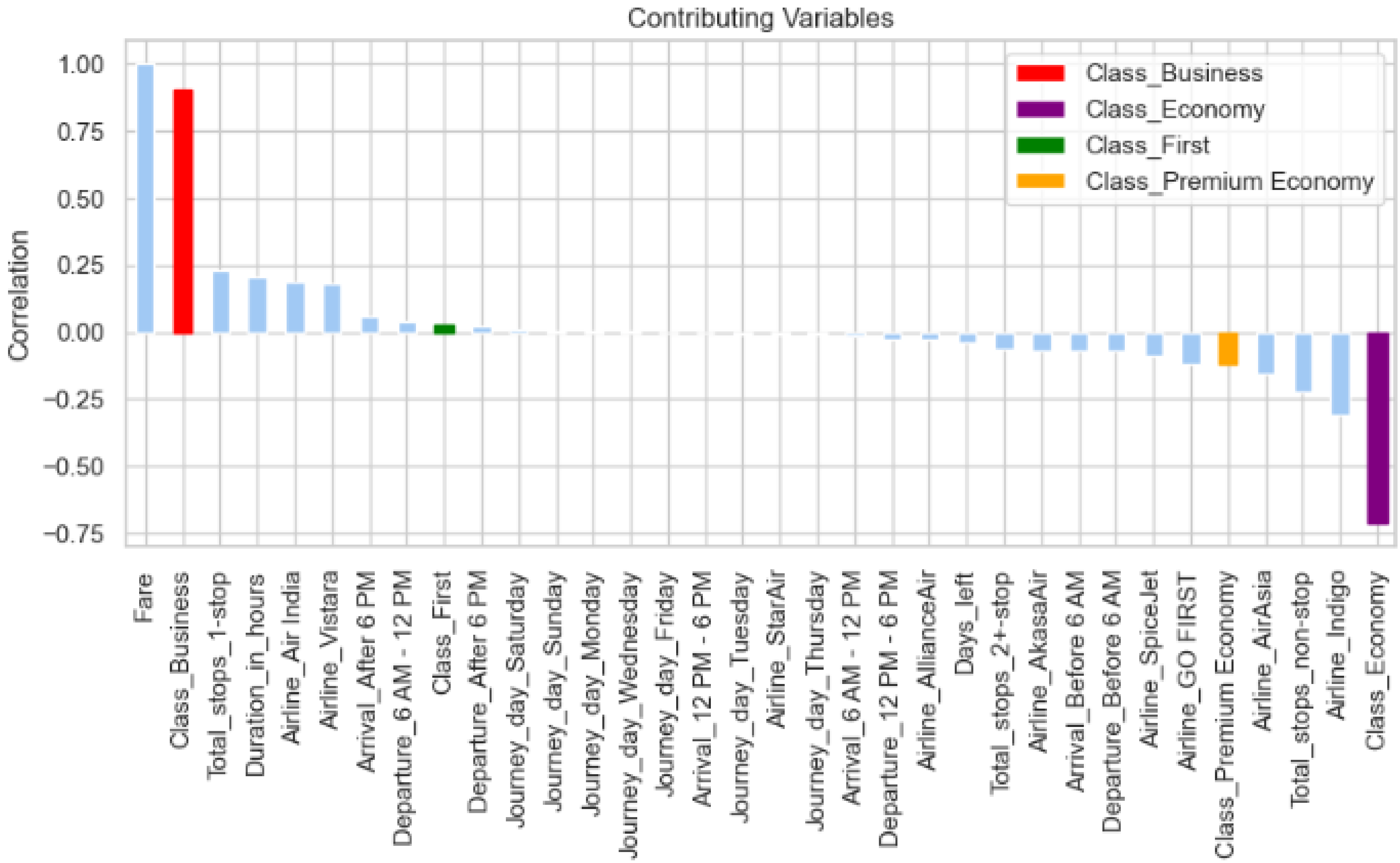
The Airfare ML dataset, which contains information on flight prices and related features, is provided in a single CSV file named Cleaned_dataset.csv. This file contains all of the available data for the dataset, making it easy to access and analyze.

Looking at the histogram and density plots, we can see that the distribution of fare prices is skewed to the right, with a majority of the values ranging from 0 to 20,000. However, there is a small peak around 40,000-60,000 which indicates that there are a few higher-priced fares in the dataset. The density plot also shows that there are some extremely high fare values lying after 100K, which can be considered outliers. While there may be some special cases where these high fares are legitimate, their count is extremely low compared to the normal data distribution.





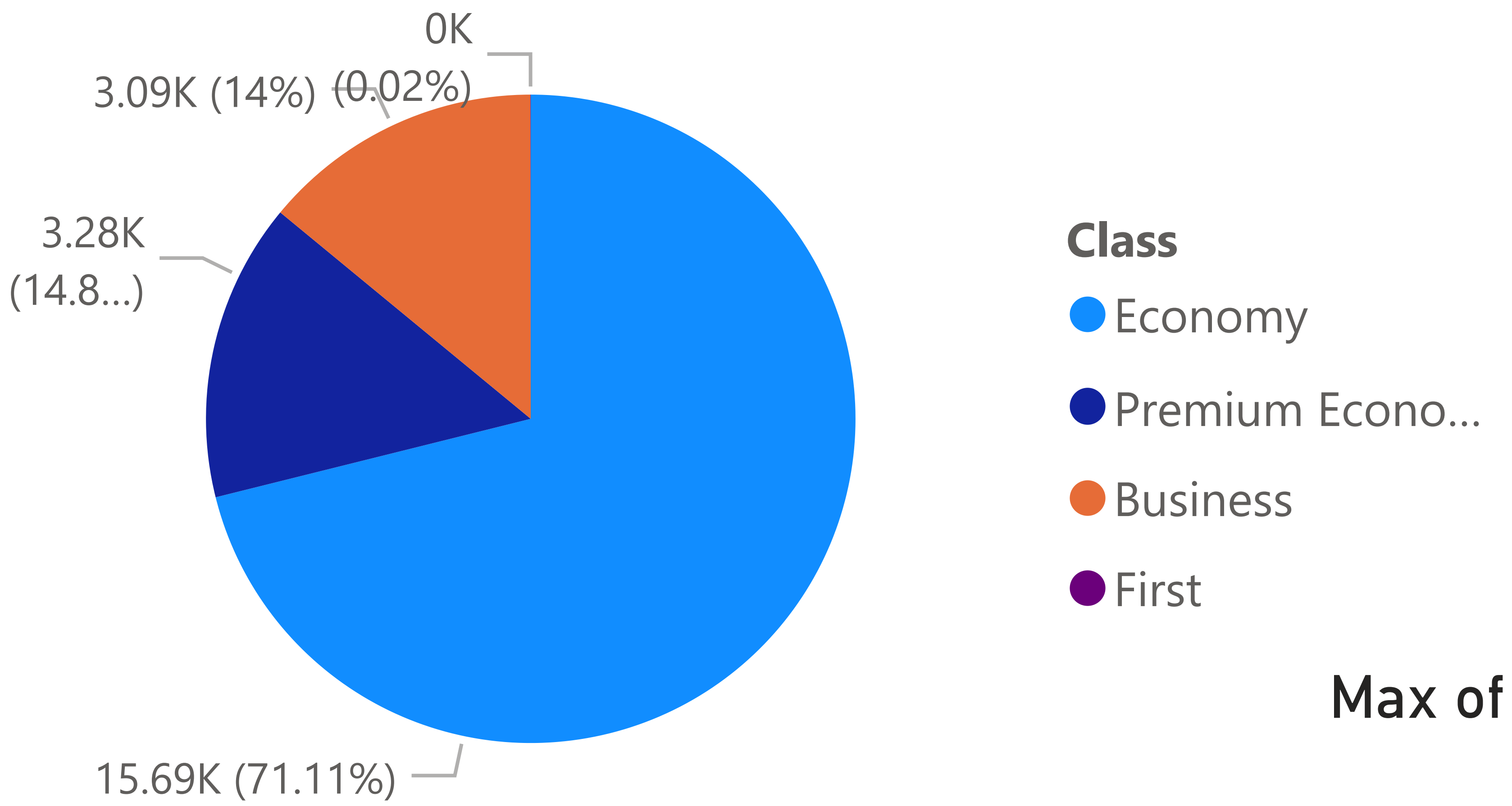




We see the highest correlation being that of Class, followed by duration in hours; but that is almost neglegable. Pointing at Class being the main contributor to fare price.

Looking at the pie chart and the bar graph, we can clearly see that economic class passengers cover almost 56% of the total passengers. Followed by business class passengers with 28% share. And after business class, there is premium economy class with a share of 16. Where is the first class passengers are extremely low. That's why I personally consider this as an outlier to the normal data set, because this really doesn't belong to the normal data distribution. The total count in percentage is lower than 0.04%

Count of Fare by Class



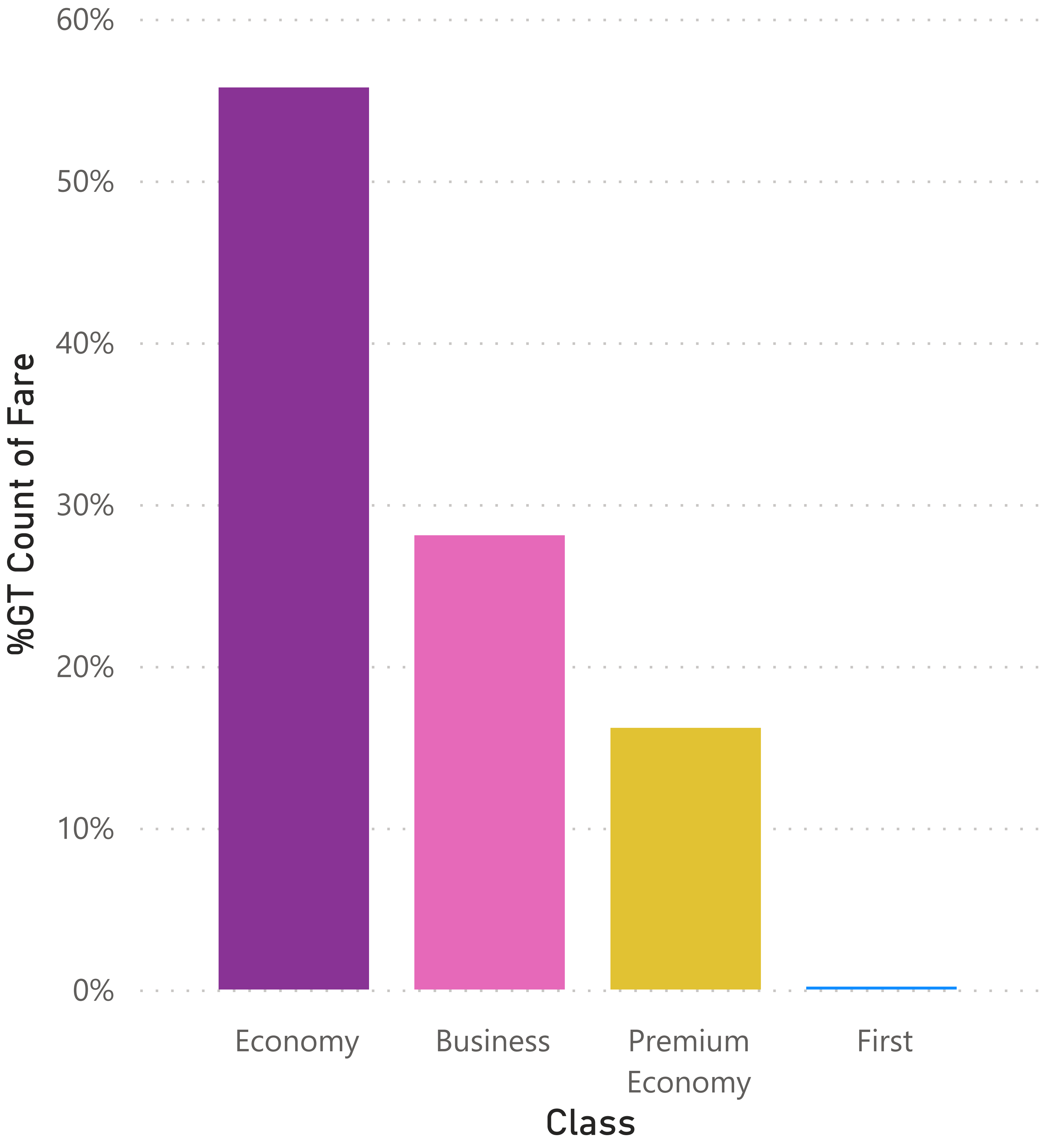
At 252,033, Economy had the highest Count of Class and was 174,922.92% higher than First, which had the lowest Count of Class at 144.

Economy had the highest Count of Class at 252,033, followed by Business, Premium Economy, and First.

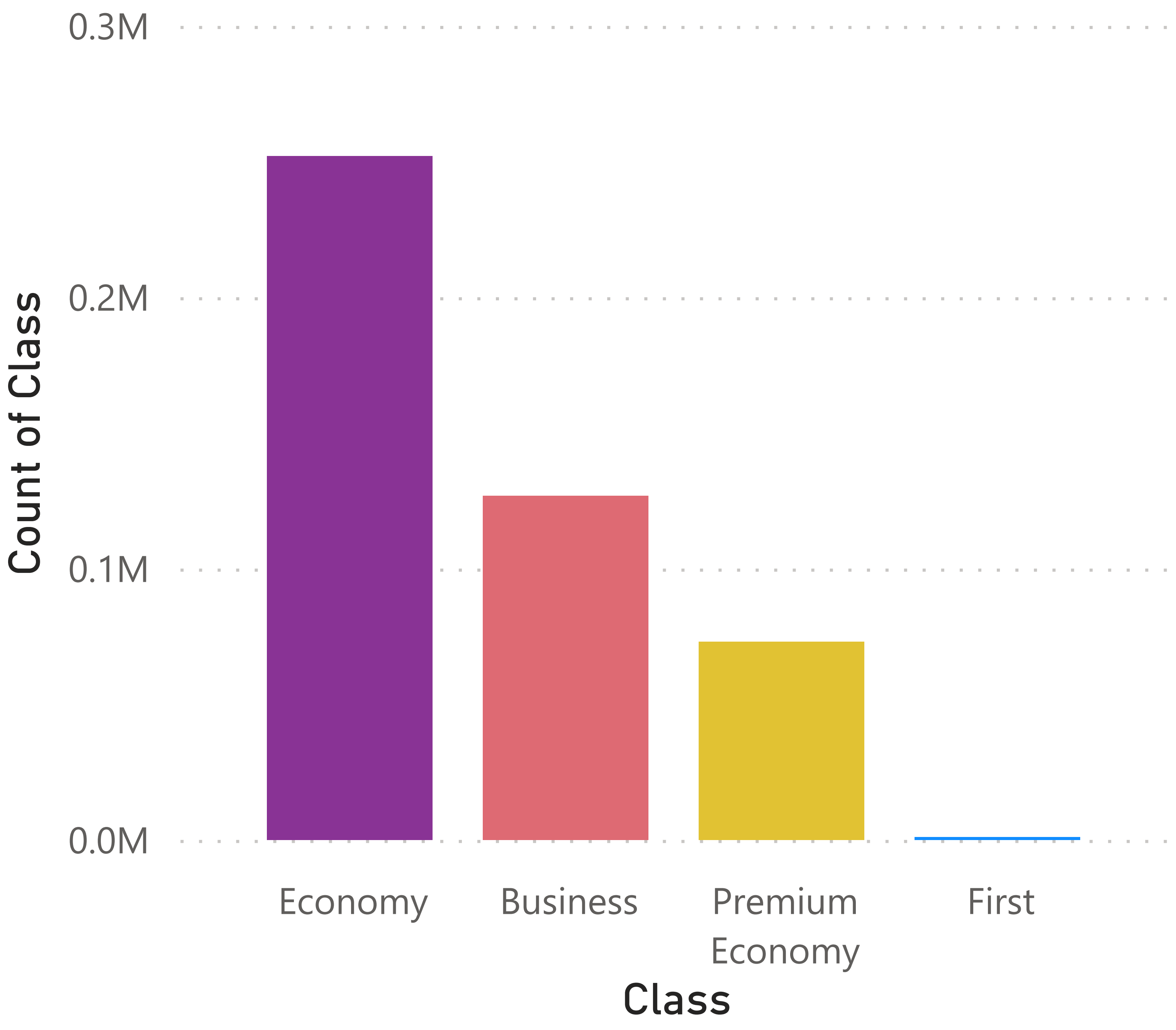
Economy accounted for 55.75% of Count of Class.

Across all 4 Class, Count of Class ranged from 144 to 252,033.

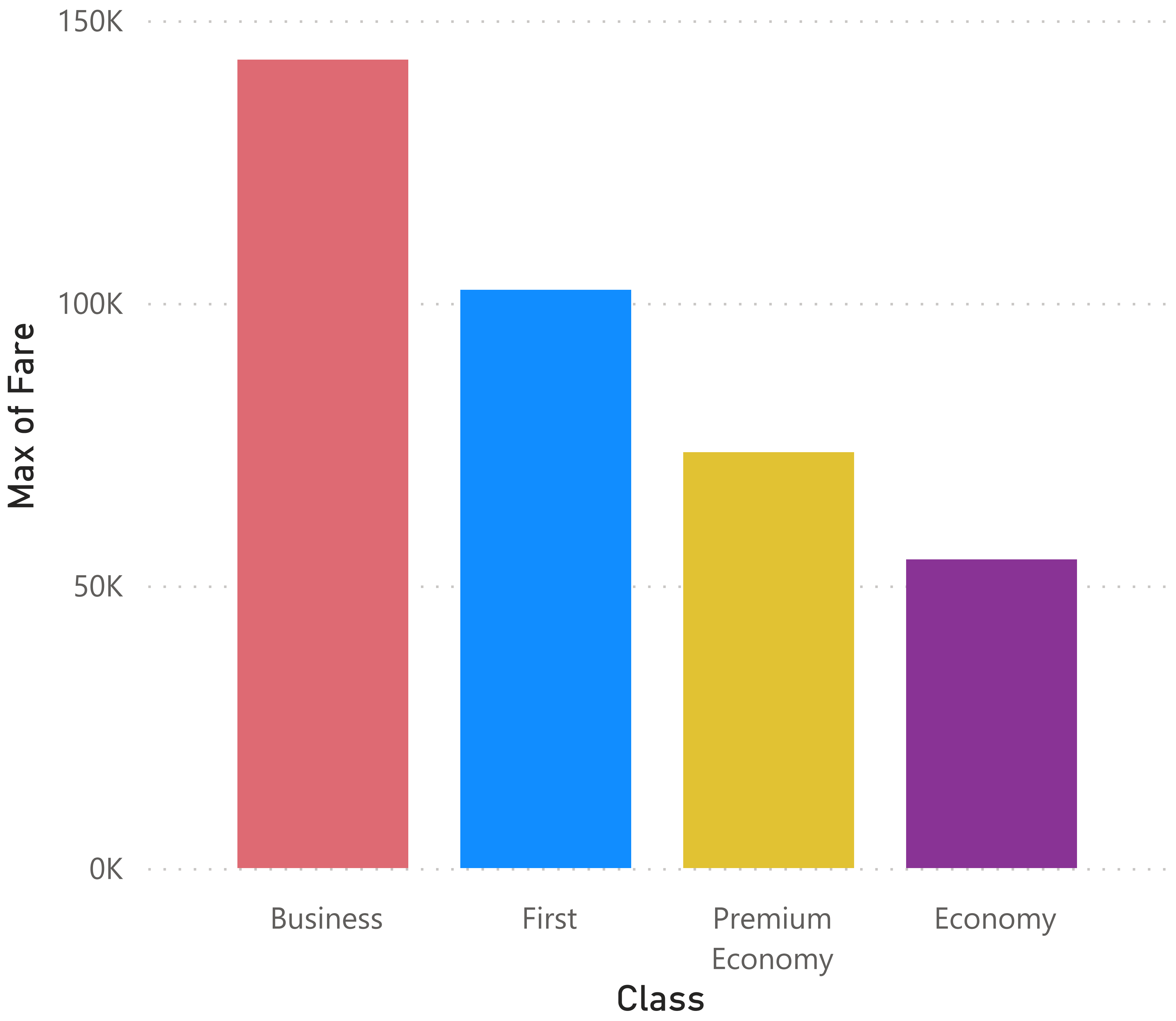
%GT Count of Fare by Class



Count of Class and First Class by Class



Max of Fare by Class

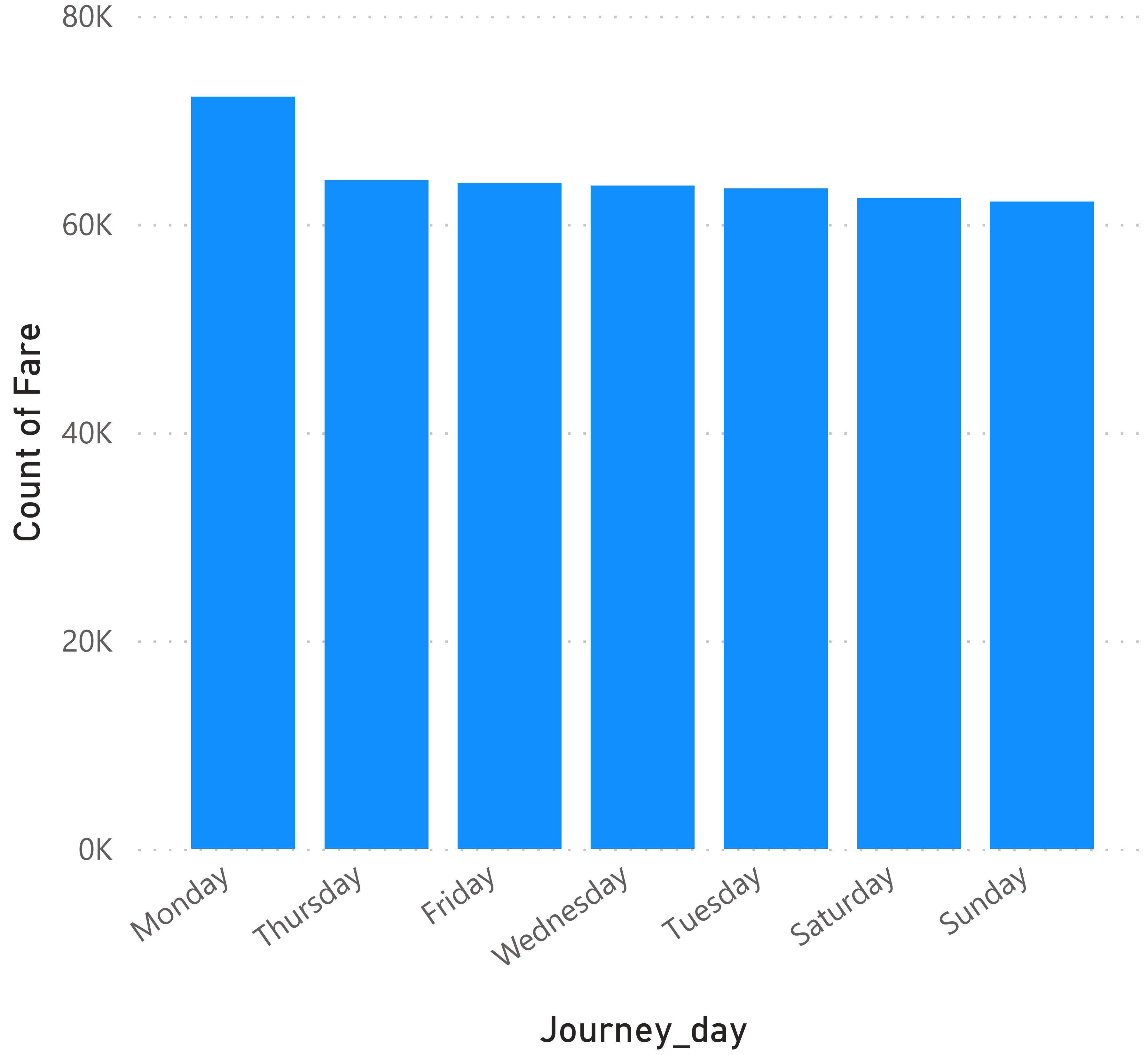


At 72,220, Monday had the highest Count of Fare and was 16.20% higher than Sunday, which had the lowest Count of Fare at 62,150.

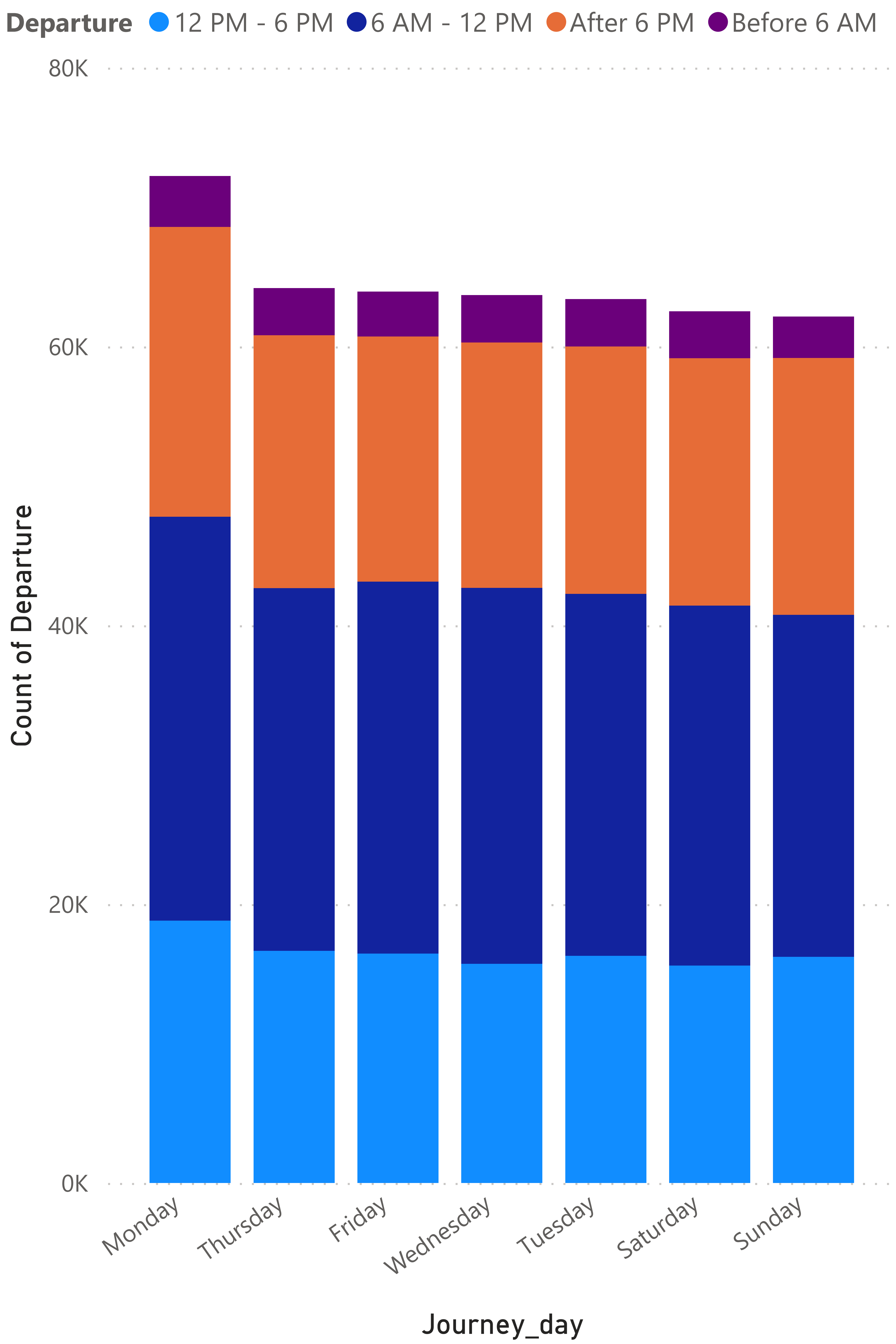
Monday accounted for 15.97% of Count of Fare.

Across all 7 Journey_day, Count of Fare ranged from 62,150 to 72,220.
This shows us that the day of the week is quite negligible in terms of fare.

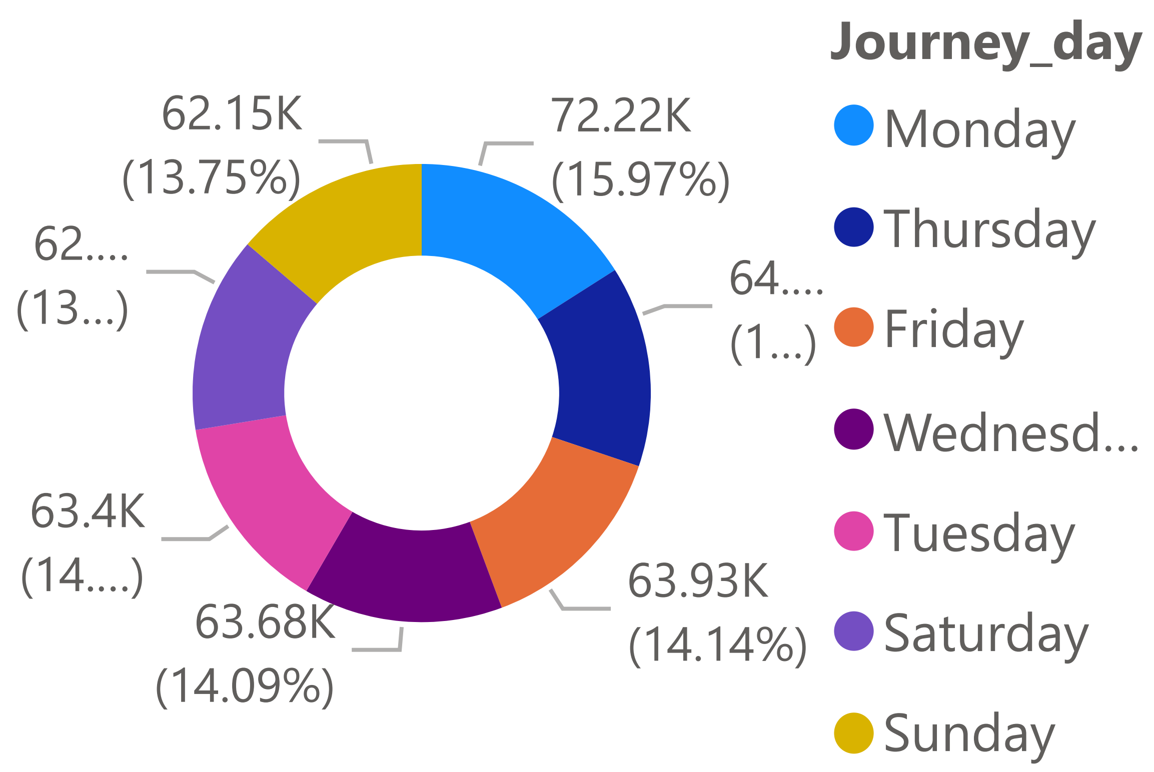
Count of Fare by Journey_day



Count of Departure by Journey_day and Departure



Count of Departure by Journey_day

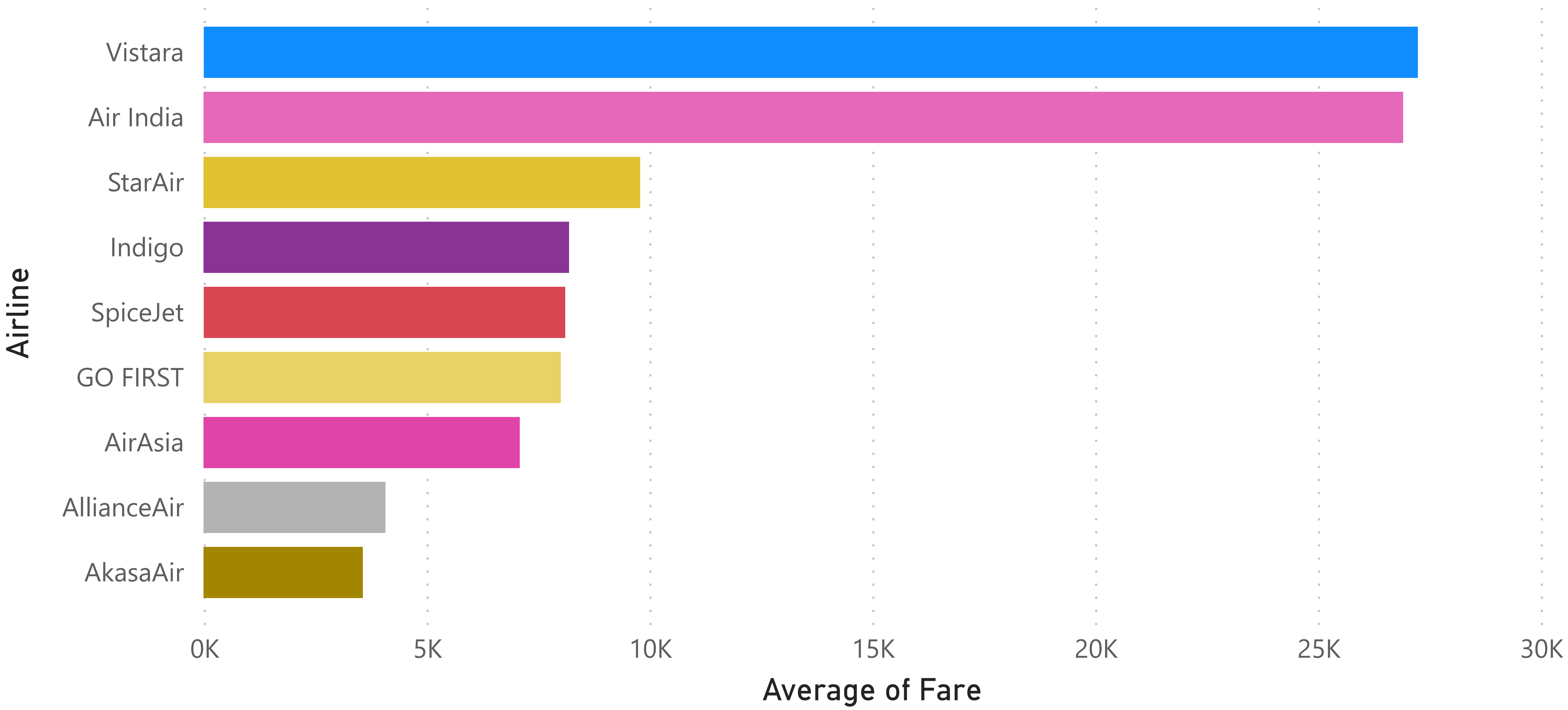


6 AM - 12 PM had the highest total Count of Departure at 184,980, followed by After 6 PM, 12 PM - 6 PM, and Before 6 AM.

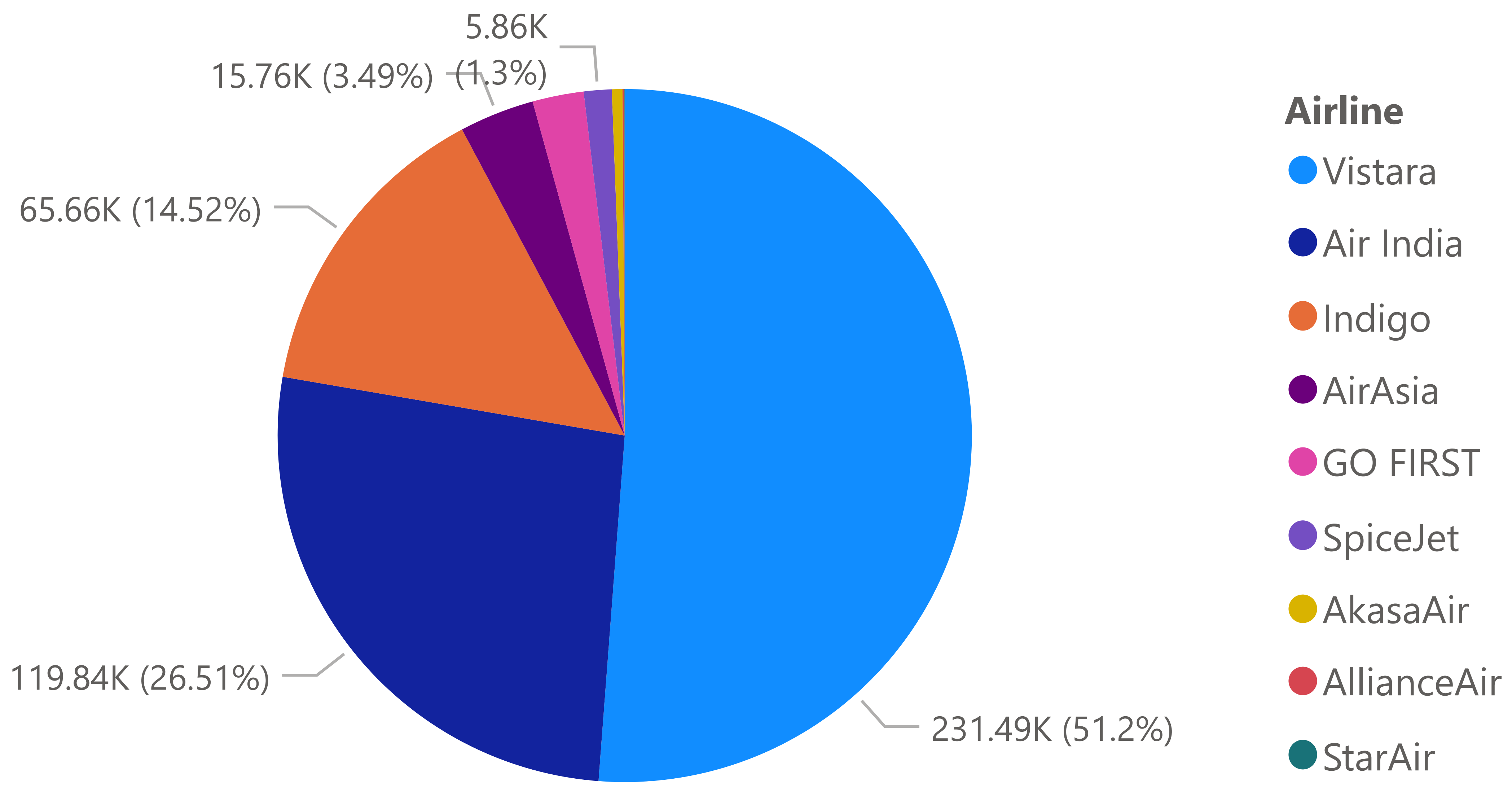
Monday in Departure made up 6.41% of Count of Departure.

6 AM - 12 PM had the highest average Count of Departure at 26,425.71, followed by After 6 PM, 12 PM - 6 PM, and Before 6 AM.

Average of Fare by Airline



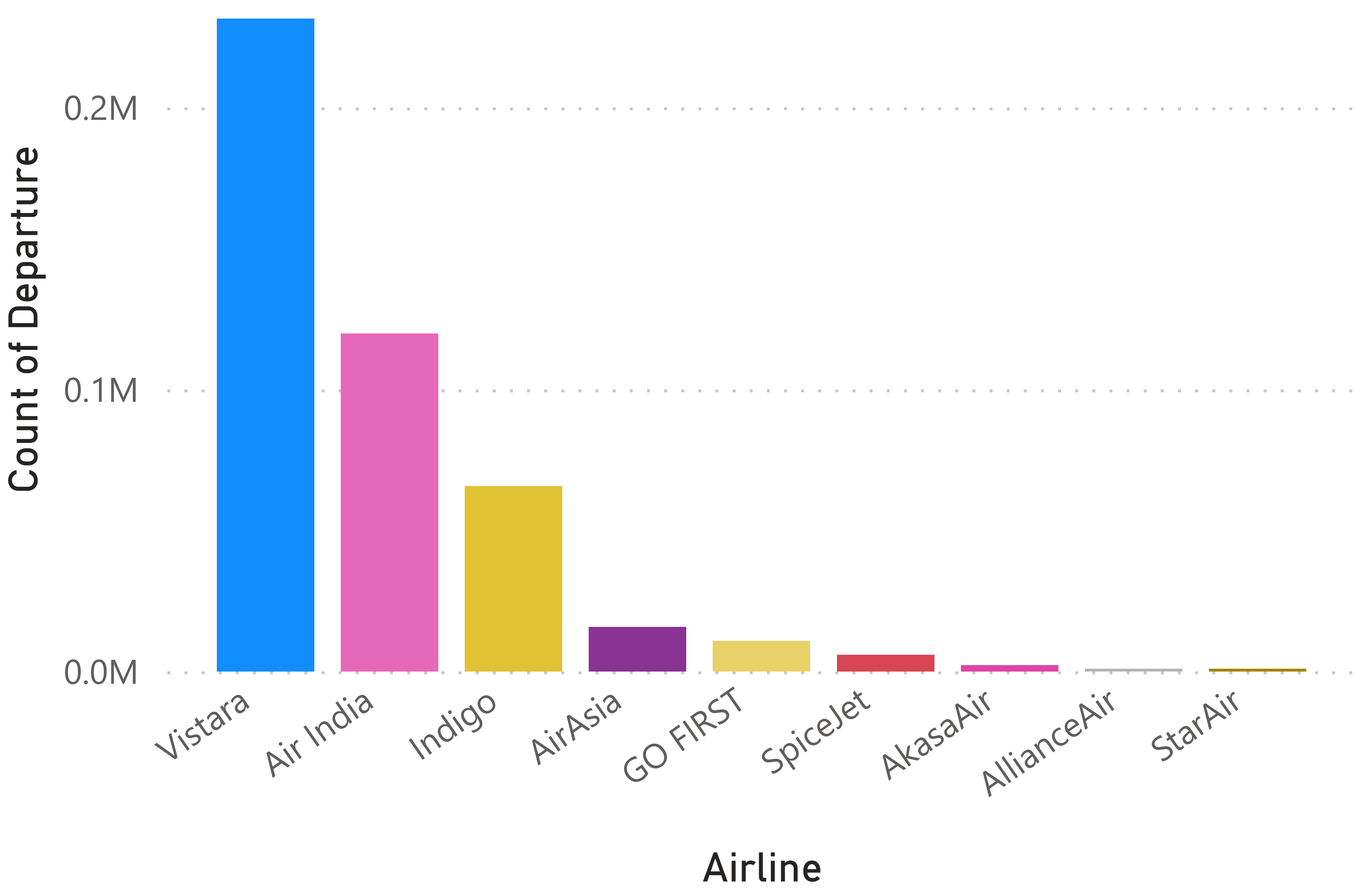
Count of Journey_day by Airline



This categorical plot clearly shows that the distribution of airlines operating is relatively consistent across all days. Regardless of the day of the week, the number of flights operated by each airline is roughly the same. This indicates that there is no significant variation in the airline industry's activity levels based on the day of the week.

Low frequency of some airlines in the dataset indicates that they are less represented and may have less influence on the target variable being studied. However, it is important to note that the impact of an airline on the target variable cannot be solely determined by its frequency in the dataset, and further analysis is required to fully understand their impact.

Count of Departure by Airline

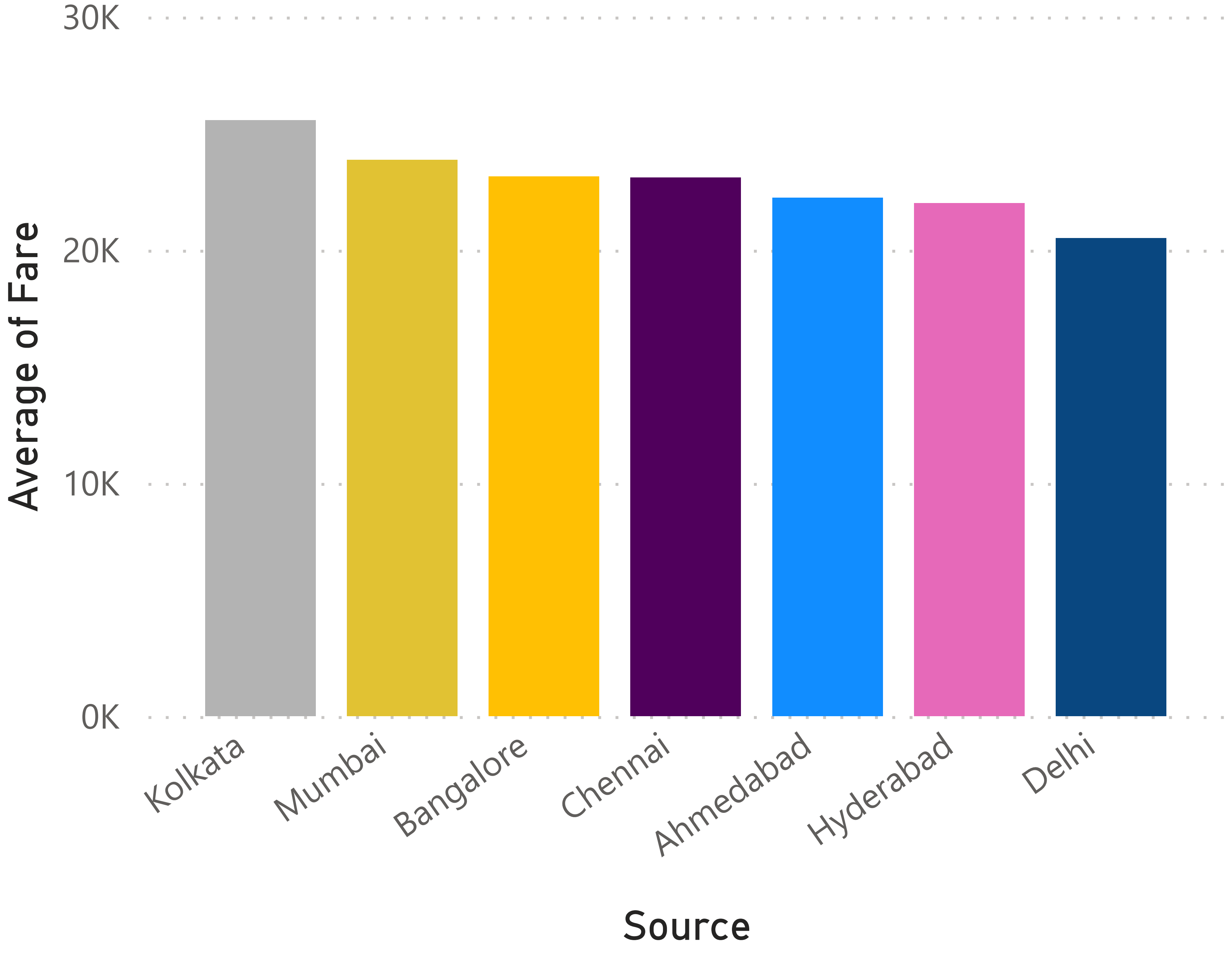


At 231,490, Vistara had the highest Count of Departure and was 373,270.97% higher than StarAir, which had the lowest Count of Departure at 62.

Vistara accounted for 51.20% of Count of Departure.

Across all 9 Airline, Count of Departure ranged from 62 to 231,490.

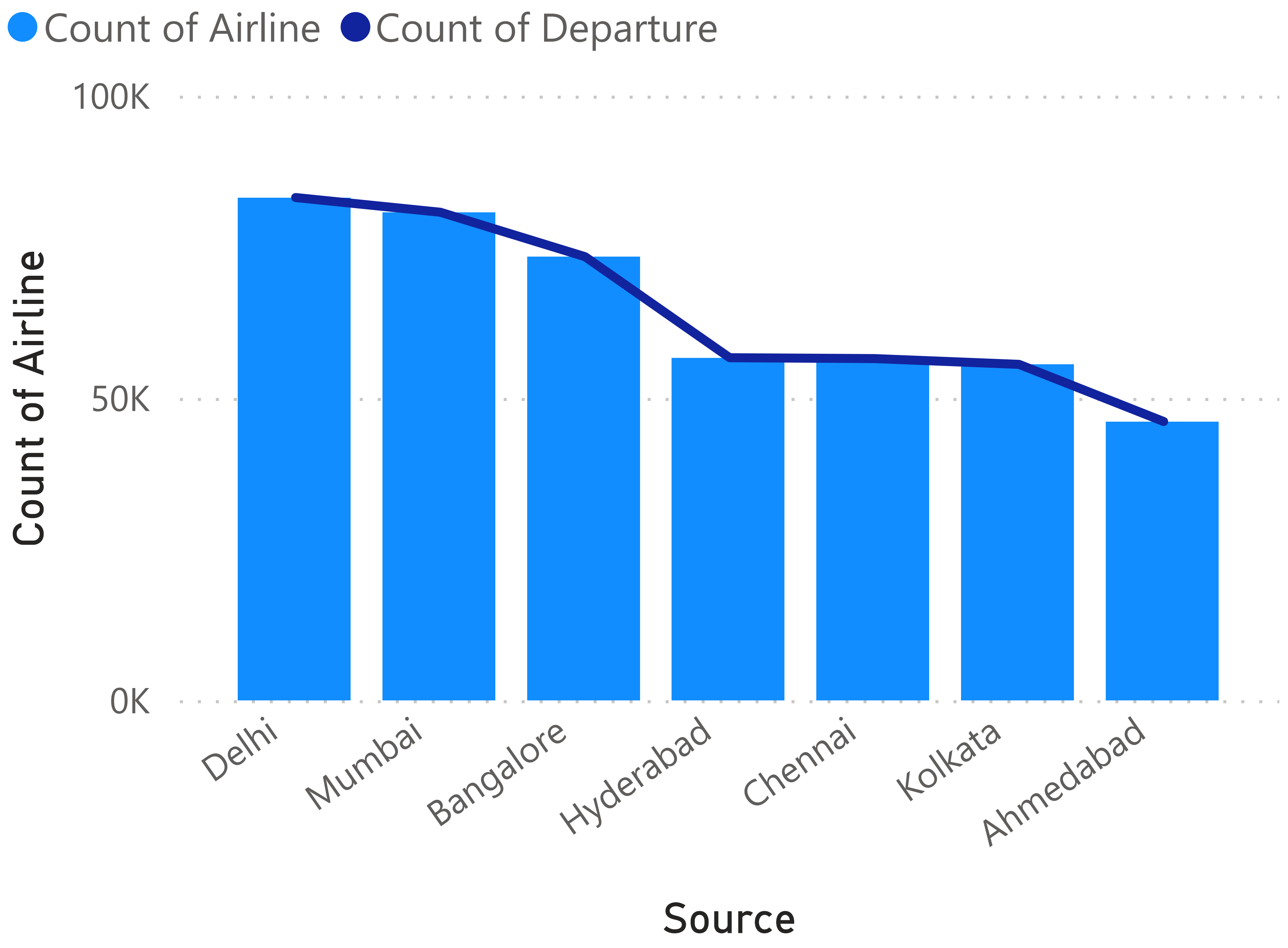
Average of Fare by Source



At 25,553.75, Kolkata had the highest Average of Fare and was 24.63% higher than Delhi, which had the lowest Average of Fare at 20,503.70.

Across all 7 Source, Average of Fare ranged from 20,503.70 to 25,553.75.

Count of Airline and Count of Departure by Source

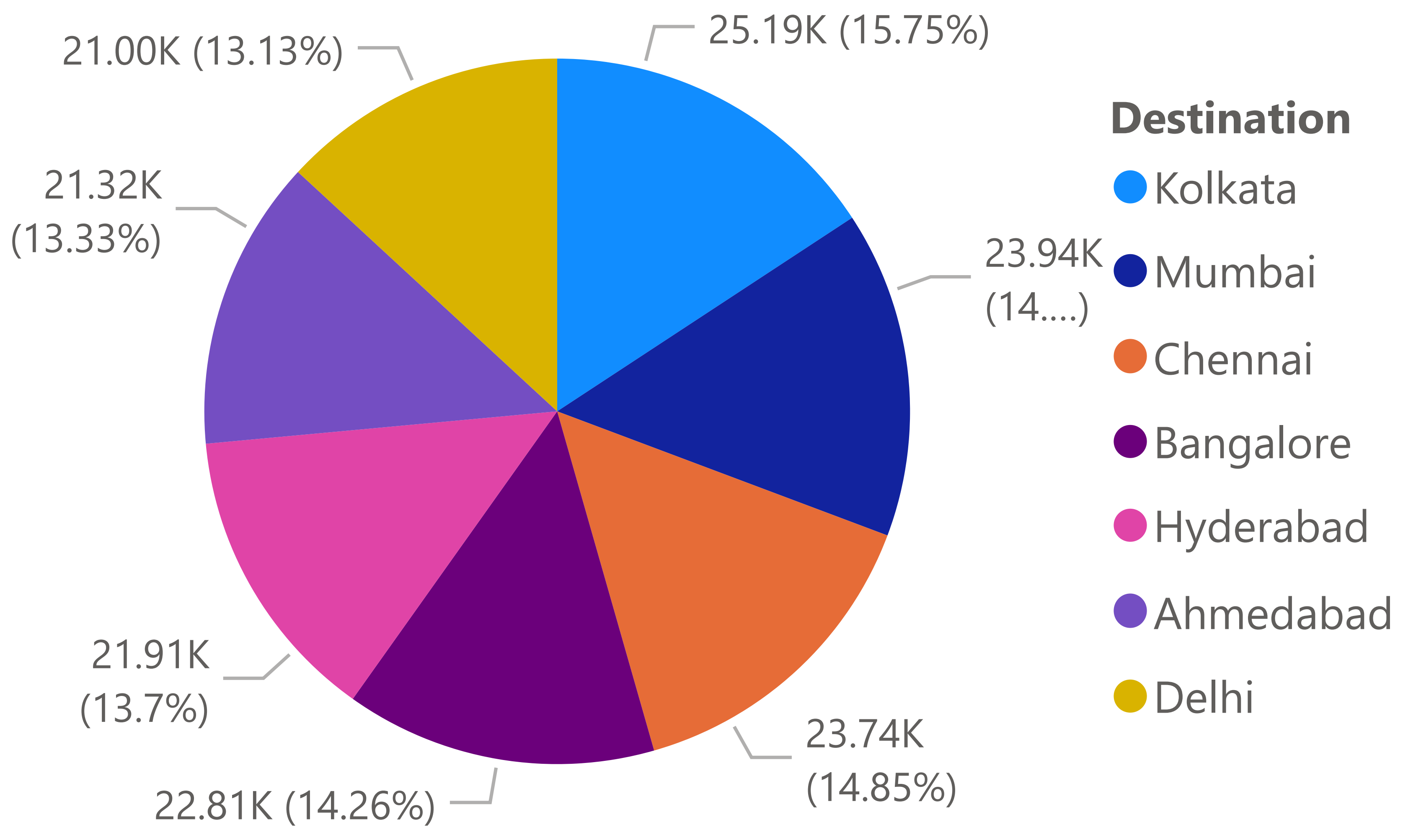


At 83,153, Delhi had the highest Count of Airline and was 80.35% higher than Ahmedabad, which had the lowest Count of Airline at 46,106.

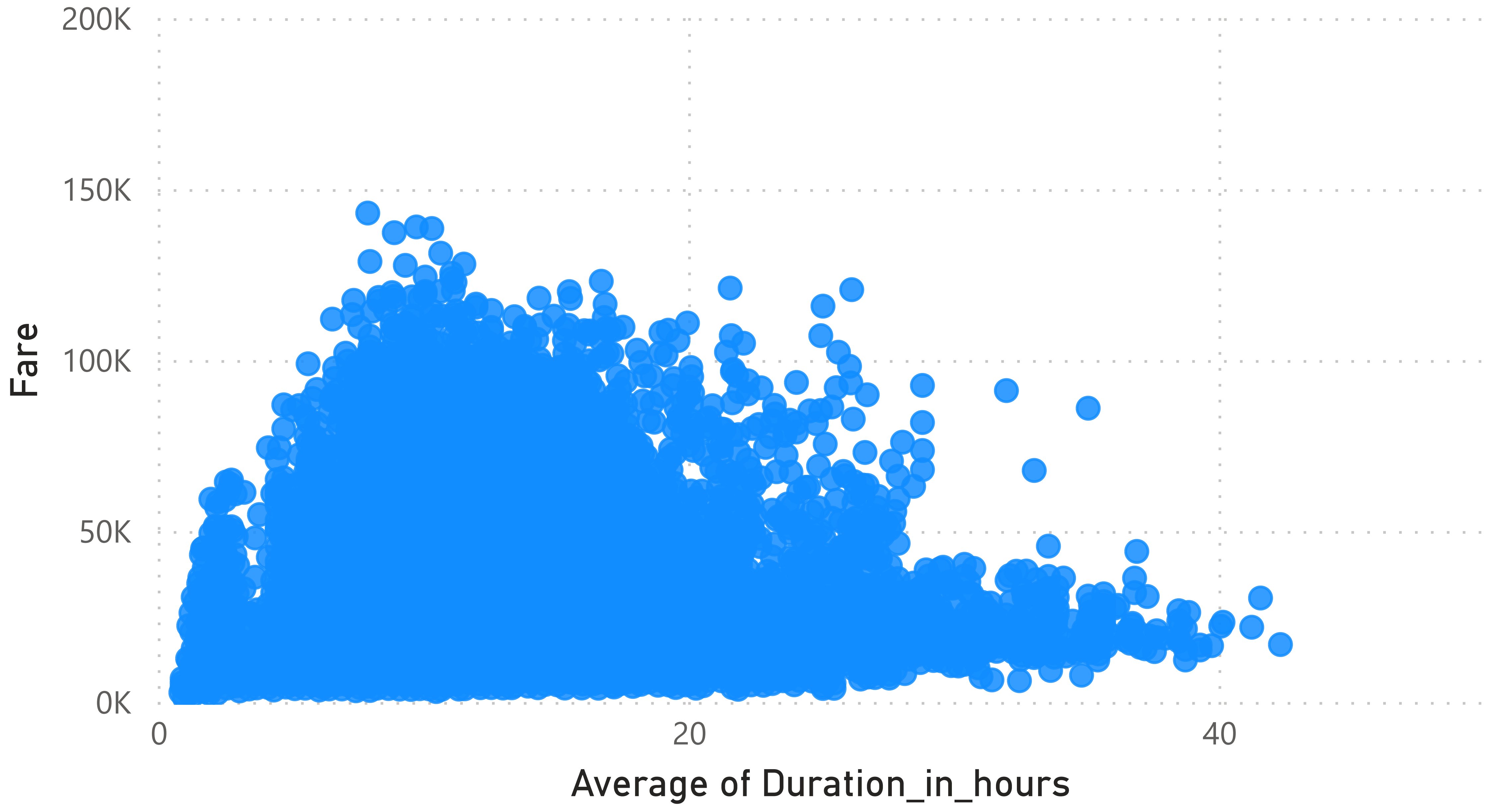
Count of Airline and total Count of Departure are positively correlated with each other.

Delhi accounted for 18.39% of Count of Airline.

Average of Fare by Destination

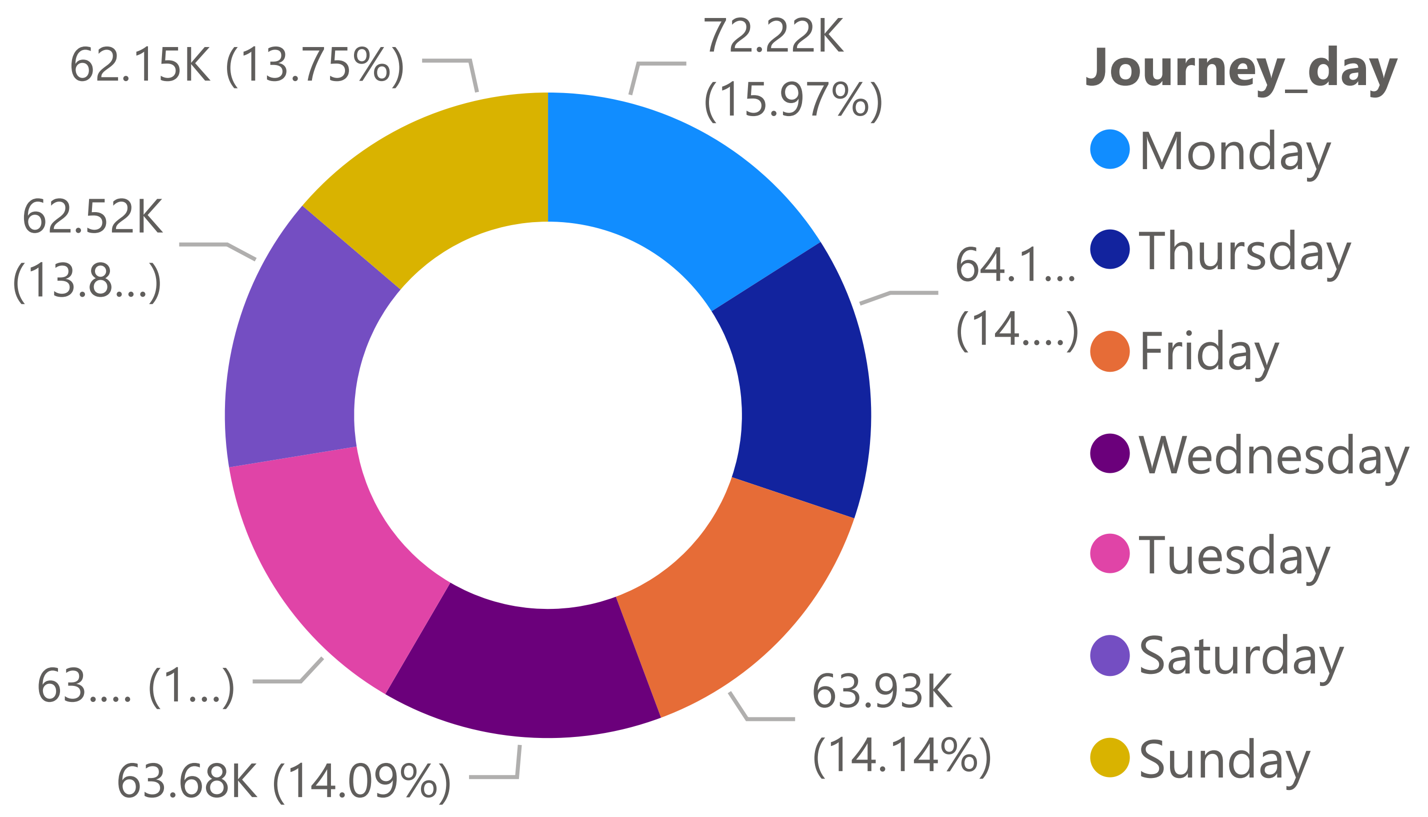


Average of Duration_in_hours by Fare

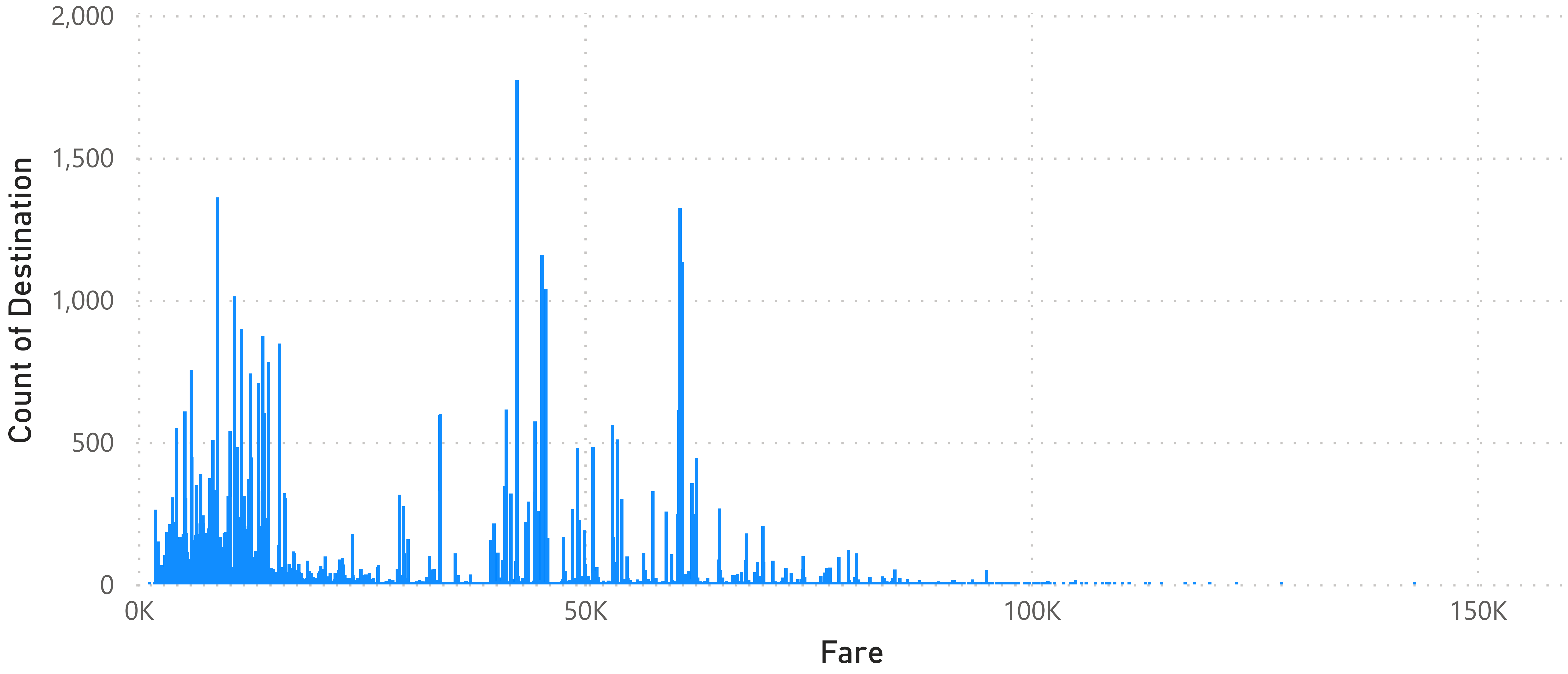


Although the average is roughly the same across the days, there is a noticeable difference in traffic volume, with Monday showing a much higher volume compared to the other days. In fact, Monday has the highest traffic volume while Sunday has the lowest. Additionally, there is a slightly decreasing pattern in traffic volume from Monday to Sunday. These insights suggest that traffic volume varies significantly across different days of the week, with Monday being the busiest day and Sunday being the least busy day.

Count of Fare by Journey_day



Count of Destination by Fare



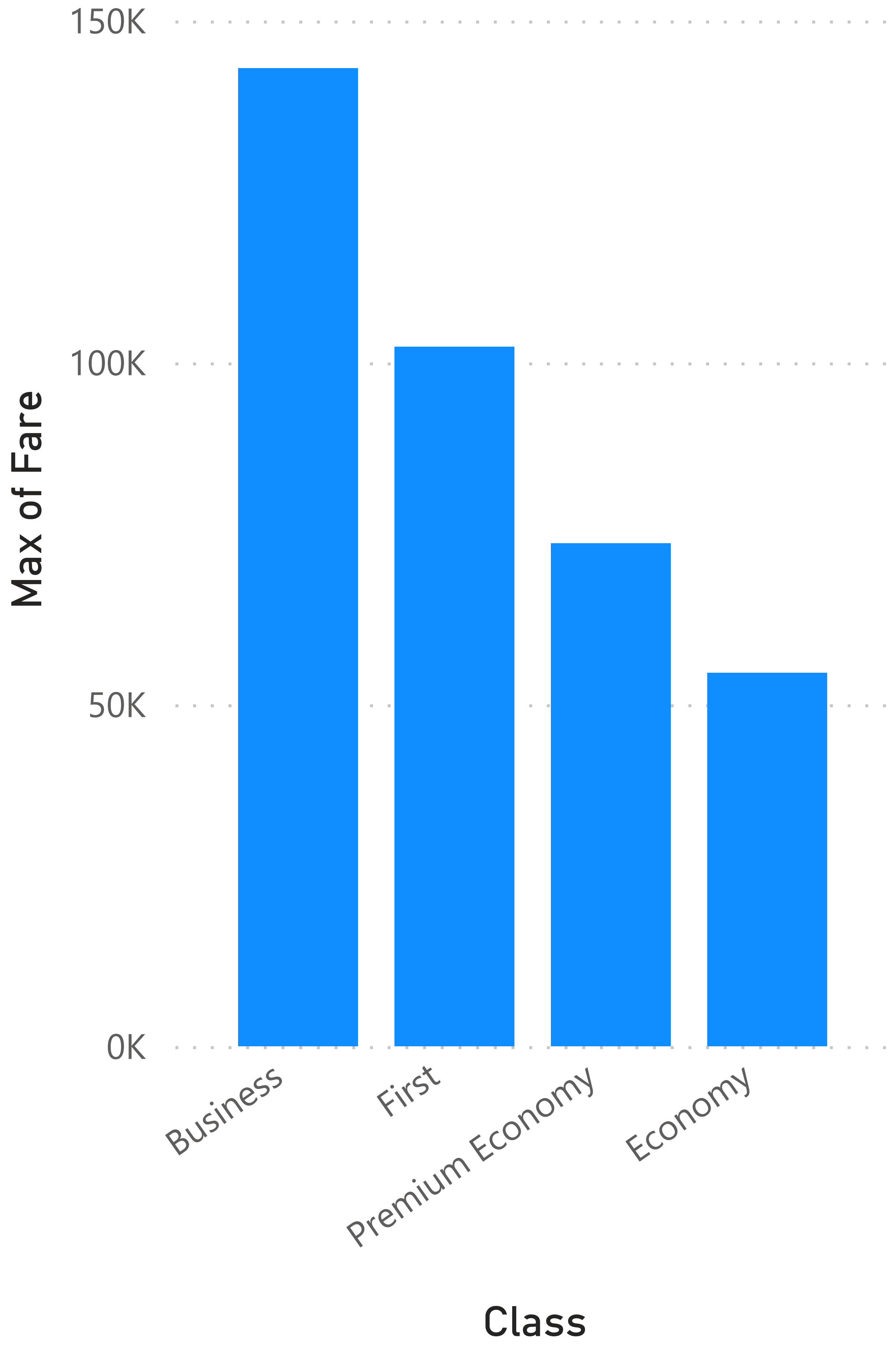
Count of Destination was highest for 54879 at 2,702, followed by 54608 and 49613.
54879 accounted for 0.60% of Count of Destination.
Across all 20,781 Fare, Count of Destination ranged from 1 to 2,702.

Training Set:
Mean Absolute Error (MAE): 4245.868239549071
Mean Squared Error (MSE): 36383774.6063297
Root Mean Squared Error (RMSE): 6031.896435312007
R-squared (R2): 0.8852171296759626

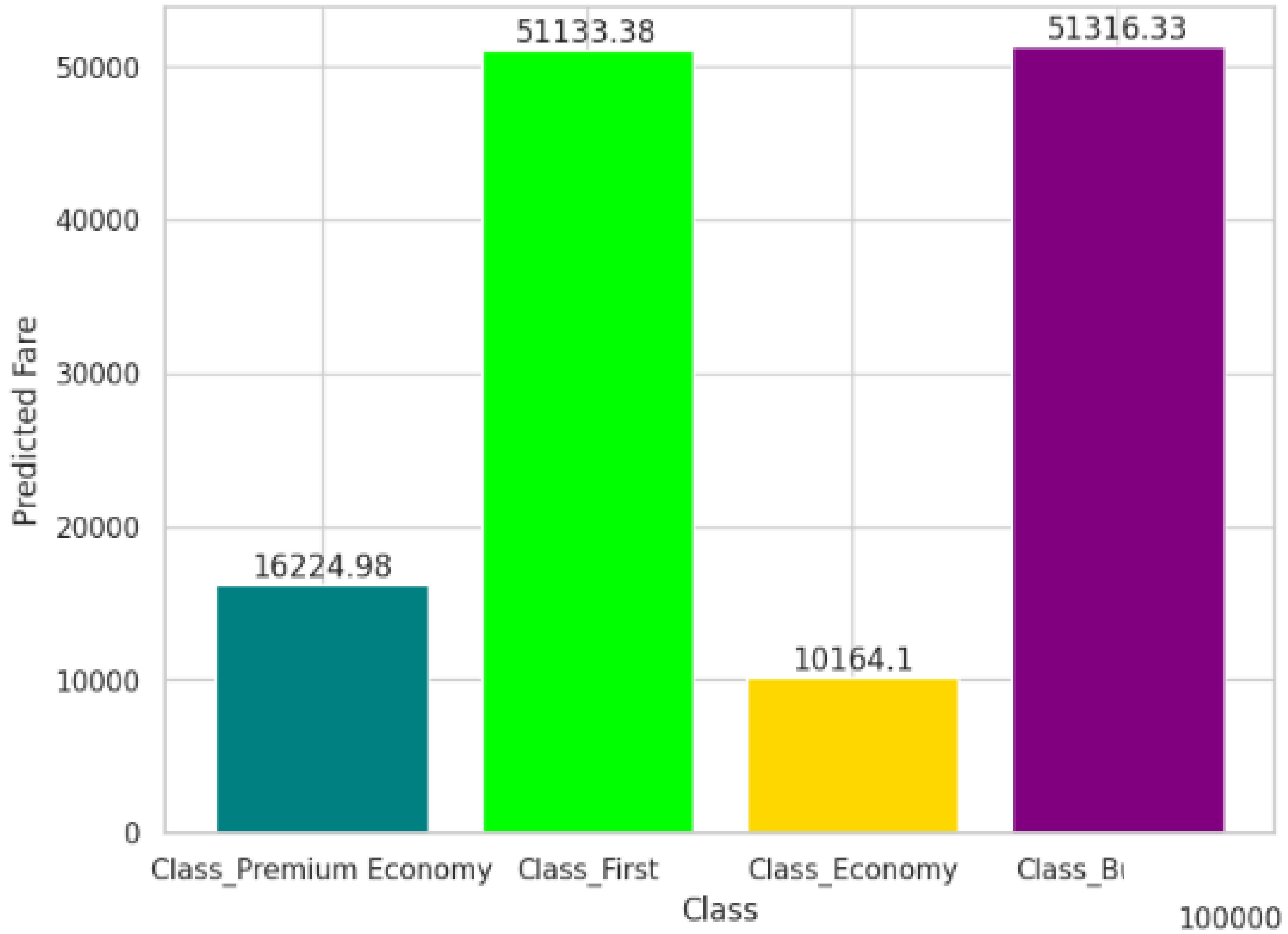
Testing Set:
Mean Absolute Error (MAE): 4231.229265985513
Mean Squared Error (MSE): 36076349.82197946
Root Mean Squared Error (RMSE): 6006.359115302669
R-squared (R2): 0.8860873199610759

Best Hyperparameters: {'regression__copy_X': True, 'regression__fit_intercept': True}
Mean Absolute Error (MAE): 4230.531494896574
Mean Squared Error (MSE): 36074011.048837915
Root Mean Squared Error (RMSE): 6006.164420729582
R-squared (R2): 0.8860947047413509

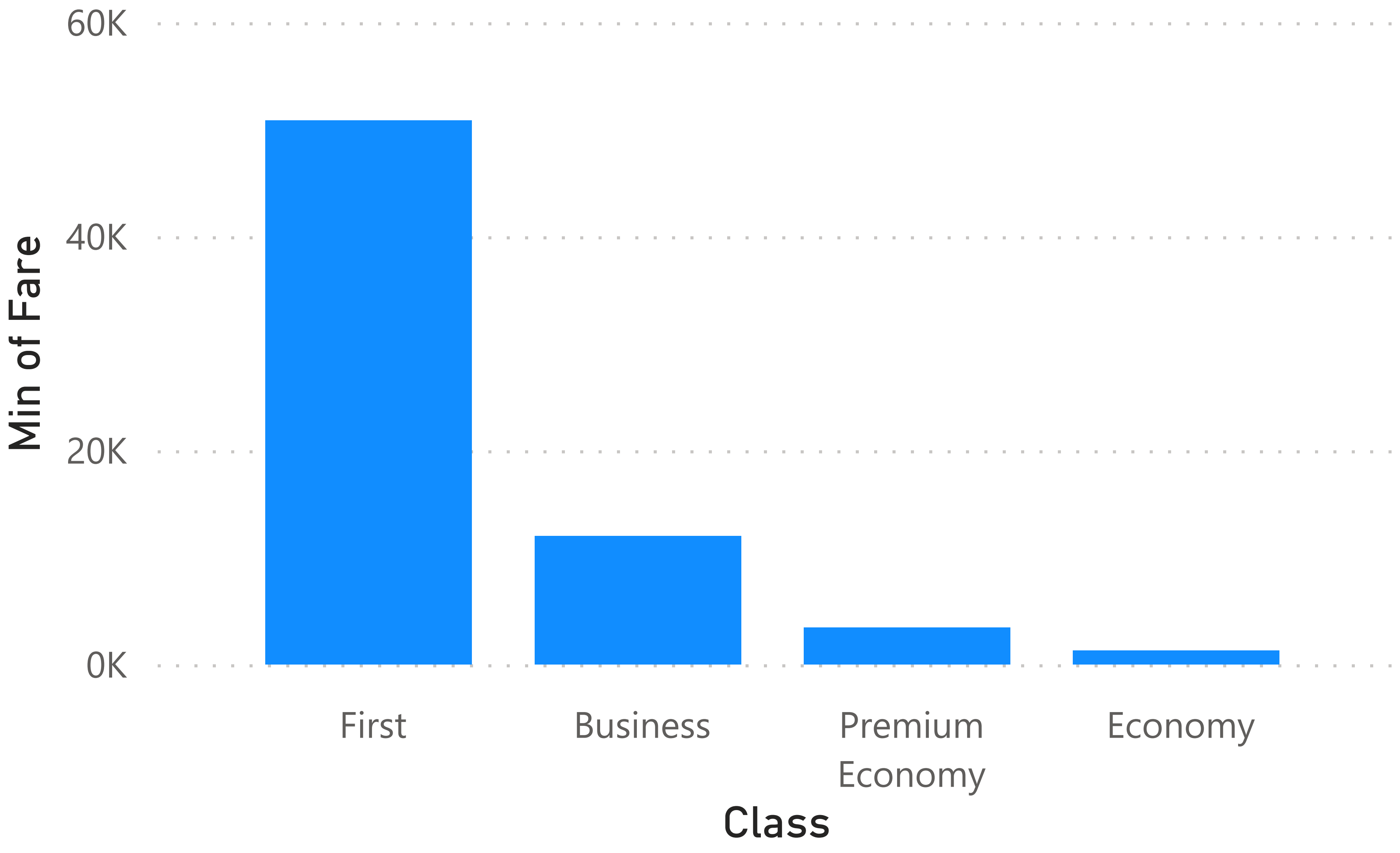
Max of Fare by Class



Predicted Fare for Each Class Combination



Min of Fare by Class



These descriptive statistics can help you understand the fare distribution within each class combination and make informed decisions regarding fare prediction. For instance, you can use these statistics to set pricing strategies, identify outliers, or assess the fairness of fares across different classes.

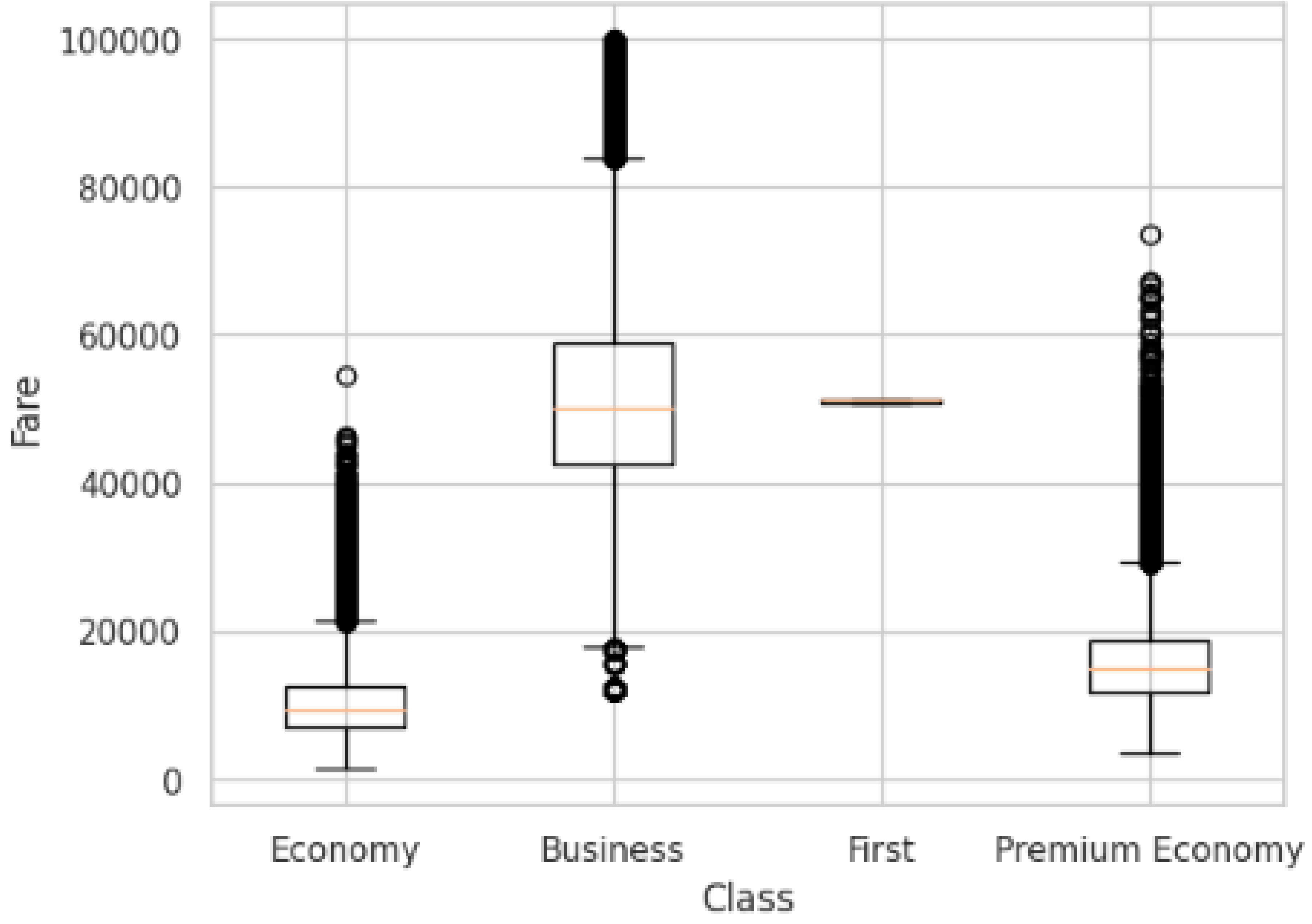
Class_Economy: Predicted Fare: $10,164.10

Class_Premium Economy: Predicted Fare: $16,224.98

Class_First: Predicted Fare: $51,133.38

Class_Business: Predicted Fare: $51,316.33

Comparison of Fare across Classes


Fare
Economy
Business
First
Premium Economy
Class