AIRLINE DATA ANALYSIS

USA

-

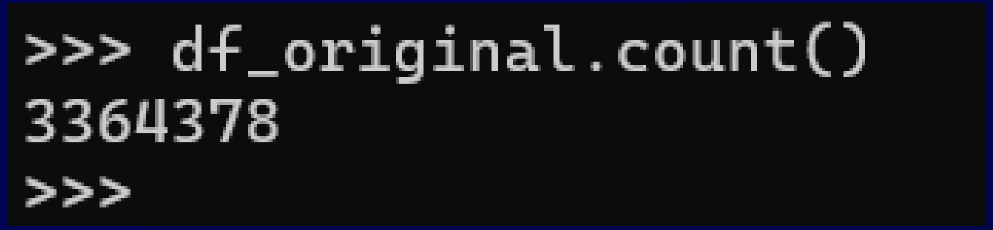
Narasimha Royal

OBJECTIVE

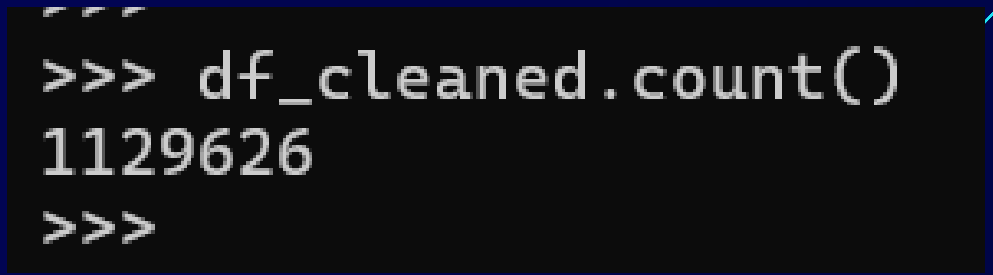
- Understand trends in international airline traffic over time
- Analyze passenger, freight, seat, and departure data
- Identify top carriers, busiest routes, and seasonal patterns
- Explore insights to help airlines or airports improve planning

ABOUT DATASET

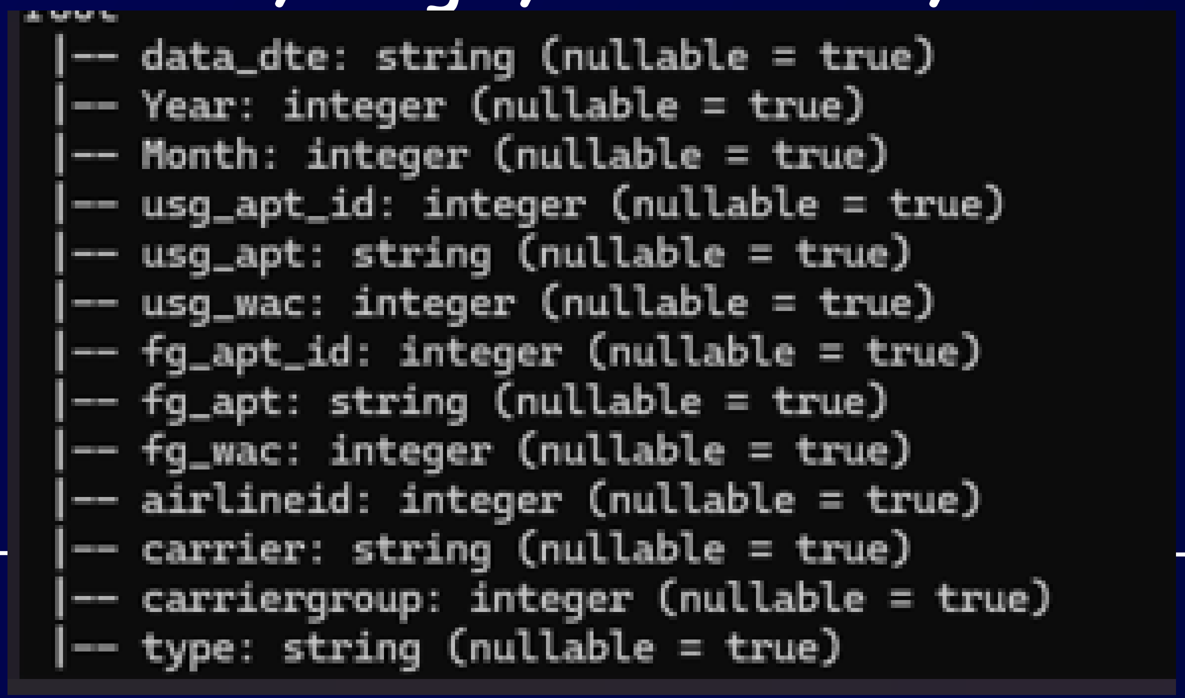
- Source: Transportation Statistics
- Dataset: Passengers Freight All Types
- Rows: 3.72 million
- Time Range: 1990-2024
- Columns: Year, Month, Carrier, Type, Total, Scheduled, Charter, Origin, Destination, etc.



```
>>> df_original.count()
3364378
>>>
```



```
>>> df_cleaned.count()
1129626
>>>
```



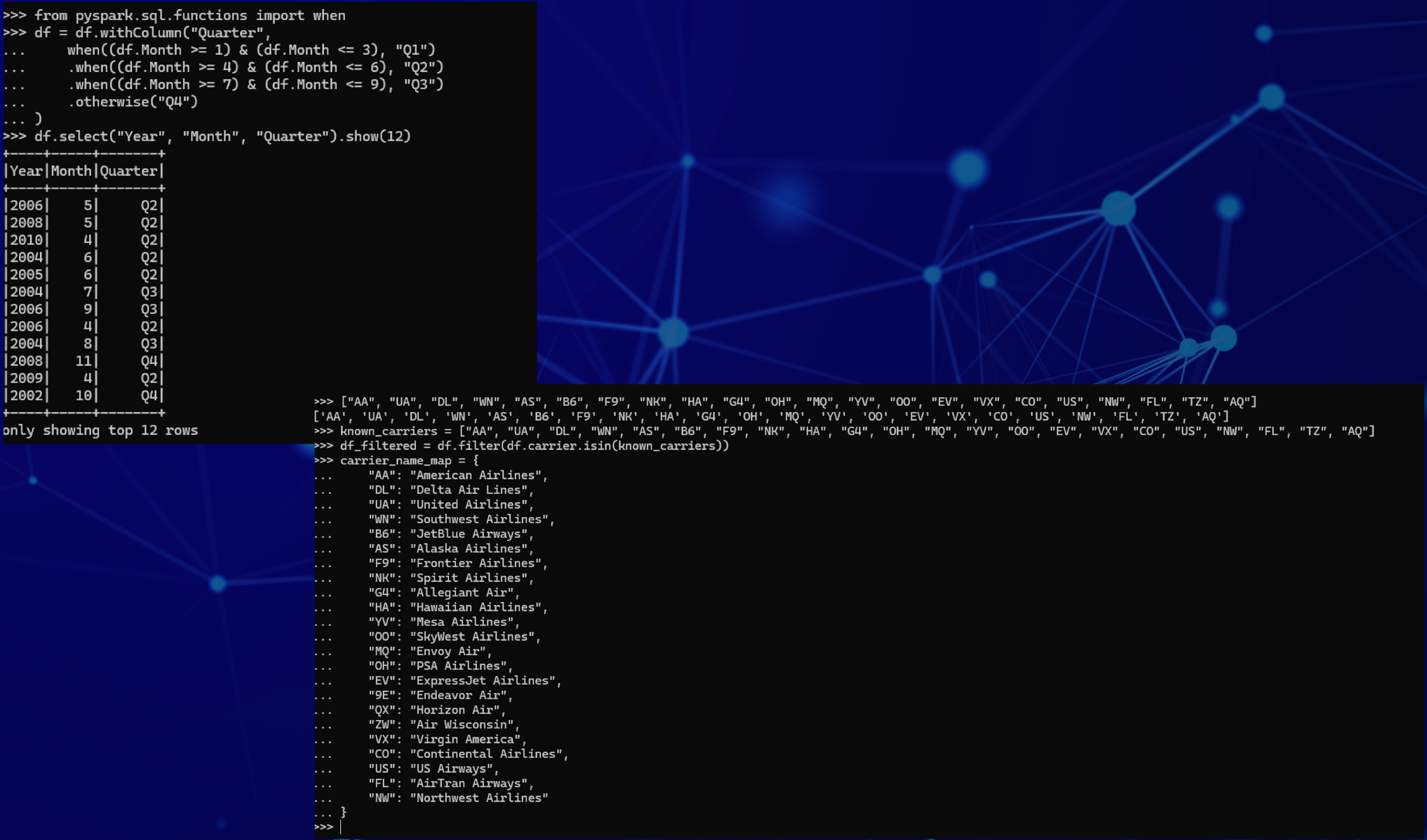
```
 |-- data_dte: string (nullable = true)
 |-- Year: integer (nullable = true)
 |-- Month: integer (nullable = true)
 |-- usg_apt_id: integer (nullable = true)
 |-- usg_apt: string (nullable = true)
 |-- usg_wac: integer (nullable = true)
 |-- fg_apt_id: integer (nullable = true)
 |-- fg_apt: string (nullable = true)
 |-- fg_wac: integer (nullable = true)
 |-- airlineid: integer (nullable = true)
 |-- carrier: string (nullable = true)
 |-- carriergroup: integer (nullable = true)
 |-- type: string (nullable = true)
```

TOOLS AND TECHNIQUES

1. **Apache Spark for cleaning and processing**
2. **Pandas + matplotlib for quick plots**
3. **Tableau Public for professional visualizations**
4. **Jupyter Notebook for model building**

DATA CLEANING & PREPROCESSING

- Dropped redundant ID and WAC columns
- Verified no null values
- Created new columns: Quarter, Total_Recalculated, carrier_full_name
- Mapped airline codes to readable names

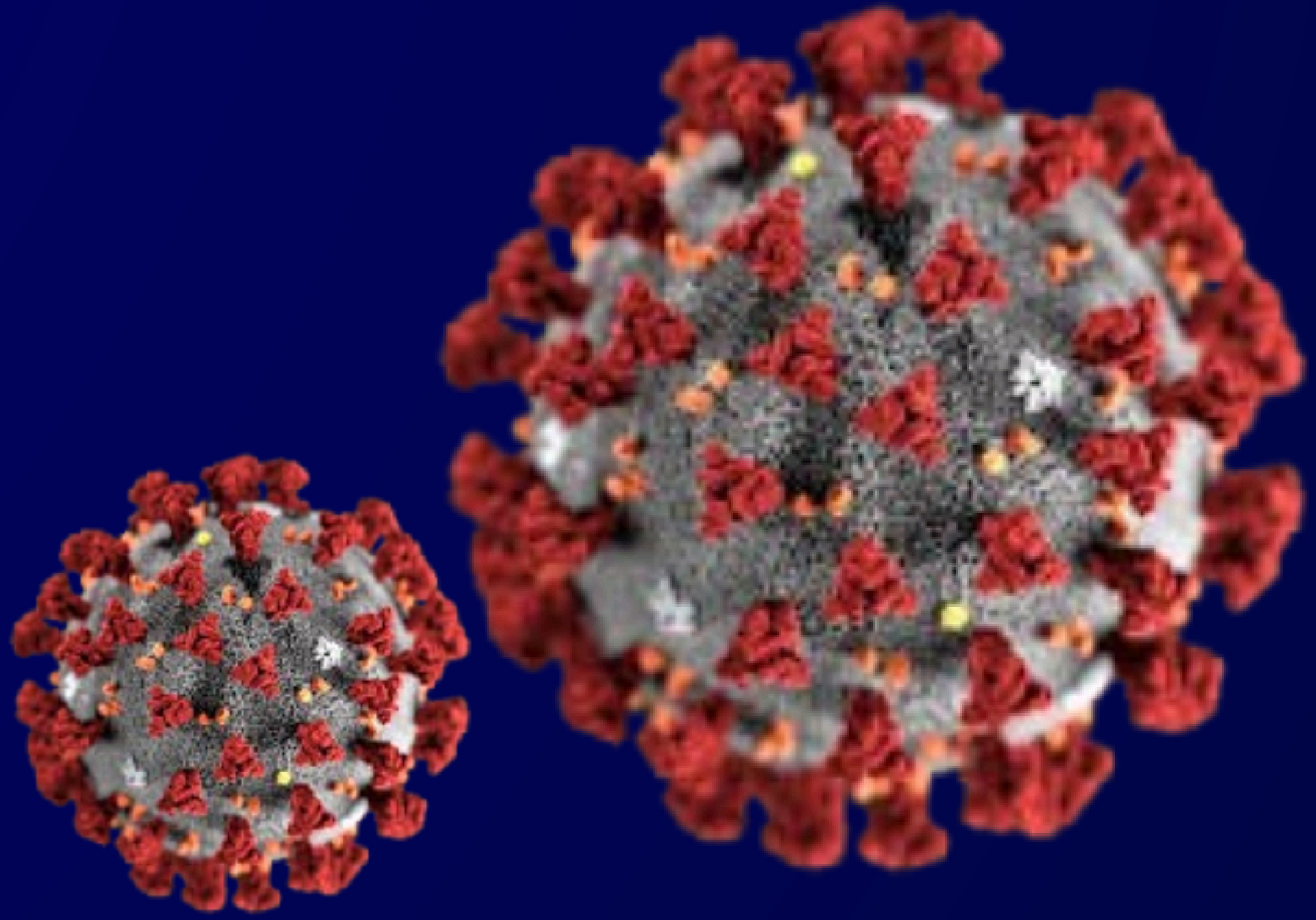


```
>>> from pyspark.sql.functions import when
>>> df = df.withColumn("Quarter",
...     when((df.Month >= 1) & (df.Month <= 3), "Q1")
...     .when((df.Month >= 4) & (df.Month <= 6), "Q2")
...     .when((df.Month >= 7) & (df.Month <= 9), "Q3")
...     .otherwise("Q4")
... )
>>> df.select("Year", "Month", "Quarter").show(12)
+----+-----+-------+
|Year|Month|Quarter|
+----+-----+-------+
|2006|    5|     Q2|
|2008|    5|     Q2|
|2010|    4|     Q2|
|2004|    6|     Q2|
|2005|    6|     Q2|
|2004|    7|     Q3|
|2006|    9|     Q3|
|2006|    4|     Q2|
|2004|    8|     Q3|
|2008|   11|     Q4|
|2009|    4|     Q2|
|2002|   10|     Q4|
+----+-----+-------+
only showing top 12 rows
```

```
>>> ["AA", "UA", "DL", "WN", "AS", "B6", "F9", "NK", "HA", "G4", "OH", "MQ", "YV", "OO", "EV", "VX", "CO", "US", "NW", "FL", "TZ", "AQ"]
['AA', 'UA', 'DL', 'WN', 'AS', 'B6', 'F9', 'NK', 'HA', 'G4', 'OH', 'MQ', 'YV', 'OO', 'EV', 'VX', 'CO', 'US', 'NW', 'FL', 'TZ', 'AQ']
>>> known_carriers = ["AA", "UA", "DL", "WN", "AS", "B6", "F9", "NK", "HA", "G4", "OH", "MQ", "YV", "OO", "EV", "VX", "CO", "US", "NW", "FL", "TZ", "AQ"]
>>> df_filtered = df.filter(df.carrier.isin(known_carriers))
>>> carrier_name_map = {
...     "AA": "American Airlines",
...     "DL": "Delta Air Lines",
...     "UA": "United Airlines",
...     "WN": "Southwest Airlines",
...     "B6": "JetBlue Airways",
...     "AS": "Alaska Airlines",
...     "F9": "Frontier Airlines",
...     "NK": "Spirit Airlines",
...     "G4": "Allegiant Air",
...     "HA": "Hawaiian Airlines",
...     "YV": "Mesa Airlines",
...     "OO": "SkyWest Airlines",
...     "MQ": "Envoy Air",
...     "OH": "PSA Airlines",
...     "EV": "ExpressJet Airlines",
...     "9E": "Endeavor Air",
...     "QX": "Horizon Air",
...     "ZW": "Air Wisconsin",
...     "VX": "Virgin America",
...     "CO": "Continental Airlines",
...     "US": "US Airways",
...     "FL": "AirTran Airways",
...     "NW": "Northwest Airlines"
... }
>>> |
```

YEARLY PASSENGER TRENDS

- Steady growth from 1990–2019
- Sharp drop in 2020 due to COVID
- Strong recovery post-2021

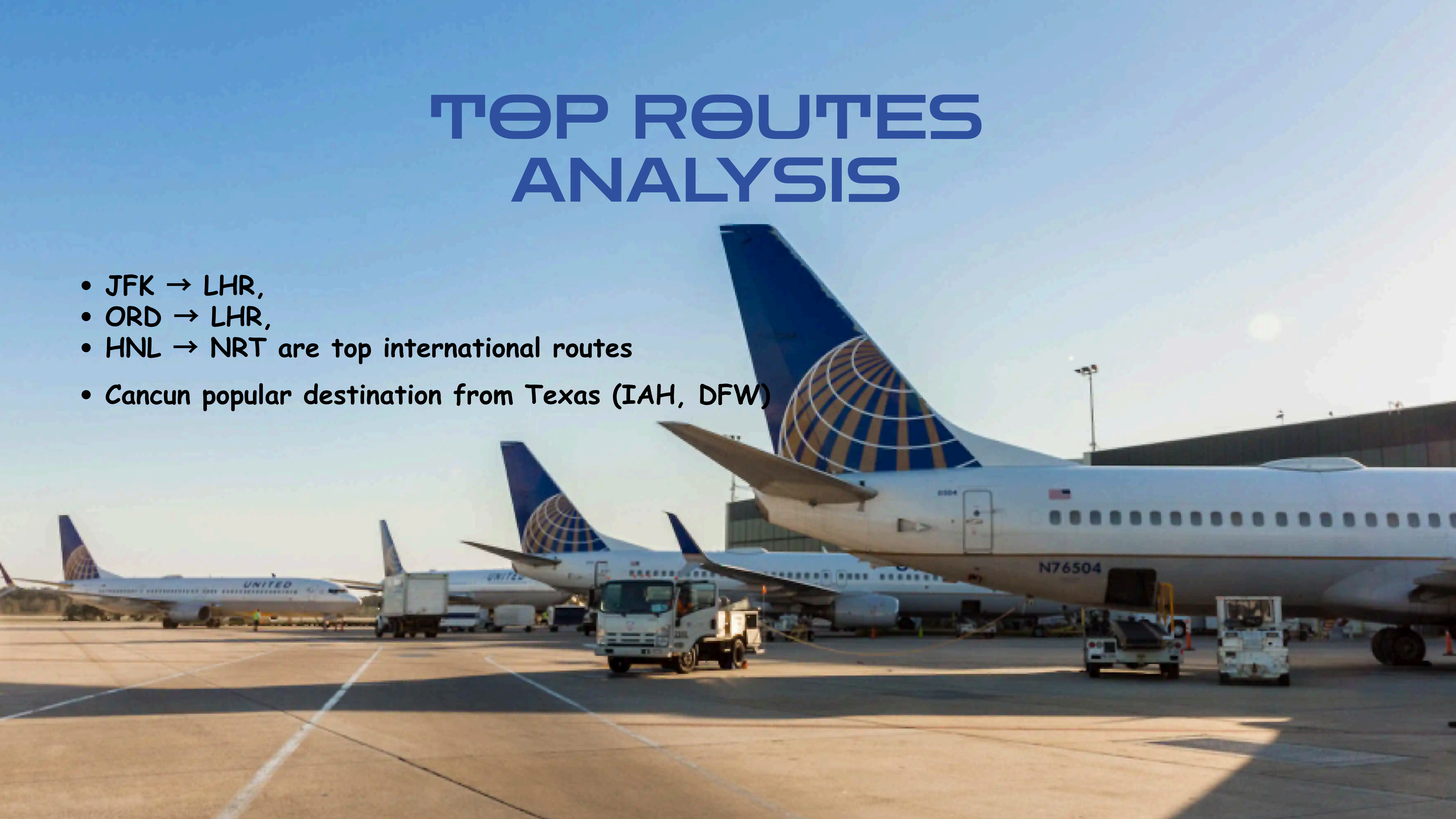


TOP AIRLINES BY PASSENGER VOLUME

- American, United, Delta lead overall
- JetBlue & Alaska hold strong in mid-tier
- Spirit & others trail in volume

TOP ROUTES ANALYSIS

- JFK → LHR,
- ORD → LHR,
- HNL → NRT are top international routes
- Cancun popular destination from Texas (IAH, DFW)



SEASONAL TRENDS (QUARTERLY)

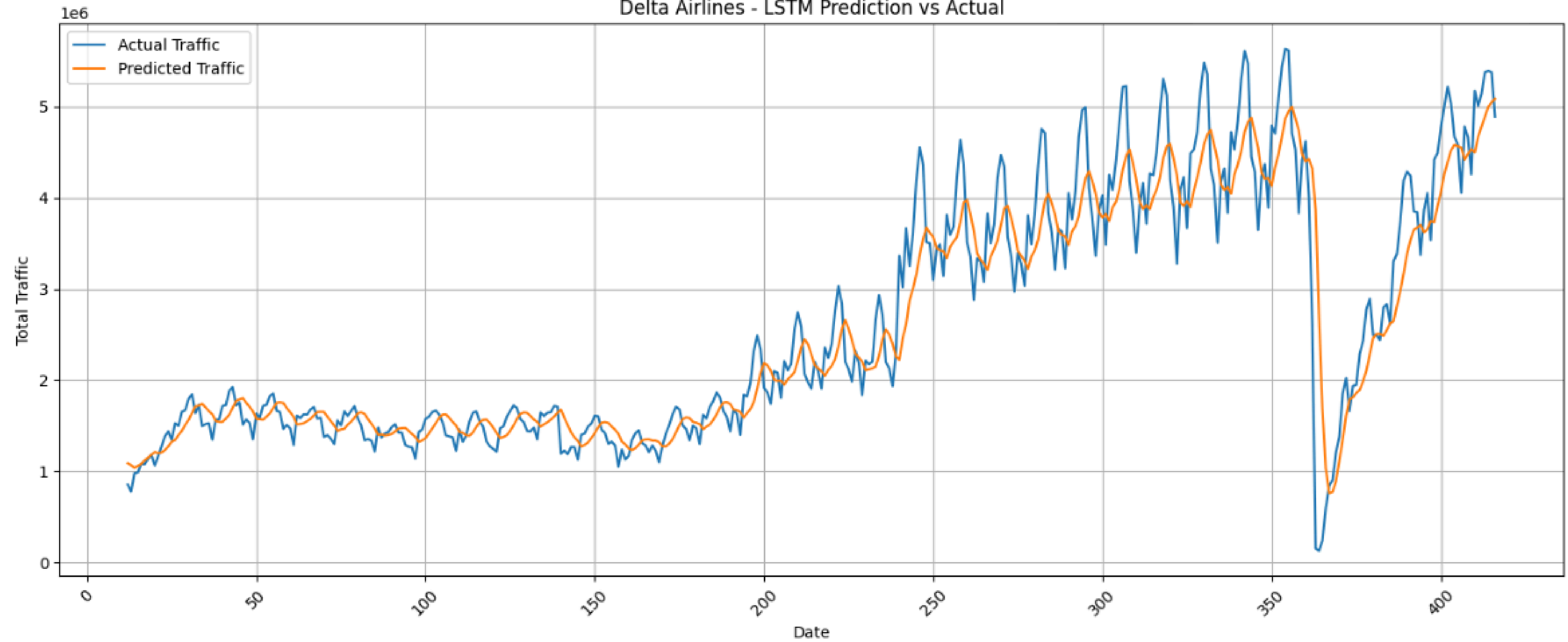
1. Q3 (summer) shows highest passenger counts
2. Q1 typically lowest
3. Patterns consistent across years

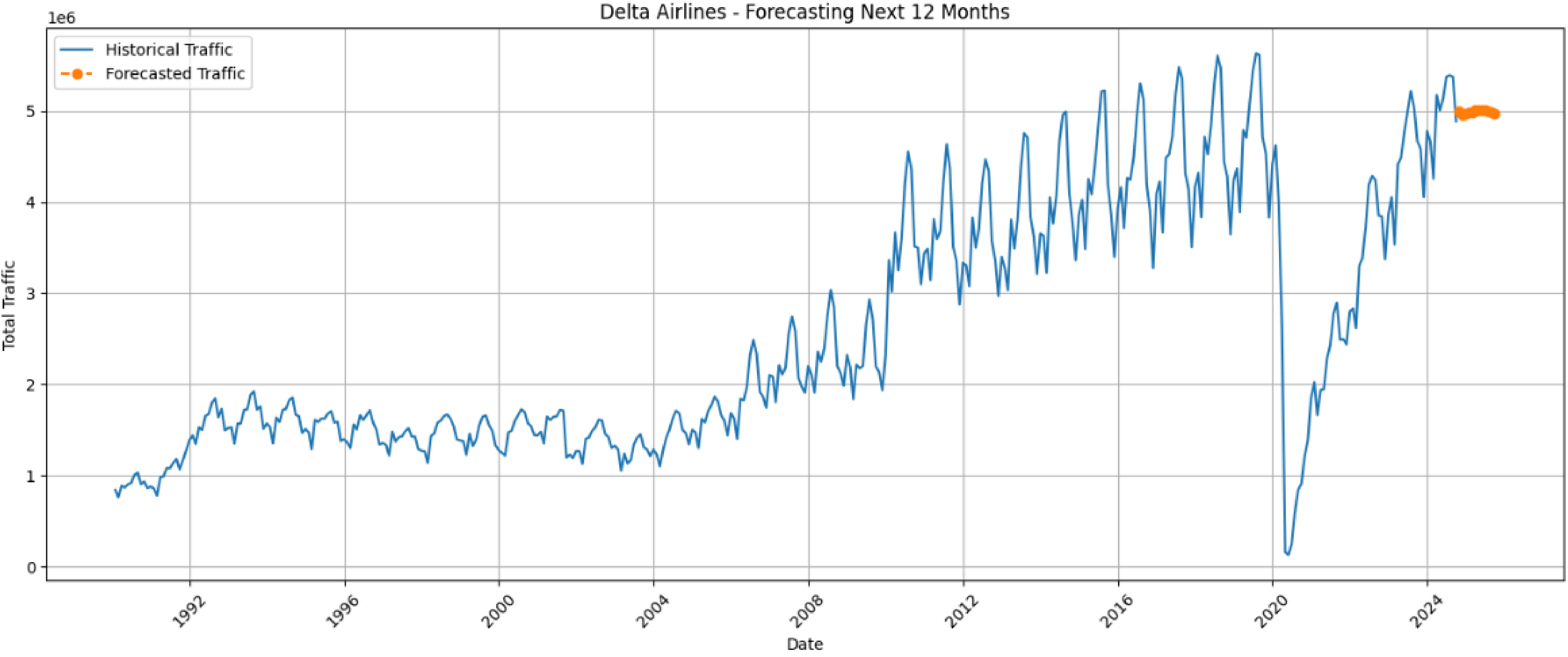
LSTM MODEL ON DELTA AIRLINES

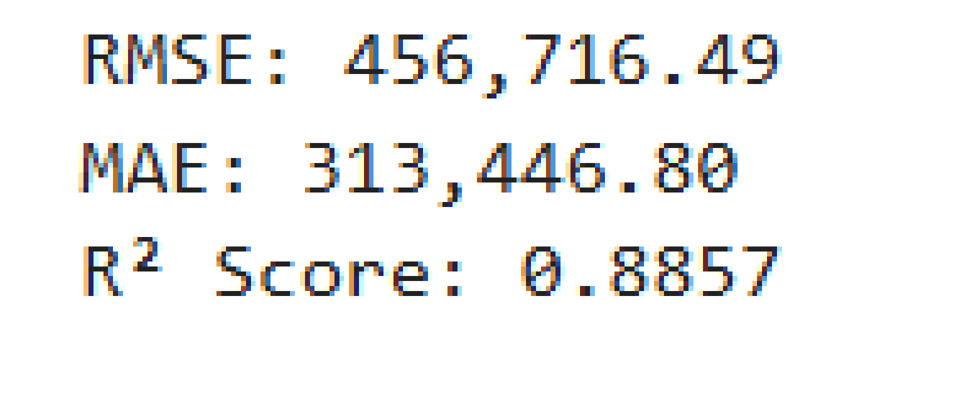
WHY LSTM???

It excels at capturing long-term dependencies, making it ideal for sequence prediction tasks.

- I fed monthly passenger traffic data from 1990 to 2024 into an LSTM model to learn sequential patterns and predict the next 12 months of future traffic for delta airlines.


Delta Airlines - LSTM Prediction vs Actual
1e6
Actual Traffic
Predicted Traffic
Total Traffic
0
1
2
3
4
5
Date
0
50
100
150
200
250
300
350
400


Delta Airlines - Forecasting Next 12 Months
1e6
Historical Traffic
Forecasted Traffic
Total Traffic
0
1
2
3
4
5
Date
1992
1996
2000
2004
2008
2012
2016
2020
2024



```
RMSE: 456,716.49
MAE: 313,446.80
R² Score: 0.8857
```

- RMSE tells us the model's average large-error sensitivity here, predictions deviate by about ±457K passengers, highlighting the worst-case error margin.

- MAE gives the average monthly prediction error, the model is typically off by ±313K passengers, regardless of direction.

- R^2 Score shows how well the model explains trends, with a score of 0.8857(88.57%), it captures ~ 89% of traffic pattern variability.

FINAL INSIGHTS

- Passenger traffic shows predictable seasonality
- Airlines with global reach (AA, UA, DL) consistently lead
- Post-COVID recovery well underway

Final Insights

THANK YOU

ANY QUESTIONS

(Suggestions also accepted)