FINAL PROJECT

LU QUOC CO

Introduction

AIRLINE SITUATION 2022

Github

Table of content

This project implements an end-to-end **data analytics and machine learning pipeline** using Python and SQL Server. The workflow starts from data extraction, continues through data preprocessing and feature enrichment, and advances to exploratory analysis and modeling. The final, cleaned dataset is stored back into the database to support efficient visualization and reporting.

1 Project overview

- **Project Objective**
- Build an end-to-end data analytics pipeline
- Transform raw data into actionable insights
- Support customer segmentation and business decisions

Key Focus

- Data preprocessing
- Exploratory Data Analysis (EDA)
- Machine Learning & RFM analysis

2 Data Source & Tools

Data Source

- SQL Server database
- Transactional & customer data

Tools & Technologies

- Python (Pandas, NumPy, Scikit-learn)
- SQL Server
- Power BI (Visualization)
- Jupyter Notebook

3 Data Preprocessing

Data Source

- SQL Server database
- Transactional & customer data

Tools & Technologies

- Python (Pandas, NumPy, Scikit-learn)
- SQL Server
- Power BI (Visualization)
- Jupyter Notebook

4 EDA & Feature Engineering

Data Preparation Steps

- Data overview & data quality check
- Missing values analysis
- Outlier detection and treatment
- Duplicate removal

Purpose

- Ensure accuracy and consistency
- Improve reliability for analysis and modeling

5 Machine Learning & RFM

Exploratory Data Analysis (EDA)

- Analyze data distributions
- Identify trends and patterns
- Explore relationships between variables

Feature Engineering

- Create meaningful variables
- Better represent customer behavior
- Enhance analytical and ML performance

6 Output & Business Value

Final Output

- Clean & enriched dataset
- Stored back into SQL Server
- Ready for Power BI dashboard

Business Value

- Clear customer segments
- Actionable insights from data
- Support data-driven decision-making

7 Recommendation

- Prioritize high-value customers using RFM-based segmentation.
- Apply targeted strategies to retain loyal customers and re-engage inactive ones.
- Use dashboards to continuously monitor customer behavior and business performance.

🌍 Global Airline Industry in 2022 – Overview

In 2022, the global airline industry continued its recovery from the COVID-19 pandemic, although it had not fully returned to pre-pandemic levels. According to industry data, **total air passenger traffic increased significantly compared to 2021**, reaching approximately **68.5% of 2019 levels** in terms of revenue passenger kilometers (RPKs). Domestic travel recovered more strongly, reaching around **79.6%** of pre-pandemic levels, while international travel grew even faster year-on-year but still remained below 2019 levels. Overall airline passenger volumes and revenue both showed strong rebound momentum as travel restrictions were eased worldwide and demand for air travel surged after two years of sharp decline.[1]

In addition, industry estimates suggest that about **4.53 billion passengers** were carried globally in 2022, and the sector generated **around $803 billion in revenue**, indicating substantial market rebound and growth compared to pandemic years.

About Dataset

Airline Dataset

File name: AirlineDataset.csv

- Description: The dataset contains 98,619 rows and is used to build analytical and machine learning models to predict whether flights at airports will be delayed. Additionally, it can be used to analyze consumer behavior, thereby identifying factors influencing tourism services that attract visitors to destinations..

About Dataset

Customer_info

Passenger ID
First Name
Last Name
Gender
Age
Nationality

Object: Identify demographic information

Service_info

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

Object: Understanding service and consumer behavior

About Dataset

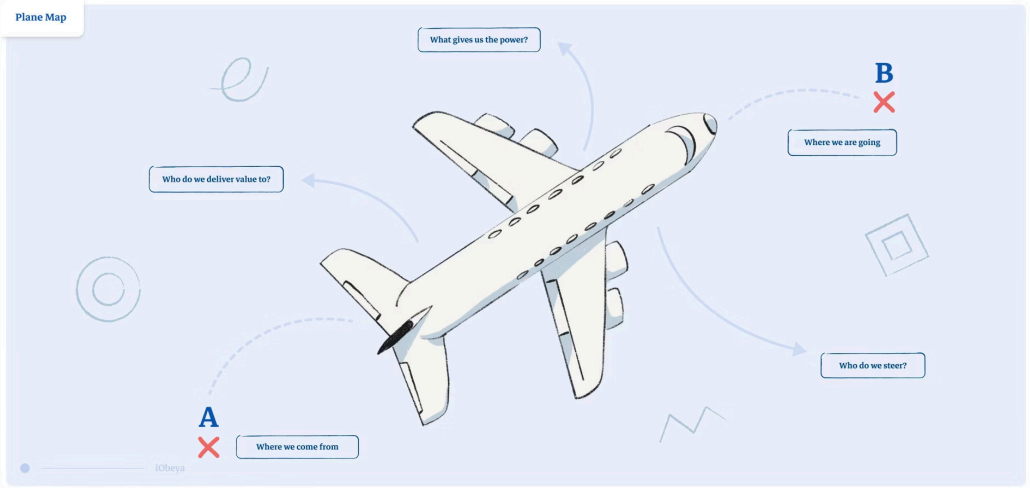
```
df
```

	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
0	ABVWIg	Edithe	Leggis	Female	62	Japan	Coldfoot Airport	US	United States	NAM	North America	6/28/2022	CXF	Fransisco Hazeldine	On Time
1	jkXXAX	Elwood	Catt	Male	62	Nicaragua	Kugluktuk Airport	CA	Canada	NAM	North America	12/26/2022	YCO	Marla Parsonage	On Time
2	CdUz2g	Darby	Felgate	Male	67	Russia	Grenoble-Isère Airport	FR	France	EU	Europe	1/18/2022	GNB	Rhonda Amber	On Time
3	BRS38V	Dominica	Pyle	Female	71	China	Ottawa / Gatineau Airport	CA	Canada	NAM	North America	9/16/2022	YND	Kacie Commucci	Delayed
4	9kvTLo	Bay	Pencost	Male	21	China	Gillespie Field	US	United States	NAM	North America	2/25/2022	SEE	Ebonee Tree	On Time
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
98614	hnGQ62	Gareth	Mugford	Male	85	China	Hasvik Airport	NO	Norway	EU	Europe	12-11-2022	HAA	Pammie Kingscote	Cancelled
98615	2omEzh	Kasey	Benedict	Female	19	Russia	Ampampamena Airport	MG	Madagascar	AF	Africa	10/30/2022	IVA	Dorice Lochran	Cancelled
98616	VUPiVG	Darrin	Lucken	Male	65	Indonesia	Albacete-Los Llanos Airport	ES	Spain	EU	Europe	09-10-2022	ABC	Gearalt Main	On Time
98617	E47NtS	Gayle	Lievesley	Female	34	China	Gagnoa Airport	CI	Côte d'Ivoire	AF	Africa	10/26/2022	GGN	Judon Chasle	Cancelled
98618	8JYEcz	Wilhelmine	Touret	Female	10	Poland	Yoshkar-Ola	RU	Russian	EU	Europe	4/16/2022	JOK	Auguste	Delayed

Project Overview

🌍 Context

The airline industry generates massive amounts of operational and customer data from flight records, bookings, customer behavior, and transactions. With increasing competition and rising customer expectations, airlines need to use data analytics and machine learning to gain deeper insights into customer behavior and improve business performance.



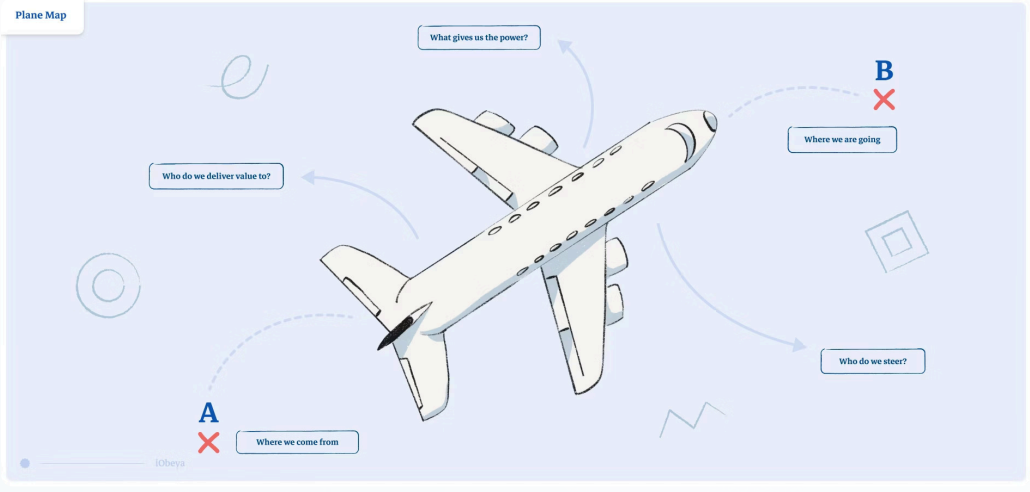
❗ Problem

Current datasets often contain inconsistencies, outliers, and duplicates that make it difficult to generate reliable insights. Additionally, many airlines struggle to segment customers effectively in order to offer personalized services, optimize loyalty programs, and improve retention.

Project Overview

⚠️ Challenges

- Large volumes of transactional and customer data with noise and irregularities
- Complex patterns in customer behavior that traditional reporting methods can’t capture
- Need for a scalable analytical workflow that integrates data cleaning, analysis, and modeling
- Translating analytical output into business decisions and visual insights



🎯 Project Objectives

- Build an end-to-end data analytics pipeline from raw data to actionable insights
- Clean and preprocess data for reliability and accuracy
- Perform Exploratory Data Analysis (EDA) to discover key patterns and trends
- Use machine learning and RFM (Recency, Frequency, Monetary) scoring to segment customers
- Prepare and store final data for visualization and business reporting

🛠️ Technologies Used

- **Python** – data manipulation, cleaning, feature engineering, modeling
- **SQL Server** – data extraction and storage
- **Power BI** – visualization and dashboarding
- **Pandas / NumPy / Scikit-Learn** – analytical libraries
- **Jupyter Notebook** – development environment

Data Preprocessing

```
df.info()
```
✓ 0.0s

```
<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  Today                 98619 non-null  object
dtypes: int64(1), object(15)
memory usage: 12.0+ MB
```

```
[6]  df.describe()
```
✓ 0.0s

	Age
count	98619.000000
mean	45.504021
std	25.929849
min	1.000000
25%	23.000000
50%	46.000000
75%	68.000000
max	90.000000

```
df.duplicated().sum()
```
✓ 0.2s

```
np.int64(0)
```

Handle Data and remove outliers

```
import pandas as pd
import numpy as np

def detect_outliers_iqr(df):
    outlier_summary = {}

    numeric_cols = df.select_dtypes(include=[np.number]).columns

    for col in numeric_cols:
        Q1 = df[col].quantile(0.25)
        Q3 = df[col].quantile(0.75)
        IQR = Q3 - Q1

        lower_bound = Q1 - 1.5 * IQR
        upper_bound = Q3 + 1.5 * IQR

        outliers = df[(df[col] < lower_bound) | (df[col] > upper_bound)]

        outlier_summary[col] = {
            "lower_bound": lower_bound,
            "upper_bound": upper_bound,
            "outlier_count": outliers.shape[0],
            "outlier_index": outliers.index.tolist()
        }

    return outlier_summary
outliers_iqr = detect_outliers_iqr(df)

for col, info in outliers_iqr.items():
    print(f"{col}: {info['outlier_count']} outliers")
```
✓ 0.0s

```
Age: 0 outliers
```

Data Preprocessing

From the Arrival_Airport data, map it with open-source data to retrieve Arrival_Country.

```
      Passenger_ID First_Name  Last_Name  Gender  Age    Nationality  \
0           zQtk5b       Arda     Molson  Female   24  United States
1           3rDKTc    Paquito    Walicki    Male   50          China
2           0z876q     Cobbie  Treadgall    Male   58         Poland
3           wk9ZwM   Maryjane   Trosdall  Female   40         Poland
4           6coA01      Raine  Skittreal  Female   32        Nigeria
...            ...        ...        ...     ...  ...            ...
98614       d5o2xU        Ric     Laroux    Male   61      Lithuania
98615       iz6cxd    Wilfrid     Nibley    Male   56         Russia
98616       x4693Z       Josi     Sanger  Female   45        Ireland
98617       6NhZJF     Karita   Petrusch  Female   28         Russia
98618       nESe6U      Goldy   Renfield  Female   14          China

                                Airport_Name Airport_Country_Code  \
0                  Maria Reiche Neuman Airport                  PE
1                          Mbuji Mayi Airport                  CD
2                          Kalkgurung Airport                  AU
3                              Binaka Airport                  ID
4           Rajiv Gandhi International Airport                  IN
...                                        ...                 ...
98614              Immokalee Regional Airport                  US
98615            Erzurum International Airport                 TR
98616  Bahías de Huatulco International Airport                  MX
98617                 Marana Regional Airport                  US
98618                        Flamingo Airport                  CR
...
98617  2026-01-24                    United States
98618  2026-01-24                       Costa Rica

[98619 rows x 17 columns]
```

Data Preprocessing

EDA

```
cat_cols = df.select_dtypes(include=["object", "category"]).columns
cat_cols = [c for c in cat_cols if c != "Passenger_ID"]

output_path = "categorical_summary_all.txt"

with open(output_path, "w", encoding="utf-8") as f:
    f.write("CATEGORICAL VARIABLE SUMMARY (Grouped by Passenger_ID)\n")
    f.write("=" * 60 + "\n")

    for col in cat_cols:
        distinct_cnt = df[col].nunique(dropna=True)

        if distinct_cnt < 20:
            f.write(f"\n=== {col} (distinct={distinct_cnt}) ===\n")
            counts = (
                df.groupby(col)["Passenger_ID"]
                  .nunique()
                  .sort_values(ascending=False)
            )

            for category, cnt in counts.items():
                f.write(f"{category}: {cnt}\n")
        else:
            f.write(f"\n--- SKIPPED {col} (distinct={distinct_cnt}) ---\n")

#Insights Cateogryrical Variables But it is hard explain here
```

Leveraging OpenAI to uncover 5 key insights, anomalies, and high-value data.

```
from openai import OpenAI

with open("categorical_summary_all.txt", "r", encoding="utf-8") as f:
    txt_content = f.read()

client = OpenAI(
    api_key=""
)

response = client.responses.create(
    model="gpt-4.1-mini",
    input=[
        {
            "role": "user",
            "content": f"""
Analyze the following categorical respondent summary.
Give 5 key insights, anomalies, and high-level observations.

TEXT:
{txt_content}
"""}
    ]
)

print(response.output_text)
```

Data Preprocessing

EDA

Movement trends

```python
# 1. #--Người dân nước nào có xu hướng di chuyển bằng máy bay nhiều nhất? #xuhuongdichuyen**
# Relevant Columns: "Nationality", "Passenger_ID"

from collections import Counter

language_counter = Counter()

for langs in df["Nationality"].dropna():
    for lang in langs.split(";"):
        language_counter[lang.strip()] += 1

# Convert to DataFrame if needed
lang_count_df = (
    pd.DataFrame(language_counter.items(), columns=["Nationality", "Appear_Count"])
    .sort_values("Appear_Count", ascending=False)
)
lang_count_df

# We observe that China, Indonesia, Russia are top 3 Foreigner go by plane.
```

✓ 0.1s

	Nationality	Appear_Count
1	China	18317
6	Indonesia	10559
8	Russia	5693
7	Philippines	5239
20	Brazil	3791
...	...	...
228	Cook Islands	2
238	Saint Helena	2
145	Jersey	1
237	Sint Maarten	1
239	Norfolk Island	1

240 rows × 2 columns

Data Preprocessing

EDA

Airport Status

```
# 2. #--Airport nào có tình trạng cancelled và delay nhiều nhất? #tinhtrangairport**
# Relevant Columns: "Airport_Name", "Flight_Status"
import pandas as pd
from collections import Counter

# ==============================
# 1. Chuẩn hóa & explode dữ liệu
# ==============================

df_clean = df[["Airport_Name", "Flight_Status"]].dropna().copy()

# Tách nhiều status nếu có dấu ;
df_clean["Flight_Status"] = (
    df_clean["Flight_Status"]
    .astype(str)
    .str.split(";")
)

df_exploded = df_clean.explode("Flight_Status")

df_exploded["Flight_Status"] = df_exploded["Flight_Status"].str.strip()

# ==============================
# 2. Đếm số lần xuất hiện (tổng quát)
# ==============================

status_count_df = (
    df_exploded
    .groupby(["Airport_Name", "Flight_Status"])
    .size()
    .reset_index(name="Appear_Count")
)
```

```
# ==============================
# 3. Top 3 airport CANCELLED nhiều nhất
# ==============================

top_cancelled_airports = (
    status_count_df[
        status_count_df["Flight_Status"].str.lower() == "cancelled"
    ]
    .sort_values("Appear_Count", ascending=False)
    .head(3)
)

# ==============================
# 4. Top 3 airport DELAY nhiều nhất
# ==============================

top_delayed_airports = (
    status_count_df[
        status_count_df["Flight_Status"].str.lower().isin(["delayed", "delay"])
    ]
    .sort_values("Appear_Count", ascending=False)
    .head(3)
)

# ==============================
# 5. Kết quả
# ==============================

print("Top 3 Airport Cancelled nhiều nhất:")
display(top_cancelled_airports)

print("Top 3 Airport Delayed nhiều nhất:")
display(top_delayed_airports)
```

✓ 0.2s

Data Preprocessing

EDA

The country is visited by many people.

```python
# 3. #--#--Nước nào được đi đến nhiều nhiều nhất? #nuocduocdennhieu**
# Relevant Columns: "Arrival_Country", "Passenger_ID"

from collections import Counter

language_counter = Counter()

for langs in df["Arrival_Country"].dropna():
    for lang in langs.split(";"):
        language_counter[lang.strip()] += 1

# Convert to DataFrame if needed
lang_count_df2 = (
    pd.DataFrame(language_counter.items(), columns=["Arrival_Country", "Appear_Count"])
    .sort_values("Appear_Count", ascending=False)
)
lang_count_df2.head(10)

# We observe that United States, Australia and Canada are top 3 Nations that Foreigners usually arrive.
```

✓ 0.0s

	Arrival_Country	Appear_Count
5	United States	21287
2	Australia	6265
28	Unknown	5804
12	Canada	5172
18	Papua New Guinea	3614
10	Brazil	3330
6	China	2865
3	Indonesia	2225
33	Russia	2175
13	Colombia	1607

Machine Learning & RFM

Machine Learning

```
##  Dự đoán Flight có bị Delay / Cancelled hay không

## Flight_Status -> chuyển thành:

## 0 = On Time
## 1 = Delayed / Cancelled
##============================
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import OneHotEncoder, StandardScaler
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, confusion_matrix

# ========================
# 1. Chọn feature & target
# ========================

df_ml = df.copy()

# Binary target: Delay / Cancelled = 1, On Time = 0
df_ml["Is_Delayed"] = df_ml["Flight_Status"].str.lower().isin(
    ["delayed", "cancelled"]
).astype(int)

X = df_ml[
    ["Age", "Gender", "Nationality", "Airport_Continent", "Arrival_Country"]
]
y = df_ml["Is_Delayed"]
```

```
# ========================
# 2. Tách numerical / categorical
# ========================

numeric_features = ["Age"]
categorical_features = [
    "Gender",
    "Nationality",
    "Airport_Continent",
    "Arrival_Country"
]

# ========================
# 3. Preprocessing pipeline
# ========================

preprocessor = ColumnTransformer(
    transformers=[
        ("num", StandardScaler(), numeric_features),
        ("cat", OneHotEncoder(handle_unknown="ignore"), categorical_features)
    ]
)

# ========================
# 4. Model
# ========================

model = LogisticRegression(max_iter=1000)

# ========================
# 5. Full pipeline
# ========================

pipeline = Pipeline(steps=[
    ("preprocess", preprocessor),
    ("model", model)
])
```

```
# 4. Model
# ========================

model = LogisticRegression(max_iter=1000)

# ========================
# 5. Full pipeline
# ========================

pipeline = Pipeline(steps=[
    ("preprocess", preprocessor),
    ("model", model)
])

# ========================
# 6. Train / Test split
# ========================

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.3, random_state=42, stratify=y
)

# ========================
# 7. Train model
# ========================

pipeline.fit(X_train, y_train)

# ========================
# 8. Predict & Evaluate
# ========================

y_pred = pipeline.predict(X_test)

print("Confusion Matrix:")
print(confusion_matrix(y_test, y_pred))

print("\nClassification Report:")
print(classification_report(y_test, y_pred))
```

Result

```
Confusion Matrix:
[[   14  9840]
 [   32 19700]]

Classification Report:
              precision    recall  f1-score   support

           0       0.30      0.00      0.00      9854
           1       0.67      1.00      0.80     19732

    accuracy                           0.67     29586
   macro avg       0.49      0.50      0.40     29586
weighted avg       0.55      0.67      0.53     29586
```

Machine Learning & RFM

RFM Score

```python
import pandas as pd

# ========================
# 1. Chuẩn hóa date
# ========================

df_rfm = df.copy()

df_rfm["Departure_Date"] = pd.to_datetime(df_rfm["Departure_Date"])
df_rfm["Today"] = pd.to_datetime(df_rfm["Today"])

# ========================
# 2. Tính R & F
# ========================

rfm = (
    df_rfm
    .groupby("Passenger_ID")
    .agg(
        Recency=("Departure_Date", lambda x: (df_rfm["Today"].max() - x.max()).days),
        Frequency=("Passenger_ID", "count")
    )
    .reset_index()
)

# ========================
# 3. Monetary (proxy)
# ========================

rfm["Monetary"] = rfm["Frequency"]
```

```python
# ========================
# 4. Chia điểm RFM (1–5)
# ========================

rfm["R_Score"] = pd.qcut(rfm["Recency"], 5, labels=[5,4,3,2,1])
rfm["F_Score"] = pd.qcut(rfm["Frequency"].rank(method="first"), 5, labels=[1,2,3,4,5])
rfm["M_Score"] = pd.qcut(rfm["Monetary"].rank(method="first"), 5, labels=[1,2,3,4,5])

# ========================
# 5. RFM Segment
# ========================

rfm["RFM_Score"] = (
    rfm["R_Score"].astype(str) +
    rfm["F_Score"].astype(str) +
    rfm["M_Score"].astype(str)
)

rfm
```

Result

	Passenger_ID	Recency	Frequency	Monetary	R_Score	F_Score	M_Score	RFM_Score
0	002Gd9	1338	1	1	3	1	1	311
1	00638T	1173	1	1	5	1	1	511
2	006RvT	1215	1	1	4	1	1	411
3	008AJe	1318	1	1	3	1	1	311
4	00Ahsk	1355	1	1	2	1	1	211
...	...	...	...	...	...	...	...	...
98614	zzmvXb	1402	1	1	2	5	5	255
98615	zzsldD	1461	1	1	1	5	5	155
98616	zzt3WK	1412	1	1	2	5	5	255
98617	zzuv5J	1233	1	1	4	5	5	455
98618	zzy9KO	1330	1	1	3	5	5	355

98619 rows × 8 columns

Output & Business Value

Insight

A total of **98,617 passengers** were recorded across **9,061 airports**, reflecting a broad aviation network and sustained travel demand.
Passenger volume exhibits clear **seasonal patterns**, peaking in **August** and reaching its lowest level in **February**.
The relatively balanced distribution of **on-time, delayed, and cancelled flights** indicates operational inefficiencies and potential areas for performance improvement.
The United States is the leading arrival destination, while passenger origins are predominantly from **China and Indonesia**, highlighting key international travel markets.

Areas for Improvement (Tourism Focus)

Enhance **on-time performance**, particularly during peak travel periods, to improve overall passenger experience and destination attractiveness.
Strengthen capacity planning and service quality for **priority tourism markets**, with a focus on routes connected to the United States, China, and Indonesia.

Count of Airport

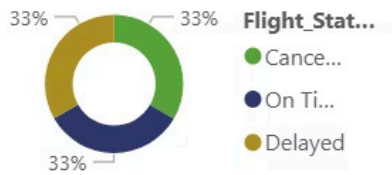
9061

Count of Passenger

98.62K

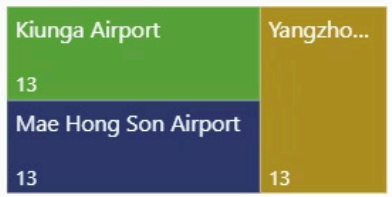
Passenger by Month

Percentage of Flight Status

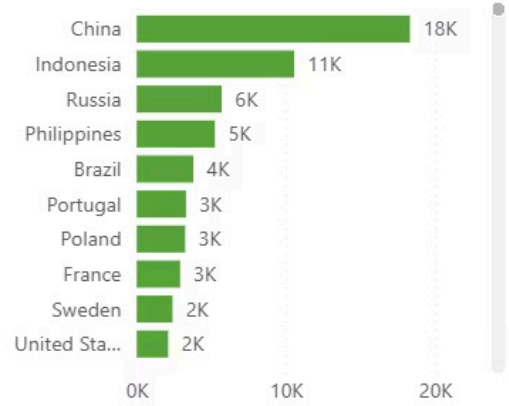


Top 3 Delay Airport

Top 3 Cancel Airport



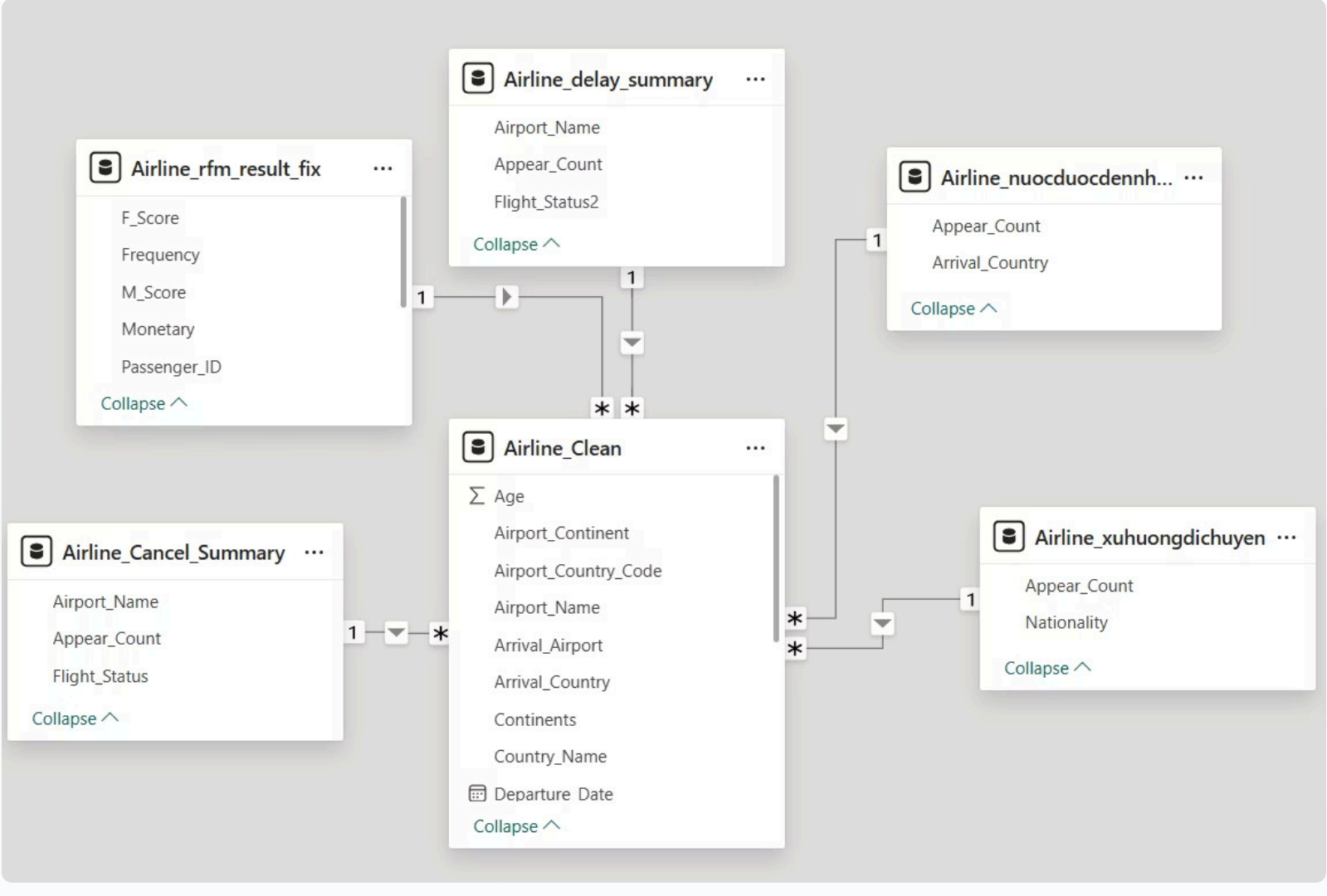
Nationality of Passenger



Arrival Country by Passenger

Arrival Country	Count of Passenger
United States	21287
Australia	6265
Unknown	5814
Canada	5172
Papua New Guinea	3614
Brazil	3330
China	2865
Indonesia	2225
Russia	2175
Colombia	1607
India	1519

Output & Business Value



Recommendation

1. Customer-Focused Segmentation Strategy

Airlines should actively leverage RFM-based customer segmentation to differentiate service strategies. High-value and loyal customers should be prioritized through personalized loyalty programs, exclusive promotions, and premium services, while low-engagement customers can be targeted with reactivation campaigns and price-sensitive offers.

2. Data-Driven Marketing Optimization

Marketing efforts should shift from mass campaigns to data-driven, targeted initiatives. By understanding customer recency, frequency, and spending behavior, airlines can optimize promotional timing, reduce marketing costs, and improve conversion rates through personalized communication.

3. Enhance Customer Retention and Loyalty

Retention strategies should focus on improving engagement among mid-tier customers, who represent strong growth potential. Incentives such as mileage bonuses, tailored offers, and improved customer experience can help convert these customers into loyal, high-value segments.

Recommendation

4. Operational and Revenue Optimization

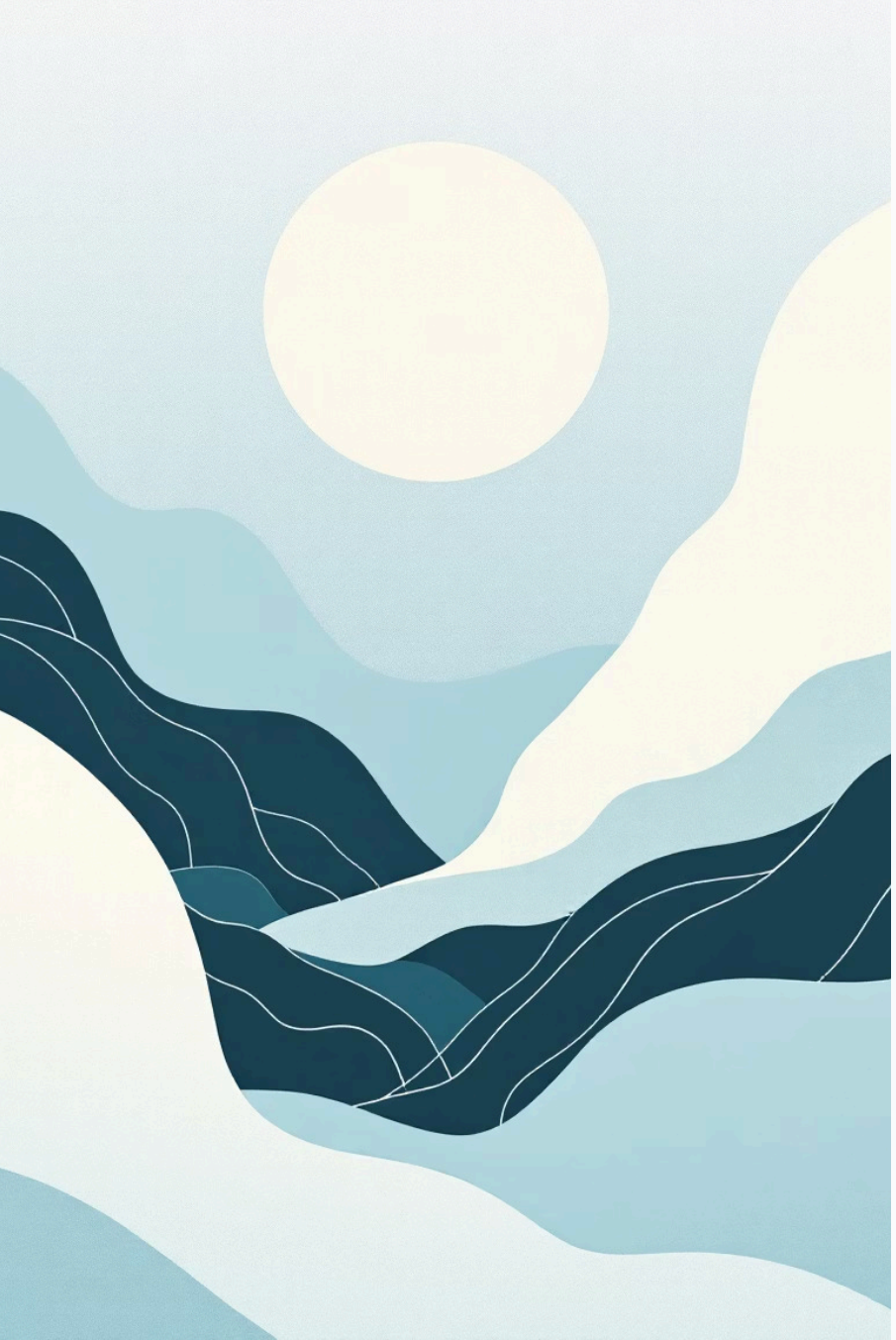
Insights derived from customer behavior analysis can be used to support revenue management and operational planning. Understanding booking patterns and spending behavior enables airlines to better align pricing strategies, route planning, and resource allocation with customer demand.

5. Scalable Analytics and Visualization Framework

The analytical pipeline built in this project should be maintained and expanded over time. Integrating updated data into Power BI dashboards allows stakeholders to continuously monitor customer trends, segment performance, and key business metrics, enabling faster and more informed decision-making.

6. Future Enhancements

Future work may include incorporating additional data sources such as customer demographics, flight satisfaction scores, or real-time booking data. Advanced machine learning models could also be applied to predict customer lifetime value (CLV), churn probability, and demand trends more accurately.



Thank you

Thank you for taking the time to listen. All your questions and feedback are greatly appreciated.

Email: quocco.mar.26@gmail.com

Moblie: +84 902320263