MGMT 59000-148 Advanced Database Management

Group Project - Team 4

Navigating the Skies

Crafting a Data-Driven Entry Strategy for India's Aviation Sector

Submitted by

Akshita Sharma
Aravind Teja Chikoti
Pooja Udayanjali Kannuri
Rahul Chowdary Kunku
Sai Mona Duvvapu
Sathwik Kanukuntla

Graduate students of

Purdue University

MS in Business Analytics & Information Management

February 2023

Table of Contents

1. Background

The Indian aviation industry stands as one of the most rapidly expanding markets globally, driven by factors such as economic growth, urbanization, and an increase in disposable income among the middle class. The proliferation of low-cost carriers, coupled with government initiatives like the UDAN scheme aimed at enhancing regional connectivity, has democratized air travel, making it accessible to a wider demographic. Nonetheless, the sector is marked by fierce competition, with established players and ambitious new entrants vying for their share of the skies.

In such a vibrant yet saturated market, a data-driven entry strategy is not just prudent but essential for any new airline. The success of an entrant in the Indian aviation sector is predicated not only on competitive pricing but also on a deep understanding of the market dynamics, traveller profiles, and journey patterns. The essence of this strategy lies in leveraging granular data to derive actionable insights that inform every aspect of the business model, from pricing and route planning to customer service and marketing.

Entering the Indian aviation market requires a new airline to navigate through intense competition and varied traveller profiles with a data-driven strategy that prioritizes demographic insights and dynamic pricing. Initially, the focus should be on underserved segments and routes, where the airline can establish itself against market challengers by offering competitive fares and superior service quality. Understanding the demographics and city characteristics is crucial; for example, city with more literacy rates demand higher frequency and flexibility, while city with low literacy might prioritize cost-efficiency deals. By leveraging advanced SQL analytics to dissect historical booking data and market trends, the airline can develop a pricing model that dynamically adapts to fluctuating demand and traveller behaviour, ensuring optimal occupancy and profitability. This holistic strategy, which blends competitive pricing with a deep understanding of customer segmentation and route viability, will pave the way for the airline to scale its operations and eventually compete with market leaders.

2. Introduction

Our consultancy objective is to devise a strategic entry price point for an emergent airline in the highly competitive Indian aviation sector. We aim to harness the power of data to inform market-aligned pricing strategies. This objective will be realized through the meticulous processing and analysis of comprehensive flight, itinerary, and airport data sourced from a premier Indian travel application (sourced from Kaggle), augmented by the demographic variables of the origin and destination locations. Our focus will be to navigate the competitive landscape by initially targeting market challengers such as Air Asia, Akasa Air, Alliance Air, and Go First, thereby securing a stable market position before scaling to take on the established market leaders.

2.1. Project Objectives:

The overarching goal of our project is to deliver a data-driven, multifaceted pricing framework and acute market insights that cater to a diverse array of traveller profiles and journey patterns. Our deliverables encompass:

- A competitive pricing strategy that is attuned to temporal fluctuations, geographic considerations, and demographic diversity. This strategy will inform not only the pricing of tickets but also ancillary revenue streams, ensuring a competitive edge.
- An analysis of potential routes and itineraries grounded in historical data, aiding operational and logistics planning. This insight will support the airline in crafting a flight schedule that aligns with market demand and optimizes resource allocation.
- A comprehensive ranking analysis of Indian airports, providing a regional perspective on market entry points. This will enable the airline to prioritize its market entries, focusing on regions with the highest potential for growth and profitability.

Through these deliverables, we aspire to equip our client with a data-backed foundation for decision-making, positioning them effectively for a successful launch and sustained growth within the Indian aviation industry.

3. Project Question

How can a new airline in the Indian aviation sector establish a competitive entry price point and market strategy using advanced SQL analytics to process flight, itinerary, and airport data, combined with demographic insights of travel origins and destinations, to effectively compete with market challengers and eventually position itself against market leaders?

4. Entity Relationship Diagram

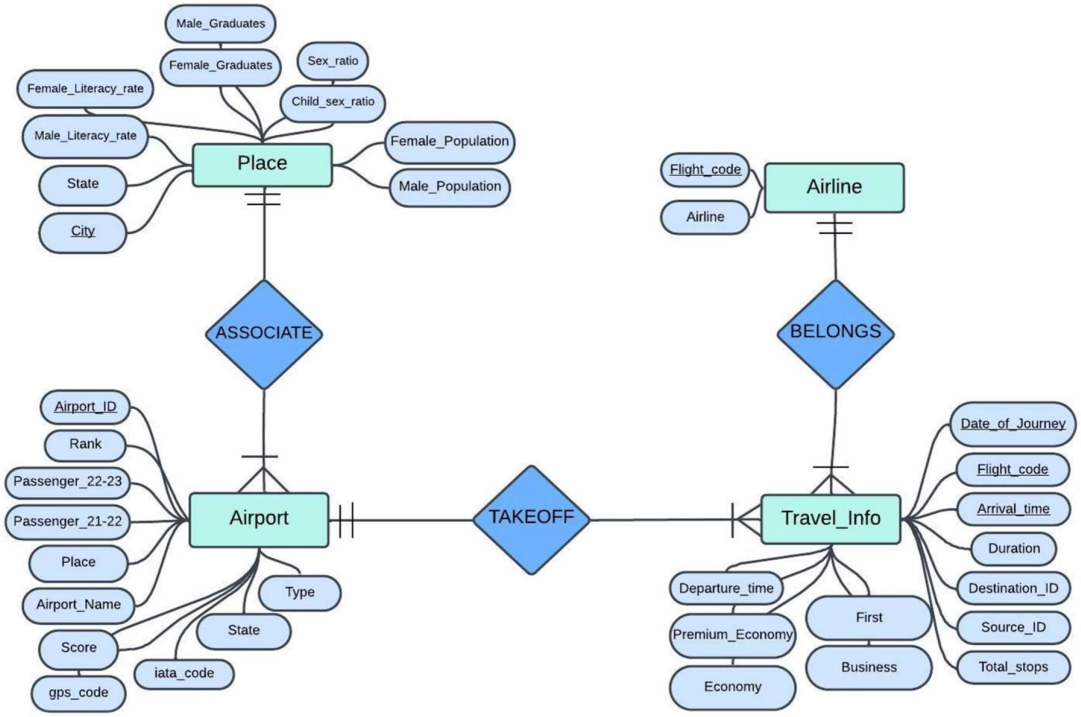

Image 1: ERD

Conceptual modelling rationale:

The above Entity-Relationship Diagram (ERD) is a visual representation of the relationships between the **entities** as follows:

1. Airline
2. Travel_Info
3. Airport
4. Place

Description:

- Airlines-Travel_Info:
    - An Airline can have many Flights taking off.
    - A Flight taking off on given date and time can be associated with single airline.
- Travel_Info-Airport:
    - A Flight can only take off from one airport on given date and time.
    - An Airport can be associated with multiple flights.
- Airport-Place:
    - An Airport can be associated with single Place.
    - A Place can have multiple airports

5. Relational Data Model

The image presents a clear and concise representation of the relational schema, complete with referential integrity constraints and functional dependencies.

- This schema serves as a foundational blueprint for organizing and structuring the database, ensuring that data relationships are maintained with precision and accuracy.
- The Referential integrity constraints guarantee the consistency and reliability of data by enforcing rules that govern the relationships between tables, while functional dependencies provide valuable insights into how attributes within the schema relate to one another.

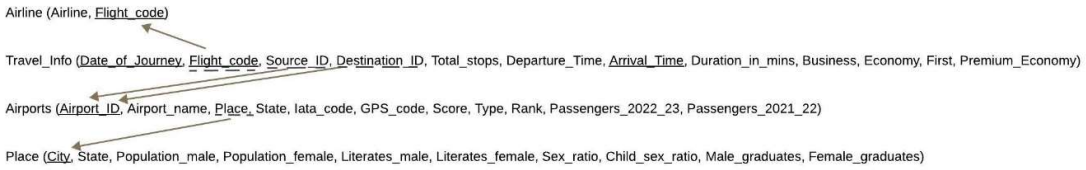

Image 2: Relational Schema

6. Normalization

Below is a snippet of the data on flights that was extracted from Kaggle.

Raw data

Date_of_journey	Journey_day	Airline	Flight_code	Class	Source	Departure	Total_stops	Arrival	Destination	Duration_in_hours	Days_left	Fare	Departure Time	Arrival Time	Duration
1/16/2023	Monday	SpiceJet	SG-8169	Economy	Delhi	After 6 PM	non-stop	After 6 PM	Mumbai	2.0833	1	5335	20:00	22:05	02h 05m
1/16/2023	Monday	Indigo	6E-2519	Economy	Delhi	After 6 PM	non-stop	Before 6 AM	Mumbai	2.3333	1	5899	23:00	01:20	02h 20m
1/16/2023	Monday	GO FIRST	G8-354	Economy	Delhi	After 6 PM	non-stop	Before 6 AM	Mumbai	2.1667	1	5801	22:30	00:40	02h 10m
1/16/2023	Monday	SpiceJet	SG-8709	Economy	Delhi	After 6 PM	non-stop	After 6 PM	Mumbai	2.0833	1	5794	18:50	20:55	02h 05m
1/16/2023	Monday	Air India	AI-805	Economy	Delhi	After 6 PM	non-stop	After 6 PM	Mumbai	2.1667	1	5955	20:00	22:10	02h 10m
1/16/2023	Monday	Air India	AI-605	Economy	Delhi	After 6 PM	non-stop	After 6 PM	Mumbai	2.25	1	5955	21:20	23:35	02h 15m
1/16/2023	Monday	Air India	AI-814	Economy	Delhi	After 6 PM	non-stop	Before 6 AM	Mumbai	2.25	1	5955	22:30	00:45	02h 15m
1/16/2023	Monday	GO FIRST	G8-330	Economy	Delhi	After 6 PM	non-stop	After 6 PM	Mumbai	2.25	1	5899	21:00	23:15	02h 15m
1/16/2023	Monday	SpiceJet	SG-2976	Economy	Delhi	After 6 PM	1-stop	6 AM - 12 PM	Mumbai	14.3333	1	5829	20:10	10:30	14h 20m
1/16/2023	Monday	GO FIRST	G8-346	Economy	Delhi	After 6 PM	non-stop	After 6 PM	Mumbai	2.0833	1	5899	20:00	22:05	02h 05m
1/16/2023	Monday	AirAsia	I5-743	Economy	Delhi	Before 6 AM	1-stop	After 6 PM	Mumbai	14.6667	1	6640	04:55	19:35	14h 40m
1/16/2023	Monday	Indigo	6E-2208	Economy	Delhi	After 6 PM	1-stop	6 AM - 12 PM	Mumbai	8.6667	1	6390	22:45	07:25	08h 40m
1/16/2023	Monday	AirAsia	I5-857	Economy	Delhi	After 6 PM	1-stop	Before 6 AM	Mumbai	5.5833	1	6872	19:05	00:40	05h 35m
1/16/2023	Monday	AirAsia	I5-773	Economy	Delhi	After 6 PM	1-stop	Before 6 AM	Mumbai	6.3333	1	6872	20:00	02:20	06h 20m

Image 3: Raw Data with flights information

There are some transitive functional dependencies that need to be eliminated to achieve 3NF. Furthermore, there are certain columns which convey the same information (Duration and duration in hours) that can be reduced.

Functional dependencies

1. **Date_of_journey** → Journey_day
2. **Flight_code** → Airline
3. **Departure_time** → Departure
4. **Arrival_time** → Arrival
5. **Departure_time, Arrival_time** → Duration

6. **Date_of_journey, Flight_code, Arrival_time** → Source, Destination, Total_stops, Fare

By removing the columns with redundant information and by separating the airlines information to a new table, our final data set looks like the below where all the non-key attributes are dependent on all the three primary keys decided for the table. Furthermore, the fare information is now split to under four different columns based on class.

Date_of_journey	Flight_code	Source_ID	Destination_ID	Total_stops	Departure_Time	Arrival_Time	Duration_in_mins	Business	Economy	First	Premium_Economy
2/1/2023	9I-301	35141	26434	non-stop	9:35	11:30	115	0	3533	0	0
2/1/2023	9I-302	26434	35141	non-stop	17:00	19:00	120	0	3423	0	0
2/1/2023	9I-517	35145	35141	non-stop	19:40	21:15	95	0	3269	0	0
2/1/2023	9I-518	35141	35145	non-stop	22:00	23:35	95	0	3423	0	0
2/1/2023	9I-695	26555	26431	1-stop	13:15	16:20	185	0	4695	0	0
2/1/2023	9I-696	26431	26555	1-stop	16:50	19:55	185	0	4642	0	0
2/1/2023	9I-893	35141	26618	non-stop	7:25	9:00	95	0	4688	0	0
2/1/2023	9I-895	35141	35145	non-stop	7:25	8:50	85	0	3953	0	0
2/1/2023	9I-896	35145	35141	non-stop	9:20	10:45	85	0	3799	0	0

Image 4: Travel_Info table in 3NF form.

Airline table has the flight_codes along with the airlines they belong to. This is in 3NF form. Airports table has Airport_ID as the sole primary key and all the attributes are dependent on this key, hence available in 3NF form. Place table has city as the primary key and all the other attributes are various metrics related specifically to each individual city. This is available in 3NF form, as well.

6. Case Scenarios and Deliverables

Topic	User Case Scenario	Client Deliverable	Key SQL techniques
Economy class Pricing analysis	To determine the average price of economy class tickets for a specific competitor airline, broken down by source location	This analysis will help our client set competitive fares for their economy class seats in different regions.	• Procedures • Aggregations • Advanced Joins
Competitor fare analysis for different airport locations	To analyze and compare the average ticket prices of competitive airlines based on various factors: source place, source airport,	This analysis will offer a panoramic view of the pricing landscape in the Indian aviation market	• Procedures • Advanced Joins

	destination place, and destination airport		
Ranking of destinations based on Departure Flight Volume from an airport	To identify and rank the top five most popular destination airports based on the number of flights departing for each	This analysis aims to highlight the busiest airports in terms of flight operations, providing insights into traffic concentration and operational scale	• Procedures • Advanced Joins
Hierarchical Analysis of Airport Popularity	To establish a hierarchy of airports based on their popularity	This analysis seeks to categorize airports into tiers, providing a structured perspective on their relative importance and influence in the aviation network	• Recursive CTE • Advanced Joins • Subqueries
Airport Ranking Based on Literacy Rates	To identify and rank the top 10 airports based on the literacy rates of males and females in their respective locations	This ranking will provide insights into the educational profile of the catchment areas for these airports, which can be crucial for targeted marketing and service design	• Subqueries
Airport Traffic analysis	To ascertain and rank the top 10 airports based on their passenger traffic	This analysis will shed light on the most frequented airports, which is indicative of high demand and potential profitability for airline routes	• Subqueries

Analysis of Airport Peak Departure Times	To identify the busiest departure times at various airports	This analysis aims to understand peak operational hours, which is vital for managing airport resources and services efficiently.	• Window • CTEs • Aggregations • Advanced Joins
Weekday Flight Frequency Analysis	To determine the average number of flights operating from a specific source to a destination airport on a given weekday	This analysis aims to provide insights into the frequency of flights on popular routes during weekdays, which can be crucial for optimizing flight schedules and understanding market demand	• Time functions • Advanced Joins • Aggregations • CTEs
Popularity Ranking of Airlines by Destination	To rank airlines based on their popularity for specific destinations	This analysis aims to identify which airlines dominate routes, providing insights into market shares and consumer preferences	• Windows • CTEs • Advanced Joins
Airline Ranking Based on Average Flight Frequency from a Specific Source Airport	To rank airlines based on the average number of flights they operate from a given source airport	This analysis aims to identify which airlines are most active at a particular airport, providing insights into their operational focus and market presence at that location.	• Partition • Window • Aggregations
Flight Duration and Pricing	To examine the relationship between	This analysis aims to understand how flight	• CTEs • Aggregations

Correlation Analysis	flight duration and ticket pricing across different airlines	length impacts fare structures, a key component in developing competitive pricing strategies	
Analysis of Traffic by Airport Category and Size	To evaluate the scalability of airports by analyzing passenger counts for domestic and international flights in relation to airport size.	This analysis will help in understanding the capacity and traffic distribution across different sizes of airports, typically with international flights associated with larger airports and domestic flights with smaller ones.	• Regexp • Case when • Aggregations

7. Conclusion

Concluding our comprehensive analysis for the strategic entry of a new airline in the Indian aviation market, we have successfully delivered key insights aligned with our project objective. By meticulously examining data on average economy pricing, flight frequencies, airport traffic, departure times, flight duration-pricing relationships, and airline popularity by destination, we have provided a detailed, data-driven foundation for our client's market entry strategy. These deliverables, grounded in advanced SQL analytics, offer a holistic view of the competitive landscape and consumer behavior, enabling our client to make informed decisions on pricing, route selection, and operational planning. This strategic approach positions the new airline not only to effectively compete with existing market challengers but also to lay a robust groundwork for future scaling to challenge the market leaders, ensuring a strong and sustainable presence in the dynamic Indian aviation sector.

8. Appendix

```
# Query 1 Flight Information by Code and Source
DELIMITER &&
CREATE PROCEDURE GetAvgEconomyPrice (IN Flight_Code VARCHAR(10), IN Source_ID INT)
BEGIN
   SELECT a.Airline AS Airline_Name, ap.Airport_name AS Source_Airport_Name, ap.Place AS Source_Place_Name,
      AVG(ti.Economy) AS Average_Economy_Price
   FROM travel_info ti
   INNER JOIN airlines a
   ON ti.Flight_code = a.Flight_code
   INNER JOIN airports ap
   ON ti.Source_ID = ap.Airport_ID
   WHERE ti.Flight_code = Flight_Code AND ti.Source_ID = Source_ID
   GROUP BY a.Airline, ap.Airport_name, ap.Place;
END &&
DELIMITER ;
CALL GetAvgEconomyPrice('9I-894', 26618);

# Query 2 Flight Information by Source and Destination IDs
DELIMITER &&
CREATE PROCEDURE GetAirlinesWithPlacesAveragePrice(IN Source_ID INT, IN Destination_ID INT)
BEGIN SELECT a.Airline AS Airline_Name, src_ap.Airport_name AS Source_Airport_Name, src_ap.Place AS Source_Place_Name, dest_ap.Airport_name AS
Destination_Airport_Name, dest_ap.Place AS Destination_Place_Name, AVG(ti.Economy) AS Average_Economy_Price
FROM travel_info ti
INNER JOIN airlines a ON ti.Flight_code = a.Flight_code
INNER JOIN airports src_ap ON ti.Source_ID = src_ap.Airport_ID
INNER JOIN airports dest_ap ON ti.Destination_ID = dest_ap.Airport_ID
WHERE ti.Source_ID = Source_ID AND ti.Destination_ID = Destination_ID
GROUP BY a.Airline, src_ap.Airport_name, src_ap.Place, dest_ap.Airport_name, dest_ap.Place;
END && DELIMITER ;

CALL GetAirlinesWithPlacesAveragePrice('35145', '35141');

# Query 3 Top 5 Destinations from Source Airport
DELIMITER &&
CREATE PROCEDURE TopDestinationFromSource(IN SourceAirportID INT)
BEGIN
   SELECT
      dest_ap.Airport_name AS Destination_Airport_Name, dest_ap.Place AS Destination_place,
      COUNT(ti.Flight_code) AS NumberOfFlights
   FROM travel_info ti
   INNER JOIN airports dest_ap ON ti.Destination_ID = dest_ap.Airport_ID
   WHERE ti.Source_ID = SourceAirportID
   GROUP BY ti.Destination_ID, dest_ap.Airport_name, dest_ap.Place
   ORDER BY NumberOfFlights DESC
   LIMIT 5;
END &&
DELIMITER ;
CALL TopDestinationFromSource('26434');

# Query 4 Airport Rank Hierarchical Structure
WITH RECURSIVE AirportHierarchy AS (
   SELECT Airport_ID, Airport_name, Place, Ranking
   FROM airports
   WHERE Ranking = (SELECT MIN(Ranking) FROM airports)
   UNION ALL
   SELECT a.Airport_ID, a.Airport_name, a.Place, a.Ranking
   FROM airports a
   INNER JOIN AirportHierarchy ah ON a.Ranking = ah.Ranking + 1)
SELECT * FROM AirportHierarchy limit 8;

# Query 5 Top Airports Ranked by Male and Female Literates
SELECT a.Airport_ID, a.Airport_name, a.Place, a.ranking,
(SELECT b.literates_male FROM place b WHERE a.Place = b.city) AS literates_male,
(SELECT b.literates_female FROM place b WHERE a.Place = b.city) AS literates_female
FROM airports a
ORDER BY literates_male DESC, literates_female DESC
LIMIT 10;
```

```
# Query 6 Top 10 Airports by Traffic & Gender Population
SELECT
a.Airport_ID, a.Airport_name, a.Place, a.Passengers_2022_23 as passengertraffic_2022_23,
(SELECT b.population_male FROM place b WHERE a.Place = b.city) AS population_male,
(SELECT b.population_female FROM place b WHERE a.Place = b.city) AS population_female
FROM airports a
ORDER BY population_male DESC, population_female DESC
LIMIT 10;

# Query 7 Airports and their busiest departure times
WITH DepartureArrivalTimes AS (
   SELECT a.Airport_name, EXTRACT(HOUR FROM t.Departure_Time) AS DepartureHour,
      COUNT(*) OVER (PARTITION BY a.Airport_ID, EXTRACT(HOUR FROM t.Departure_Time)) AS DepartureCount,
      COUNT(*) OVER (PARTITION BY a.Airport_ID, EXTRACT(HOUR FROM t.Arrival_Time)) AS ArrivalCount
   FROM travel_info t
   JOIN Airports a ON t.Source_ID = a.Airport_ID OR t.Destination_ID = a.Airport_ID
),
RankedDepartureTimes AS (
   SELECT Airport_name, DepartureHour, DepartureCount,
      RANK() OVER (PARTITION BY Airport_name ORDER BY DepartureCount DESC) AS DepartureRank
   FROM DepartureArrivalTimes
)
SELECT DISTINCT d.Airport_name, CONCAT(d.DepartureHour,":00") AS BestDepartureHour
FROM RankedDepartureTimes d
WHERE d.DepartureRank = 1
ORDER BY d.Airport_name;

# Query 8 Average #Flights Source to Destination on Weekday
WITH DailyFlights AS (
   SELECT DAYNAME(STR_TO_DATE(t.Date_of_journey, '%d-%m-%Y')) AS Weekday, s.Airport_name AS SourceAirport,
      d.Airport_name AS DestinationAirport, COUNT(*) AS FlightsOnDay
   FROM travel_info t
   JOIN airports s ON t.Source_ID = s.Airport_ID
                  JOIN airports d ON t.Destination_ID = d.Airport_ID
   GROUP BY Weekday, SourceAirport, DestinationAirport
),
RankedFlights AS ( SELECT Weekday, SourceAirport, DestinationAirport, AVG(FlightsOnDay) AS AvgFlights,
      RANK() OVER (PARTITION BY Weekday ORDER BY AVG(FlightsOnDay) DESC) AS Day_Rank
   FROM DailyFlights
   GROUP BY Weekday, SourceAirport, DestinationAirport
)
SELECT Weekday, SourceAirport, DestinationAirport, AvgFlights
FROM RankedFlights
WHERE Day_Rank <= 3 -- Filter to show only top 3 ranks per weekday
ORDER BY Weekday, Day_Rank;

# Query 9 Ranking Airlines by Popularity per Destination
WITH FlightCounts AS (
SELECT a.Airline AS airline, ti.destination_id AS destination, ap.Place, COUNT(*) AS flights_count
FROM travel_Info ti
JOIN airlines a ON ti.flight_code = a.flight_code
JOIN airports ap ON ap.airport_ID = ti.destination_id
GROUP BY a.Airline, ti.destination_id, ap.Place),

RankedAirlines AS
(SELECT airline, destination, Place, flights_count,
RANK() OVER(PARTITION BY destination ORDER BY flights_count DESC) AS rank_destination
FROM FlightCounts)

SELECT destination, Place, airline, flights_count, rank_destination
FROM RankedAirlines
WHERE rank_destination <= 3;

# Query 10 Airline Ranking by Scheduled Trips per Airport
WITH AirlineFlightCounts AS (
  SELECT ti.Source_ID, a.Airline, COUNT(ti.Flight_code) AS FlightCount,
                  AVG(ti.Economy) AS AvgEconomyPrice
  FROM travel_info ti
```

```
  JOIN airlines a ON ti.Flight_code = a.Flight_code
  GROUP BY ti.Source_ID, a.Airline
),
RankedAirlines AS (
  -- Rank airlines by flight count for each airport
  SELECT Source_ID, Airline, FlightCount, AvgEconomyPrice,
    RANK() OVER (PARTITION BY Source_ID ORDER BY FlightCount DESC) AS rank_airline
  FROM AirlineFlightCounts
)
SELECT ra.Source_ID, ap.Airport_name AS Airport_Name, ra.Airline, ra.FlightCount, ra.AvgEconomyPrice, ra.rank_airline
FROM RankedAirlines ra
JOIN airports ap ON ra.Source_ID = ap.Airport_ID
ORDER BY ra.Source_ID, ra.rank_airline;

# Query 11 Analyzing Flight Duration and Pricing Relationship
CREATE VIEW DurationPricing AS
SELECT ti.destination_ID, AVG(ti.economy) AS average_price, AVG(duration_in_mins) AS average_duration
FROM travel_Info ti
GROUP BY ti.destination_ID; -- Query the view

SELECT dp.destination_ID AS Destination_ID, ap.place AS Destination, dp.average_price, dp.average_duration
FROM DurationPricing dp
JOIN airports ap ON ap.airport_ID = dp.destination_id
ORDER BY average_duration DESC;

# Query 12 Analysis of Traffic by Airport Category and Size
SELECT
    CASE
        WHEN Airport_name REGEXP '[Ii]nternational' THEN 'International'
        ELSE 'Domestic'
    END AS Airport_Category,
    type AS Size_Category,
    SUM(Passengers_2022_23) AS Total_Passengers_2022_23,
    SUM(Passengers_2021_22) AS Total_Passengers_2021_22,
    ROUND(
        ((SUM(Passengers_2022_23) - SUM(Passengers_2021_22)) / SUM(Passengers_2021_22)) * 100, 2
    ) AS Percentage_Growth
FROM airports
GROUP BY Airport_Category, Size_Category
ORDER BY Airport_Category, Size_Category;
```