Airline Flight Analysis

Project 3 - Group 6

Camden Beck, Kelvin Osei Assibey, Jennifer Vega, Zilan Yidil

January 2025

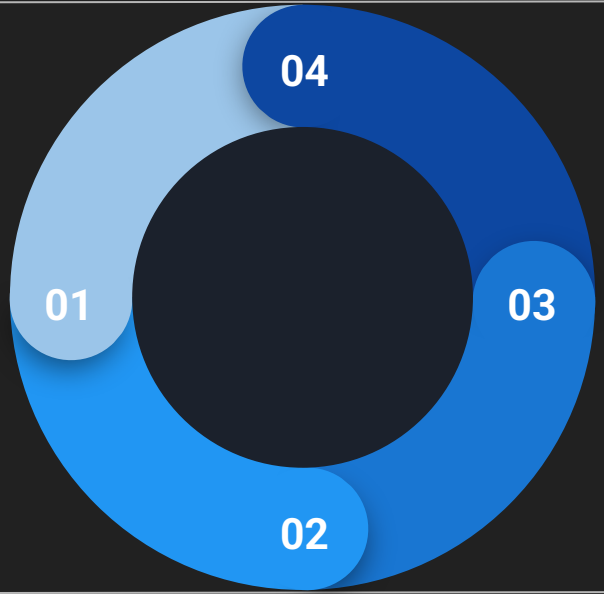
Group 6

Jennifer

Zilan

Camden

Kelvin



Overview

Design and implement a comprehensive ETL pipeline for integrating and analyzing airline flight data to provide accurate, unified datasets for analytical and reporting purposes.

Understanding the problems

Passenger Demand & Booking Trends

How does flight demand fluctuate across different seasons?
Booking trends could be used to boost flight times/ costs?

Ticket Sales & Pricing

What correlation exists between demand for specific airlines and ticket fees?

Financial Insights

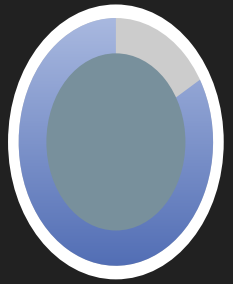
Which flight paths yield the highest profits?

Project objective

1. Extract data from airline schedules, and flight tracking datasets.
2. Transform the data by cleaning and normalizing it.
3. Load the data into a centralized dataset.
4. Develop analytical visualizations for insights.

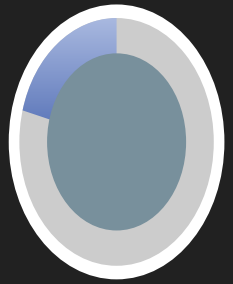
Flight Dataset Cleaning

Data Extraction



Remove Duplicates, N/A Values, and fix Inconsistencies

Data Transformation



Validate and Save Cleaned Dataset

Database

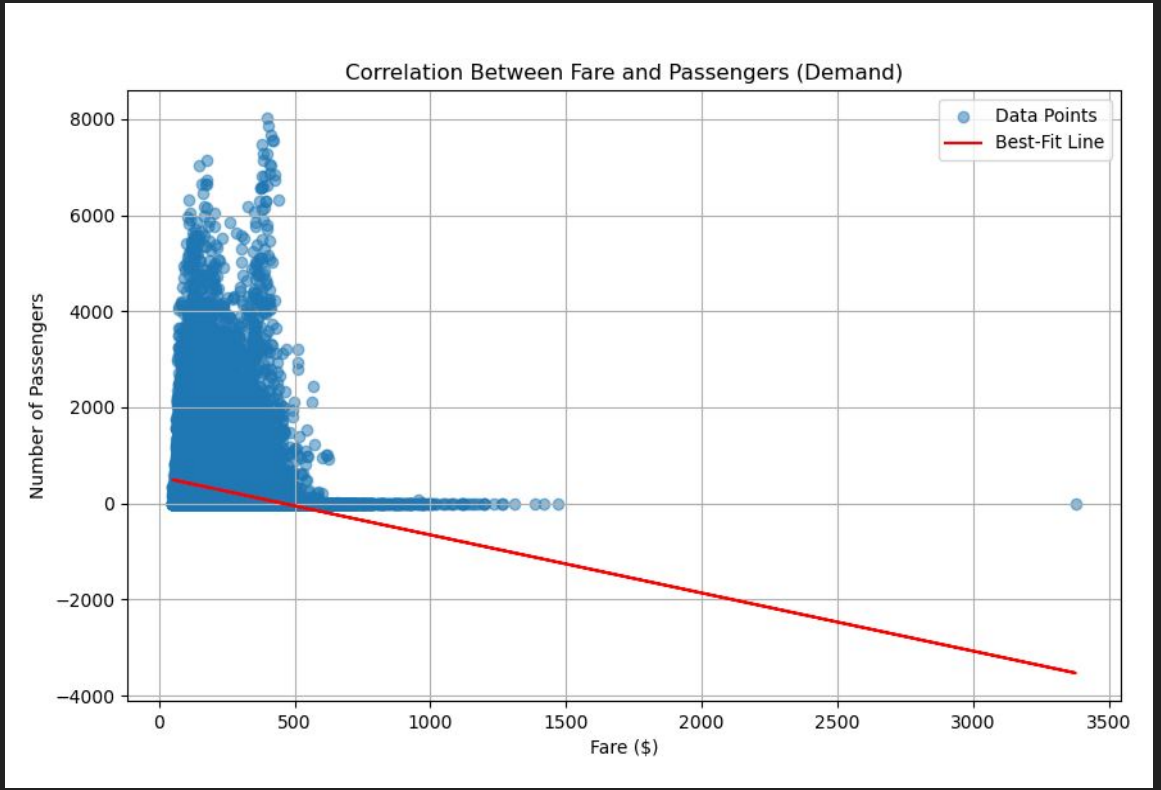
- The cleaned data was then saved to a csv file.
- Then we imported it into MongoDB in order to get a better understanding of it.
- The reason we chose MongoDB is that it displays data in the JSON format by default.
- The data was then further analysed and we added two collections:
  - flights
  - locations
- The two collections were then saved into JSON files which were used for the web-application.

Correlation between Fare and Passengers

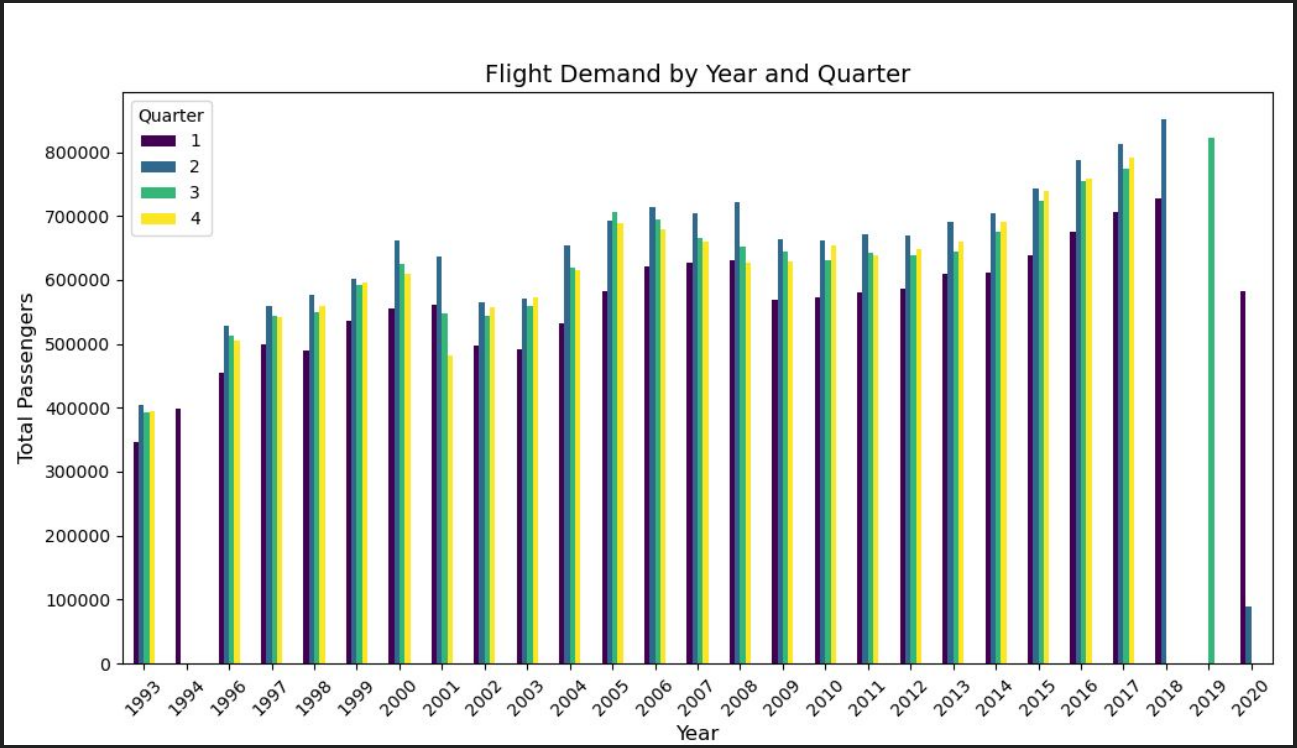
Correlation: -0.19278
P-value: 0.00000

The Correlation being -0.192 indicates that there is a weak negative correlation—as fares increase, passenger demand tends to decrease slightly.

The p-value of 0.00000 suggests that the correlation is statistically significant, meaning the trend is unlikely due to random chance.



Overview of Flight demand by Year and Quarter



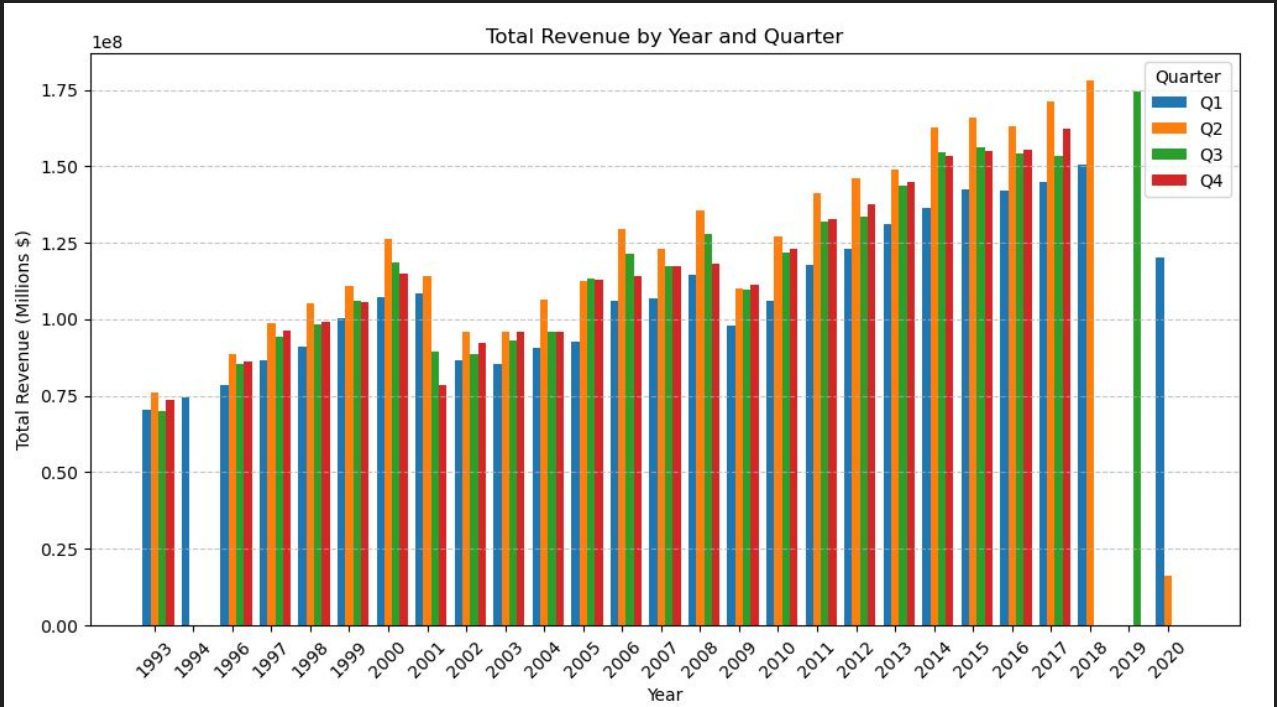
Flight Demand Analysis:

Year with Highest Flight Demand: 2017 with 3,086,637 passengers.

Quarter with Highest Flight Demand in 2017: Q2 (April - June) with 813,003 passengers.

Statistical Insight:
The highest flight demand in 2017 occurred in Q2, which could be due to increased travel during summer vacation months, business travel, or public holidays. These periods often see a surge in passengers.

Overview of Total Revenue by Year and Quarter



Revenue Analysis:

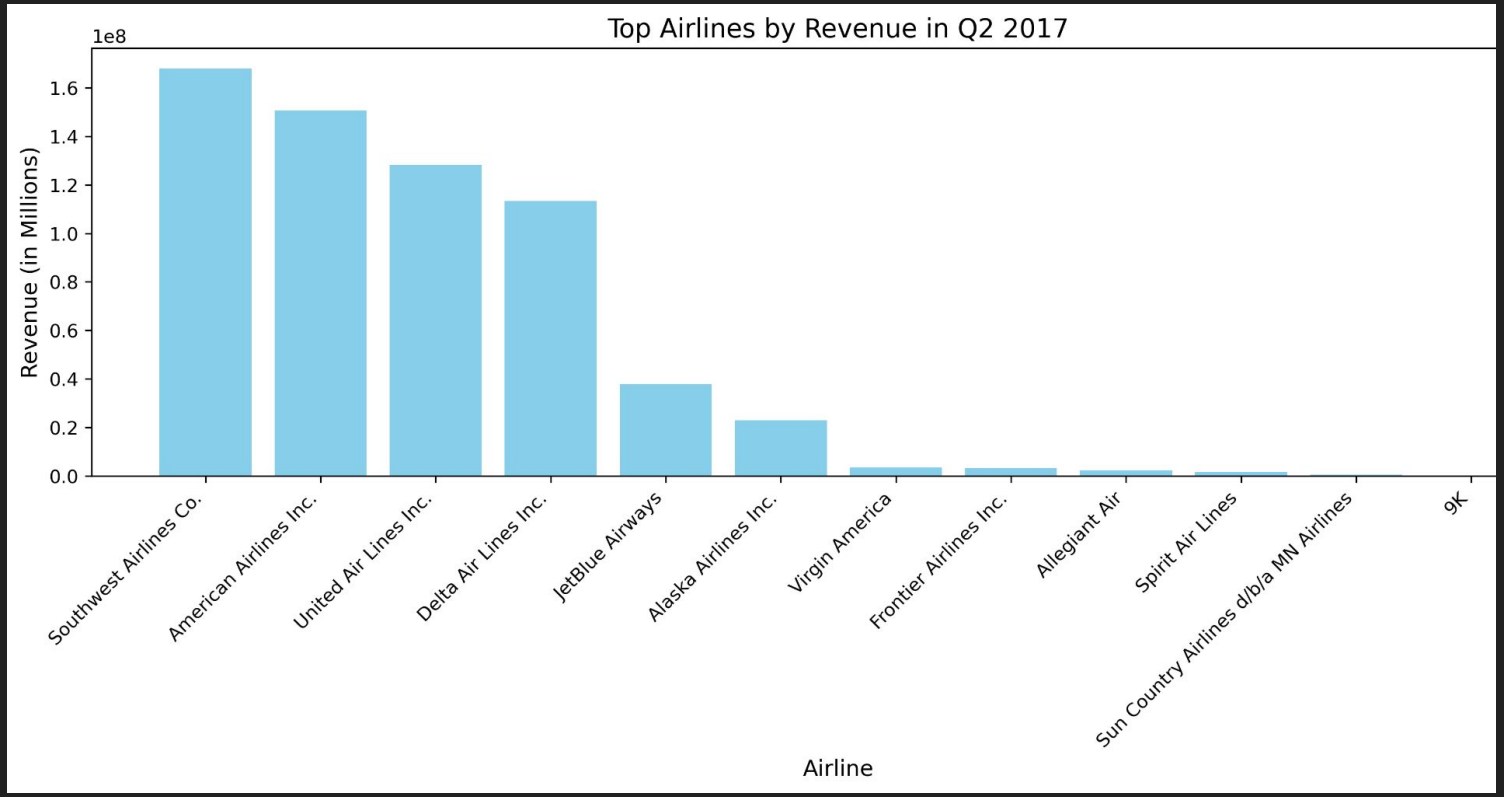
Year with the Best Revenue: 2017 with a total revenue of $632,009,964.43.

Quarter with the Best Revenue within 2017: **Q2** (April - June) with a revenue of $171,378,275.99.

Correlation between Flight Demand and Revenue:

0.9084048843871347

Top Airlines by Revenue in Q2



Q2 of 2017 generated the highest revenue, likely due to increased flight demand during the summer travel season.

Between 1993 and 2020, a total of 205,189 flights were operated within the United States.

The top 10 cities with the highest number of departures and arrivals, categorized by year, are as follows:

This version clarifies that both departures and arrivals are being considered and organizes the sentence in a clearer manner.

Los Angeles, CA (Metropolitan Area) 21527

New York City, NY (Metropolitan Area) 19718

Boston, MA (Metropolitan Area) 19674

Chicago, IL 15546

Dallas/Fort Worth, TX 12142

Houston, TX 11317

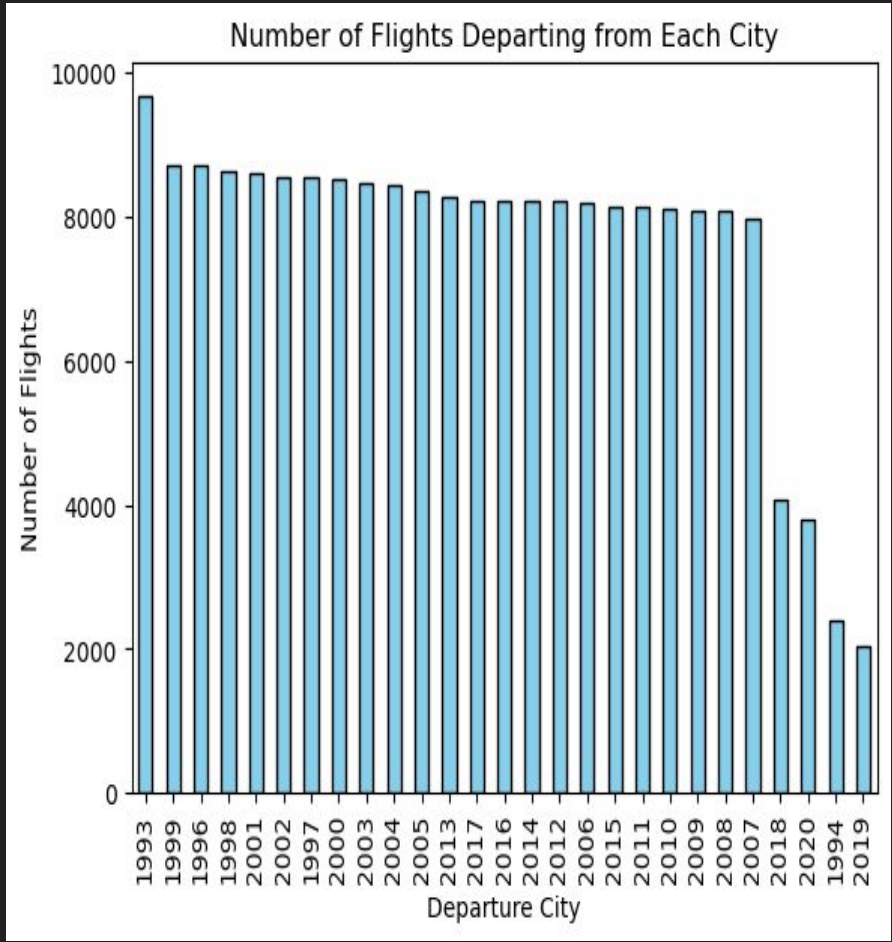
Cleveland, OH (Metropolitan Area) 7697

Miami, FL (Metropolitan Area) 6746

Atlanta, GA (Metropolitan Area) 3516

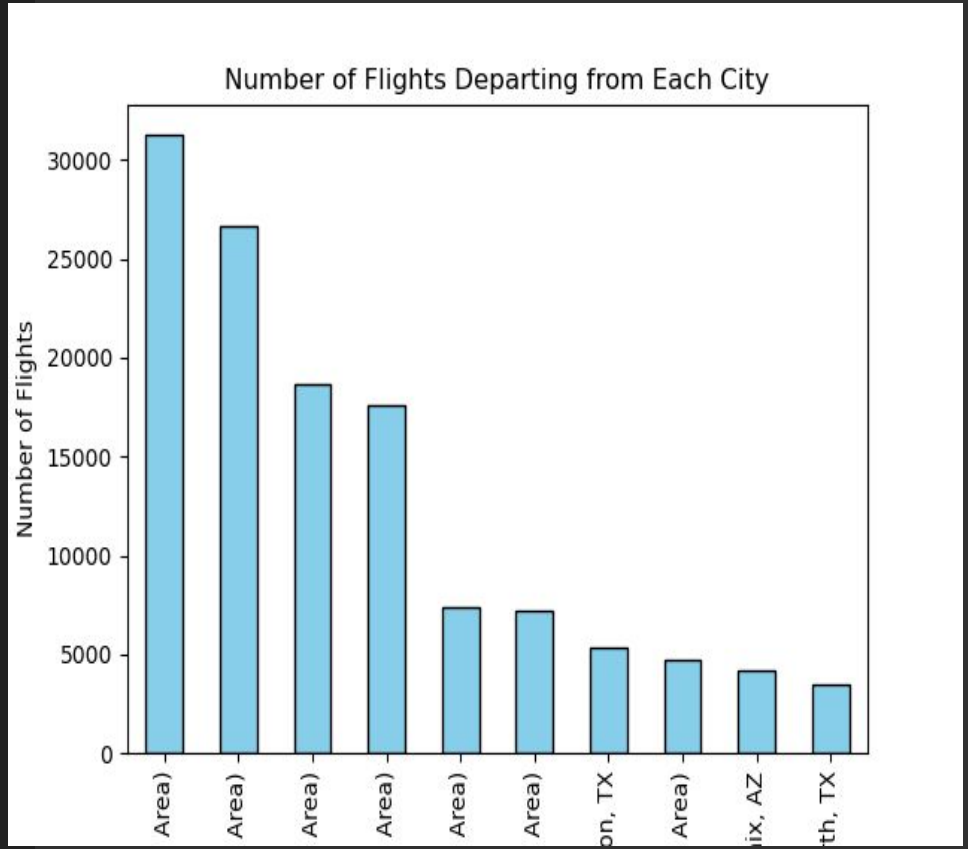
Detroit, MI 3390

The departure_city column contains a total of 136 distinct cities.



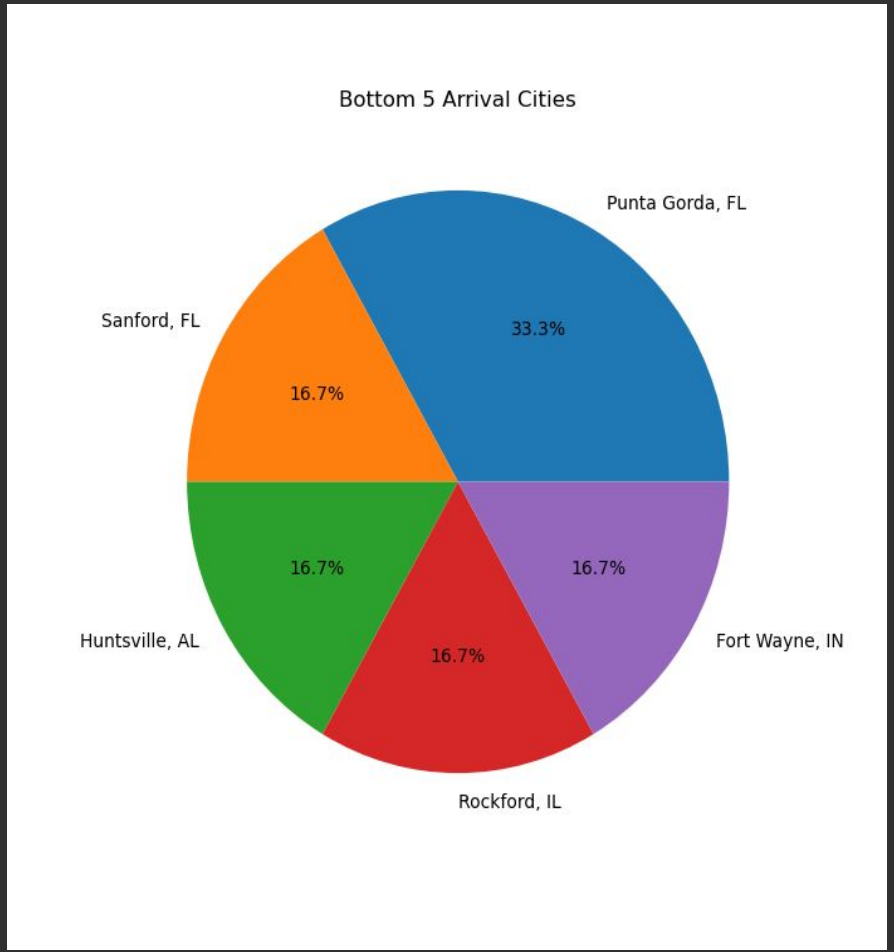
New York City, NY (Metropolitan Area) 31247
Washington, DC (Metropolitan Area) 26680
San Francisco, CA (Metropolitan Area) 18625
Los Angeles, CA (Metropolitan Area) 17555
Tampa, FL (Metropolitan Area) 7352
Miami, FL (Metropolitan Area) 7238
Houston, TX 5321
Norfolk, VA (Metropolitan Area) 4686
Phoenix, AZ 4158
Dallas/Fort Worth, TX 3443

The **arrival_city** column contains a total of 131 distinct cities.



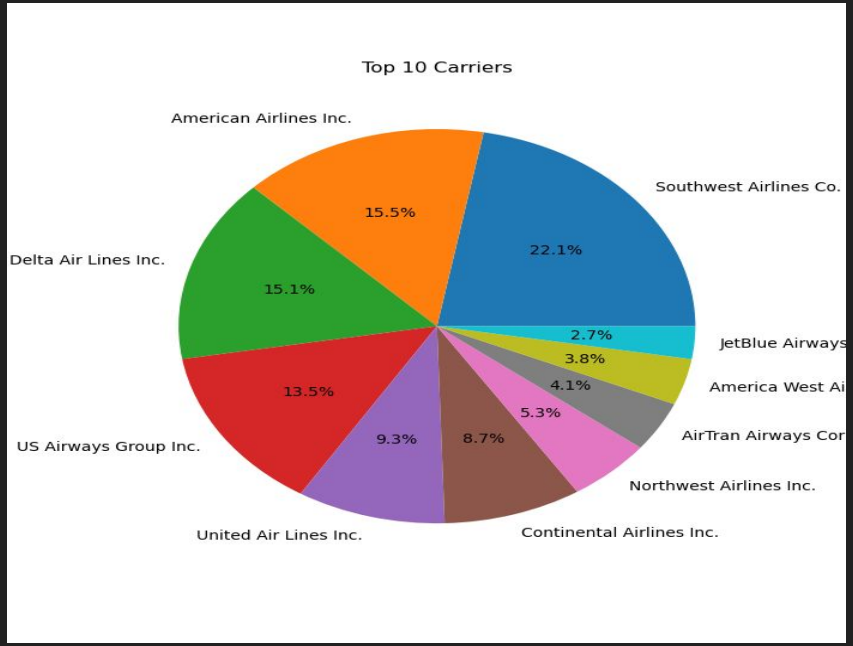
The five cities with the fewest flights during this period are as follows:

Punta Gorda, FL	2
Sanford, FL	1
Huntsville, AL	1
Rockford, IL	1
Fort Wayne, IN	1



The ten most frequently used airline carriers and the number of times each was preferred are as follows:

Southwest Airlines Co.	40663
American Airlines Inc.	28446
Delta Air Lines Inc.	27748
US Airways Group Inc.	24811
United Air Lines Inc.	17141
Continental Airlines Inc.	15936
Northwest Airlines Inc.	9728
AirTran Airways Corporation	7544
America West Airlines Inc.	6973
JetBlue Airways	4922



The top 10 lowest-cost flights, along with their departure cities, carriers, and low-fare ticket details, are as follows:

29392	Dallas/Fort Worth, TX	Southwest Airlines Co.
174885	New York City, NY (Metropolitan Area)	American Airlines Inc.
35670	Chicago, IL	United Air Lines Inc.
138939	Dallas/Fort Worth, TX	Southwest Airlines Co.
123097	Chicago, IL	Southwest Airlines Co.
52238	Cleveland, OH (Metropolitan Area)	DH
186280	Los Angeles, CA (Metropolitan Area)	JetBlue Airways
96861	New York City, NY (Metropolitan Area)	Continental Airlines Inc.
44548	Dallas/Fort Worth, TX	Southwest Airlines Co.
119875	Dallas/Fort Worth, TX	Southwest Airlines Co.

29392	50.00
174885	50.10
35670	50.40
138939	50.41
123097	50.50
52238	50.50
186280	50.60
96861	50.60
44548	50.65
119875	50.72

The most preferred company is Southwest Airlines Co., 40,663 times

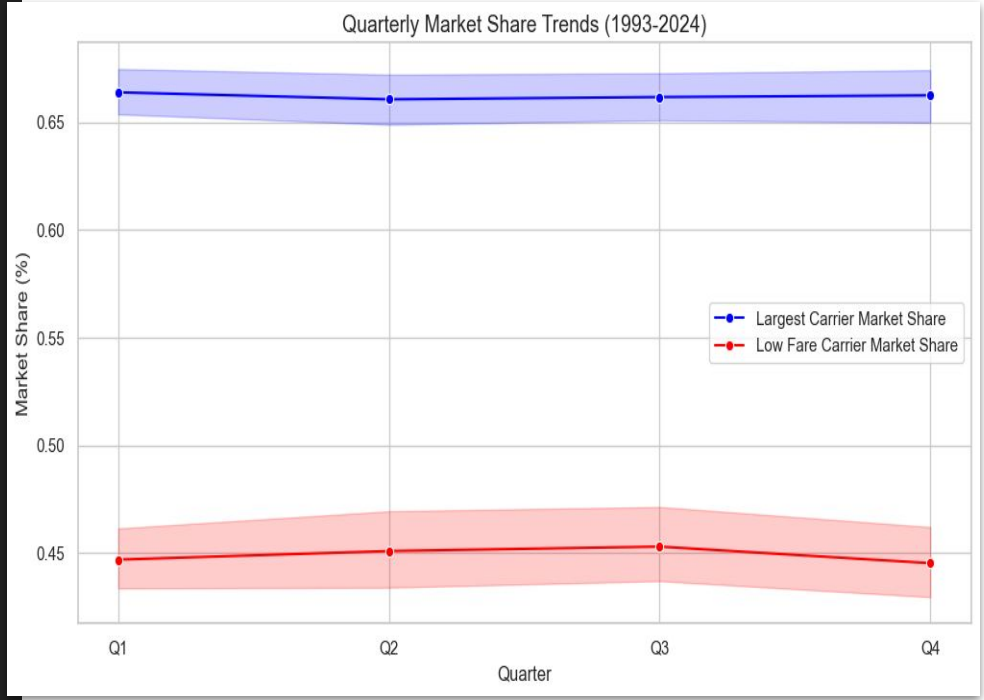
Quarterly Market share trends (1993-2024)

Observations

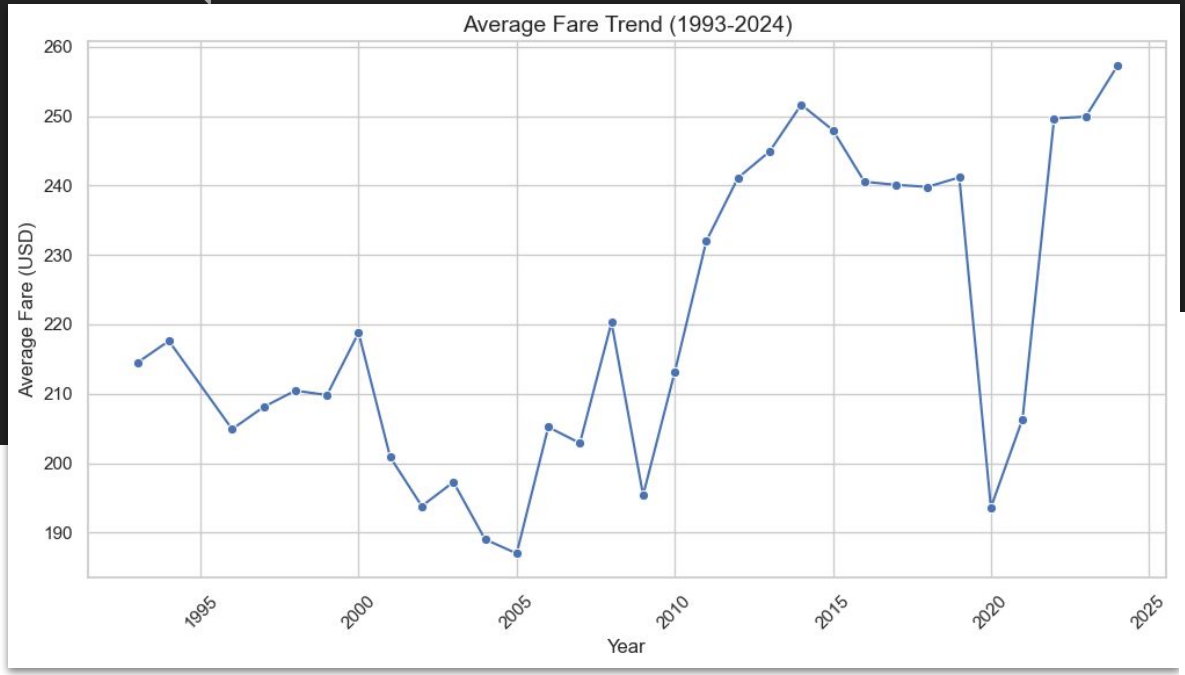
- Stability in market share of the largest carrier.
- Fluctuations in low-fare carrier market share.
- Competitive landscape & seasonal trends

Implications

- Highly stable market for the dominant airline
- Demand for strategic pricing, fleet planning and competitive position in different quarters.



Analysis of average fare trend (1993-2024)



Observations

- Early Decline (1993 - Early 2000s)
- Early 2000s - 2010: Volatility & Recovery
- Strong Growth (2010 - Mid-2010s)
- Stabilization & Pre-Pandemic Trends (2015 - 2019)
- Pandemic Disruption (2020 - 2021)
- Rapid Post-Pandemic Rebound (2022 - 2024)

Conclusions & Future outlook

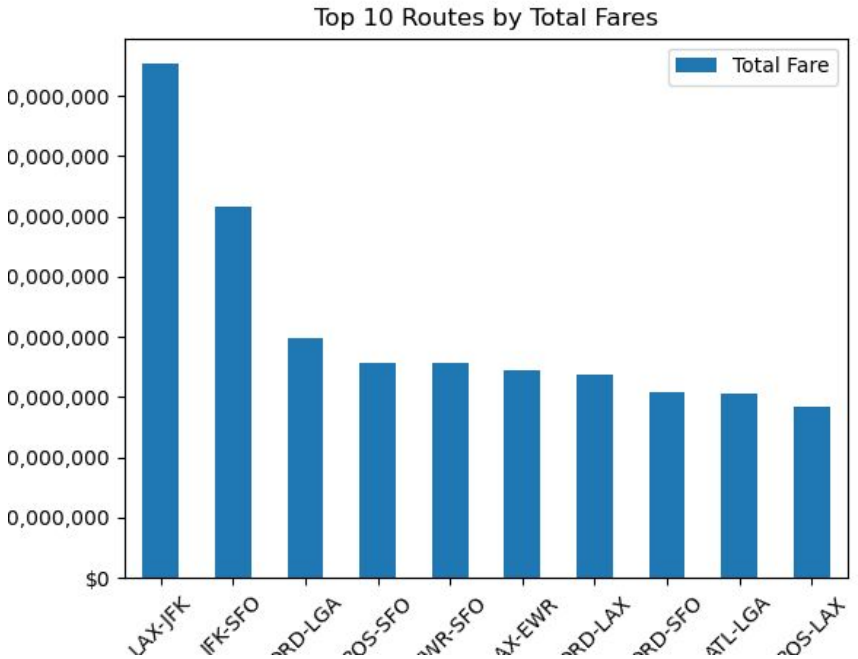
- Fares have reached their highest levels in 2024.
- If trends continue, we may see further increases.
- Low-cost carriers may still disrupt pricing

Most Popular Flight Paths

LAX-JFK:	$17,069,366,187.48
JFK-SFO:	$12,342,920,693.84
ORD-LGA:	$7,952,891,598.28
BOS-SFO:	$7,157,493,317.36
EWR-SFO:	$7,114,535,897.52
LAX-EWR:	$6,911,172,558.72
ORD-LAX:	$6,745,578,207.84
ORD-SFO:	$6,177,967,330.08
ATL-LGA:	$6,111,064,595.95
BOS-LAX:	$5,692,233,083.40

Least Popular Flight Paths

FLL-TSS
BZN-SWF
FMY-BDL
FMY-DCA
FMY-JFK
DAL-XNA
CAE-LGB
FTW-SFO
MHT-SWF
ORD-HFD



Visualizations in the Web-Application

Libraries included in the webpage include:

- Bootstrap (for the layout)
- Plotly (for the display)

Features include:

- Interactive plots
- Filtering by airline
- Total flights per airline
- Filtering by year
- Total passengers per airline
- Quarterly trends

References:

Datasets used:

- https://www.kaggle.com/datasets/bhavikjikadara/us-airline-flight-routes-and-fares-1993-2024?resource=download

In developing this project, we have taken several steps to ensure ethical considerations are addressed. The data used in this project is sourced from publicly available datasets, ensuring that all information is in the public domain and complies with data privacy regulations.

Thank you!