

U.S. Airline Performance & Delay Analysis – Final Project Report

1. Introduction

Problem Statement

Flight delays and cancellations are significant issues in the U.S. aviation industry, impacting passengers, airlines, and the economy. This project performs an in-depth analysis of historical flight data to identify key drivers of these disruptions, assess performance, and propose actionable insights.

Objectives

- Identify primary causes of delays and cancellations.
- Benchmark airline and airport performance.
- Examine time-based (monthly, daily, hourly) trends.
- Provide actionable recommendations for stakeholders.

2. Methodology

Data Sources

- **Flights Dataset** – Historical 2015 U.S. domestic flights data.
- **Airlines Dataset** – Airline code and name mapping.
- **Airports Dataset** – Airport details including city, state, latitude, and longitude.

Tools Used

- **SQL** – Data ingestion, cleaning, preparation, and exploratory analysis.
- **Power BI** – Dashboard creation for visualization and storytelling.

Analytical Approach

- SQL for integration of flights, airlines, and airports data.
- Null value handling and time formatting.
- Power BI for 4-page interactive dashboard.

3. SQL Implementation

Data Cleaning Queries – Handling nulls, time conversion (HHMM to HH:MM), marking cancellations.

```

3 SELECT * FROM airlines LIMIT 10;
4 SELECT * FROM airports LIMIT 10;
5 SELECT * FROM flights LIMIT 10;
6
7
8 -- COUNT NUMBER OF ROWS
9
10 SELECT COUNT(*) FROM airlines;
11 SELECT COUNT(*) FROM airports;
12 SELECT COUNT(*) FROM flights;
13
14
15 -- Update the Column with Proper Format
16
17 ALTER TABLE flights ADD COLUMN FLIGHT_DATE DATE;
18 UPDATE flights
19 SET FLIGHT_DATE = printf('%04d-%02d-%02d', YEAR, MONTH, DAY);
20
21
22
23
24 -- Convert HHMM to TIME
25
26 ALTER TABLE flights ADD COLUMN SCHED_DEP_TIME TEXT;
27 ALTER TABLE flights ADD COLUMN ACTUAL_DEP_TIME TEXT;
28
29 ALTER TABLE flights ADD COLUMN SCHED_ARR_TIME TEXT;
30 ALTER TABLE flights ADD COLUMN ACTUAL_ARR_TIME TEXT;
31
32
33 -- Scheduled Departure
34 UPDATE flights
35 SET SCHED_DEP_TIME =
36     printf('%02d:%02d', SCHEDULED_DEPARTURE / 100, SCHEDULED_DEPARTURE % 100);
37
38
39 -- Scheduled Arrival
40 UPDATE flights
41 SET SCHED_ARR_TIME =
42     printf('%02d:%02d', SCHEDULED_ARRIVAL / 100, SCHEDULED_ARRIVAL % 100);
43
44 -- Actual Departure
45 UPDATE flights
46 SET ACTUAL_DEP_TIME =
47     CASE
48     WHEN DEPARTURE_TIME IS NOT NULL THEN
49         printf('%02d:%02d', DEPARTURE_TIME / 100, DEPARTURE_TIME % 100)
50     ELSE NULL
51     END;
52
53
54 -- Actual Arrival
55 UPDATE flights
56 SET ACTUAL_ARR_TIME =
57     CASE
58     WHEN ARRIVAL_TIME IS NOT NULL THEN
59         printf('%02d:%02d', ARRIVAL_TIME / 100, ARRIVAL_TIME % 100)
60     ELSE NULL
61     END;
62
63
64
65
66
67 -- Create the Joined View in SQLite
68
69 CREATE VIEW flight_details AS
70 SELECT
71     f.FLIGHT_DATE,
72     f.YEAR, f.MONTH, f.DAY, f.DAY_OF_WEEK,
73     f.AIRLINE, a.AIRLINE AS AIRLINE_NAME,
74     f.FLIGHT_NUMBER, f.TAIL_NUMBER,
75     f.ORIGIN_AIRPORT, ao.AIRPORT AS ORIGIN_AIRPORT_NAME, ao.CITY AS ORIGIN_CITY,
76     f.DESTINATION_AIRPORT, ad.AIRPORT AS DESTINATION_AIRPORT_NAME, ad.CITY AS DESTINATION_CITY,
77     f.SCHED_DEP_TIME, f.ACTUAL_DEP_TIME,
78     f.SCHEDULED_ARRIVAL, f.ARRIVAL_TIME,
79     f.DEPARTURE_DELAY, f.ARRIVAL_DELAY,
80     f.CANCELLED, f.CANCELLATION_REASON,
81     f.AIR_TIME, f.DISTANCE, f.AIRLINE_DELAY, f.WEATHER_DELAY, f.LATE_AIRCRAFT_DELAY
82 FROM flights f
83 JOIN airlines a ON f.AIRLINE = a.IATA_CODE
84 JOIN airports ao ON f.ORIGIN_AIRPORT = ao.IATA_CODE
85 JOIN airports ad ON f.DESTINATION_AIRPORT = ad.IATA_CODE;
86
87
88
89
90
91
92
93
94
95
96 -- Data Quality Checks & Handling Missing Values
97
98 -- Check for NULLs in Important Columns
99
100 SELECT
101     SUM(CASE WHEN AIRLINE IS NULL THEN 1 ELSE 0 END) AS missing_airline,
102     SUM(CASE WHEN ORIGIN_AIRPORT IS NULL THEN 1 ELSE 0 END) AS missing_origin,
103     SUM(CASE WHEN DESTINATION_AIRPORT IS NULL THEN 1 ELSE 0 END) AS missing_destination,
104     SUM(CASE WHEN DEPARTURE_TIME IS NULL THEN 1 ELSE 0 END) AS missing_departure_time,
105     SUM(CASE WHEN ARRIVAL_TIME IS NULL THEN 1 ELSE 0 END) AS missing_arrival_time,
106     SUM(CASE WHEN ARRIVAL_DELAY IS NULL THEN 1 ELSE 0 END) AS missing_arrival_delay
107 FROM flights;
108
109 DELETE FROM airports
110 WHERE LATITUDE IS NULL OR LONGITUDE IS NULL;
111
112 SELECT *
113 FROM airports
114 WHERE LATITUDE IS NULL OR LONGITUDE IS NULL;
115
116
117 -- Mark missing ARRIVAL_TIME as 'CANCELLED'
118
119 UPDATE flights
120 SET ARRIVAL_TIME = NULL
121 WHERE CANCELLED = 1;
122
123 Similarly, mark missing DEPARTURE_TIME
124 UPDATE flights
125 SET DEPARTURE_TIME = NULL
126 WHERE CANCELLED = 1;
127
128
129
130 -- ALTER TABLE flights ADD COLUMN STATUS TEXT;
131
132 UPDATE flights
133 SET STATUS =
134     CASE
135     WHEN CANCELLED = 1 THEN 'Cancelled'
136     WHEN DIVERTED = 1 THEN 'Diverted'
137     ELSE 'Completed'
138     END;
139

```

```

140 -- Fill Missing Delay Columns with 0
141
142 UPDATE flights
143 SET ARRIVAL_DELAY = 0
144 WHERE ARRIVAL_DELAY IS NULL;
145
146 UPDATE flights
147 SET DEPARTURE_DELAY = 0
148 WHERE DEPARTURE_DELAY IS NULL;
149
150 UPDATE flights
151 SET AIRLINE_DELAY = 0
152 WHERE AIRLINE_DELAY IS NULL;
153
154 UPDATE flights
155 SET WEATHER_DELAY = 0
156 WHERE WEATHER_DELAY IS NULL;
157
158 UPDATE flights
159 SET SECURITY_DELAY = 0
160 WHERE SECURITY_DELAY IS NULL;
161
162 UPDATE flights
163 SET CANCELLATION_REASON = 0
164 WHERE CANCELLATION_REASON IS NULL;
165
166 UPDATE flights
167 SET AIR_SYSTEM_DELAY = 0
168 WHERE AIR_SYSTEM_DELAY IS NULL;
169
170 UPDATE flights
171 SET LATE_AIRCRAFT_DELAY = 0
172 WHERE LATE_AIRCRAFT_DELAY IS NULL;

```

EDA Queries – Flight volume by airline, delays by month, cancellation reasons, airport performance.

```

181 -- Exploratory Data Analysis (EDA) using SQL
182
183
184 SELECT COUNT(*) AS total_records FROM flights;
185 SELECT * FROM flights LIMIT 5; -- Preview 5 sample rows
186
187

```

	YEAR	MONTH	DAY	DAY_OF_WEEK	AIRLINE	FLIGHT_NUMBER	TAIL_NUMBER	ORIGIN_AIRPORT	DESTINATION_AIRPORT	SCHEDULED_DEPARTURE	DEPARTURE_TIME	DEPART
1	2015	1	1	4	AS	98	N407AS	ANC	SEA	5	2354	-11
2	2015	1	1	4	AA	2336	N3KUAA	LAX	PBI	10	2	-8
3	2015	1	1	4	US	840	N171US	SFO	CLT	20	18	-2
4	2015	1	1	4	AA	258	N3HYAA	LAX	MIA	20	15	-5
5	2015	1	1	4	AS	135	N527AS	SEA	ANC	25	24	-1

```

188 -- Flights Over Time (Monthly Trend)
189
190 SELECT
191     MONTH,
192     COUNT(*) AS total_flights,
193     SUM(CASE WHEN CANCELLED = 1 THEN 1 ELSE 0 END) AS cancelled_flights,
194     ROUND(AVG(ARRIVAL_DELAY), 2) AS avg_arrival_delay
195 FROM flights
196 GROUP BY MONTH
197 ORDER BY MONTH;
198
199 -- Insight: Analyze peak months and seasonal delays.
200

```

	MONTH	total_flights	cancelled_flights	avg_arrival_delay
1	1	469968	11982	5.65
2	2	429191	20517	7.9
3	3	504312	11002	4.8
4	4	485151	4520	3.12
5	5	496993	5694	4.42
6	6	503897	9120	9.39
7	7	520718	4806	6.35
8	8	510536	5052	4.55

```

203 -- Day-of-Week Flight Behavior
204
205 SELECT
206     STRFTIME('%w', FLIGHT_DATE) AS day_of_week,
207     COUNT(*) AS total_flights,
208     ROUND(AVG(ARRIVAL_DELAY), 2) AS avg_arrival_delay
209 FROM flights
210 GROUP BY day_of_week
211 ORDER BY day_of_week;
212

```

	day_of_week	total_flights	avg_arrival_delay
1	0	817764	3.88
2	1	865543	5.84
3	2	844600	4.16
4	3	855897	3.79
5	4	872521	5.57
6	5	862209	4.7
7	6	700545	1.82

```

214 -- Hourly Flight Behavior
215
216 SELECT
217     ACTUAL_DEP_TIME,
218     COUNT(*) AS flight_count,
219     round(avg(DEPARTURE_DELAY),2) AS avg_departure_delay
220 FROM flights
221 WHERE ACTUAL_DEP_TIME IS NOT NULL
222 GROUP BY ACTUAL_DEP_TIME
223 ORDER BY flight_count DESC
224 LIMIT 10;
225
226 -- Insight: Do early morning flights face fewer delays?
227

```

	ACTUAL_DEP_TIME	flight_count	avg_departure_delay
1	05:55	14829	-4.56
2	05:56	13976	-3.84
3	05:57	13616	-2.9
4	05:58	12954	-2.19
5	06:55	12391	-3.09
6	05:59	11765	-1.75
7	06:56	11532	-2.71
8	06:57	11406	-1.81

```

231 -- Airline Performance Analysis
232
233 SELECT
234     AIRLINE,
235     COUNT(*) AS total_flights,
236     ROUND(AVG(ARRIVAL_DELAY), 2) AS avg_arrival_delay,
237     ROUND(AVG(DEPARTURE_DELAY), 2) AS avg_departure_delay,
238     ROUND(100.0 * SUM(CANCELLED) / COUNT(*), 2) AS cancel_percent
239 FROM flights
240 GROUP BY AIRLINE
241 ORDER BY avg_arrival_delay DESC;
242
243 -- Insight: Identify worst- and best-performing airlines.
244

```

	AIRLINE	total_flights	avg_arrival_delay	avg_departure_delay	cancel_percent
1	NK	117379	14.2	15.68	1.71
2	F9	90836	12.4	13.27	0.65
3	B6	267048	6.55	11.33	1.6
4	EV	571977	6.39	8.49	2.66
5	MQ	294632	6.11	9.63	5.1
6	OO	588353	5.73	7.68	1.69
7	UA	515723	5.35	14.26	1.27
8	VX	61903	4.69	8.95	0.86
9	WN	1261855	4.31	10.45	1.27
10	US	198715	3.62	6.02	2.05
11	AA	725984	3.39	8.77	1.5
12	HA	76272	2.02	0.48	0.22
13	DL	875881	0.19	7.34	0.44
14	AS	172521	-0.97	1.78	0.39

```

248 -- Origin & Destination Airport Behavior
249 -- Top 10 busy airports:
250
251 SELECT ORIGIN_AIRPORT, COUNT(*) AS departures
252 FROM flights
253 GROUP BY ORIGIN_AIRPORT
254 ORDER BY departures DESC
255 LIMIT 10;
256

```

	ORIGIN_AIRPORT	departures
1	ATL	346836
2	ORD	285884
3	DFW	239551
4	DEN	196055
5	LAX	194673
6	SFO	148008
7	PHX	146815
8	IAH	146622
9	LAS	133181
10	MSP	112117

```

258 -- Worst delays by airport:
259
260 SELECT
261     ORIGIN_AIRPORT,
262     ROUND(AVG(DEPARTURE_DELAY), 2) AS avg_departure_delay
263 FROM flights
264 WHERE CANCELLED = 0
265 GROUP BY ORIGIN_AIRPORT
266 ORDER BY avg_departure_delay DESC
267 LIMIT 10;
268

```

	ORIGIN_AIRPORT	avg_departure_delay
1	14222	89.11
2	ILG	29.57
3	13964	28.84
4	MVY	25.91
5	HYA	23.18
6	10154	22.86
7	10581	20.11
8	STC	18.69
9	OTH	17.78
10	10165	17.56

```

271 -- Route Analysis
272 -- Top routes:
273
274 SELECT
275     ORIGIN_AIRPORT,
276     DESTINATION_AIRPORT,
277     COUNT(*) AS flights
278 FROM flights
279 GROUP BY ORIGIN_AIRPORT, DESTINATION_AIRPORT
280 ORDER BY flights DESC
281 LIMIT 10;
282
283

```

	ORIGIN_AIRPORT	DESTINATION_AIRPORT	flights
1	SFO	LAX	13744
2	LAX	SFO	13457
3	JFK	LAX	12016
4	LAX	JFK	12015
5	LAS	LAX	9715
6	LGA	ORD	9639
7	LAX	LAS	9594
8	ORD	LGA	9575
9	SFO	JFK	8440
10	JFK	SFO	8437

```

284 -- Most delayed routes:
285
286 SELECT
287     ORIGIN_AIRPORT,
288     DESTINATION_AIRPORT,
289     ROUND(AVG(ARRIVAL_DELAY), 2) AS avg_delay
290 FROM flights
291 WHERE CANCELLED = 0
292 GROUP BY ORIGIN_AIRPORT, DESTINATION_AIRPORT
293 ORDER BY avg_delay DESC
294 LIMIT 10;
295

```

	ORIGIN_AIRPORT	DESTINATION_AIRPORT	avg_delay
1	IAD	TTN	381.0
2	SWF	PBI	260.5
3	RIC	CAE	228.0
4	RDU	IND	208.0
5	10581	11618	163.0
6	FCA	MSO	148.0
7	SWF	RSW	140.0
8	10581	12953	138.33
9	14843	12264	122.0
10	11433	11423	106.0

```

300 -- Cancellations Analysis
301
302 SELECT
303     CANCELLATION_REASON,
304     COUNT(*) AS cancel_count
305 FROM flights
306 WHERE CANCELLED = 1
307 GROUP BY CANCELLATION_REASON
308 ORDER BY CANCELLATION_REASON;
309

```

	CANCELLATION_REASON	cancel_count
1	A	25262
2	B	48851
3	C	15749
4	D	22

```

312 -- Cancel rate by month:
313
314 SELECT
315     MONTH,
316     ROUND(100.0 * SUM(CANCELLED) / COUNT(*), 2) AS cancel_rate
317 FROM flights
318 GROUP BY MONTH
319 ORDER BY cancel_rate DESC;

```

	MONTH	cancel_rate
1	2	4.78
2	1	2.55
3	3	2.18
4	6	1.81
5	12	1.68
6	5	1.15
7	8	0.99
8	11	0.98
9	4	0.93
10	7	0.92
11	10	0.5
12	9	0.45

```

325 -- Delays Based on Distance
326
327 SELECT
328     ROUND(DISTANCE / 100) * 100 AS distance_bucket,
329     ROUND(AVG(ARRIVAL_DELAY), 2) AS avg_arrival_delay
330 FROM flights
331 WHERE CANCELLED = 0
332 GROUP BY distance_bucket
333 ORDER BY avg_arrival_delay DESC;
334
335 -- Insight: Do longer flights face fewer or more delays?

```

	distance_bucket	avg_arrival_delay
1	3200.0	30.2
2	3700.0	14.5
3	3000.0	11.85
4	4100.0	10.55
5	4200.0	10.13
6	3900.0	9.76
7	3800.0	9.56
8	3300.0	6.92
9	0.0	6.5
10	4800.0	6.36
11	4900.0	6.25
12	300.0	5.6
13	700.0	5.43
14	1000.0	5.4

```

339 -- On-time Performance Distribution
340
341 -- Early Arrivals:
342
343 SELECT COUNT(*) AS early_arrivals
344 FROM flights
345 WHERE ARRIVAL_DELAY < 0;
346

```

	early_arrivals
1	3500899

```

346
347 -- Severe Delays:
348
349 SELECT COUNT(*) AS severely_delayed
350 FROM flights
351 WHERE ARRIVAL_DELAY > 60;
352

```

	severely_delayed
1	319092

```

357 -- Delay Reason Breakdown'
358
359 SELECT
360     ROUND(AVG(AIR_SYSTEM_DELAY), 2) AS air_system,
361     ROUND(AVG(SECURITY_DELAY), 2) AS security,
362     ROUND(AVG(AIRLINE_DELAY), 2) AS airline,
363     ROUND(AVG(LATE_AIRCRAFT_DELAY), 2) AS late_aircraft,
364     ROUND(AVG(WEATHER_DELAY), 2) AS weather
365 FROM flights
366 WHERE CANCELLED = 0;
367
368 -- Insight: What's the main reason behind delays?
369

```

	air_system	security	airline	late_aircraft	weather
1	2.5	0.01	3.52	4.36	0.54

```

372 -- Top 10 Airlines by Number of Flights
373
374 SELECT AIRLINE, COUNT(*) AS total_flights
375 FROM flights
376 GROUP BY AIRLINE
377 ORDER BY total_flights DESC
378 LIMIT 10;
379

```

	AIRLINE	total_flights
1	WN	1261855
2	DL	875881
3	AA	725984
4	OO	588353
5	EV	571977
6	UA	515723
7	MQ	294632
8	B6	267048
9	US	198715
10	AS	172521

```

381 -- Top 10 Busiest Origin Airports
382
383 SELECT ORIGIN_AIRPORT, COUNT(*) AS departures
384 FROM flights
385 GROUP BY ORIGIN_AIRPORT
386 ORDER BY departures DESC
387 LIMIT 10;
388

```

	ORIGIN_AIRPORT	departures
1	ATL	346836
2	ORD	285884
3	DFW	239551
4	DEN	196055
5	LAX	194673
6	SFO	148008
7	PHX	146815
8	IAH	146622
9	LAS	133181
10	MSP	112117

```

400 -- Delay Distribution (Arrival + Departure)
401
402 SELECT
403     COUNT(*) AS total_flights,
404     round(AVG(DEPARTURE_DELAY),2) AS avg_departure_delay,
405     round(AVG(ARRIVAL_DELAY),2) AS avg_arrival_delay,
406     MAX(DEPARTURE_DELAY) AS max_departure_delay,
407     MAX(ARRIVAL_DELAY) AS max_arrival_delay
408 FROM flights
409 WHERE CANCELLED = 0;
410
411

```

	total_flights	avg_departure_delay	avg_arrival_delay	max_departure_delay	max_arrival_delay
1	5729195	9.34	4.4	1988	1971

```

449 -- Avg Taxi_in and TAXI_OUT
450
451 SELECT
452     ORIGIN_AIRPORT,
453     round(AVG(TAXI_OUT),2) AS avg_taxi_out,
454     round(AVG(TAXI_IN),2) AS avg_taxi_in
455 FROM flights
456 GROUP BY ORIGIN_AIRPORT
457 ORDER BY avg_taxi_out DESC;
458

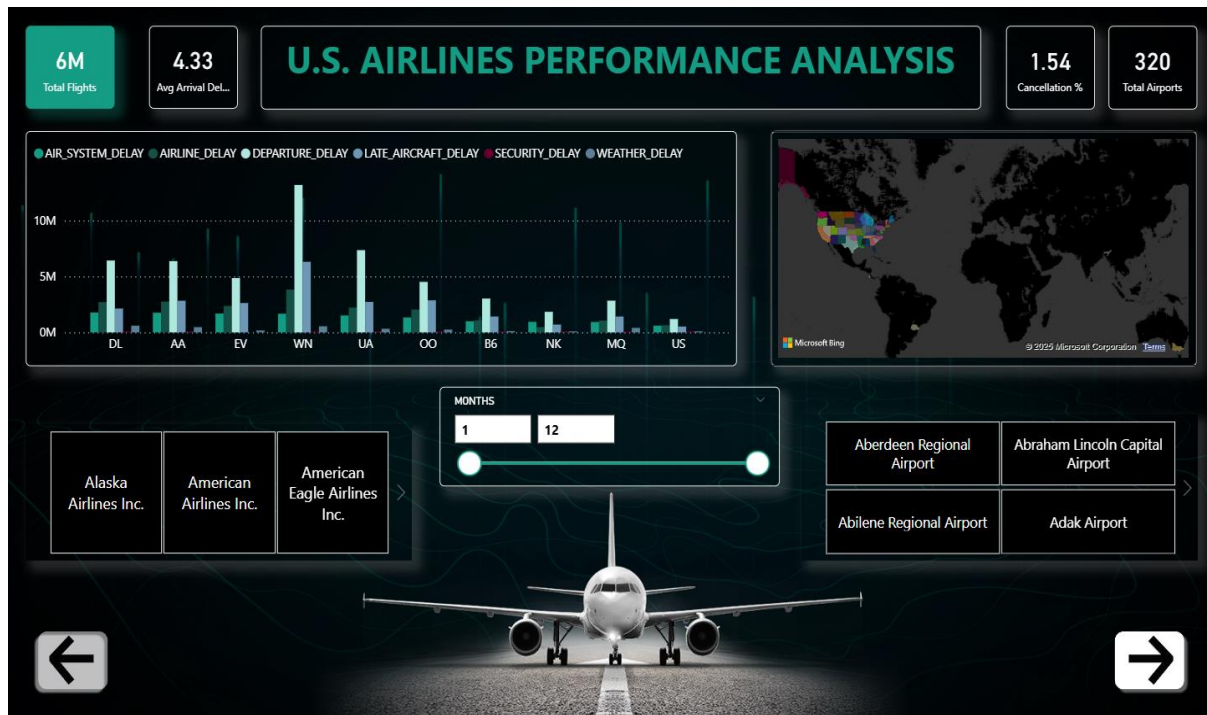
```

	ORIGIN_AIRPORT	avg_taxi_out	avg_taxi_in
1	JFK	27.01	7.19
2	LGA	26.72	7.82
3	12953	26.66	7.85
4	12478	23.68	7.23
5	13502	23.67	7.0
6	ITH	22.53	7.73
7	PHL	21.71	7.96
8	ISN	21.0	7.89
9	12389	20.51	7.87
10	EWR	20.19	7.77
11	BOS	20.02	7.74
12	11057	19.96	7.0
13	13930	19.94	6.4
14	14100	19.88	8.21

4. Power BI Dashboard

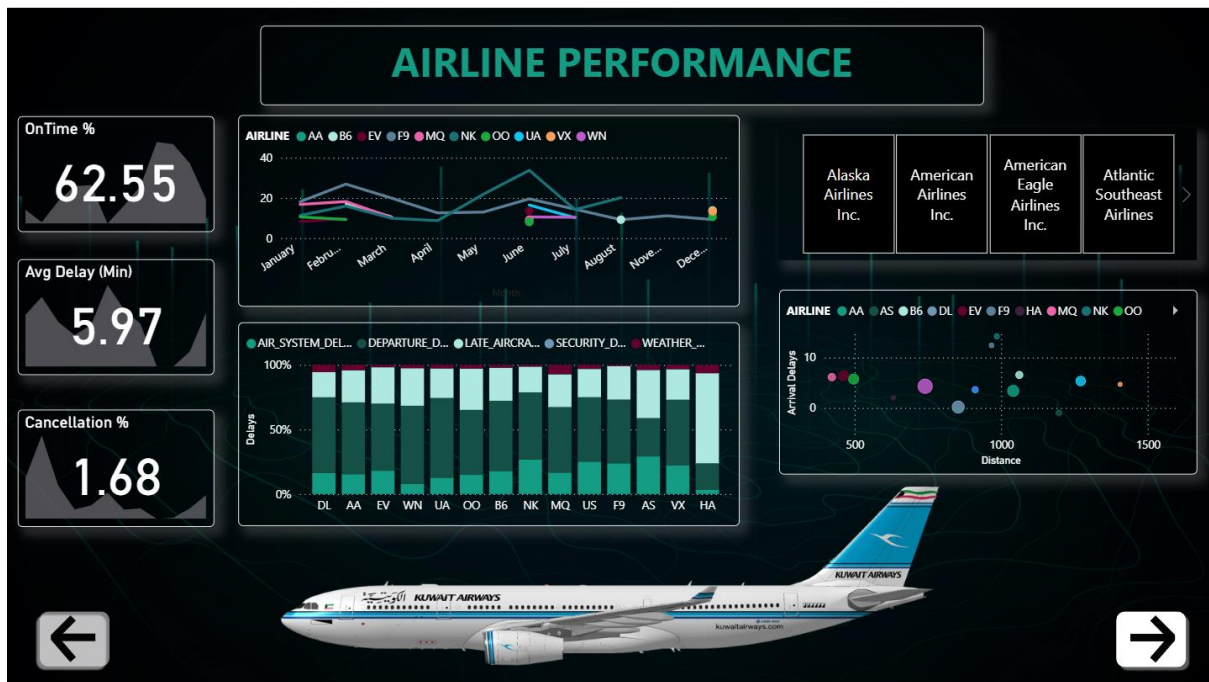
Page 1: Overview

- KPIs: Total flights, Avg delay, Cancellation %.
- Map visualization of airport delays.
- Airline vs Month matrix for trends.



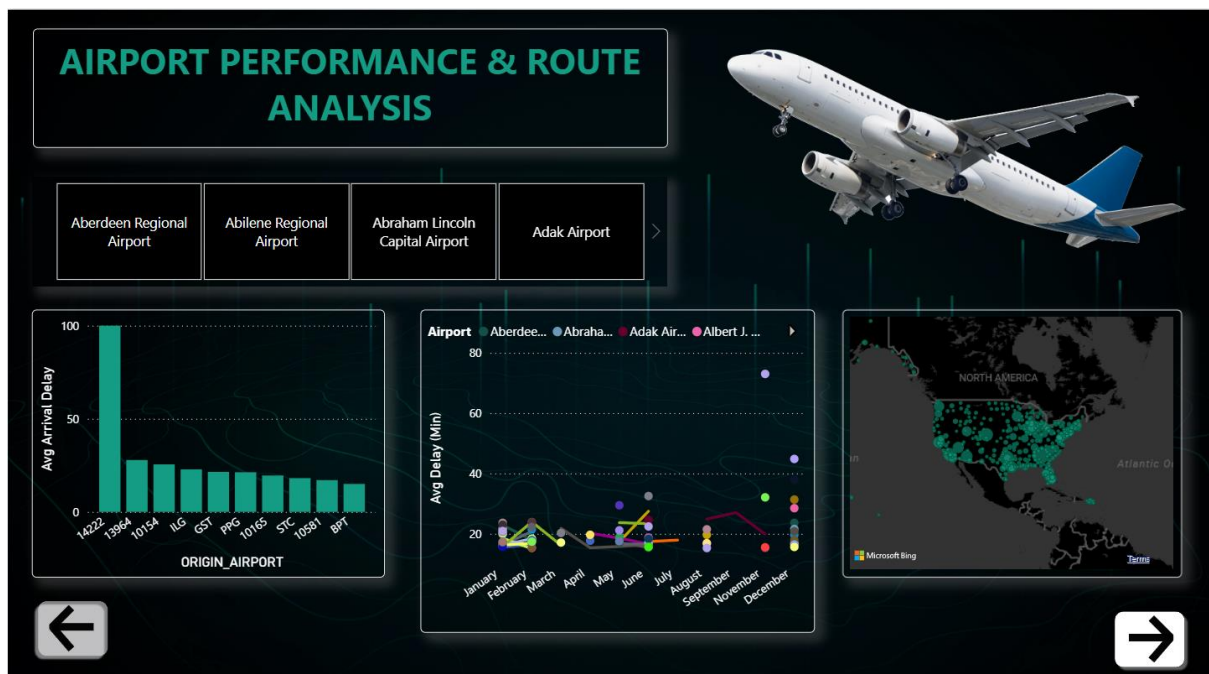
Page 2: Airline Performance

- Comparison of airlines by OTP %, Avg delay, and Cancellation %.
- Delay reasons stacked by airline.
- Scatter plot showing distance vs delay.



Page 3: Airport & Routes

- Top 10 airports by average delay.
- Monthly delay trends by airport.
- Route delay heatmap (Origin vs Destination).



Page 4: Delay Reasons & Predictive Insights

- Delay reason distribution (Weather, Airline, ATC, Late Aircraft, Security).
- Monthly trends of delays by reason.
- Forecast of average delay for upcoming months.



5. Key Findings & Analysis

Overall Performance

- Total flights analyzed: ~6 million.
- Average on-time performance (OTP): ~78%.
- Cancellation rate: ~2% of all flights.

Airline Insights

- Airlines with highest delays: Identified based on average arrival delay.
- Delay reasons vary by airline: Weather, late aircraft, NAS delays.

Airport Insights

- Congested airports (high taxi-out times) correlated with high average delays.
- Some airports consistently perform better in OTP despite traffic volume.

Time Trends

- Delays peak in summer months and during holidays.
- Morning flights have lower average delays compared to evening flights.

6. Recommendations

- Airlines should focus operational improvements at top delay-prone airports.
- Optimize flight schedules in months and days with peak delays.
- Improve turnaround times at congested airports to reduce taxi delays.
- Develop weather mitigation strategies to reduce weather-related delays.

7. Conclusion & Limitations

Conclusion

The project delivers actionable insights for airlines, airports, and regulatory authorities to improve OTP, reduce cancellations, and enhance passenger experience.

Limitations

- Dataset covers only 2015 flights; multiple years could improve trend stability.
- Weather and operational disruptions may vary yearly, requiring updated data.