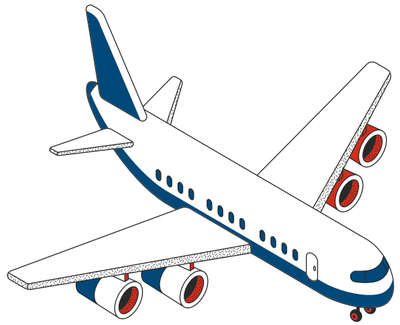
Technical Data Storytelling
Sample Work

Airline Data Challenge



The Problem Statement

- **Objective:** Enter the United States domestic airline market.
- **Initial Plan:** Launch 5 round trip routes between medium and large US airports.
- **Example Roundtrip Route:** JFK to ORD and ORD to JFK.
- **Investment:** Acquire 5 new airplanes, each costing $90 million.
- **Brand Motto:** "On time, for you" – emphasizing punctuality as a key brand value.

Available Data

- File Name: Flights.csv Loaded into Dataframe **Flights_df**
- Number of Rows: 1,915,886
- Number of Fields: 16

- File Name: Tickets.csv Loaded into Dataframe **Tickets_df**
- Number of Rows: 1,167,285
- Number of Fields: 12

- File Name: Airport_Codes.csv Loaded into Dataframe **Airport_Codes_df**
- Number of Rows: 55,369
- Number of Fields: 8

How Did I Approach?

Step 1:
Data Cleaning

Flights_filtered_df

Step 2:
10 Busiest Round Trip Routes

top_10_round_trips

Step 3:
10 Profitable Round Trip Routes

top_profitable_routes

Step 4:
5 Recommended Round Trips

Recommended_Paths

Step 5:
Round Trips to Break Even

Recommended_Paths_Filtered_df

Step 6:
Recommended KPI's

Data Cleaning

Step 1:

Libraries Used:

Numpy Pandas Plotly Matplotlib Seaborn Datetime

Reusable Functions Created:

1. **Lookup:** Merges dataframes on specified columns and Renames
2. **Clean_Numeric_Column:** Cleans input, converts to float, returns NaN if conversion fails.
3. **Clean_Percentile:** Removed top and bottom 1% outliers

How Did I Approach?

Quality Checks:

- **Data Completeness**: Filtered out rows with missing values
- **Data Consistency:** Removed non-numeric characters and corrected data type
- **Data Validation**: Excluded canceled flights, non-zero air time and distance
- **Handling Outliers**: Removed extreme values (top and bottom 1%) for key metrics.

Data Visualization:

- **Box Plots:** To check outliers and skewness in key metrics
- **Histograms:** To analyze data before and after outlier treatment

Data Cleaning

1) Lookup

Flights_df + Airport_Codes_df

ME	DEP_DELAY	ARR_DELAY	CANCELLED	AIR_TIME	DISTANCE	OCCUPANCY_RATE	ORIGIN_TYPE	DESTINATION_TYPE
OH	-8.00	-6.00	0.00	143.00	1,025.00	0.97	large_airport	large_airport
OH	1.00	5.00	0.00	135.00	930.00	0.55	large_airport	large_airport
OH	0.00	4.00	0.00	132.00	930.00	0.91	large_airport	large_airport
OH	11.00	14.00	0.00	136.00	930.00	0.67	large_airport	large_airport
TX	0.00	-17.00	0.00	151.00	1,005.00	0.62	large_airport	large_airport

Filtering Conditions:

- Large or Medium Airports.
- Flights were not cancelled.
- Has non- zero Air time and distance data.
- Occupancy rate data is available.

Number of Rows Originally: 1,915,886 ~ 1.91M

Number of rows after filtering= 1,844,131 ~ 1.84M

Data Cleaning

2) Clean_Numeric_Column

AIR_TIME	DISTANCE
$$$	****
$$$	****
$$$	*213
$$$	****

AIR_TIME	DISTANCE
NaN	NaN
NaN	NaN
NaN	213
NaN	NaN

Filtering Conditions:

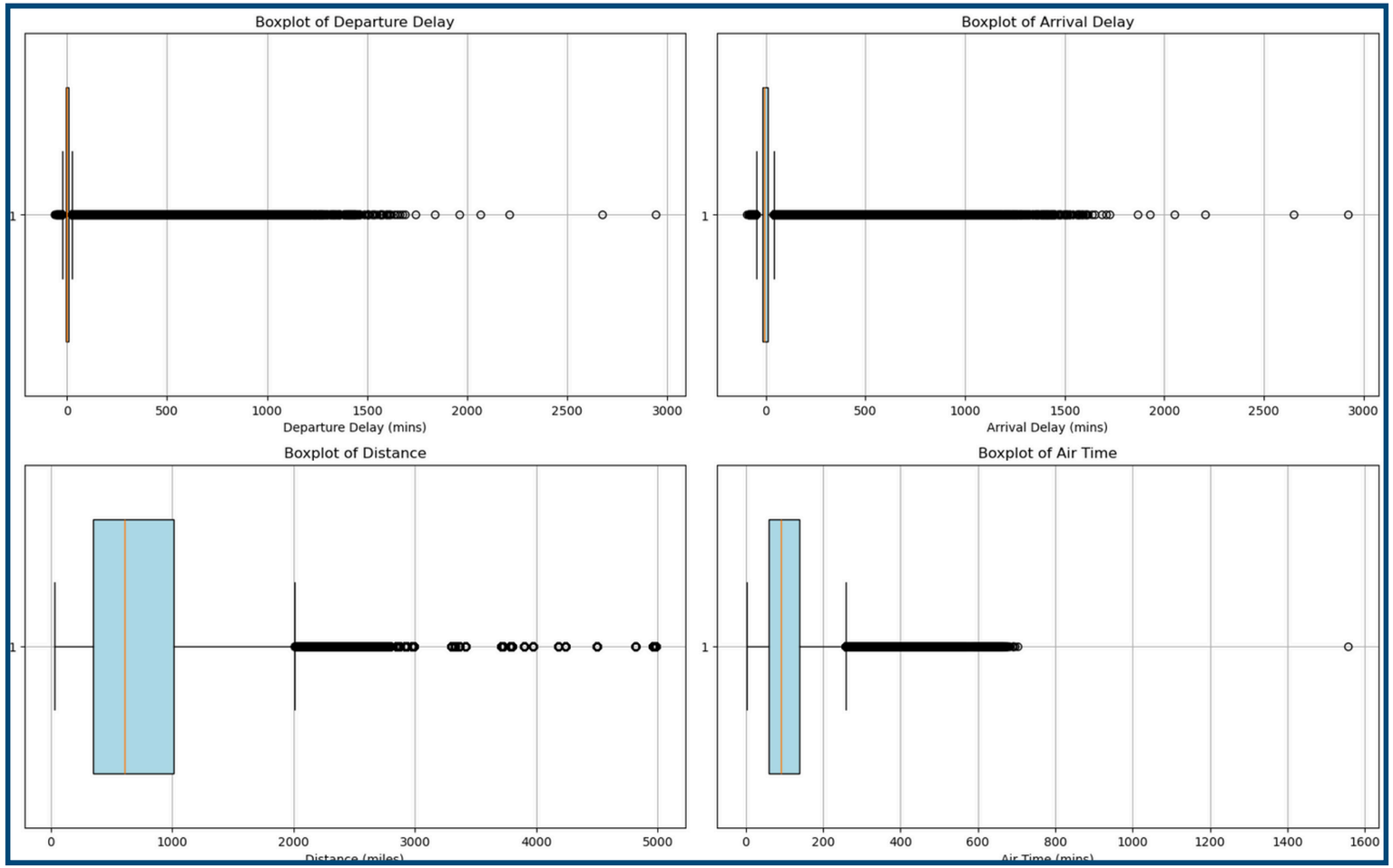
- Converting datatypes.
- Remove Invalid Data (NaN).
- Excluding the non departed flights (AIR_TIME or DISTANCE is ≤ 0)

Number of Rows Originally: 1,915,886 ~ 1.91M

Number of rows after filtering= 1,840,541 ~ 1.84M

Data Cleaning

Plotting Box Plot:



Matplotlib

Flights_filtered_df

Data Visualization

It is evident that each of these four Box Plots indicates a significant number of **outliers** that require attention.

Data Cleaning

Matplotlib

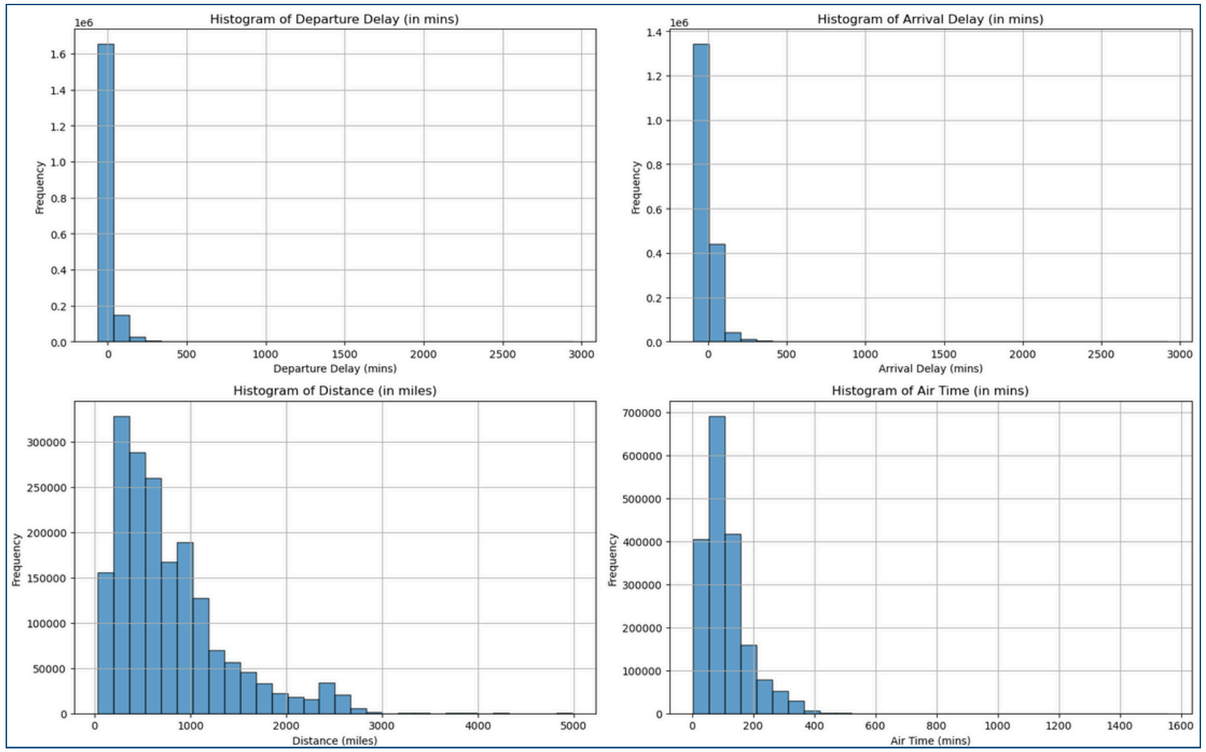
The data is right-skewed, with outliers at the higher end needing attention.

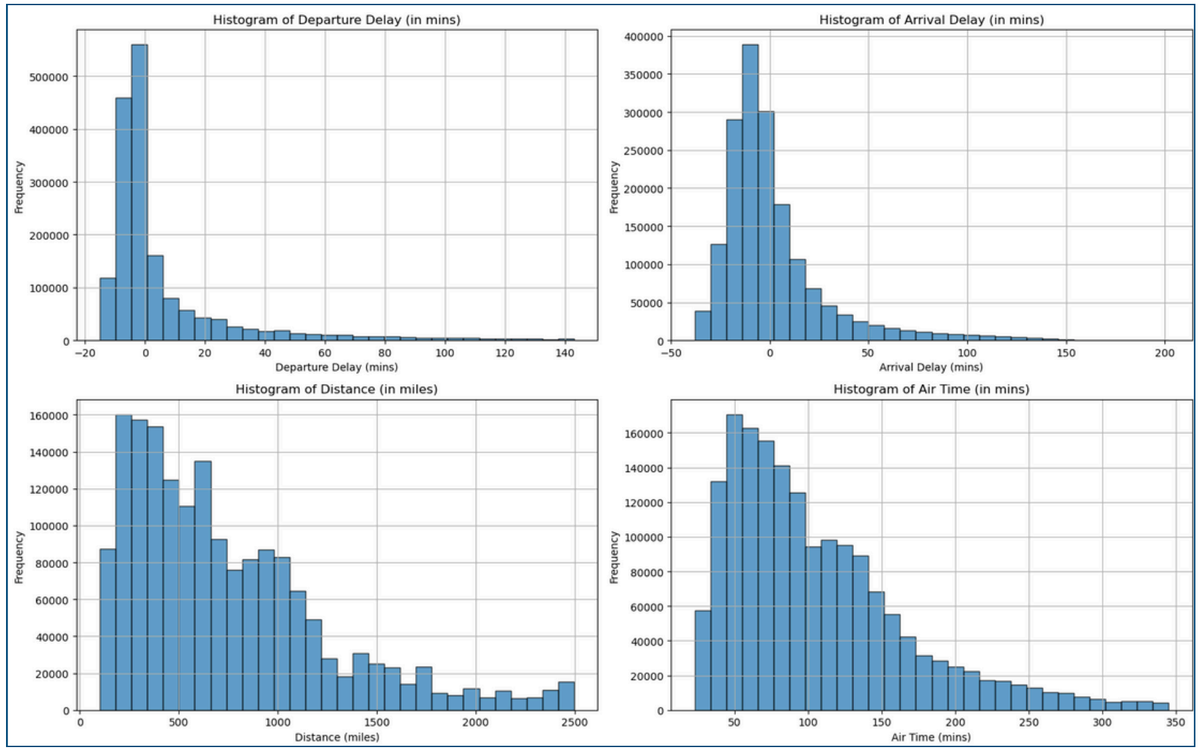
Plotting Histogram:

3) Clean_Percentile

Removed top and bottom 1% outliers of right-skewed data

Flights_filtered_df





Data Cleaning

Tickets_df

1) Lookup

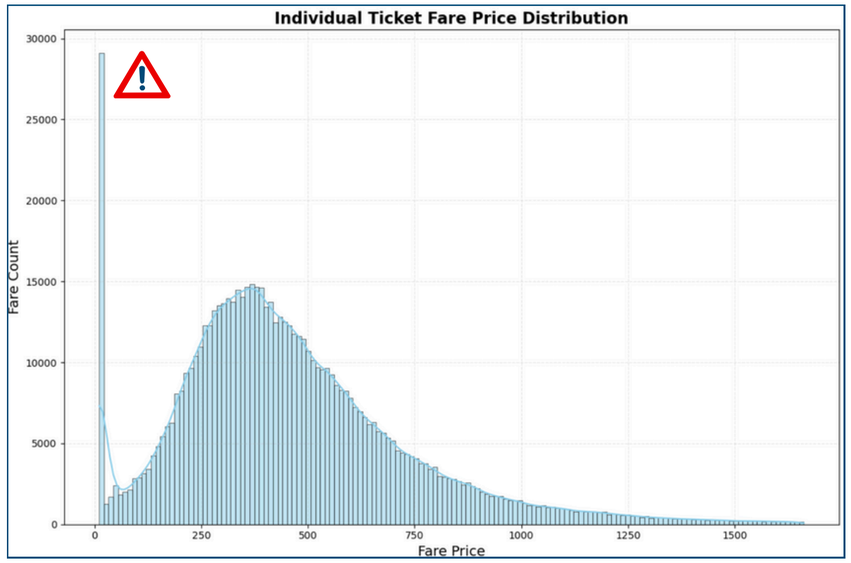
2) Clean_Numeric_Column

Filtering Conditions:

- Q1 2019 **round-trip** flights between **large** and **medium** US airports.
- Include passenger count and fares.

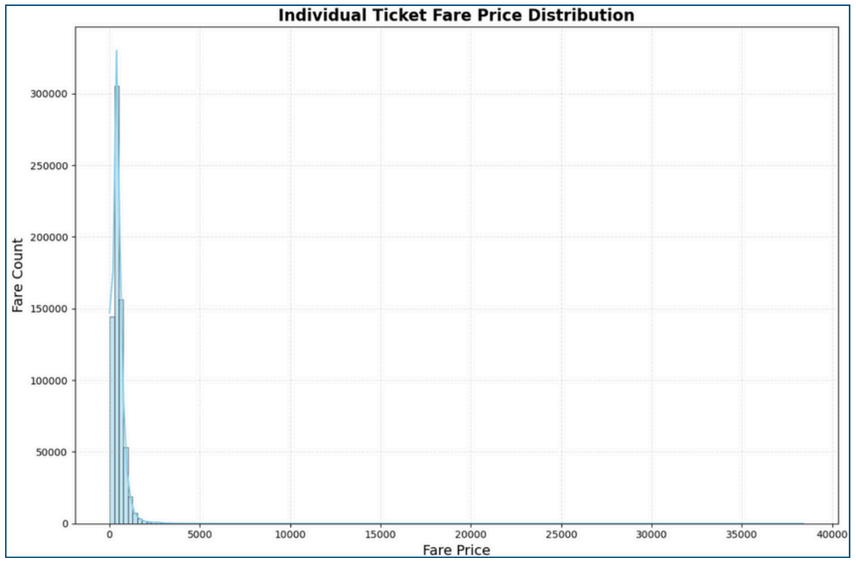
Tickets_filtered_df

Detected anomaly in 28,771 tickets priced at $11 across diverse flight routes.



Removed top and bottom 1% outliers of right-skewed data

3) Clean_Percentile



Data Cleaning Summary

- **Functions Created:** Lookup, Clean_Numeric_Column, Clean_Percentile
- **Flights.csv Rows Cleaned: 1.9M** to **1.7M** (**10.76%** of the rows cleaned)
- **Tickets.csv Rows Cleaned: 1.1M** to **681K** (**41.63%** of the **sample** data cleaned)
- **New Fields Created:** **ORIGIN_TYPE** and **DESTINATION_TYPE** created in both Flights and Tickets dataframes
- **Names:** Flights_df -> **Flights_filtered_df** and Tickets_df -> **Tickets_filtered_df**

Busiest Round Trip Routes

Step 2

Problem Statement:

Identify the 10 busiest round trip routes in terms of number of round trip flights in 1Q2019.

Reusable Functions Created:

4. **Round_Trip:** Concatenates origin and destination to find flight path regardless of order.

Busiest Round Trip Routes

4) Round Trip

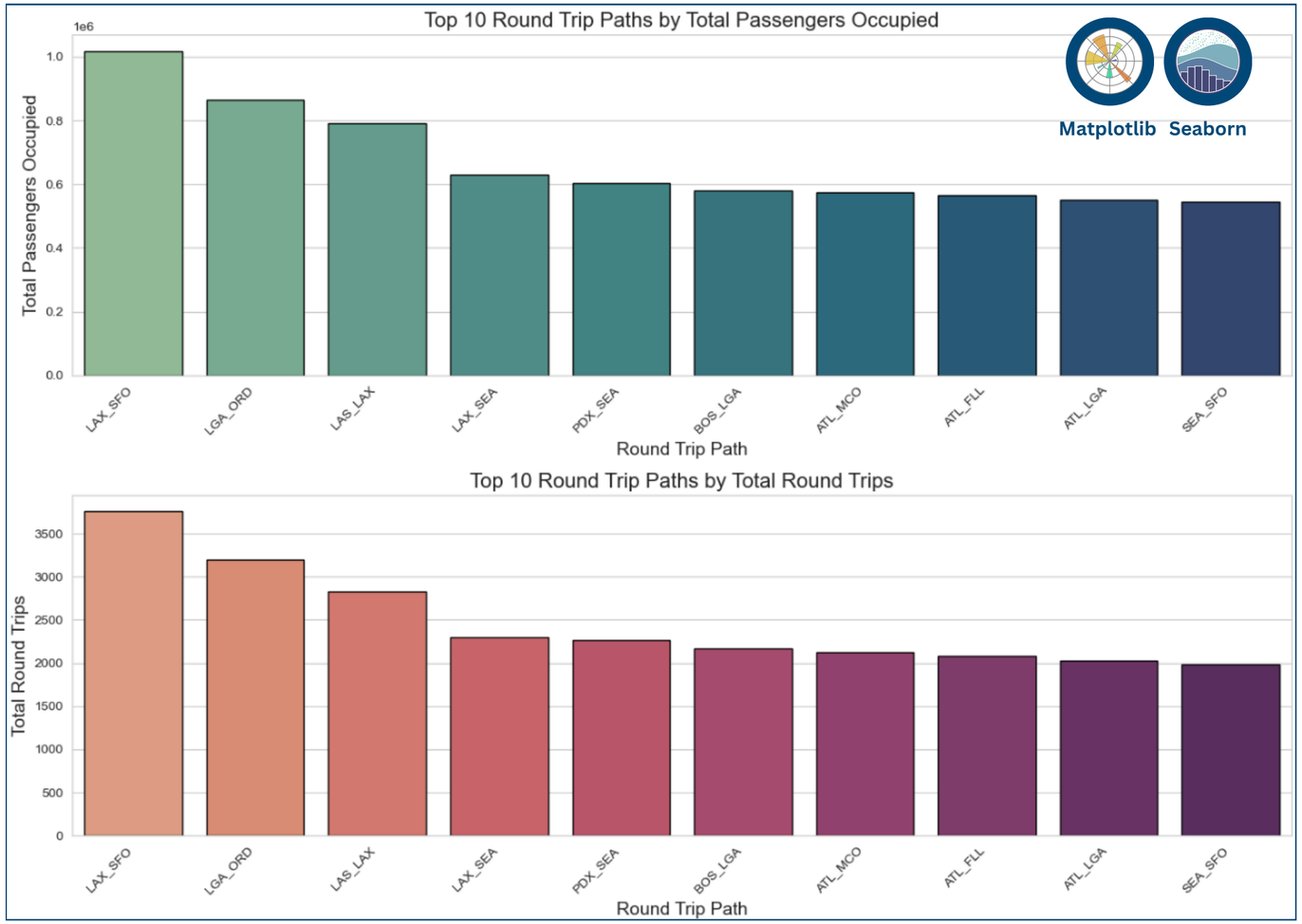
ORIGIN	DESTINATION
RSW	CLE
RSW	CMH
CMH	RSW
JFK	CMH

ORIGIN	DESTINATION	Flight_Path
RSW	CLE	RSW_CLE
RSW	CMH	RSW_CMH
CMH	RSW	RSW_CMH
JFK	CMH	JFK_CMH

Busiest Round Trip Routes

Calculate Passenger Occupancy	PASSENGERS_OCCUPIED = OCCUPANCY_RATE X 200
Round Trip Identification	Group by: Flight_Path and TAIL_NUM Filter: Flight_Count ≥ 2
Round Trip Calculation and Summarize	Round_Trip_Count = Flight_Count // 2 Total_Round_Trips = ∑ (Round_Trip_Count)
Sorting	Total_Passengers_Occupied: Descending Total_Round_Trips: Descending

Busiest Round Trip Routes



Top 10 Routes

By Total Passengers Occupied:

- **Identifies high-demand routes** based on passenger occupancy.
- **Optimize service delivery** by prioritizing high-demand routes.

By Total Round Trips:

- Highlight routes with **frequent round trip patterns**.
- **Allocate resources efficiently** based on repetitive travel demand.

Busiest Round Trip Routes

Final Output:

Out[52]:

	Flight_Path	Total_Round_Trips	Total_Passengers_Occupied
2077	LAX_SFO	3758	1,018,254.00
2118	LGA_ORD	3202	866,108.00
1985	LAS_LAX	2829	791,502.00
2076	LAX_SEA	2296	629,440.00
2515	PDX_SEA	2266	603,760.00
525	BOS_LGA	2170	581,536.00
178	ATL_MCO	2121	574,294.00
146	ATL_FLL	2085	565,112.00
174	ATL_LGA	2031	551,090.00
2672	SEA_SFO	1987	545,578.00

Top 10 Busiest Round Trip Routes

1. Los Angeles, CA (LAX) - San Francisco, CA (SFO)
2. New York, NY (LGA) - Chicago, IL (ORD)
3. Las Vegas, NV (LAS) - Los Angeles, CA (LAX)
4. Los Angeles, CA (LAX) - Seattle, WA (SEA)
5. Portland, OR (PDX) - Seattle, WA (SEA)
6. Boston, MA (BOS) - New York, NY (LGA)
7. Atlanta, GA (ATL) - Orlando, FL (MCO)
8. Atlanta, GA (ATL) - Fort Lauderdale, FL (FLL)
9. Atlanta, GA (ATL) - New York, NY (LGA)
10. Seattle, WA (SEA) - San Francisco, CA (SFO)

Profitable Round Trip Routes

Step 3

Problem Statement:

Identify the 10 most profitable round trip routes (without considering the upfront airplane cost) in 1Q2019.

Profitable Round Trip Routes

Calculate Average Ticket Price

Avg_Ticket_Price = Mean (ITIN_FARE) for each Flight_Path

Flight_Path	Avg_Ticket_Price
ABE_ABI	758.00
ABE_ABQ	534.00
ABE_AGS	391.00
ABE_AMA	654.00
ABE_ASE	742.00

Combining Average Ticket Prices with Flights Data on Flight_Path

TAIL_NUM	ORIGIN	DESTINATION	DEP_DELAY	ARR_DELAY	Flight_Path	PASSENGERS_OCCUPIED	Avg_Ticket_Price
N955WN	RSW	CLE	-8.00	-6.00	CLE_RSW	194.00	289.26
N754SW	RSW	CLE	-7.00	-22.00	CLE_RSW	126.00	289.26
N14249	CLE	RSW	-10.00	-23.00	CLE_RSW	152.00	289.26
N14240	RSW	CLE	-1.00	8.00	CLE_RSW	72.00	289.26
N14240	CLE	RSW	-4.00	-23.00	CLE_RSW	132.00	289.26

Profitable Round Trip Routes

Revenue

Ticket Revenue =
Occupancy Rate X 200 X Avg Ticket Price

Baggage Revenue =
Occupancy Rate X 200 X 0.5 X $70

Total Revenue =
Ticket Revenue + Baggage Revenue

Cost

Mileage Cost (Fuel maintenance and Depreciation Insurance) **=**
Distance X ($8 + $1.18)

Airport Cost =
$5,000 for medium airports and $10,000 for large airports

Delay Cost =
$75 X MAX(0, Delay Minutes - 15), applied only if Delay Minutes > 0
Delay costs apply at $75 per minute after a 15-minute grace period, for late flights only.

Total Cost = Mileage Cost + Airport Cost + Delay Cost

Profitable Round Trip Routes

Total Profit = **Total Revenue** − **Total Cost**

FL_DATE	TAIL_NUM	ORIGIN	DESTINATION	DEP_DELAY	ARR_DELAY	AIR_TIME	DISTANCE	OCCUPANCY_RATE	ORIGIN_TYPE	...	Airport_Cost	Dep_Delay_Cost	Arr_Delay_Cost	Total_Delay_Cost	Total_Cost	Ticket_Revenue	Baggage_Fee_Revenue	Total_Revenue	Profit
2019-03-02	N955WN	RSW	CLE	-8.00	-6.00	143.00	1,025.00	0.97	large_airport	...	20000	0.00	0.00	0.00	29,409.50	56,116.01	6,790.00	62,906.01	33,496.51
2019-03-09	N754SW	RSW	CLE	-7.00	-22.00	137.00	1,025.00	0.63	large_airport	...	20000	0.00	0.00	0.00	29,409.50	36,446.48	4,410.00	40,856.48	11,446.98
2019-03-24	N14249	CLE	RSW	-10.00	-23.00	136.00	1,025.00	0.76	large_airport	...	20000	0.00	0.00	0.00	29,409.50	43,967.18	5,320.00	49,287.18	19,877.68
2019-03-11	N14240	RSW	CLE	-1.00	8.00	138.00	1,025.00	0.36	large_airport	...	20000	0.00	0.00	0.00	29,409.50	20,826.56	2,520.00	23,346.56	-6,062.94
2019-03-11	N14240	CLE	RSW	-4.00	-23.00	130.00	1,025.00	0.66	large_airport	...	20000	0.00	0.00	0.00	29,409.50	38,182.02	4,620.00	42,802.02	13,392.52

Calculating Average Daily Flights (ADF) as it helps in capacity planning and operational efficiency.

Group Daily Flights:

Counting daily flights by grouping data by Flight_Path, TAIL_NUM, and FL_DATE

Calculate Average Daily Flights:

Computing average daily flights for each TAIL_NUM within each Flight_Path

Compute Overall ADF:

Calculate overall average daily flights (ADF) across all tail numbers for each Flight_Path

Profitable Round Trip Routes

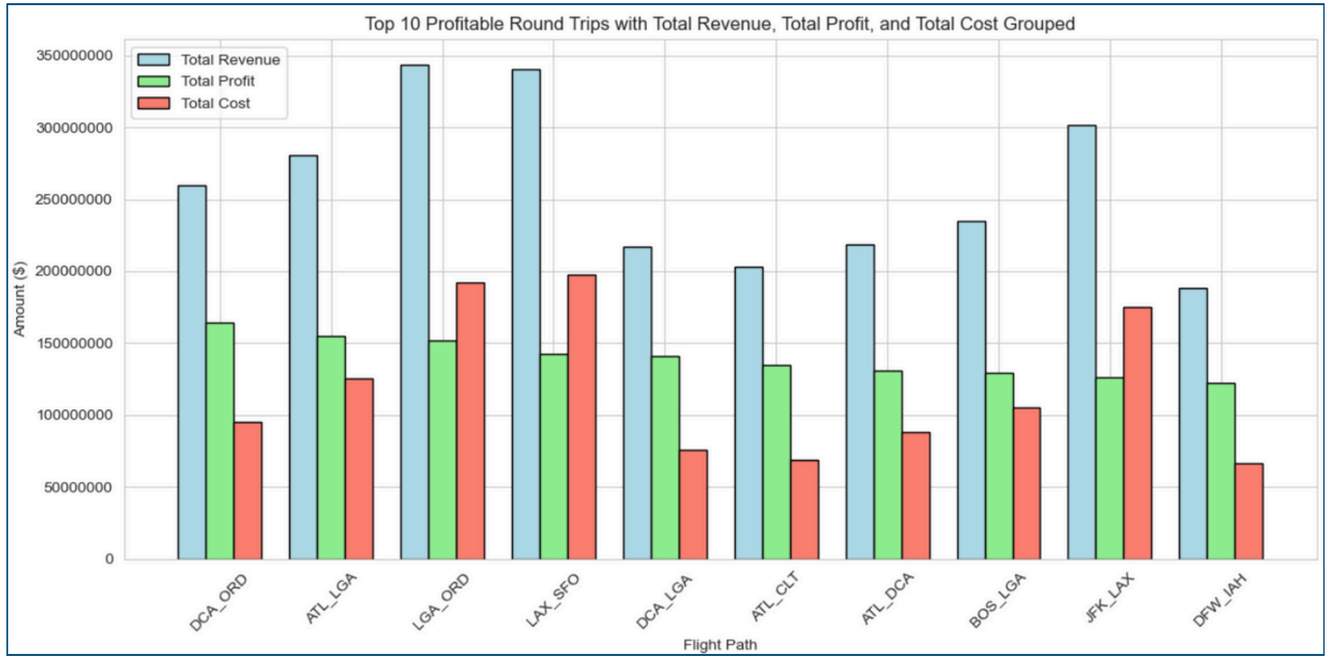
Top 10 Profitable Routes:

Flight_Path	Total_Profit	Total_Revenue	Total_Cost	Total_Round_Trip_Flights	Total_Passengers_Occupied	Avg_Ticket_Price	Total_Distance	Total_Mileage_Cost	Total_Airport_Cost	Total_Dep_Delay_Cost	Total_Baggage_Fee_Revenue	Average_Average_Daily_Flights
DCA_ORD	164,101,597.10	259,434,989.38	95,333,392.28	3583	464,304.00	523.76	2,192,796.00	20,129,867.28	71660000	1,656,300.00	16,250,640.00	1.39
ATL_LGA	155,286,624.99	280,795,942.59	125,509,317.60	4485	584,148.00	445.69	3,417,570.00	31,373,292.60	89700000	2,113,650.00	20,445,180.00	1.24
LGA_ORD	151,796,332.99	343,956,966.03	192,160,633.04	6766	877,430.00	357.01	4,959,478.00	45,528,008.04	135320000	5,044,275.00	30,710,050.00	1.72
LAX_SFO	142,849,235.23	340,836,858.17	197,987,622.94	8009	1,040,274.00	292.64	2,699,033.00	24,777,122.94	160180000	6,025,875.00	36,409,590.00	2.06
DCA_LGA	141,081,382.54	216,952,218.50	75,870,835.96	3248	424,656.00	475.89	695,072.00	6,380,760.96	64960000	1,866,900.00	14,862,960.00	2.22
ATL_CLT	134,437,957.99	203,026,334.87	68,588,376.88	3041	394,584.00	479.53	687,266.00	6,309,101.88	60820000	748,875.00	13,810,440.00	1.47
ATL_DCA	130,866,045.90	218,848,993.30	87,982,947.40	3440	447,004.00	454.59	1,881,680.00	17,273,822.40	68800000	879,450.00	15,645,140.00	1.27
BOS_LGA	129,525,785.68	234,820,051.36	105,294,265.68	4539	591,240.00	362.17	835,176.00	7,666,915.68	90780000	3,266,700.00	20,693,400.00	2.02
JFK_LAX	126,523,370.14	302,041,924.64	175,518,554.50	4049	528,704.00	536.29	10,021,275.00	91,995,304.50	80980000	1,489,800.00	18,504,640.00	1.19
DFW_IAH	122,226,010.86	188,611,944.62	66,385,933.76	2893	376,898.00	465.43	648,032.00	5,948,933.76	57860000	1,193,175.00	13,191,430.00	1.43

Key Insights:

- **Top Performing Routes**: **DCA_ORD** and **ATL_LGA** are the most profitable, generating **over $150 million** in profit each. They also have **high total revenues** and **relatively low costs** compared to other routes.
- **High Passenger Volume**: **LGA_ORD** and **LAX_SFO** have the **highest total passengers** occupied (~877K and ~1M, respectively), indicating **strong demand** and **high utilization**.
- **Strategic Opportunities in High-Density Short-Haul Routes**: Routes like **LAX_SFO** and **BOS_LGA** leverage **lower average ticket prices** with **high total passenger volume** which drives substantial **baggage fee revenue**.

Profitable Round Trip Routes



Top 10 Profitable Routes

1. Washington, D.C. (DCA) - Chicago, IL (ORD)
2. Atlanta, GA (ATL) - New York, NY (LGA)
3. New York, NY (LGA) - Chicago, IL (ORD)
4. Los Angeles, CA (LAX) - San Francisco, CA (SFO)
5. Washington, D.C. (DCA) - New York, NY (LGA)
6. Atlanta, GA (ATL) - Charlotte, NC (CLT)
7. Atlanta, GA (ATL) - Washington, D.C. (DCA)
8. Boston, MA (BOS) - New York, NY (LGA)
9. New York, NY (JFK) - Los Angeles, CA (LAX)
10. Dallas, TX (DFW) - Houston, TX (IAH)

Recommended Routes

Step 4

Problem Statement:

5 round trip routes that you recommend to invest in based on any factors that you choose.

Recommended Routes

Defining KPIs

On-Time Performance (OTP):
% of flights departing and arriving on schedule

$$\frac{\text{On-time flights}}{\text{Total flights}} \times 100$$

On-time flight:
DEP_DELAY & ARR_DELAY ≤ 0

Available Seat Miles (ASM):
Total seat miles available for passengers

200 X Flight Distance

Cost per Available Seat Mile (CASM):
Cost efficiency of operations

$$\frac{\text{Total Cost}}{\text{ASM}}$$

Revenue per Available Seat Mile (RASM):
Revenue generation relative to capacity

$$\frac{\text{Total Revenue}}{\text{ASM}}$$

Profit Per Passenger:
Profitability per occupied seat

$$\frac{\text{Profit}}{\text{Number of Occupied Passengers}}$$

Recommended Routes

Defining KPI

Breakeven Round Trips:
Number of round-trip flights needed to recover the **$90 million** airplane cost for each route (CapEx)

For Each Flight Path:

$$\frac{\text{Airplane Cost (\$90M)}}{\text{Total Profit}} \text{ X Total Round Trip Flights}$$

Summary Statistics:

	Total_Revenue	Total_Cost	Total_Profit	Average_Flight_Path_OTP	Average_Daily_Flights	Average_ASM	Average_CASM	Average_RASM	Average_Occupancy_Rate	Average_Profit_Per_Passenger	Breakeven_Round_Trips
count	2,734.00	2,734.00	2,734.00	2,734.00	2,734.00	2,734.00	2,734.00	2,734.00	2,734.00	2,734.00	2,734.00
mean	37,341,838.13	16,762,039.28	20,579,798.85	55.79	1.42	171,093.66	0.22	0.53	0.65	221.66	-21,588.36
std	40,000,112.94	19,309,060.72	22,213,210.33	11.46	0.33	107,658.25	0.14	0.46	0.02	119.15	1,570,416.45
min	42,288.00	19,030.20	-2,293,305.80	0.00	1.00	20,400.00	0.08	0.08	0.42	-130.82	-81,753,554.50
25%	9,383,030.30	4,544,280.13	4,595,480.24	49.65	1.12	89,050.00	0.13	0.24	0.64	144.54	2,079.02
50%	25,456,953.81	10,703,802.36	13,590,470.43	56.16	1.39	148,900.00	0.18	0.38	0.65	212.52	2,828.97
75%	50,533,284.80	21,523,836.70	28,510,100.86	63.26	1.67	224,200.00	0.26	0.65	0.66	299.54	3,931.70
max	343,956,966.03	197,987,622.94	164,101,597.10	100.00	3.19	499,200.00	1.02	3.76	0.95	683.00	6,115,530.09

Recommended Routes

Filtering Conditions

KPI	Threshold (Percentile)	Rationale
On Time Performance OTP	< 63.26 (75th Percentile)	**Improve Punctuality:** Targeting routes with **potential for better punctuality**
Available Seat Miles ASM	> 89,050 (25th Percentile)	**Maximize Revenue:** Focusing on routes with **significant seating capacity** and **distance**
Revenue per ASM RASM	> 0.24 (25th Percentile)	**Boost Revenue Efficiency:** Choosing routes **generating high revenue**
Occupancy Rate	> 0.64 (25th Percentile)	**Optimize Utilization**: Prefering routes with **higher seat occupancy**
Profit Per Passenger	> 144.54 (25th Percentile)	**Enhance Profit Margins**: Selecting routes with **high profit per passenger**
Breakeven Round Trips	< 3931.70 (75th Percentile)	**Minimize Financial Risk**: Ensuring **fewer flights needed to cover costs**

Recommended Routes

Competitive Advantage:

- Punctuality
- High Revenue

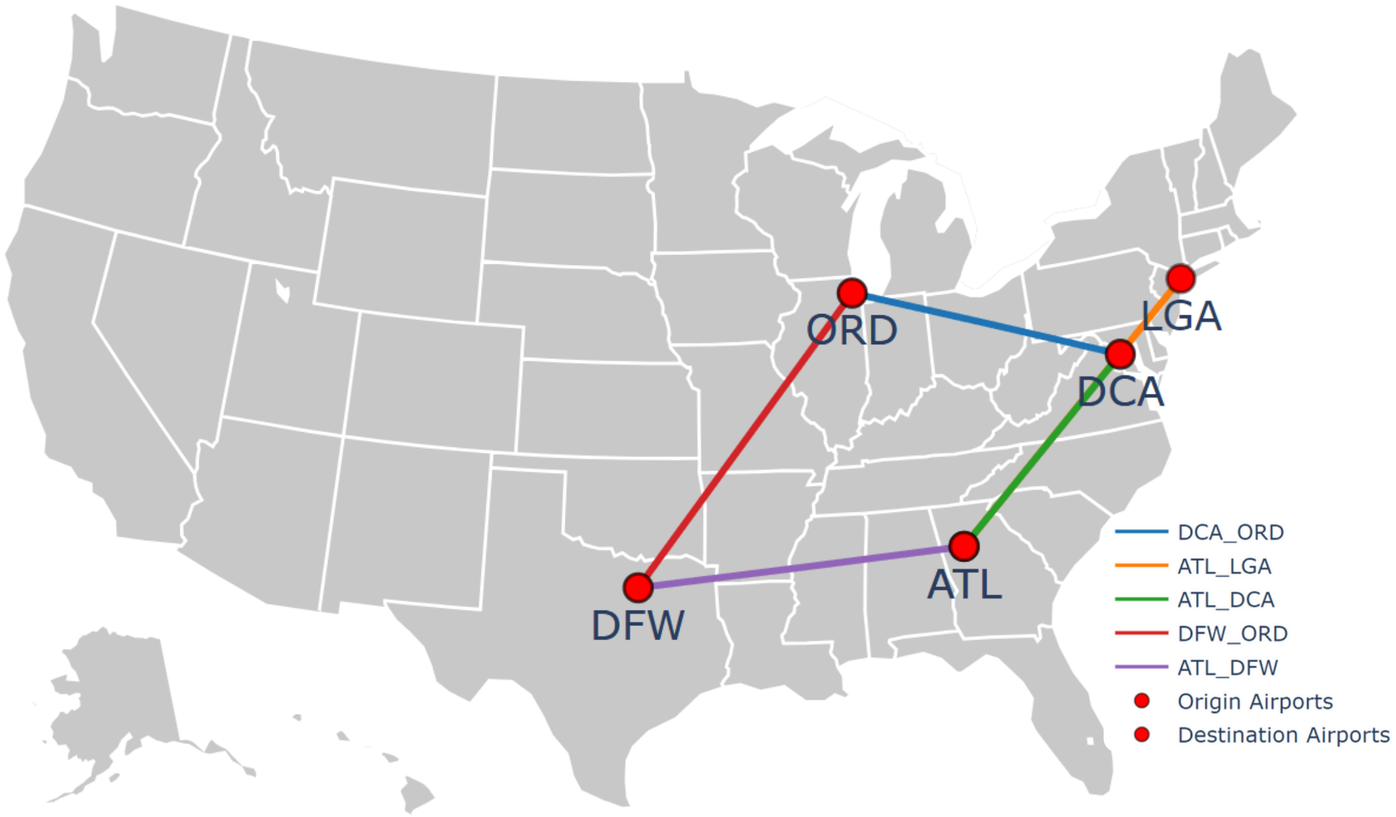
Operational Efficiency:

- Optimized Utilization
- Minimized Financial Risk

Top 5 Profitable Routes Meeting Filtering Conditions:

Flight_Path	Total_Profit	Average_Flight_Path_OTP	Average_Daily_Flights	Average_ASM	Average_CASM	Average_RASM	Average_Occupancy_Rate	Average_Profit_Per_Passenger
DCA_ORD	164,101,597.10	60.20	1.39	122,400.00	0.22	0.59	0.65	328.89
ATL_LGA	155,286,624.99	51.62	1.24	152,400.00	0.18	0.41	0.65	240.38
ATL_DCA	130,866,045.90	58.84	1.27	109,400.00	0.23	0.58	0.65	269.42
DFW_ORD	118,221,449.86	48.62	1.13	160,203.87	0.18	0.41	0.66	259.30
ATL_DFW	107,836,823.55	53.68	1.50	146,200.00	0.19	0.42	0.65	232.64

Recommended Routes



Reccomended Round Trip Routes

1. Washington, D.C. (DCA) - Chicago, IL (ORD)
2. Atlanta, GA (ATL) - New York, NY (LGA)
3. Atlanta, GA (ATL) - Washington, D.C. (DCA)
4. Dallas/Fort Worth, TX (DFW) - Chicago, IL (ORD)
5. Atlanta, GA (ATL) - Dallas/Fort Worth, TX (DFW)

Cyclic Route Advantages:

- Optimizes network efficiency and asset utilization
- Enhances passenger connectivity (fewer transfers)
- Improves scheduling flexibility, market penetration
- Provides competitive differentiation from point-to-point carriers

Breakeven

Step 5

Problem Statement:

Find the number of round trip flights it will take to breakeven on the upfront airplane cost. Print key summary components fo these routes.

Breakeven

Defining KPI

Break-Even Point (BEP in Days):
Days needed to reach the break-even

$$\frac{\text{Breakeven Round Trips}}{\text{Average Daily Flights}}$$

Final Output:

Flight_Path	Breakeven_Round_Trips	Average_Daily_Flights	Total_Cost	Total_Revenue	Total_Profit	Total_Passengers_Occupied	Average_Ticket_Price	BEP_in_Days
DCA_ORD	1,965.06	1.39	95,333,392.28	259,434,989.38	164,101,597.10	464,304.00	523.76	1,412.29
ATL_LGA	2,599.39	1.24	125,509,317.60	280,795,942.59	155,286,624.99	584,148.00	445.69	2,096.57
ATL_DCA	2,365.78	1.27	87,982,947.40	218,848,993.30	130,866,045.90	447,004.00	454.59	1,866.82
DFW_ORD	2,400.33	1.13	89,772,461.52	207,993,911.38	118,221,449.86	414,808.00	466.42	2,133.35
ATL_DFW	2,675.71	1.50	87,824,194.48	195,661,018.03	107,836,823.55	418,232.00	432.83	1,779.18

We can recover our capital in approximately **3 years and 10 months** to **5 years and 10 months**, after which we will start generating profit.High Revenue

Recommended KPIs

Step 6

Problem Statement:

Key Performance Indicators (KPI's) that you recommend tracking in the future to measure the success of the round trip routes that you recommend

Recommended KPIs

Why Track These KPIs?

- **Our Motto "On time, for you":** Ensuring punctuality and reliability.
- **Maximize Profits:** Focus on profitability, revenue, and cost-efficiency.
- **Operational Efficiency:** High resource utilization and minimal risk.
- **Customer-Centric Approach:** Enhancing customer satisfaction and loyalty.
- **Competitive Advantage:** Staying ahead of competitors with superior service and efficiency.
- **Sustainability:** Minimizing environmental impact.

Recommended KPIs

Primary KPIs

Financial Metrics:

- **Profit:** Net income from airline operations.
- **Break-Even Point (BEP in Days):** Days needed to reach the breakeven point.

Operational Metrics:

- **On-Time Performance (OTP):** Percentage of flights departing and arriving on time.
- **Aircraft Utilization:** Time an aircraft is generating revenue.
- **Operational Reliability:** Flights operating as scheduled without major issues.

Customer-Centric Metrics:

- **Customer Satisfaction Score:** Passenger satisfaction via surveys.
- **Customer Retention Rate:** Percentage of repeat bookings.

Recommended KPIs

Secondary KPIs

- **Profit Per Passenger:** Profit earned per passenger.
- **Route Profitability:** Profitability of each flight route.
- **Available Seat Miles (ASM):** Seat miles available for passengers.
- **Cost per Available Seat Mile (CASM):** Cost per seat mile available for passengers.
- **Revenue per Available Seat Mile (RASM):** Revenue per seat mile available for passengers.
- **Average Daily Flights:** Average number of flights operated per day.
- **Market Share:** Seat miles or passengers compared to competitors.
- **Baggage Handling Efficiency:** Percentage of mishandled baggage.
- **Customer Loyalty Index:** Measurement of loyalty and repeat bookings.
- **Environmental Impact:** Airline’s environmental footprint.

Thank You