AIRLINE OPERATIONS ANALYSIS

FINAL PROJECT

Author: Nguyen Thi Phuong Trang

INTRODUCTION

This project analyzes an airline dataset to uncover key patterns in passenger demographics, flight performance, and geographic distribution. Through data cleaning, exploration, and visualization, the analysis reveals trends that support data-driven decisions aimed at improving operational efficiency, customer experience, and strategic planning within the aviation industry.

- **GitHub link:** *https://github.com/trangntp37/Airline_Analysis*
- **Dataset source:** *https://www.kaggle.com/datasets/iamsouravbanerjee/airline-dataset/data*

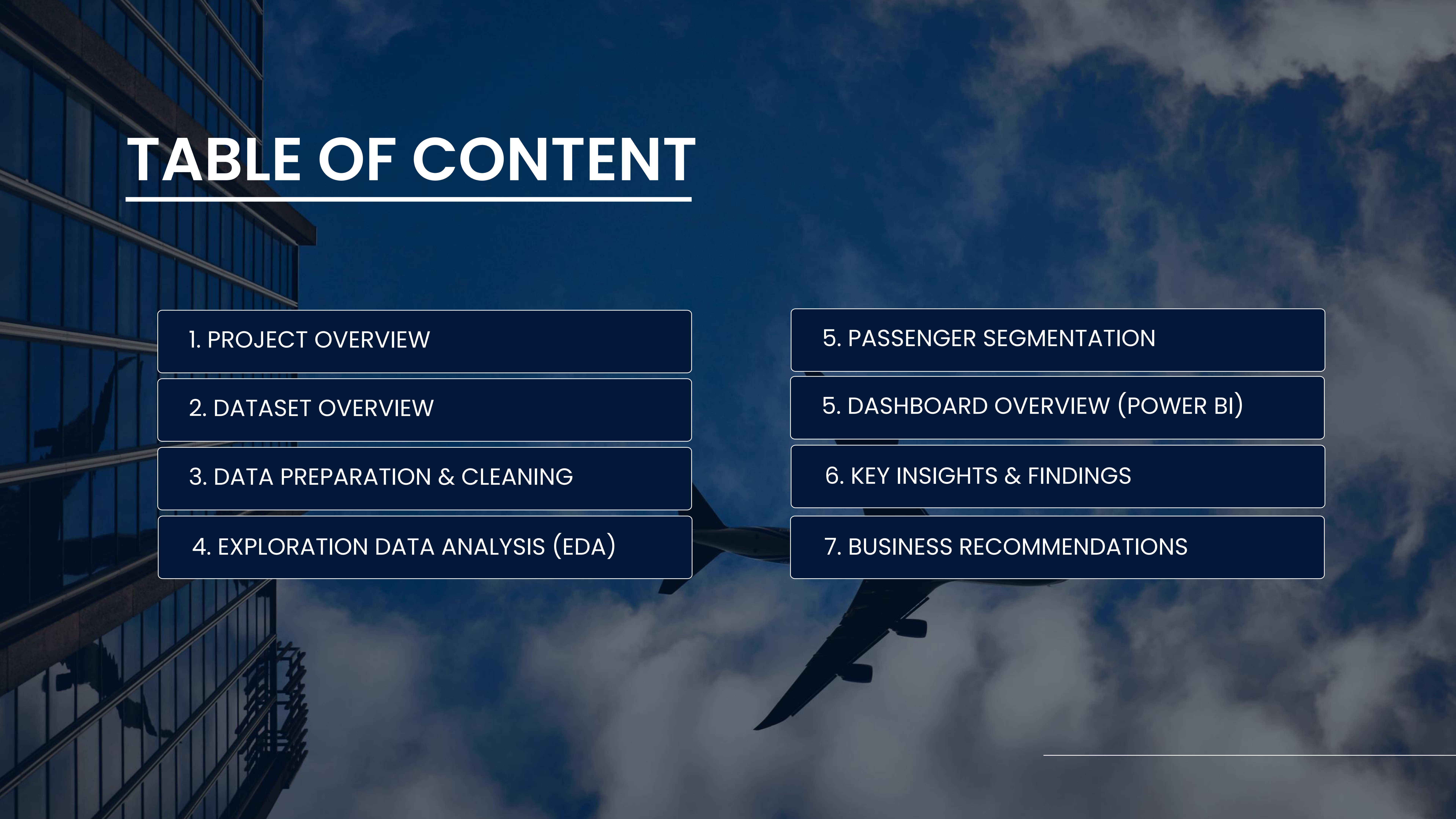


TABLE OF CONTENT

1. PROJECT OVERVIEW

2. DATASET OVERVIEW

3. DATA PREPARATION & CLEANING

4. EXPLORATION DATA ANALYSIS (EDA)

5. PASSENGER SEGMENTATION

5. DASHBOARD OVERVIEW (POWER BI)

6. KEY INSIGHTS & FINDINGS

7. BUSINESS RECOMMENDATIONS

1. Project Overview

Context

- Airlines generate large volumes of passenger data across different regions and routes, but translating this data into actionable customer insights remains challenging.

Problem

- Passenger populations are often treated as homogeneous, limiting the ability to understand demographic patterns and identify meaningful customer groups.

Challenges

- Limited behavioral variables in the dataset
- High-cardinality categorical features (e.g. nationality, airports)
- Potentially pre-processed or synthetic data with low natural variance

Objectives

- Analyze passenger demographics and geographic distribution
- Identify high-level passenger segments using unsupervised learning
- Summarize insights through interactive dashboards

Technology Used

- SQL Server (database)
- Python (Pandas, Scikit-learn) for data processing and clustering
- Power BI for visualization and dashboarding

2. Dataset Overview

❖ **Scope:** The dataset provides insights into various aspects of airline operations, covering passenger demographics, travel details, flight routes, crew information, and flight statuses.

❖ **Number of records:** 98,619

❖ **Key Columns:**

Passenger Information

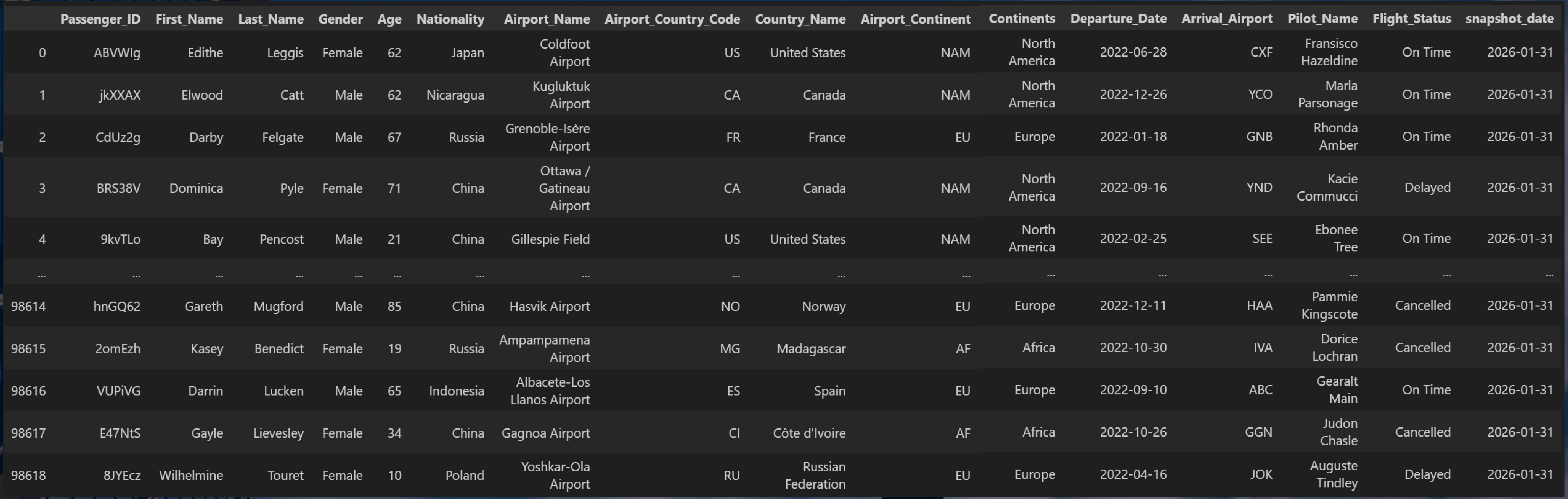
- Passenger ID
- First Name
- Last Name
- Gender
- Age
- Nationality

Flight & Airport Information

- Airport Name
- Airport Country Code
- Country Name
- Airport Continent
- Continents
- Departure Date
- Arrival Airport
- Pilot Name
- Flight Status

2. Dataset Overview

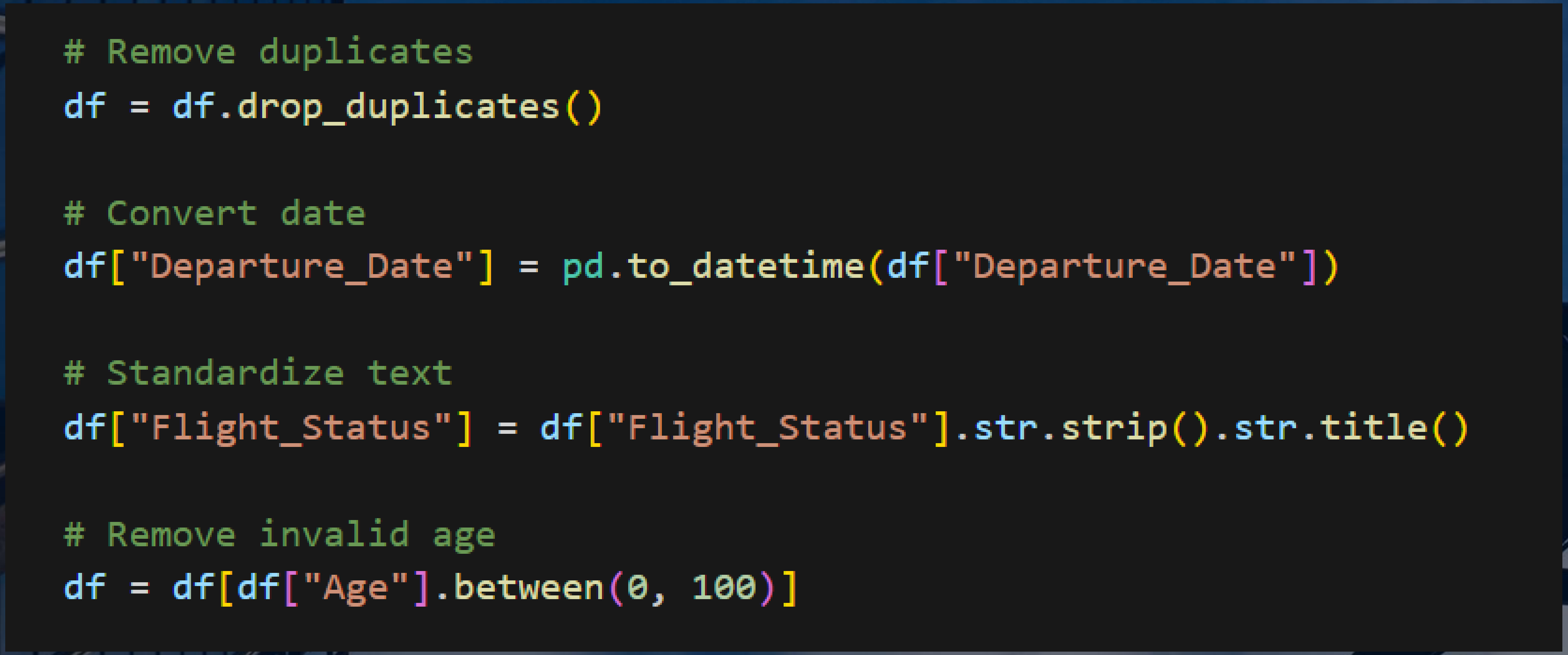
Data Frame Preview:



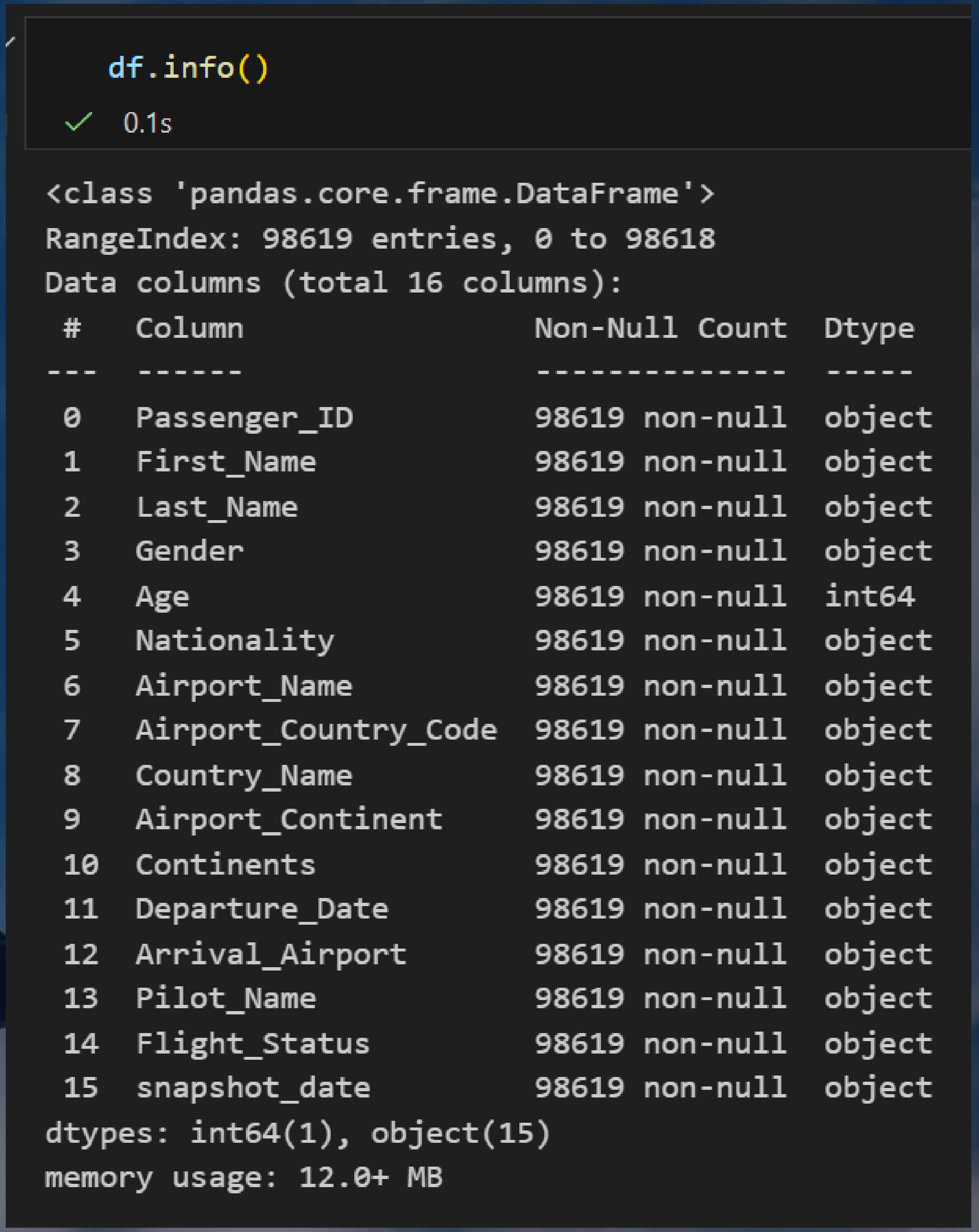
	Passenger_ID	First_Name	Last_Name	Gender	Age	Nationality	Airport_Name	Airport_Country_Code	Country_Name	Airport_Continent	Continents	Departure_Date	Arrival_Airport	Pilot_Name	Flight_Status	snapshot_date
0	ABVWIg	Edithe	Leggis	Female	62	Japan	Coldfoot Airport	US	United States	NAM	North America	2022-06-28	CXF	Fransisco Hazeldine	On Time	2026-01-31
1	jkXXAX	Elwood	Catt	Male	62	Nicaragua	Kugluktuk Airport	CA	Canada	NAM	North America	2022-12-26	YCO	Marla Parsonage	On Time	2026-01-31
2	CdUz2g	Darby	Felgate	Male	67	Russia	Grenoble-Isère Airport	FR	France	EU	Europe	2022-01-18	GNB	Rhonda Amber	On Time	2026-01-31
3	BRS38V	Dominica	Pyle	Female	71	China	Ottawa / Gatineau Airport	CA	Canada	NAM	North America	2022-09-16	YND	Kacie Commucci	Delayed	2026-01-31
4	9kvTLo	Bay	Pencost	Male	21	China	Gillespie Field	US	United States	NAM	North America	2022-02-25	SEE	Ebonee Tree	On Time	2026-01-31
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
98614	hnGQ62	Gareth	Mugford	Male	85	China	Hasvik Airport	NO	Norway	EU	Europe	2022-12-11	HAA	Pammie Kingscote	Cancelled	2026-01-31
98615	2omEzh	Kasey	Benedict	Female	19	Russia	Ampampamena Airport	MG	Madagascar	AF	Africa	2022-10-30	IVA	Dorice Lochran	Cancelled	2026-01-31
98616	VUPiVG	Darrin	Lucken	Male	65	Indonesia	Albacete-Los Llanos Airport	ES	Spain	EU	Europe	2022-09-10	ABC	Gearalt Main	On Time	2026-01-31
98617	E47NtS	Gayle	Lievesley	Female	34	China	Gagnoa Airport	CI	Côte d'Ivoire	AF	Africa	2022-10-26	GGN	Judon Chasle	Cancelled	2026-01-31
98618	8JYEcz	Wilhelmine	Touret	Female	10	Poland	Yoshkar-Ola Airport	RU	Russian Federation	EU	Europe	2022-04-16	JOK	Auguste Tindley	Delayed	2026-01-31

3. Data preparation & Cleaning

The dataset required minimal cleaning, as the raw data was already well-structured, with no missing values, duplicates, or obvious inconsistencies. Most preprocessing focused on feature extraction rather than data correction.



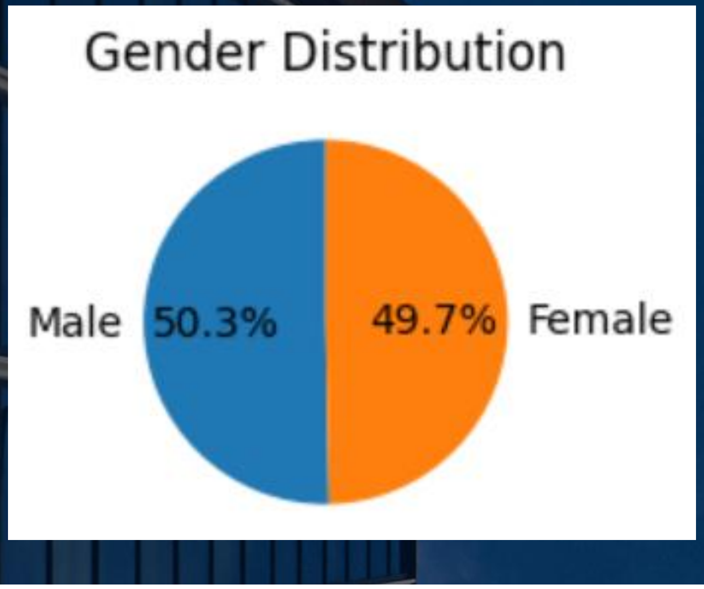
```
# Remove duplicates
df = df.drop_duplicates()

# Convert date
df["Departure_Date"] = pd.to_datetime(df["Departure_Date"])

# Standardize text
df["Flight_Status"] = df["Flight_Status"].str.strip().str.title()

# Remove invalid age
df = df[df["Age"].between(0, 100)]
```

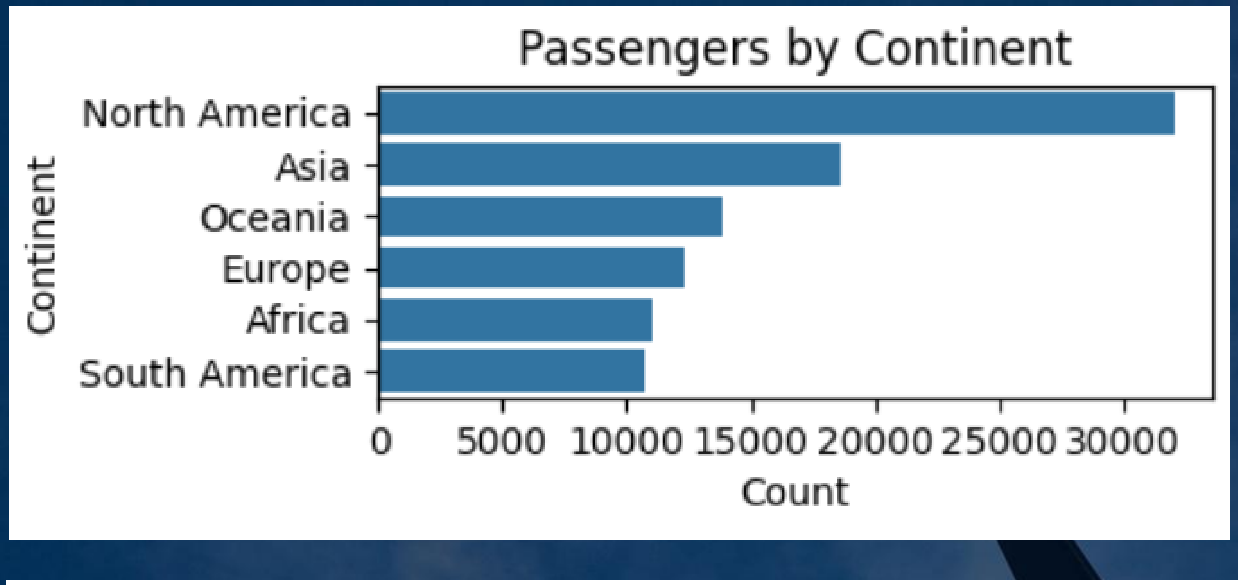


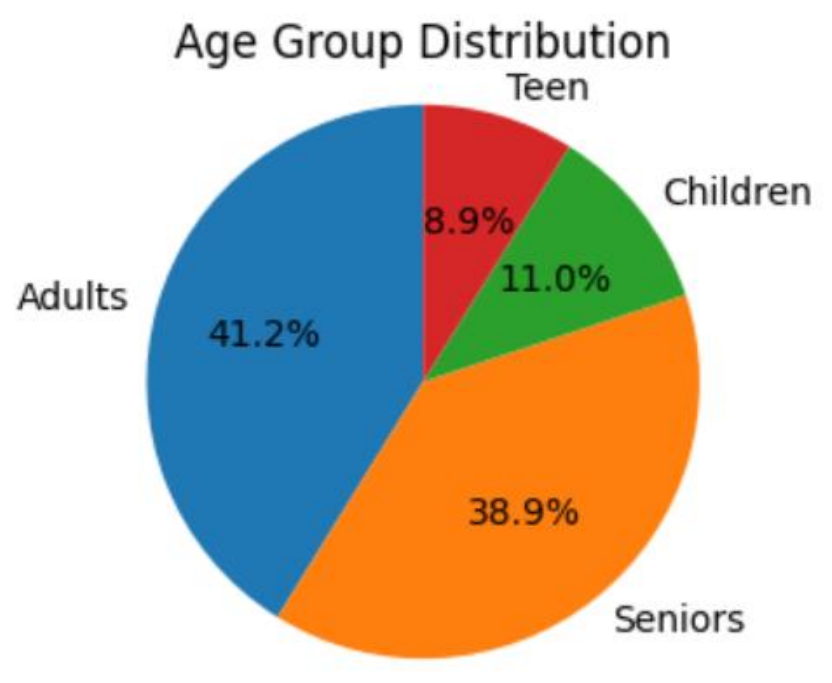
```
df.info()
✓  0.1s

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 98619 entries, 0 to 98618
Data columns (total 16 columns):
 #   Column                Non-Null Count  Dtype 
---  ------                --------------  ----- 
 0   Passenger_ID          98619 non-null  object
 1   First_Name            98619 non-null  object
 2   Last_Name             98619 non-null  object
 3   Gender                98619 non-null  object
 4   Age                   98619 non-null  int64 
 5   Nationality           98619 non-null  object
 6   Airport_Name          98619 non-null  object
 7   Airport_Country_Code  98619 non-null  object
 8   Country_Name          98619 non-null  object
 9   Airport_Continent     98619 non-null  object
 10  Continents            98619 non-null  object
 11  Departure_Date        98619 non-null  object
 12  Arrival_Airport       98619 non-null  object
 13  Pilot_Name            98619 non-null  object
 14  Flight_Status         98619 non-null  object
 15  snapshot_date         98619 non-null  object
dtypes: int64(1), object(15)
memory usage: 12.0+ MB
```

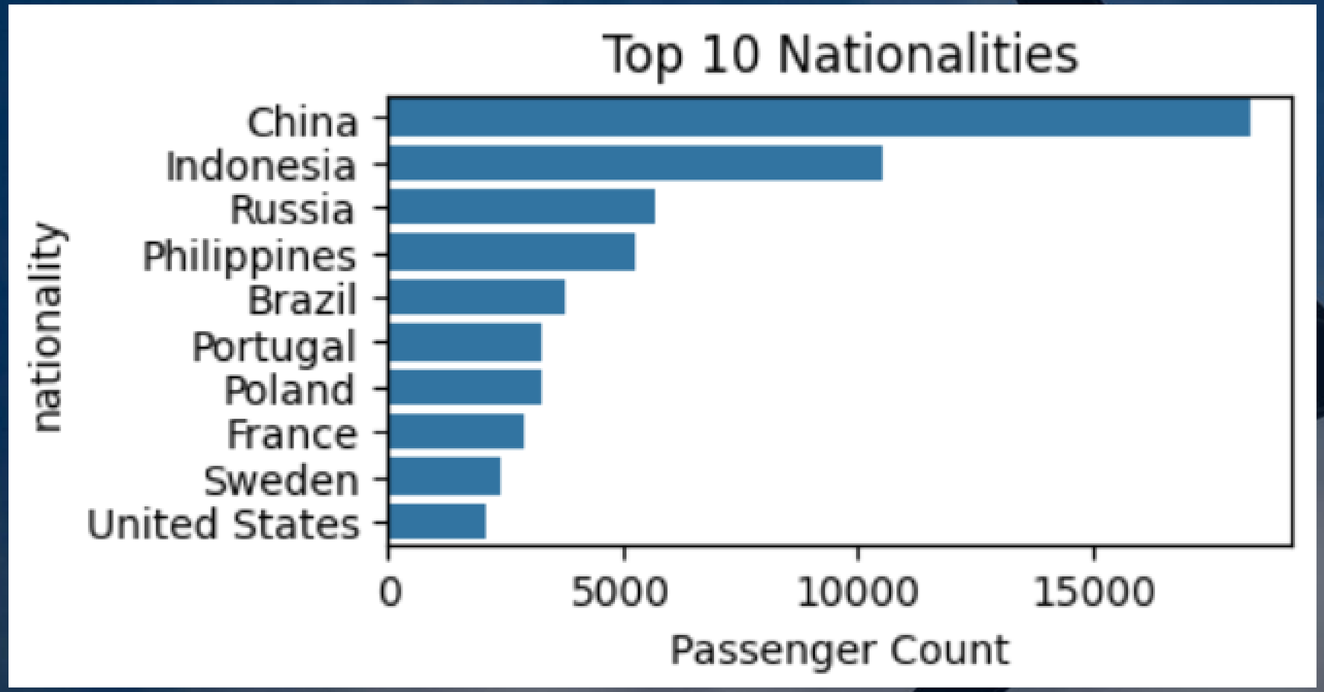
4. Exploration Data Analysis (EDA)

- **Passengers Demographic**







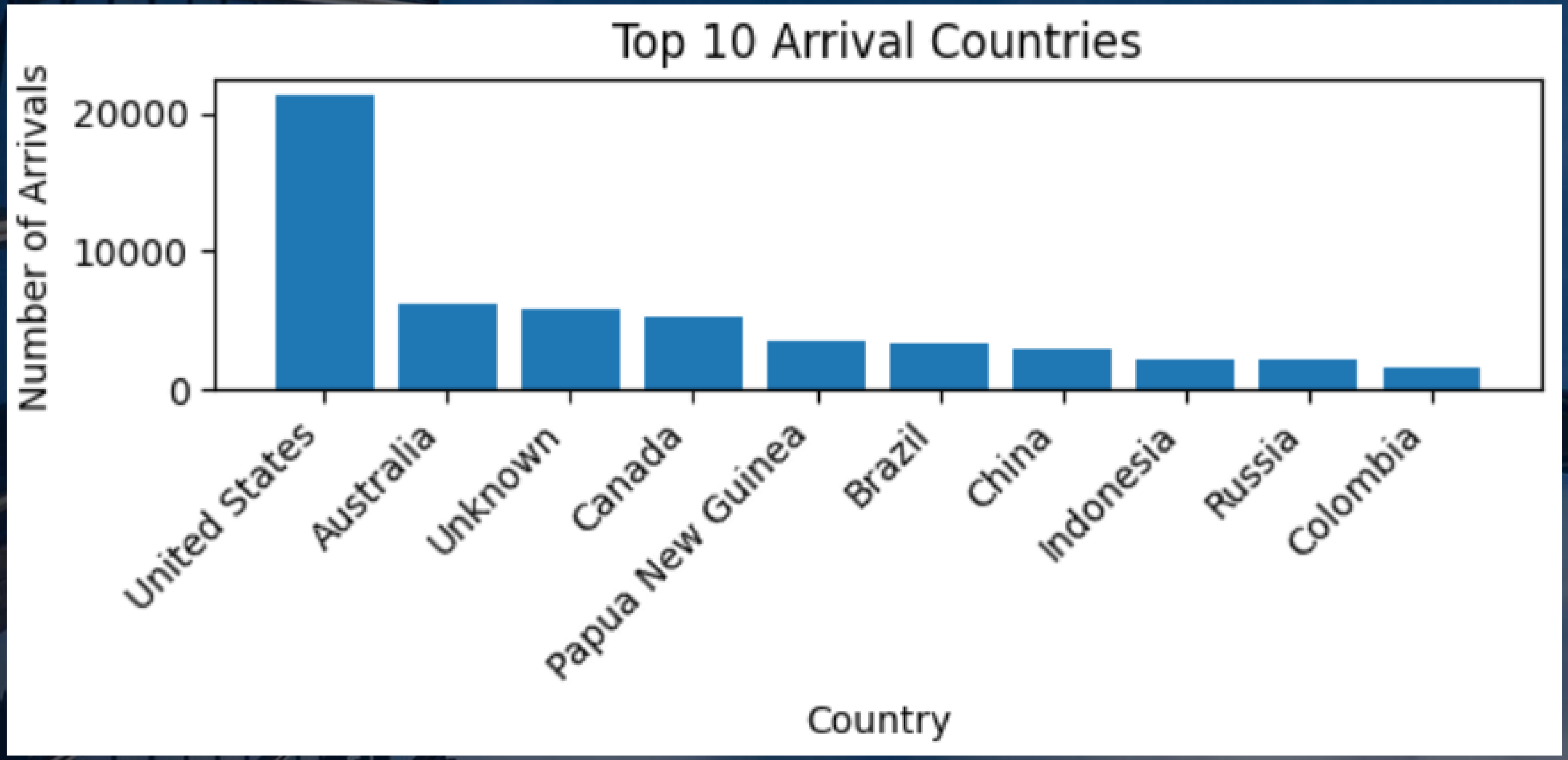


- Passenger gender distribution is nearly **balanced**, indicating no strong gender bias in travel demand.
- **Adults and seniors dominate** the passenger base, while children and teens form a smaller share.
- **North America and Asia** lead in passenger volume, reflecting strong travel demand.
- Passenger traffic is concentrated among a limited number of nationalities, with **China** as the top contributor.
- Overall, the demographic profile points to a **stable, mature customer base focused on key regions and markets**.

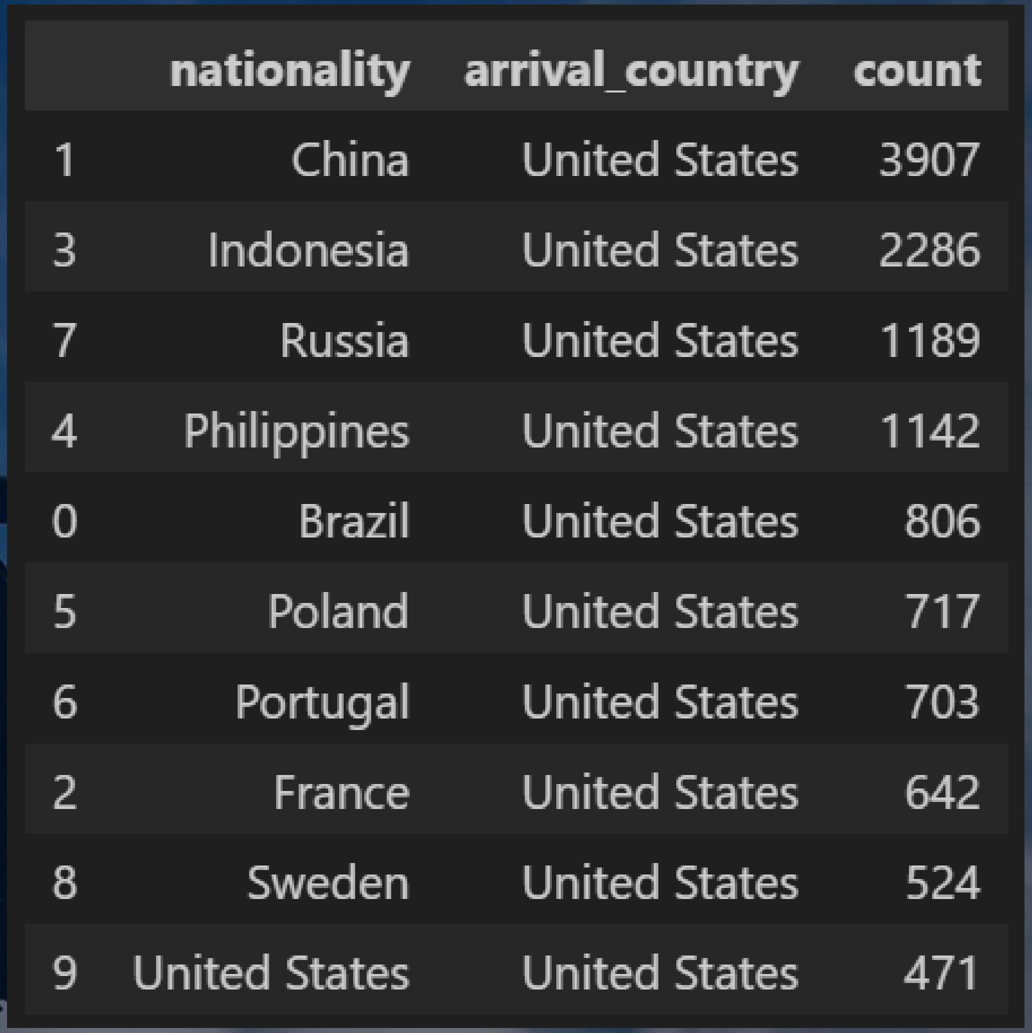
4. Exploration Data Analysis (EDA)

- Geographics

- **The United States** receives the highest number of arriving passengers, with a figure more than **three times higher** than that of the second-ranked country, **Australia**.



Top Nationality to United States

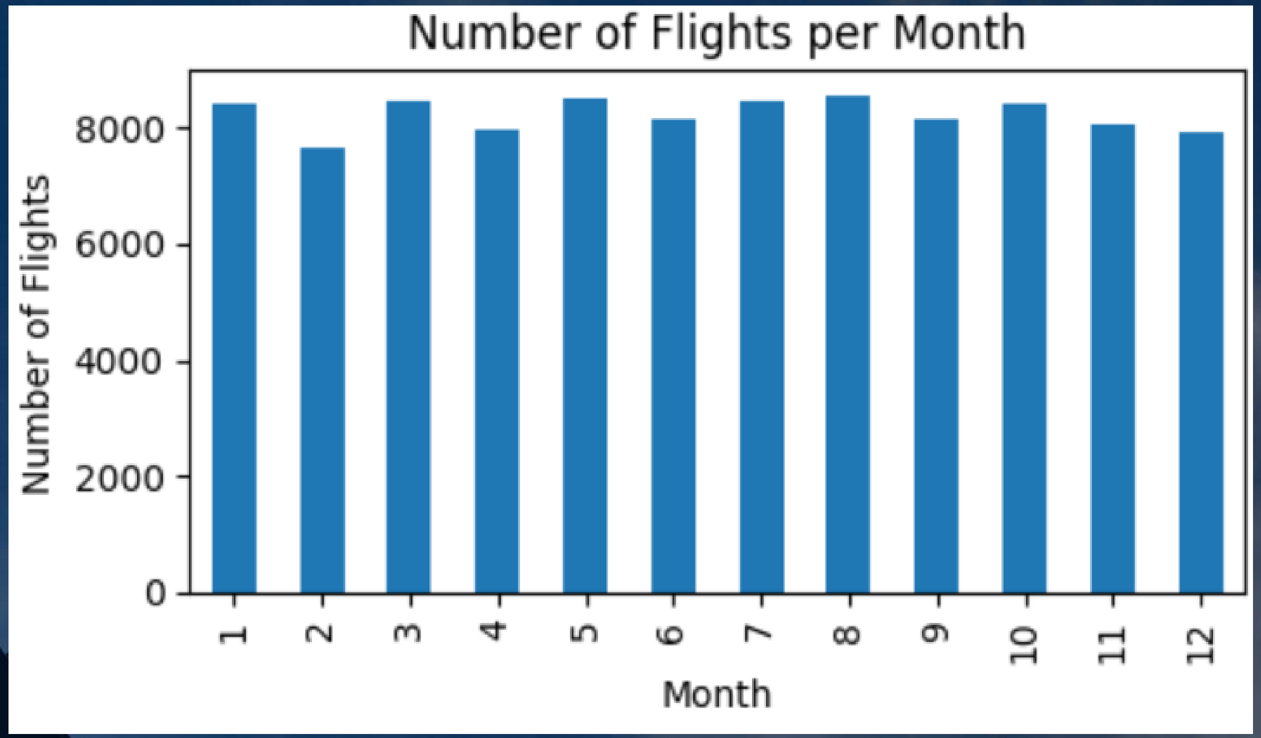


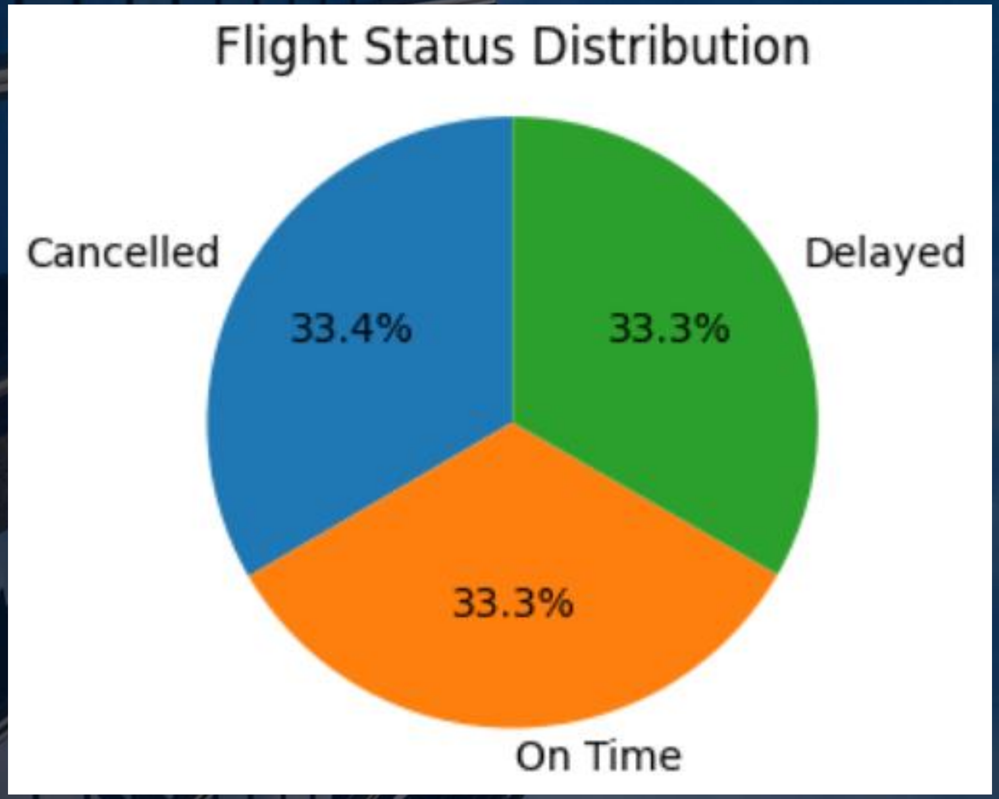
	nationality	arrival_country	count
1	China	United States	3907
3	Indonesia	United States	2286
7	Russia	United States	1189
4	Philippines	United States	1142
0	Brazil	United States	806
5	Poland	United States	717
6	Portugal	United States	703
2	France	United States	642
8	Sweden	United States	524
9	United States	United States	471

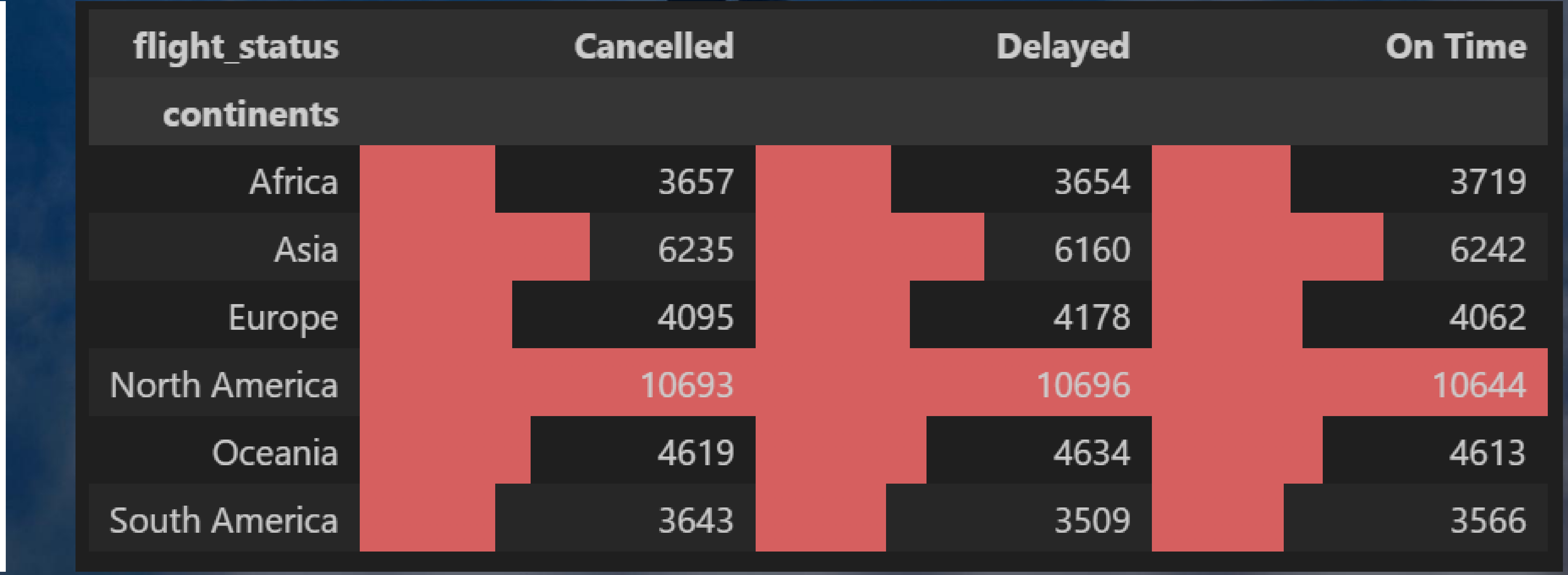
4. Exploration Data Analysis (EDA)

- ## Flights Operations

- **A high proportion of problem flights (68%) is observed.** However, unlike real-world airline data—where on-time flights typically dominate—the balanced status distribution indicates limited realism, reducing the usefulness of this metric for operational decision-making.
- Monthly flight volumes remain largely stable throughout the year, with a slight dip observed in February.







flight_status	Cancelled	Delayed	On Time
continents			
Africa	3657	3654	3719
Asia	6235	6160	6242
Europe	4095	4178	4062
North America	10693	10696	10644
Oceania	4619	4634	4613
South America	3643	3509	3566

5. Passenger Segmentation

Objective

- Segment passengers using demographic and geographic features to identify high-level customer groups.

Data & Method

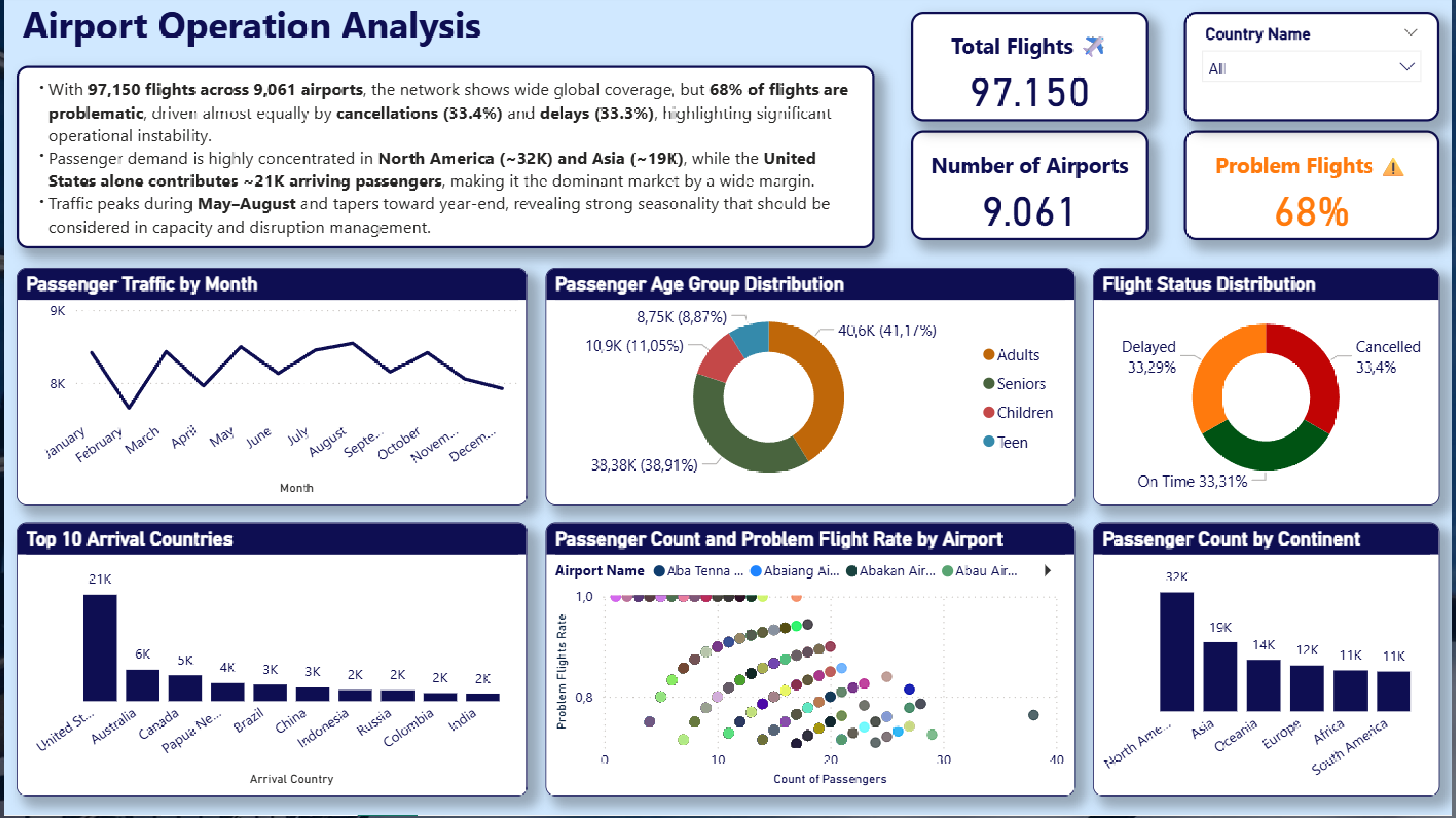
- Features: Age, Gender, Nationality, Departure Country, Arrival Airport
- Applied **K-Means clustering (K = 2)** after encoding and standardization

Output

- Each passenger assigned to a customer segment
- Segment profiles created based on dominant attributes and passenger volume

Metric	Segment 0	Segment 1	Key Insight
Avg Travel Frequency	1	1	No difference
Dominant Age Group	**Seniors**	**Adults**	**Main differentiator**
Dominant Nationality	China	China	Identical
Top Destination	United States	United States	Identical
Passenger Count	47,736	50,883	Balanced distribution

5. Dashboard Overview (Power BI)



Airport Operation Analysis

- With **97,150 flights across 9,061 airports**, the network shows wide global coverage, but **68% of flights are problematic**, driven almost equally by **cancellations (33.4%)** and **delays (33.3%)**, highlighting significant operational instability.
- Passenger demand is highly concentrated in **North America (~32K) and Asia (~19K)**, while the **United States alone contributes ~21K arriving passengers**, making it the dominant market by a wide margin.
- Traffic peaks during **May–August** and tapers toward year-end, revealing strong seasonality that should be considered in capacity and disruption management.

Total Flights
97.150

Number of Airports
9.061

Country Name: All

Problem Flights ⚠
68%

Dashboard Overview

- This dashboard provides a high-level overview of flight operations across time and geography.

Key questions addressed

- What is the distribution of flight statuses?
- How does flight volume change across months?
- How are flight outcomes distributed by continent?

Key visuals

- Flight status distribution
- Monthly flight volume
- Flight status by continent

Note: Due to the balanced nature of the dataset, findings should be interpreted as illustrative rather than representative of real airline performance.

6. Key Insights & Findings

Passenger & Geographic Distribution

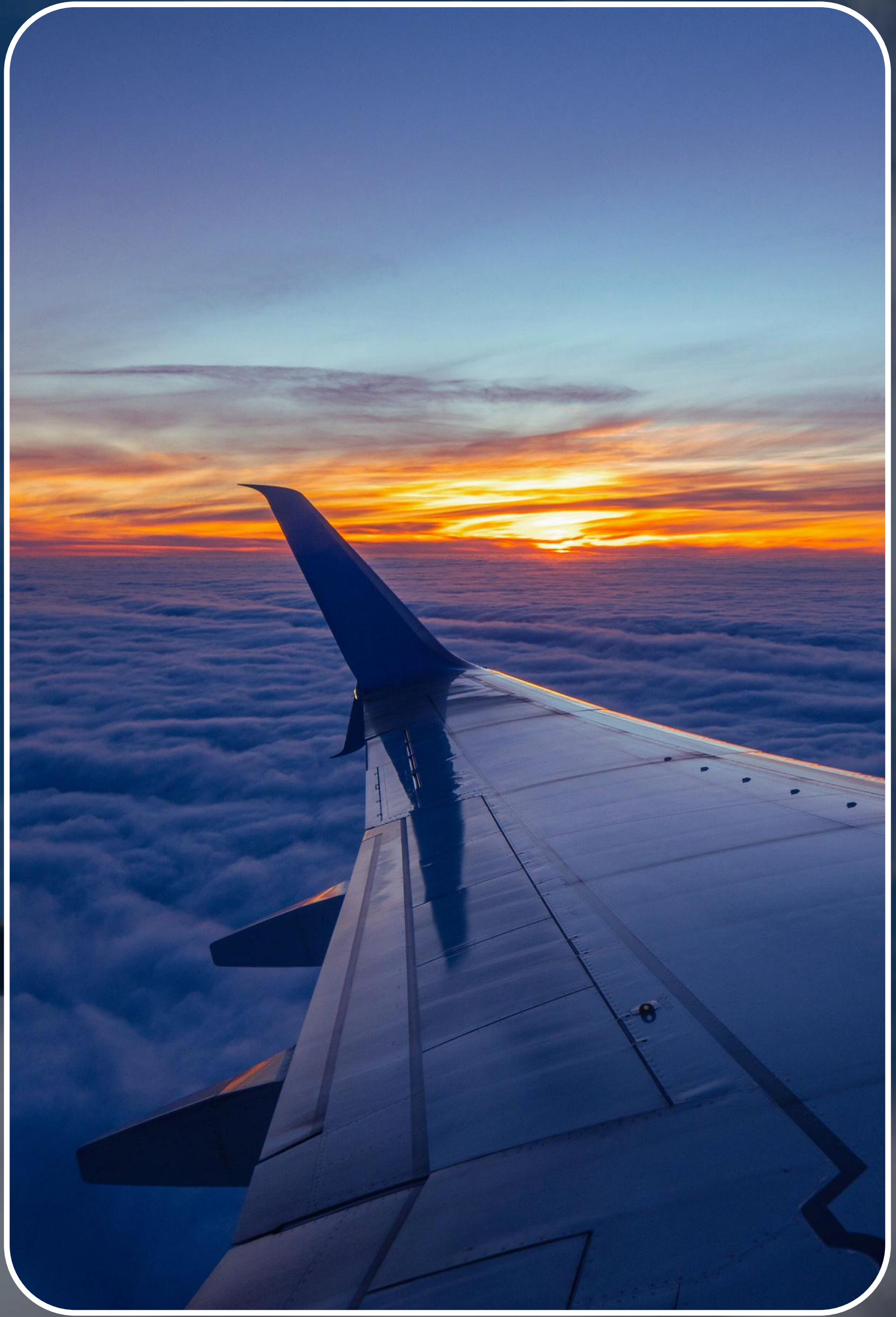
- The network covers **97,150 flights across 9,061 airports**, indicating broad global coverage.
- Passenger demand is **concentrated in North America (~32K) and Asia (~19K)**, with the United States emerging as the dominant arrival country.
- Other continents contribute more evenly distributed but smaller passenger volumes, suggesting limited geographic skew beyond major hubs.

Demographic Composition

- **Adults and seniors account for the majority of passengers**, together representing nearly 80% of total traffic.
- Children and teens form a smaller share, indicating a predominantly adult travel population within the dataset.

Flight Operations & Temporal Patterns

- Approximately **68% of flights are classified as problematic**, driven almost equally by delays (~33%) and cancellations (~33%).
- Passenger traffic shows **moderate mid-year increases** and a softer decline toward year-end, suggesting some seasonality but no extreme peaks.
- Problem flight rates remain **consistently high across airports**, regardless of passenger volume.



7. Recommendation

Improve Operational Reliability

- Prioritize initiatives to **reduce delays and cancellations**, focusing first on high-traffic regions such as **North America and Asia**, where passenger impact is greatest.
- Introduce tighter **schedule buffers and contingency planning** to mitigate disruption during peak travel periods.

Capacity & Seasonal Planning

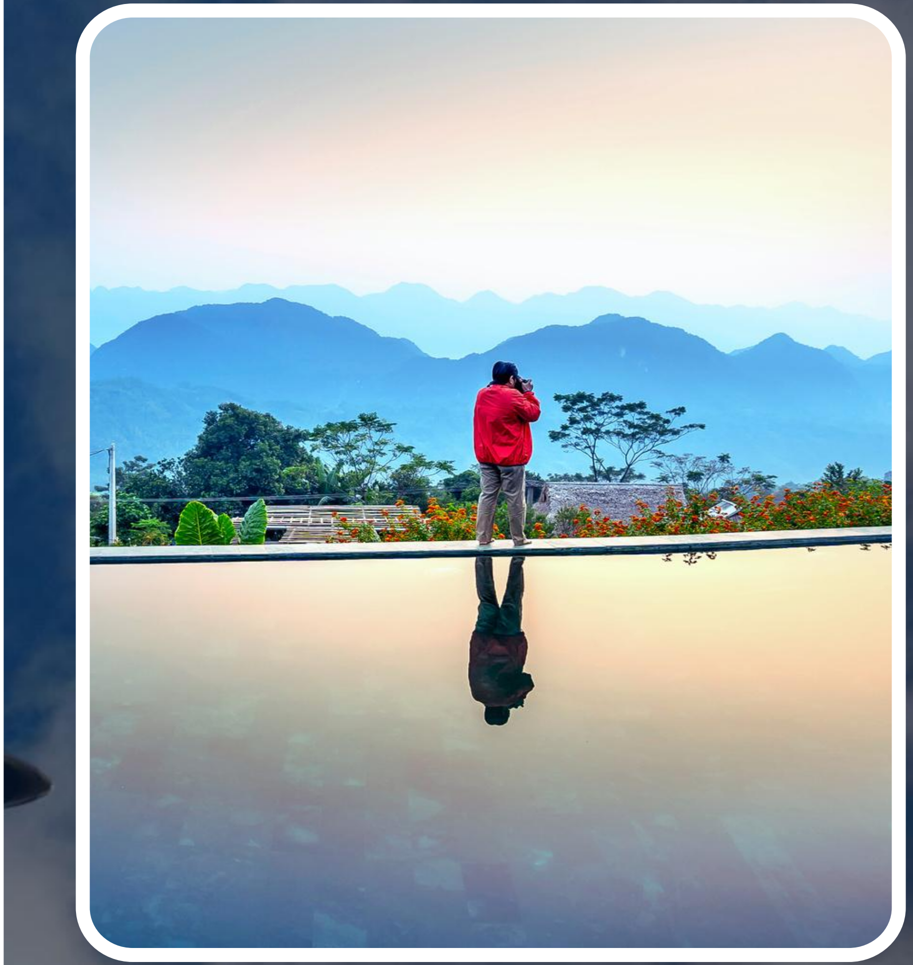
- Allocate additional resources during **mid-year peak months (May–August)** to manage increased passenger demand.
- Adjust staffing and ground operations toward year-end to maintain service levels as demand tapers.

Airport & Network Optimization

- Monitor airports with **high problem-flight rates**, even at moderate passenger volumes, to identify operational bottlenecks.
- Strengthen coordination with **major hub airports**, particularly in the United States, to reduce network-wide disruption.

Passenger-Centric Strategies

- Enhance communication and rebooking support for **adult and senior travelers**, who represent the majority of passengers.
- Offer targeted compensation or service recovery programs to retain customer trust during service disruptions.



THANK YOU!