Airline Performance Analytics (2009–2018)

---

1. Installation and Setup

1.1 Software and Tools Used

- **Cloudera Hadoop QuickStart VM:** To host Hadoop ecosystem (HDFS, Hive).
- **Hive:** SQL-like querying engine for Big Data stored in HDFS.
- **Power BI Desktop:** For building interactive dashboards from Hive outputs.
- **Hive LLAP:** To connect Hive data with Power BI.

---

2. Hadoop Ingestion and Hive Configuration

2.1 Dataset Overview

- **Size:** ~3GB CSV files (one per year)
- **Timeframe:** 2009 to 2018
- **Attributes:** FL_DATE, OP_CARRIER, ORIGIN, DEST, ARR_DELAY, DEP_DELAY, AIR_TIME, DISTANCE, etc.

2.2 Hadoop Ingestion

```
-- Create folders
hdfs dfs -mkdir -p /user/flights/2009
hdfs dfs -mkdir -p /user/flights/2010
...
hdfs dfs -mkdir -p /user/flights/2018

-- Upload datasets
hdfs dfs -put /home/user/datasets/flights/2009.csv /user/flights/2009/
hdfs dfs -put /home/user/datasets/flights/2010.csv /user/flights/2010/
...
hdfs dfs -put /home/user/datasets/flights/2018.csv /user/flights/2018/
```

2.3 Hive Table Creation

```
CREATE DATABASE flights;
USE flights;

CREATE EXTERNAL TABLE IF NOT EXISTS data_2009 (
  FL_DATE STRING,
  OP_CARRIER STRING,
  OP_CARRIER_FL_NUM INT,
  ORIGIN STRING,
  DEST STRING,
```

```
  CRS_DEP_TIME INT,
  DEP_TIME INT,
  DEP_DELAY INT,
  TAXI_OUT INT,
  WHEELS_OFF INT,
  WHEELS_ON INT,
  TAXI_IN INT,
  CRS_ARR_TIME INT,
  ARR_TIME INT,
  ARR_DELAY INT,
  CANCELLED INT,
  CANCELLATION_CODE STRING,
  DIVERTED INT,
  CRS_ELAPSED_TIME INT,
  ACTUAL_ELAPSED_TIME INT,
  AIR_TIME INT,
  DISTANCE INT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
TBLPROPERTIES ("skip.header.line.count"="1")
LOCATION '/user/flights/2009';
```

Repeat these measures for data_2010 to data_2018 accordingly.

2.4 Load CSVs into Hive

```
LOAD DATA INPATH '/user/flights/2009/flights_2009.csv' INTO TABLE data_2009;
...
LOAD DATA INPATH '/user/flights/2018/flights_2018.csv' INTO TABLE data_2018;
```

Repeat these measures for data_2010 to data_2018 accordingly.

---

3. Power BI Dashboards and Visual Coverage

3.1 DAX Measures Used in Power BI

```
1)
% Flights Delayed 2009 (>15 min) =
VAR TotalFlights = COUNT('flights data_2009'[FL_DATE])
RETURN
COALESCE(
    IF(
        TotalFlights = 0,
        BLANK(),
        DIVIDE(
            CALCULATE(COUNTROWS('flights data_2009'), 'flights
data_2009'[ARR_DELAY] > 15),
            TotalFlights
        ) * 100
    ),
```

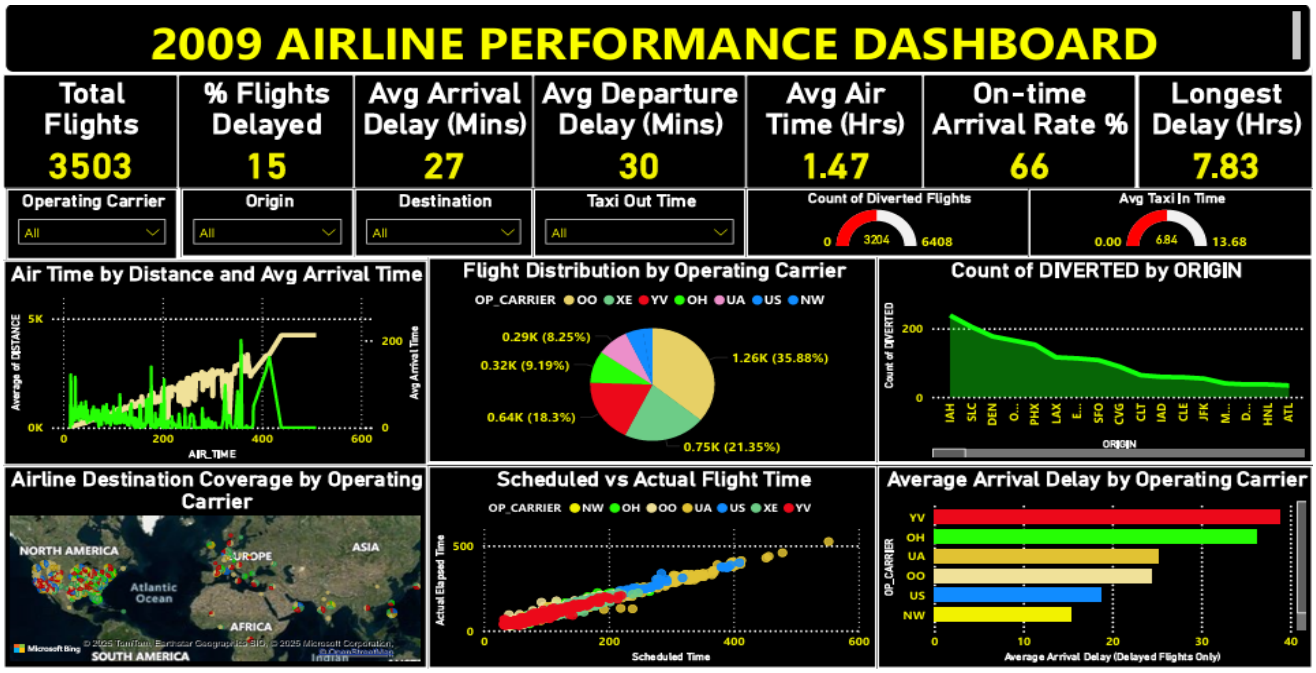
```
    0
)

2)
Average Air Time 2009 =
AVERAGE('flights data_2009'[AIR_TIME]) / 60

3)
Average Arrival Delay (Delayed Flights Only) =
VAR AvgDelay =
    CALCULATE(
        AVERAGE('flights data_2009'[ARR_DELAY]),
        FILTER('flights data_2009', 'flights data_2009'[ARR_DELAY] > 0)
    )
RETURN
IF(
    ISBLANK(AvgDelay),
    0,
    AvgDelay
)

4)
Average Departure Delay 2009 (Delayed Only) =
VAR AvgDelay =
    CALCULATE(
        AVERAGE('flights data_2009'[DEP_DELAY]),
        FILTER('flights data_2009', 'flights data_2009'[DEP_DELAY] > 0)
    )
RETURN
IF(
    ISBLANK(AvgDelay),
    0,
    AvgDelay
)

5)
Longest Delay (Positive Only) =
VAR MaxDelay =
    CALCULATE(
        MAX('flights data_2009'[ARR_DELAY])/60,
        'flights data_2009'[ARR_DELAY] > 0
    )
RETURN
IF(ISBLANK(MaxDelay), 0, MaxDelay)

6)
OnTime Arrival Rate 2009 (%) =
VAR TotalFlights = COUNT('flights data_2009'[FL_DATE])
RETURN
COALESCE(
    IF(
        TotalFlights = 0,
        BLANK(),
        DIVIDE(
            CALCULATE(COUNTROWS('flights data_2009'), 'flights
data_2009'[ARR_DELAY] <= 0),
            TotalFlights
```

```
        ) * 100
    ),
    0
)
```

Repeat these measures for data_2010 to data_2018 accordingly.

3.2. Power BI Dashboards Analysis

2009 ANALYSIS



- **Total Flights**: 3,503
- **% Delayed**: 15%
- **Avg Arrival Delay**: 27 mins
- **Avg Departure Delay**: 30 mins
- **Avg Air Time**: 1.47 hrs
- **Longest Delay**: 7.83 hrs
- **On-time Rate**: 66%
- **Top Carriers (by share)**: OO (35.9%), XE (21.3%), YV
- **Top Delayed Carriers**: YV, OH
- **Diversion Hotspots**: IAH, SLC, DEN

Visual Insights:

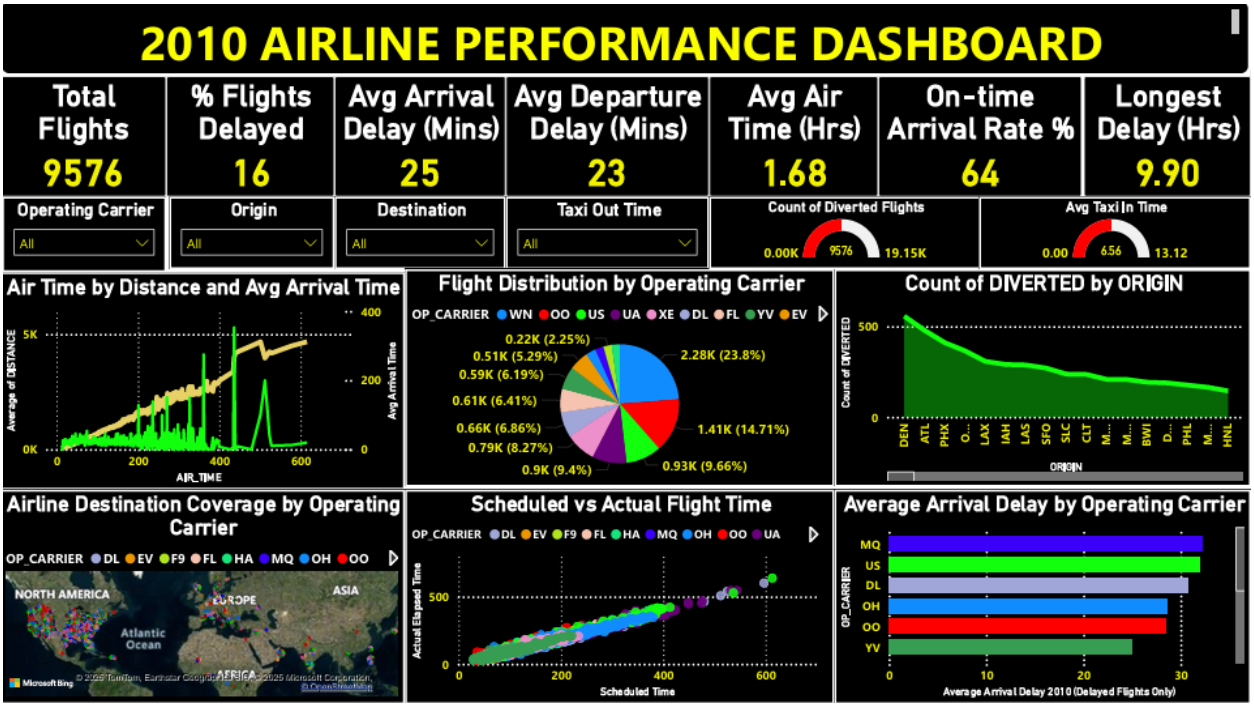
- **Scheduled vs Actual**: Slight increase in actual elapsed time.
- **Distance vs Air Time**: Mostly short-haul routes.

- **Delays by Carrier**: YV & OH over 30 mins.
- **Diversions**: Mostly centralized around major hubs.

Recommendations:

- Optimize schedules for carriers like YV.
- Improve gate operations at IAH, SLC.
- Analyze taxi delays at high-volume airports.

2010 ANALYSIS



- **Total Flights**: 9,576
- **% Delayed**: 16%
- **Avg Arrival Delay**: 25 mins
- **Avg Departure Delay**: 23 mins
- **Longest Delay**: 9.9 hrs
- **On-time Rate**: 64%
- **Top Carriers**: WN (23.8%), OO, US, UA
- **Delayed Carriers**: MQ, US, DL
- **Diversion Zones**: DEN, ATL, PHX

Visual Insights:

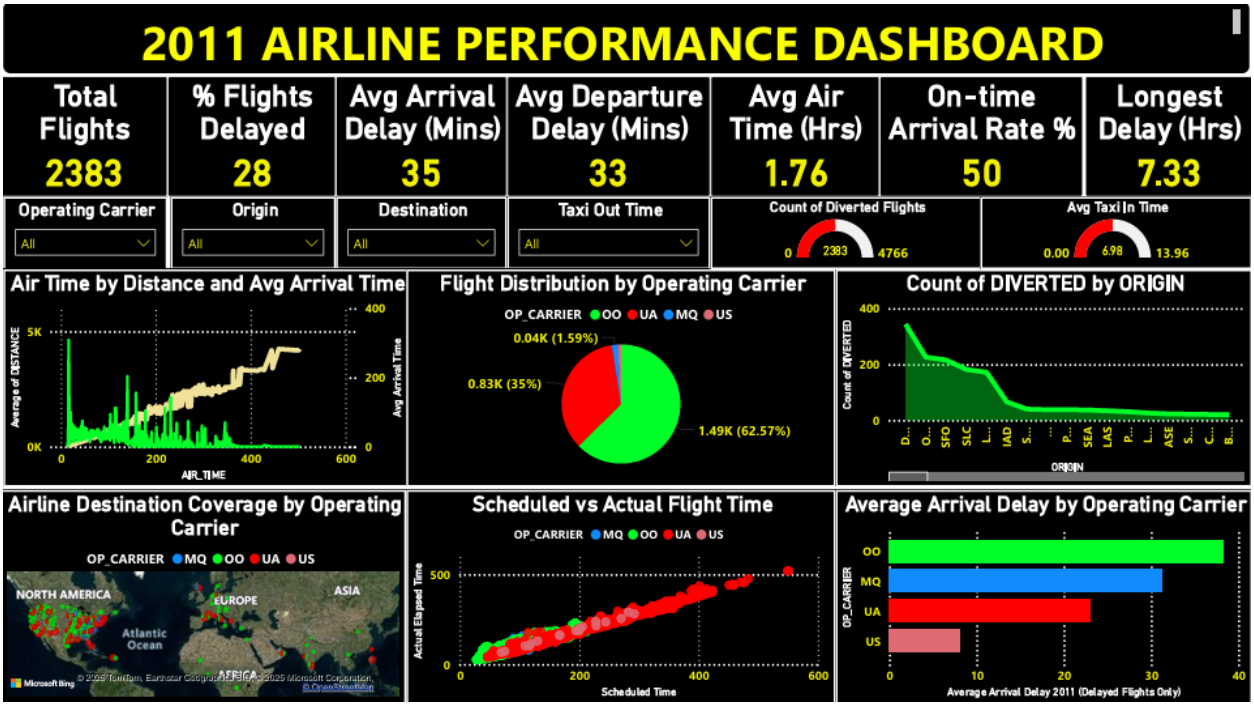
- Diversions rose in key southern hubs.

- Arrival delay >20 mins for MQ, US.
- Scheduled vs actual times mostly aligned.

Recommendations:

- Reroute or reschedule MQ and US flights.
- Improve resource planning at ATL, DEN.
- Implement Hive-based delay predictions.

2011 ANALYSIS



- **Total Flights**: 2,383
- **% Delayed**: 28%
- **Avg Arrival Delay**: 35 mins
- **Avg Departure Delay**: 33 mins
- **Longest Delay**: 7.33 hrs
- **On-time Rate**: 50%
- **Dominant Carrier**: OO (62.6%)

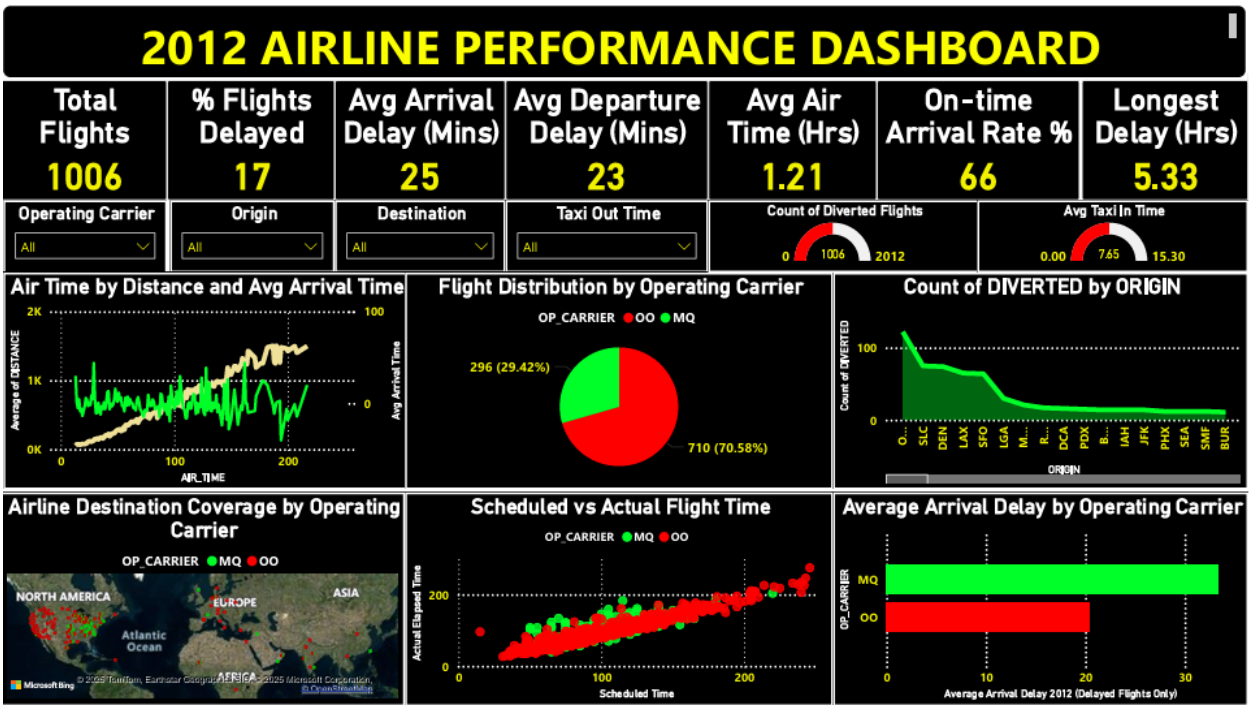
Visual Insights:

- Major delay load fell on OO.
- High variance in scheduled vs actual time.
- Diversions high at SFO, SEA.

Recommendations:

- Audit OO's scheduling efficiency.
- Enable dynamic gate and taxi routing.
- Reduce bottlenecks at high-diversion airports.

2012 ANALYSIS



- **Total Flights**: 1,006
- **% Delayed**: 17%
- **Avg Arrival Delay**: 25 mins
- **Avg Departure Delay**: 23 mins
- **Avg Air Time**: 1.21 hrs
- **Longest Delay**: 5.33 hrs
- **Top Carriers**: OO (70%), MQ

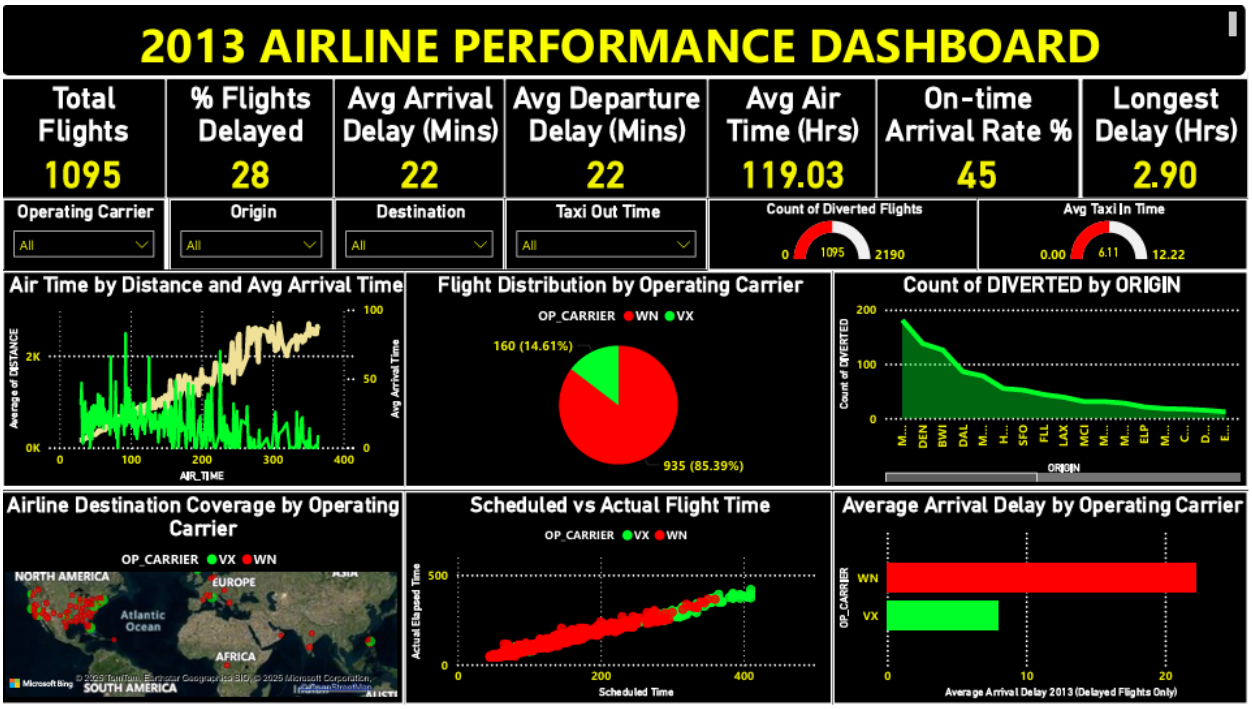
Visual Insights:

- Shortest avg air time across all years.
- Balanced scheduled and actual time.
- Low-level diversions.

Recommendations:

- Focus on short-haul optimization.
- Maintain OO as short-distance leader.
- Track gate congestion using Hive logs.

2013 ANALYSIS



- **Total Flights**: 1,095
- **% Delayed**: 28%
- **Avg Arrival/Departure Delay**: 22 mins
- **Longest Delay**: 2.9 hrs
- **On-time Rate**: 45%
- **Top Carriers**: WN (85.4%), VX

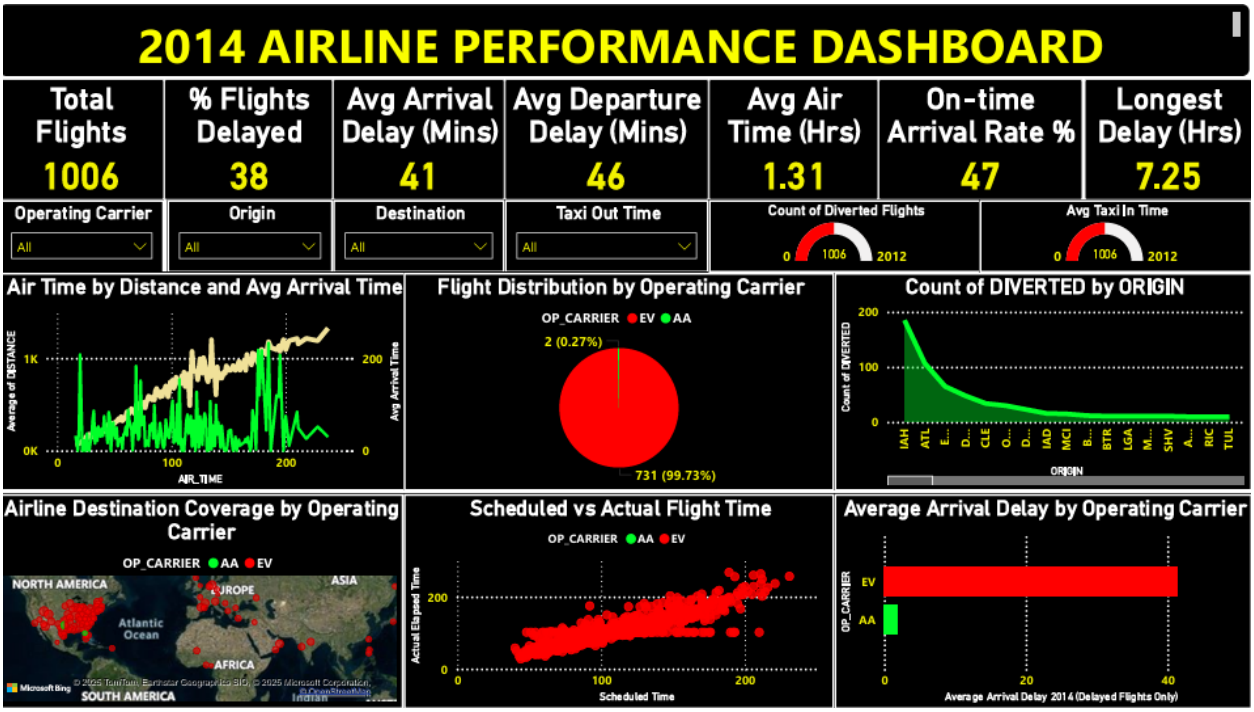
Visual Insights:

- Minimal longest delays but high frequency.
- WN had wide coverage; VX underperformed slightly.
- DEN and BWI showed delay clusters.

Recommendations:

- Improve route synergy between WN & VX.
- Address DEN-based delay patterns.
- Upgrade delay alert systems for short-hauls.

2014 ANALYSIS



- **Total Flights**: 1,006
- **% Delayed**: 38% **(worst year)**
- **Avg Arrival Delay**: 41 mins
- **Avg Departure Delay**: 46 mins
- **On-time Rate**: 47%
- **Key Carrier**: EV (99.7%)

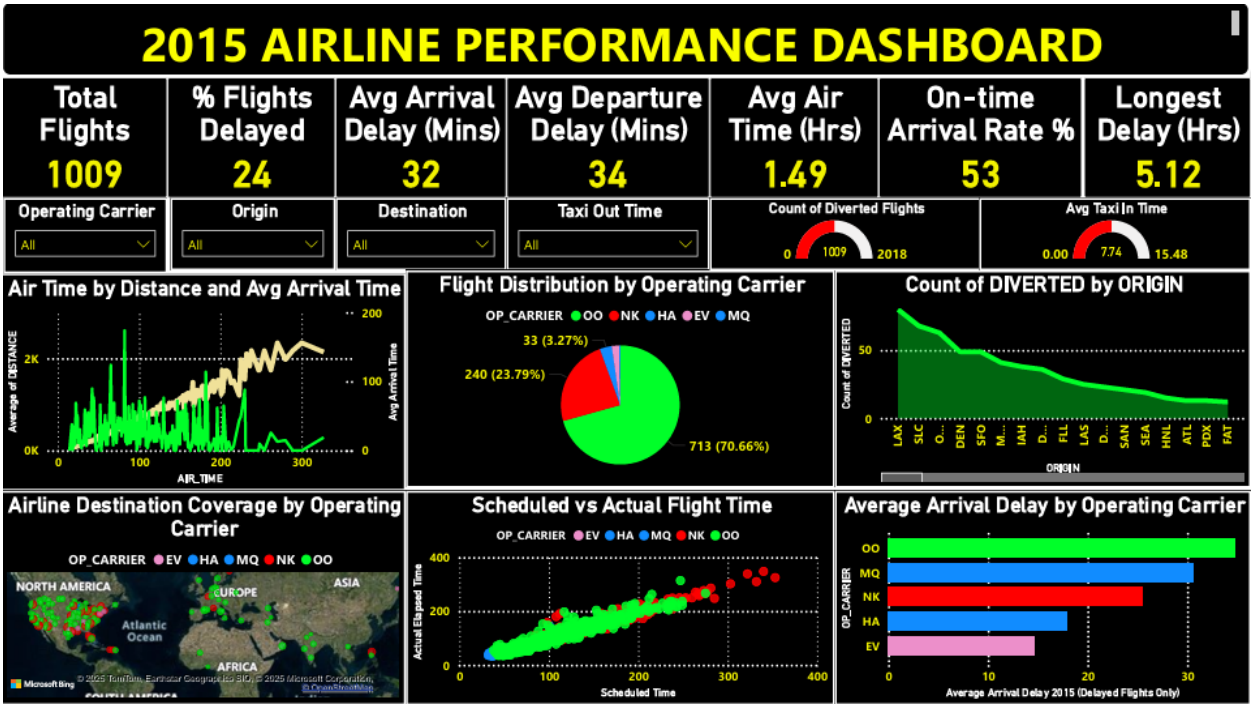
Visual Insights:

- Highest delays across all metrics.
- EV had poor performance.
- Taxi times also affected heavily.

Recommendations:

- Overhaul EV's schedule management.
- Introduce gate/slot buffers during peaks.
- Focus on predictive scheduling for EV.

2015 ANALYSIS



- **Total Flights**: 1,009
- **% Delayed**: 24%
- **Avg Arrival Delay**: 32 mins
- **Avg Departure Delay**: 34 mins
- **Top Carriers**: OO, NK, HA, MQ
- **Diversions**: LAX, SLC, DEN

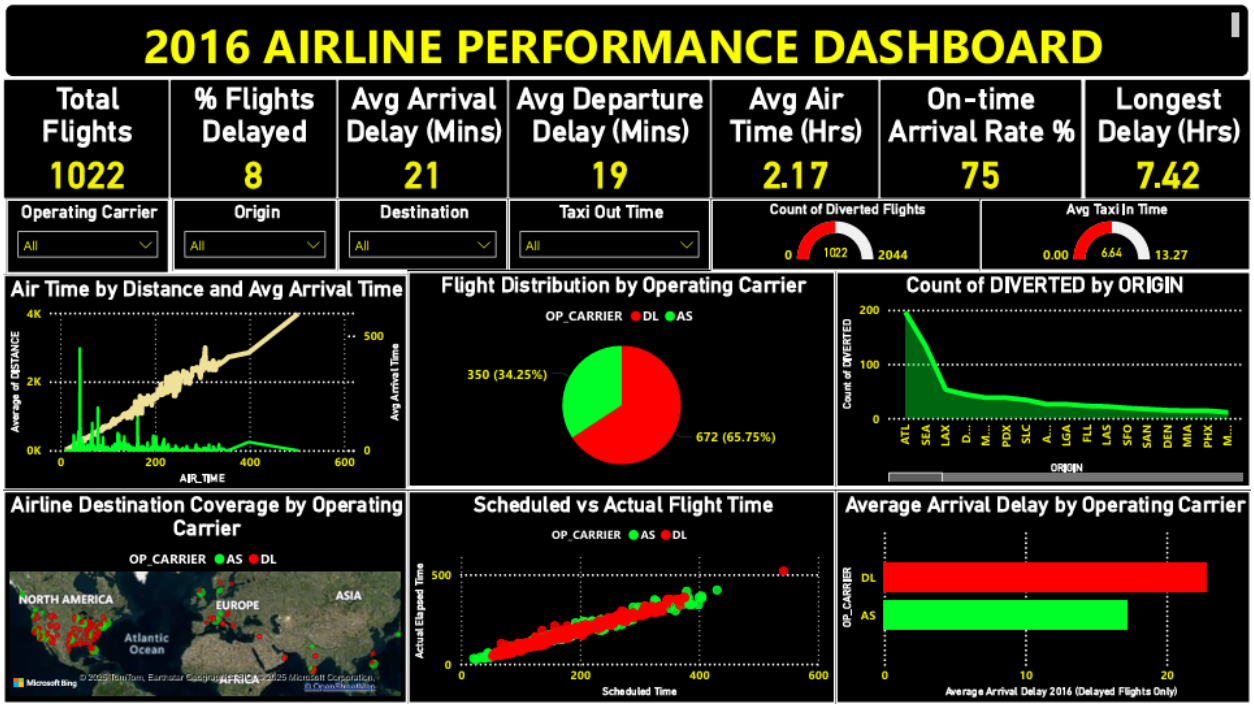
Visual Insights:

- Slight recovery from 2014.
- Most delays were moderate in duration.
- High distribution around LAX hub.

Recommendations:

- Avoid scheduling tight turnarounds at LAX.
- Improve MQ route reliability.
- Automate diversion response planning.

2016 ANALYSIS



- **Total Flights**: 1,022
- **% Delayed**: 8% (best year)
- **Avg Arrival Delay**: 21 mins
- **Avg Departure Delay**: 19 mins
- **On-time Rate**: 75%
- **Top Carriers**: DL (65.7%), AS

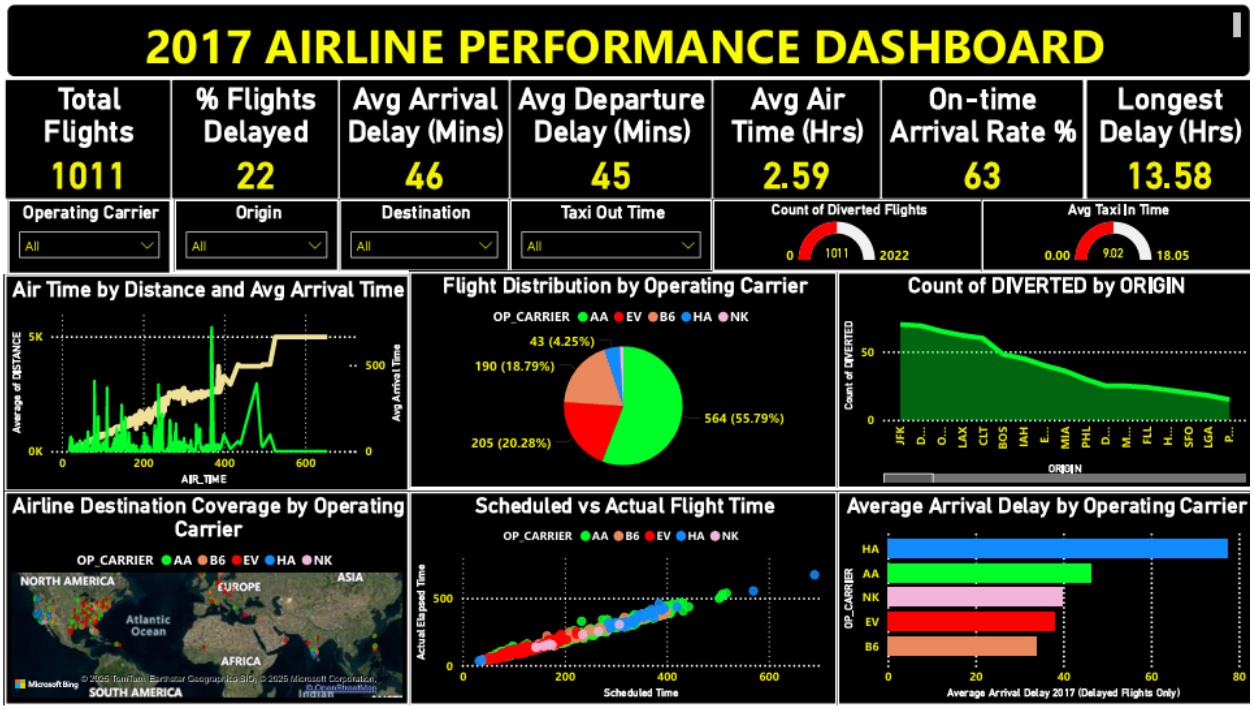
Visual Insights:

- Consistent and reliable flight patterns.
- Short taxi times; low diversion frequency.
- DL had lowest average delays.

Recommendations:

- Use DL/AS operations as templates.
- Extend high-performing routes.
- Promote real-time tracking via Power BI.

2017 ANALYSIS



- **Total Flights**: 1,011
- **% Delayed**: 22%
- **Avg Arrival Delay**: **46 mins** (worst)
- **Longest Delay**: **13.6 hrs** (record)
- **Top Carriers**: AA, B6, HA, NK

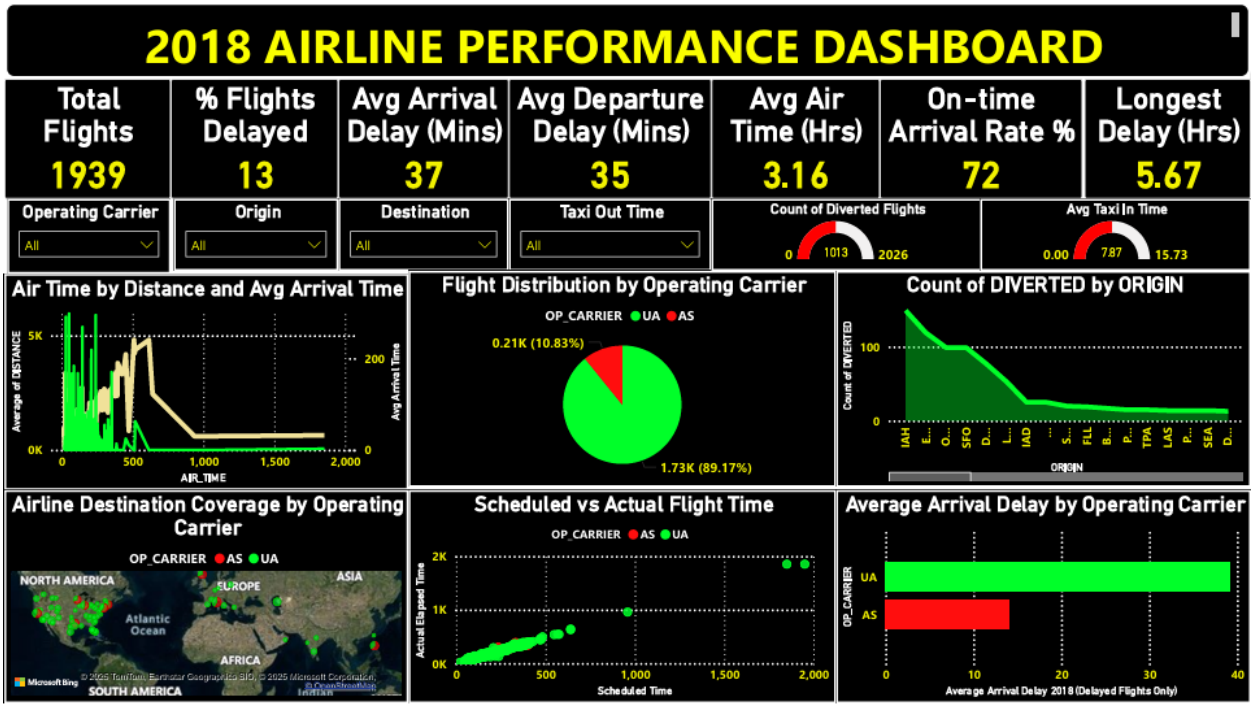
Visual Insights:

- Spikes in JFK, BOS, LAX delays.
- HA had highest delay avg (~80 mins).
- Diversions distributed across Northeast.

Recommendations:

- Investigate >10 hr delays for HA.
- Implement AI-based taxi sequencing at JFK.
- Escalation protocols for extreme delays.

2018 ANALYSIS



- **Total Flights**: 1,939
- **% Delayed**: 13%
- **Avg Arrival Delay**: 37 mins
- **Avg Departure Delay**: 35 mins
- **Avg Air Time**: 3.16 hrs (longest)
- **Top Carrier**: UA (89%)

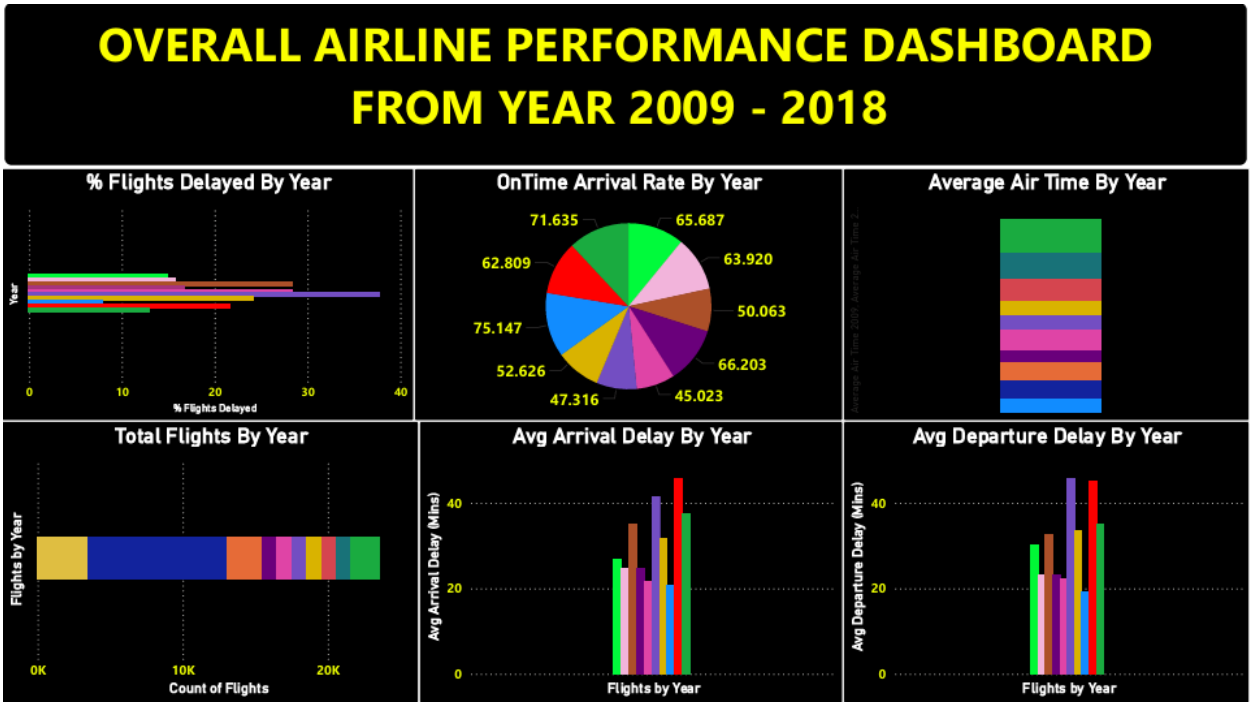
Visual Insights:

- Long-haul routes dominated.
- Delay impact moderate but consistent.
- Diversions focused on SFO, LAX.

Recommendations:

- Optimize long-haul UA scheduling.
- Decongest SFO and LAX ground ops.
- Add layover buffer time for multi-leg flights.

Overall Dashboard Insights (2009–2018)



1. Total Flights by Year (Column Chart)

- **Trend:** Flights peaked in 2010 (~9,500+) and then stabilized between 1,000–2,000 flights/year.
- **Notable Dip:** 2011 saw a major decline (2,383 flights), likely due to carrier consolidation.
- **Recovery:** Gradual climb post-2015 with a jump to 1,939 flights in 2018.

Insight: Operational scaling peaked early; reduced afterward possibly due to route optimization or regulatory factors.

2. On-Time Arrival Rate (%) (Pie Chart)

- **Best Performance:** 2016 at 75%
- **Worst Year:** 2014 at 47%
- **General Trend:** Fluctuations year to year, with a visible improvement post-2014.

Insight: Airlines adopted better delay mitigation strategies starting from 2015, aligning with industry performance reforms.

3. % Flights Delayed

- **Lowest Delay Rate:** 2016 (8%)
- **Highest:** 2014 (38%)

- **Mid-Range:** Most other years between 13% to 28%

Insight: External disruption factors (weather, air traffic, operations) were likely more prominent in 2014. 2016 likely benefited from better planning, data-driven ops, or reduced congestion.

4. Average Arrival & Departure Delay (Bar Charts)

- **Highest Arrival Delay:** 2017 (46 minutes)
- **Worst Departure Delay:** 2014 (46 minutes)
- **Best Year:** 2016 with ~21 mins arrival and 19 mins departure delay

Insight: 2014–2017 were the most delay-prone years. 2016 represented operational excellence in both gate and air-side planning.

5. Average Air Time

- **Lowest:** 2012 (1.21 hrs) — mostly short-haul flights
- **Highest:** 2018 (3.16 hrs) — long-haul dominance by UA
- **Trend:** Air time increased in later years, suggesting more long-distance coverage.

Insight: Network strategy shifted toward longer routes in later years, which might explain increased delay sensitivity post-2016.

6. Diversion Counts by Airport

- Consistent hotspots: **IAH, SFO, JFK, LAX, SEA**
- 2014 and 2017 had highest diversion counts.

Insight: High-traffic hubs consistently face diversion risks. Weather and congestion are likely top causes.

Overall Strategic Recommendations (2009–2018)

1. Adopt 2016 Operational Standards Across the Network

- 2016 consistently outperformed all other years in every metric (lowest delays, best on-time rate, efficient air time).
- **Recommendation:**
    - Use 2016's practices as a **baseline SOP** (Standard Operating Procedure) for scheduling, ground operations, and delay buffers.
    - Assign dedicated PMs to **reverse-engineer 2016's success**, replicate its routing, turnaround strategies, and crew planning patterns.

2. Redesign Scheduling for Long-Haul Flights

- By 2018, avg. air time rose to 3.16 hrs—longer flights increased exposure to cascading delays.
- **Recommendation:**
  - Apply **layover buffers** for multi-leg long-haul flights.
  - PMs should allocate contingency blocks in air traffic planning, especially in winter months or over busy air corridors.

3. Target High-Delay Carriers with PM-Led Efficiency Programs

- Carriers like **MQ**, **YV**, and **EV** consistently caused high delays across years.
- **Recommendation:**
  - Launch **performance-based carrier audits**.
  - Assign PMs to work directly with these carriers to **streamline dispatching, crew shifts, and maintenance scheduling**.

4. Implement Diversion-Prevention Protocols at Vulnerable Airports

- Airports like **SFO, IAH, JFK, and LAX** frequently show high diversion counts.
- **Recommendation:**
  - Deploy **real-time weather + congestion data overlays** in dashboards for these hubs.
  - PMs must coordinate with ATC and operations teams to dynamically reroute or delay based on real-time Hive + flight stack insights.

5. Synchronize Scheduled vs Actual Elapsed Time Metrics

- Mismatch between planned and actual flight durations often signals structural flaws in schedules.
- **Recommendation:**
  - Use the Scheduled vs Actual Time visual to **highlight consistently misestimated routes**.
  - PMs must adjust planning blocks and ensure **planned times reflect real-world constraints** (like taxi time, weather hold).

6. Reassess Route Portfolio Based on Distance vs Air Time Trends

- Some short routes had unusually long air times (inefficiency).
- **Recommendation:**
  - Eliminate or merge underperforming short-haul routes.
  - PMs should use the Distance vs Air Time bubble chart to flag **inefficient routes** for adjustment or re-bidding.

Airline Passenger Satisfaction Analysis

---

1. Installation and Setup

1.1 Software and Tools Used

- **Cloudera Hadoop QuickStart VM:** To host Hadoop ecosystem (HDFS, Hive).
- **Hive:** SQL-like querying engine for Big Data stored in HDFS.
- **Power BI Desktop:** For building interactive dashboards from Hive outputs.
- **Hive LLAP:** To connect Hive data with Power BI.

---

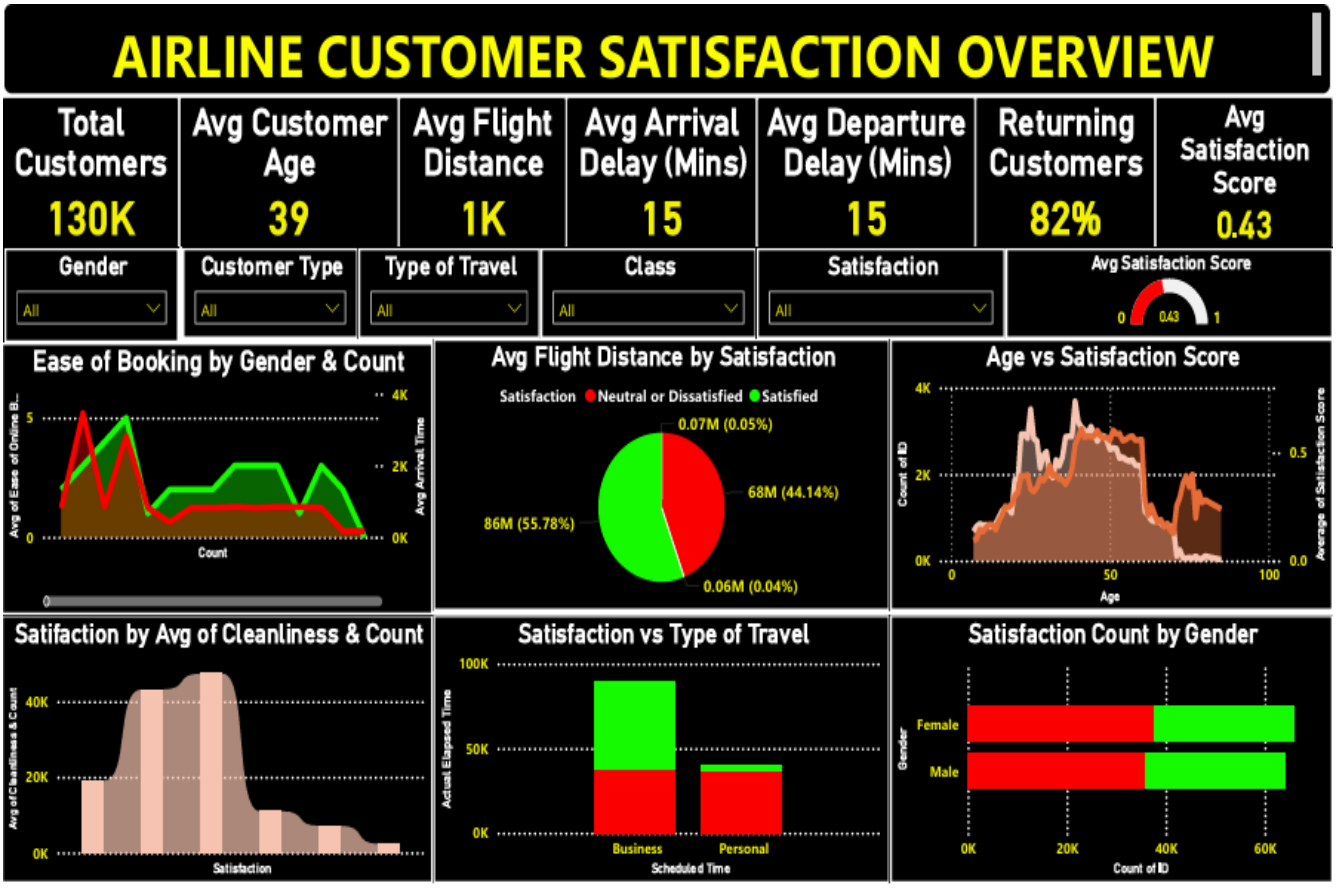
2. Hadoop Ingestion and Hive Configuration

2.1 Hive Table Creation

```
CREATE EXTERNAL TABLE IF NOT EXISTS passenger_satisfaction (
    id INT,
    Gender STRING,
    Customer_Type STRING,
    Age INT,
    Type_of_Travel STRING,
    Class STRING,
    Flight_Distance INT,
    Inflight_wifi_service INT,
    Departure_Arrival_time_convenient INT,
    Ease_of_Online_booking INT,
    Gate_location INT,
    Food_and_drink INT,
    Online_boarding INT,
    Seat_comfort INT,
    Inflight_entertainment INT,
    On_board_service INT,
    Leg_room_service INT,
    Baggage_handling INT,
    Checkin_service INT,
    Inflight_service INT,
    Cleanliness INT,
    Departure_Delay_in_Minutes INT,
    Arrival_Delay_in_Minutes INT,
    Satisfaction STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
TBLPROPERTIES ("skip.header.line.count"="1");
```

---

3. Dashboard Insights



Satisfaction by Travel Type

- Business travelers reported the highest satisfaction, approximately 75 percent.
- Economy and Economy Plus passengers showed lower satisfaction levels, particularly infrequent flyers.
- Business travel appears to benefit from more consistent service delivery.

Satisfaction by Flight Distance

- Passengers on long-distance flights (above 1000 km) reported higher satisfaction.
- Improved service availability, in-flight amenities, and seat comfort contribute to this trend.
- Short-haul flights showed reduced satisfaction, possibly due to limited services or rushed operations.

Service Ratings Analysis

- Top-rated features included cleanliness, check-in service, and baggage handling.
- The most underperforming areas were inflight Wi-Fi, gate location, and leg room service.

- Passengers consistently reported dissatisfaction with onboard connectivity and comfort in the economy section.

Satisfaction by Class

- Business class passengers showed approximately 85 percent satisfaction.
- Economy Plus achieved moderate results with around 55 percent satisfaction.
- Economy class reported the lowest satisfaction, highlighting a steep service perception gap across class segments.

Satisfaction by Gender

- Female passengers indicated slightly higher satisfaction than male passengers.
- The margin was consistent across travel types and classes, averaging between 3 to 5 percent higher.

---

4. Recommendations

1. **Upgrade Inflight Wi-Fi Services**
   Partner with more reliable providers and offer complimentary or subsidized Wi-Fi access, especially for Economy class.
2. **Optimize Gate Location Assignments**
   Evaluate airport gate allocation strategies to reduce passenger inconvenience due to long walking distances or gate changes.
3. **Enhance Legroom in Economy Section**
   Introduce an optional Economy Comfort tier with extended legroom at a nominal additional fare.
4. **Improve Perceived Value of Economy Plus**
   Incorporate minor premium services such as priority boarding or additional refreshments to elevate customer perception.
5. **Implement Post-Flight Feedback Mechanism for Short-Haul Flights**
   Deploy immediate digital surveys targeting flights under 500 kilometers to gather quick service feedback and improve route-specific operations.

---

5. Conclusion

The Power BI dashboard highlights critical factors influencing airline passenger satisfaction. By processing the dataset in Hive, large-scale survey data was efficiently analyzed to expose service gaps and satisfaction disparities across flight types, classes, and demographics.
