King Khalid International Airport Flight Analysis

Introduction :

This report analyzes a dataset of flights arriving and departing from King Khalid International Airport (RUH).
The dataset includes detailed information about each flight, such as airline, aircraft type, flight number,
scheduled times, origin and destination airports, and flight status.

The goal of this analysis is to provide insights into airline operations, aircraft usage, flight scheduling,
and airport traffic patterns. The report will help identify the busiest airlines, most frequently used aircraft,
peak flight hours, and the most active airports, which can support operational decision-making and strategic planning .

Key Questions:

1. What are the basic statistics of the flights in the dataset, including total number of flights, unique airline companies, and different aircraft types used?
2. Which airlines have the highest number of flights, and what are the top 10 airlines by flight count?
3. How many flights of each status does each airline have?
4. Which airlines operate the highest average number of flights per day, and what are the top 10 airlines by average daily flights?
5. What are the top 10 most frequently used aircraft models?
6. Which hours of the day have the highest number of flights, and what are the top 10 busiest hours?
7. What is the average number of flights per day?
8. How do flight statuses vary across different hours of the day?
9. What are the top 5 days with the highest number of flights?
10. Which airports have the highest number of departing flights?
11. Which destination airports receive the highest number of flights?
12. Which destination airports are served by the highest number of airlines, and what are the top 10 busiest airports by airline diversity?
13. How are flights distributed across different airport time zones?
14. Which airline–origin airport combinations have the most flights?
15. What are the top 10 destination airports by number of flights?

Initial Data Overview

```python
import pandas as pd

# Load the Parquet file
df = pd.read_parquet("flights_RUH.parquet")

# Show the first 5 rows and the shape of the DataFrame
print("First 5 rows:\n", df.head())
print("\nData shape (rows, columns):", df.shape)
print("\nColumn names:\n", df.columns)

# General information about columns
print("\nColumn info, data types, and non-null counts:\n")
print(df.info())
print("\nStatistical summary of numeric columns:\n")
print(df.describe())

# Number of unique values per column with handling complex columns
print("\nNumber of unique values per column:")
for col in df.columns:
    try:
        print(f"{col}: {df[col].nunique()}")
    except TypeError:
        print(f"{col}: contains non-countable values -> converting to string")
        df[col] = df[col].astype(str)
        print(f"{col} (after conversion): {df[col].nunique()}")
```

```
First 5 rows:
   flight_number  ... movement.scheduledTime.local
0        PF 769  ...       2025-03-15 00:01+03:00
1        XY 333  ...       2025-03-15 00:05+03:00
2        QP 568  ...       2025-03-15 00:05+03:00
3        F3 161  ...       2025-03-15 00:10+03:00
4        KL 423  ...       2025-03-15 00:15+03:00

[5 rows x 23 columns]

Data shape (rows, columns): (153308, 23)

Column names:
 Index(['flight_number', 'aircraft.model', 'aircraft.reg', 'aircraft.modeS',
       'airline.name', 'airline.iata', 'airline.icao', 'status', 'flight_type',
       'codeshareStatus', 'isCargo', 'callSign', 'origin_airport_name',
       'origin_airport_icao', 'origin_airport_iata', 'movement.terminal',
       'movement.quality', 'destination_airport_icao',
       'destination_airport_iata', 'destination_airport_name',
       'movement.airport.timeZone', 'movement.scheduledTime.utc',
       'movement.scheduledTime.local'],
      dtype='object')

Column info, data types, and non-null counts:

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 153308 entries, 0 to 153307
Data columns (total 23 columns):
 #   Column                       Non-Null Count   Dtype
---  ------                       --------------   -----
 0   flight_number                153308 non-null  object
```

```
Statistical summary of numeric columns:

       flight_number  ... movement.scheduledTime.local
count         153308  ...                       153308
unique          1369  ...                        54364
top           LH 622  ...       2025-09-04 16:00+03:00
freq             403  ...                           16

[4 rows x 23 columns]

Number of unique values per column:
flight_number: 1369
aircraft.model: 48
aircraft.reg: 1431
aircraft.modeS: 1612
airline.name: 68
airline.iata: 62
airline.icao: 63
status: 5
flight_type: 2
codeshareStatus: 2
isCargo: 1
callSign: 790
origin_airport_name: 1
origin_airport_icao: 1
origin_airport_iata: 1
movement.terminal: 5
movement.quality: contains non-countable values -> converting to string
movement.quality (after conversion): 1
destination_airport_icao: 120
```

What are the basic statistics of the flights in the dataset, including total number of flights, unique airline companies, and different aircraft types used?

```python
total_flights = len(df)
print("Total flights:", total_flights)

num_airlines = df['airline.name'].nunique()
print("Total airlines companies:",num_airlines)

num_aircraft_types = df['aircraft.model'].nunique()
print("Total of aircraft types:",num_aircraft_types)
```

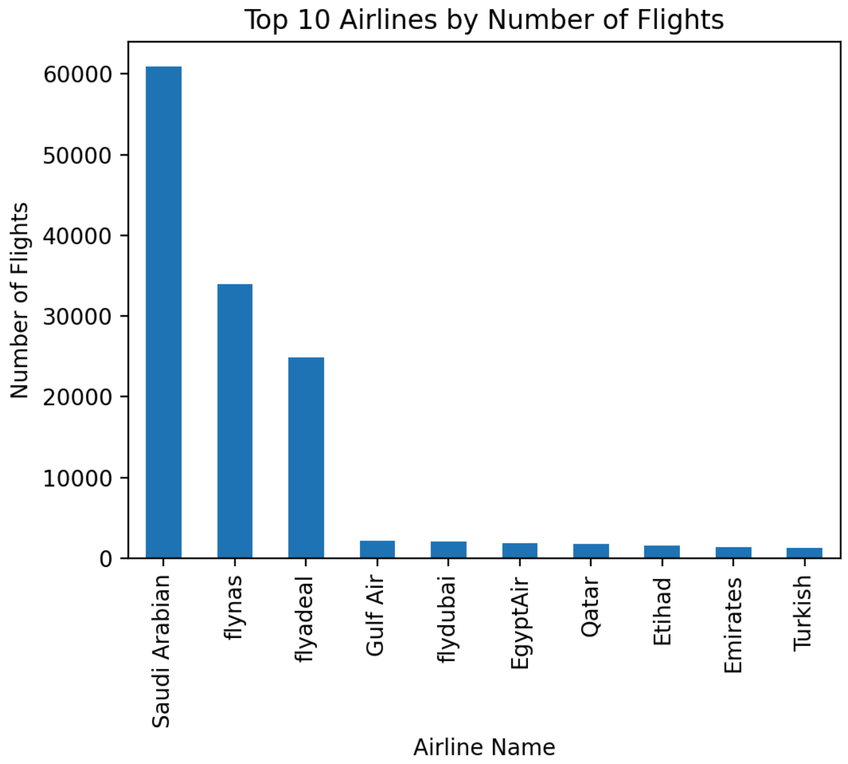
```
Total flights: 153308
Total airlines company: 68
Total of aircraft types: 48
```

Which airlines have the highest number of flights, and what are the top 10 airlines by flight count?

```python
flights_per_airline = df['airline.name'].value_counts()
print(flights_per_airline)

# Visualization
flights_per_airline.head(10).plot(kind='bar', figsize=(10,6),
title='Top 10 Airlines by Number of Flights')
plt.xlabel('Airline Name')
plt.ylabel('Number of Flights')
plt.show()
```

```
airline.name
Saudi Arabian    60886
flynas           33935
flyadeal         24835
Gulf Air          2198
flydubai          2059
                 ...
QQE                  4
VistaJet             4
Malindo Air          2
AIR X Charter        1
Atlas Air            1
Name: count, Length: 68, dtype: int64
```



How many flights of each status does each airline have?

```
status_by_airline = df.groupby(['airline.name','status']).size().unstack(fill_value=0)
print(status_by_airline)
```

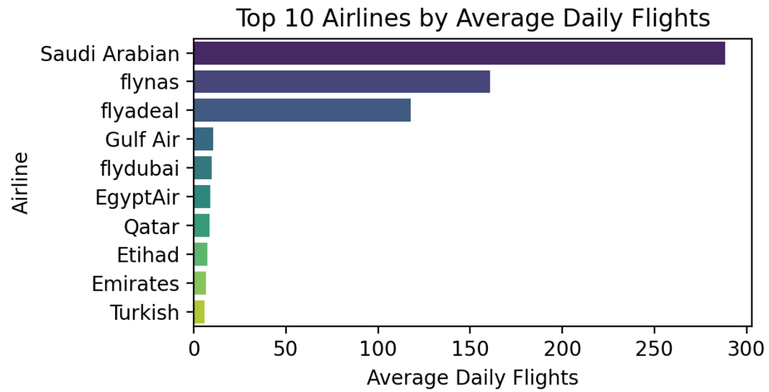
status	Canceled	CanceledUncertain	Departed	Expected	Unknown
airline.name					
AIR X Charter	0	0	0	1	0
AJet	0	0	0	0	401
AZAL Azerbaijan	0	0	0	0	155
Aegean	2	0	115	98	0
Air Arabia	0	0	0	0	965
...	...	...	...	...	...
Wizz Air Malta	1	0	3	4	0
ZanAir	0	0	0	0	47
flyadeal	8	8	1526	1641	21652
flydubai	0	0	184	267	1608
flynas	1	8	2420	2831	28675

Which airlines operate the highest average number of flights per day, and what are the top 10 airlines by average daily flights?

```python
df['departure_date'] = pd.to_datetime(df['movement.scheduledTime.utc']).dt.date
daily_counts = df.groupby(['airline.name', 'departure_date']).size().reset_index(name='daily_flights')
avg_daily_flights = daily_counts.groupby('airline.name')['daily_flights'] \
                                .mean() \
                                .round(2) \
                                .sort_values(ascending=False) \
                                .reset_index(name='avg_daily_flights')
print(avg_daily_flights.head(10))

# Visualization
plt.figure(figsize=(12,6))
sns.barplot(x='avg_daily_flights', y='airline.name',
            data=avg_daily_flights.head(10), palette='viridis')
plt.title('Top 10 Airlines by Average Daily Flights')
plt.xlabel('Average Daily Flights')
plt.ylabel('Airline')
plt.show()
```

	airline.name	avg_daily_flights
0	Saudi Arabian	288.56
1	flynas	160.83
2	flyadeal	117.70
3	Gulf Air	10.47
4	flydubai	9.76
5	EgyptAir	8.94
6	Qatar	8.59
7	Etihad	7.33
8	Emirates	6.47
9	Turkish	5.97



What are the top 10 most frequently used aircraft models ?

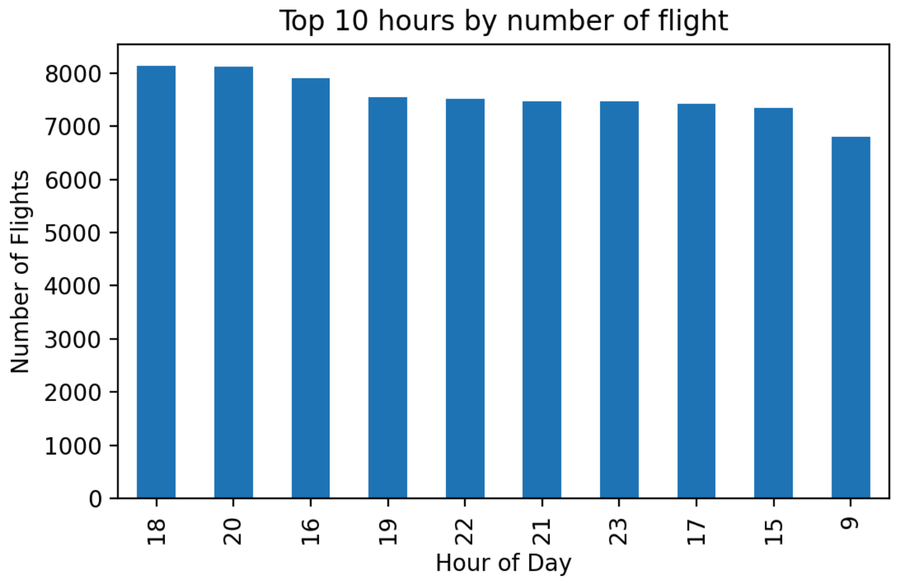
```
aircraft_model_counts = df['aircraft.model'].value_counts()
print(aircraft_model_counts.head(10))
```

```
aircraft.model
Airbus A320          59394
Airbus A320 NEO      35518
Airbus A321          20185
Airbus A330           8108
Boeing 777            4687
Airbus A330-300       3655
Boeing 737-800        2823
Airbus A321 NEO       2465
Boeing 777-300        2188
Boeing 737            1933
Name: count, dtype: int64
```

Which hours of the day have the highest number of flights, and what are the top 10 busiest hours?

```python
df['movement.scheduledTime.local'] = pd.to_datetime(df['movement.scheduledTime.local'], errors='coerce')
df['hour'] = df['movement.scheduledTime.local'].dt.hour
hourly_distribution = df['hour'].value_counts()
top_10_hours = hourly_distribution.nlargest(10)
print(top_10_hours)

top_10_hours.plot(kind='bar', figsize=(12,5), title='Top 10 hours by number of flight')
plt.xlabel('Hour of Day')
plt.ylabel('Number of Flights')
plt.show()
```



hour	
18	8142
20	8131
16	7904
19	7545
22	7516
21	7479
23	7477
17	7419
15	7356
9	6811

What is the average number of flights per day ?

```
df['movement.scheduledTime.local'] = pd.to_datetime(df['movement.scheduledTime.local'], errors='coerce')
df['date'] = df['movement.scheduledTime.local'].dt.date
daily_avg = df.groupby('date').size().mean()
print("Average flights per day:", round(daily_avg))
```

```
Average flights per day: 730
```

How do flight statuses vary across different hours of the day?

```
df['movement.scheduledTime.local'] = pd.to_datetime(
    df['movement.scheduledTime.local'], errors='coerce'
)

df = df.dropna(subset=['movement.scheduledTime.local'])

df['hour'] = df['movement.scheduledTime.local'].dt.hour

status_by_hour = df.groupby(['hour', 'status']).size().unstack(fill_value=0)
print(status_by_hour)
```

status	Canceled	CanceledUncertain	Departed	Expected	Unknown
hour					
0	5	0	250	728	4617
1	7	0	474	645	3432
2	3	1	1081	349	3559
3	2	1	77	550	3054
4	1	0	61	101	4516
5	1	0	295	218	4073
6	2	1	402	729	4830
7	3	2	789	277	4690
8	5	1	712	202	5362
9	2	0	758	66	5985
10	2	4	573	432	5658
11	1	2	638	697	5134
12	1	1	404	194	5051
13	2	3	295	441	5308
14	0	0	120	728	5731

What are the top 5 days with the highest number of flights ?

```
df['departure_date'] = pd.to_datetime(df['movement.scheduledTime.local']).dt.date
daily_flights = df.groupby('departure_date').size().reset_index(name='num_flights')
top5_days = daily_flights.sort_values(by='num_flights', ascending=False).head(5)
print(top5_days)
```

	departure_date	num_flights
149	2025-08-11	805
156	2025-08-18	803
163	2025-08-25	799
128	2025-07-21	799
142	2025-08-04	798

Which airports have the highest number of departing flights?

```
origin_counts = df['origin_airport_name'].value_counts()
print(origin_counts)
```

```
origin_airport_name
Riyadh    153308
Name: count, dtype: int64
```

Which airports receive the highest number of flights?

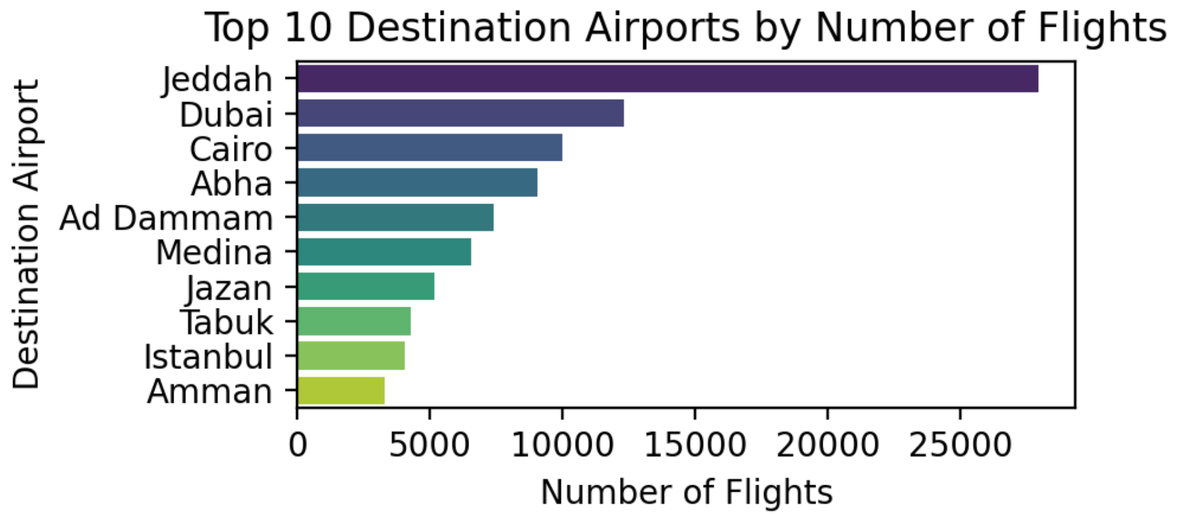
```
dest_counts = df['destination_airport_name'].value_counts()
print(dest_counts.head(10))
```

```
destination_airport_name
Jeddah        27938
Dubai         12333
Cairo         10003
Abha           9043
Ad Dammam      7393
Medina         6563
Jazan          5196
Tabuk          4295
Istanbul       4083
Amman          3325
Name: count, dtype: int64
```

Which destination airports receive the highest number of flights, and what are the top 10 destination airports by flight volume?

```
top10_dest = df.groupby('destination_airport_name').size().reset_index(name='num_flights')
top10_dest = top10_dest.sort_values(by='num_flights', ascending=False).head(10)
print(top10_dest)

# Visualization
plt.figure(figsize=(12,6))
sns.barplot(x='num_flights', y='destination_airport_name', data=top10_dest, palette='viridis')
plt.title('Top 10 Destination Airports by Number of Flights')
plt.xlabel('Number of Flights')
plt.ylabel('Destination Airport')
plt.show()
```



	destination_airport_name	num_flights
57	Jeddah	27938
39	Dubai	12333
28	Cairo	10003
0	Abha	9043
2	Ad Dammam	7393
77	Medina	6563
55	Jazan	5196
112	Tabuk	4295
53	Istanbul	4083
10	Amman	3325

Which airline–origin airport combinations have the most flights?

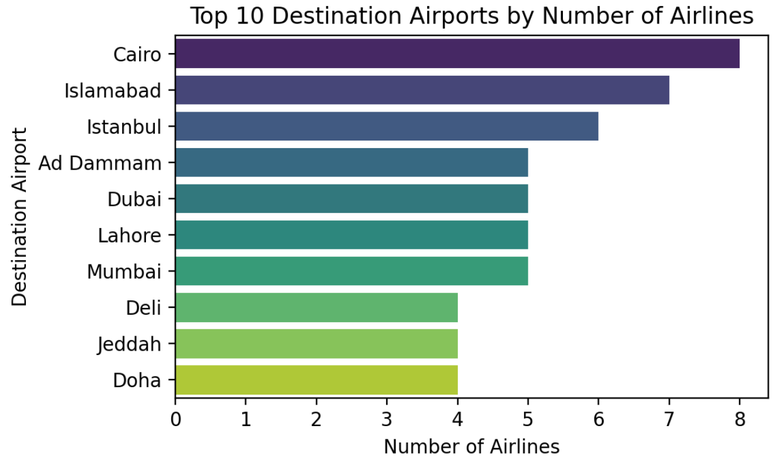
```
flights_airline_origin = df.groupby(['airline.name','origin_airport_name']).size()
print(flights_airline_origin.head(10))
```

```
airline.name      origin_airport_name
AIR X Charter     Riyadh                1
AJet              Riyadh              401
AZAL Azerbaijan   Riyadh              155
Aegean            Riyadh              215
Air Arabia        Riyadh              965
Air Arabia Egypt  Riyadh              724
Air Cairo         Riyadh              641
Air China         Riyadh              186
Air France        Riyadh              168
Air India         Riyadh              654
dtype: int64
```

Which destination airports are served by the highest number of airlines, and what are the top 10 busiest airports by airline diversity?

```
airport_airlines = df.groupby("destination_airport_name")['airline.name'] \
                    .nunique() \
                    .reset_index(name='num_airlines') \
                    .sort_values(by='num_airlines', ascending=False)

print(airport_airlines.head(10))
# Visualization
plt.figure(figsize=(12,6))
sns.barplot(x='num_airlines', y='destination_airport_name',
            data=airport_airlines.head(10), palette='viridis')
plt.title('Top 10 Destination Airports by Number of Airlines')
plt.xlabel('Number of Airlines')
plt.ylabel('Destination Airport')
plt.show()
```



	destination_airport_name	num_airlines
28	Cairo	8
52	Islamabad	7
53	Istanbul	6
2	Ad Dammam	5
39	Dubai	5
70	Lahore	5
81	Mumbai	5
36	Deli	4
57	Jeddah	4
38	Doha	4

How are flights distributed across different airport time zones?

```
timezone_counts = df['movement.airport.timeZone'].value_counts()
print(timezone_counts)
```

```
movement.airport.timeZone
Asia/Riyadh            79413
Asia/Dubai             16164
Africa/Cairo           11726
Europe/Istanbul         4775
Asia/Kolkata            4017
Asia/Karachi            3909
Asia/Amman              3320
Asia/Kuwait             3081
Asia/Bahrain            3020
Asia/Qatar              2921
Europe/London           2280
Asia/Muscat             1555
Asia/Shanghai           1194
Asia/Dhaka              1040
Europe/Rome              973
Asia/Beirut              877
Africa/Addis_Ababa       796
```