DATA ENGINEERING and APPLIED DATA SCIENCE assignment

TOPIC

A Data Pipeline for Flight Schedules - Chhatrapati Shivaji Maharaj International Airport (IATA: BOM): From Raw data to Insights.

ROLL NO: 12

MSc. DATA SCIENCE PART 2

Chhatrapati Shivaji Maharaj International Airport (IATA: BOM) is an international airport serving Mumbai, Maharashtra. It is India's second busiest airport in terms of total and international passengers followed by Delhi. In 2023–24, it was ranked ninth in Asia and 25th worldwide by passenger traffic.

URLs used –

1. Flight schedule for incoming and outgoing flights:
   https://api.aviationstack.com/v1/timetable?iataCode=BOM&type=departure&access_key=26f51262be2cbdb8562d8ebb4f998136
   https://api.aviationstack.com/v1/timetable?iataCode=BOM&type=arrival&access_key=26f51262be2cbdb8562d8ebb4f998136
2. Airports:
   https://raw.githubusercontent.com/jpatokal/openflights/master/data/airports.dat

Note - Due to limited requests I have used an excel sheet with Airport IATA, Airport name and Country as columns. Besides, the data is also limited.

Flight Schedule Data Analysis is important for enhancing operational efficiency and improving customer experience in the aviation industry. By studying this information, airlines can improve flight routes, reduce delays, and better manage their resources such as aircrafts, staff, etc. It also aids in demand forecasting, allowing airlines to better manage capacity and adjust pricing strategies. Tracking performance also shows where they can do better and ensures they follow safety rules. Overall, this analysis leads to more informed decision-making, contributing to smoother operations and greater passenger satisfaction.

DATA PIPELINE -

1. **Import libraries:**

```
import datetime as dt
from datetime import timedelta
from airflow import DAG
from airflow.operators.bash_operator import BashOperator
from airflow.operators.python_operator import PythonOperator
import requests as req
import pandas as pd
```

2. **Get Data from necessary sources –**

```
def getData():
    data_dep=req.get(“https://api.aviationstack.com/v1/timetable?iataCode=BOM&type=departure&access_key=26f51262be2cbdb8562d8ebb4f998136”)
    dep_df=pd.read_csv(data_dep.json()[‘data’])
    dep_df.to_csv(‘/home/akshata/Flight_schedules/depart_dir/departure.csv”)
    data_arr=req.get(“https://api.aviationstack.com/v1/timetable?iataCode=BOM&type=arrival&access_key=26f51262be2cbdb8562d8ebb4f998136”)
    arr_df=pd.read_csv(data_arr.json()[‘data’])
    arr_df.to_csv(‘/home/akshata/Flight_schedules/arrival_dir/arrival.csv”)
```

3. **Cleaning the Flights data for departure as well as arrival:**

```
def clean_Flights():
    #for departure
    dep_flight=pd.read_csv("/home/akshata/Flight_schedules/depart_dir/depature.csv")

    columns=['codeshared' ,'airline.icaoCode', 'arrival.icaoCode', 'arrival.actualRunway',
    'arrival.actualTime', 'arrival.baggage', 'arrival.estimatedRunway', 'arrival.gate',
    'departure.iataCode', 'departure.icaoCode', 'departure.actualRunway',
    'departure.actualTime', 'departure.baggage', 'departure.estimatedRunway',
    'departure.gate',' flight.icaoNumber', 'codeshared.airline.iataCode',
    codeshared.airline.icaoCode', 'codeshared.airline.name', 'codeshared.flight.iataNumber',
    'codeshared.flight.icaoNumber', 'codeshared.flight.number']

    dep_flight_df=dep_flight.drop(columns=columns)
    dep_flight_df.to_csv("/home/akshata/Flight_schedules/depart_dir/clean_dep.csv")

    #for arrival
    arr_flight=pd.read_csv("/home/akshata/Flight_schedules/arrival_dir/arrival.csv")

    columns= ['codeshared', 'airline.icaoCode', 'arrival.iataCode', 'arrival.icaoCode',
    'arrival.actualRunway', 'arrival.actualTime', 'arrival.baggage', 'arrival.delay',
    'arrival.estimatedRunway', 'arrival.gate', 'departure.icaoCode', 'departure.actualRunway',
    'departure.actualTime', 'departure.baggage', 'departure.estimatedRunway', 'departure.gate',
    'flight.icaoNumber', 'codeshared.airline.iataCode', 'codeshared.airline.icaoCode',
    ‘codeshared.airline.name', 'codeshared.flight.iataNumber',
    'codeshared.flight.icaoNumber', 'codeshared.flight.number']

    arr_flight_df=arr_flight.drop(columns=columns)
    arr_flight_df.to_csv("/home/akshata/Flight_schedules/arrival_dir/clean_arr.csv")
```

4. **Preprocessing the data**

```
def preprocess_data():

    final_dep=pd.read_csv("/home/akshata/Flight_schedules/depart_dir/clean_dep.csv")
    final_arr=pd.read_csv("/home/akshata/Flight_schedules/arrival_dir/celan_arr.csv")

    #for departure
    #converting string to datetime
    final_dep['departure.scheduledTime']=pd.to_datetime(final_dep['departure.scheduledTime'])
    final_dep['departure.estimatedTime']=pd.to_datetime(final_dep['departure.scheduledTime'])
    final_dep['arrival.scheduledTime']=pd.to_datetime(final_dep['arrival.scheduledTime'])
    final_dep['arrival.estimatedTime']=pd.to_datetime(final_dep['arrival.scheduledTime'])

    #handling missing data
    final_dep['departure.delay']=final_dep['departure.delay'].fillna(0)
    final_dep['arrival.delay']=final_dep['arrival.delay'].fillna(0)

    #derive columns
    final_dep["flight_duration"]=final_dep["arrival.scheduledTime"]-
    final_dep["departure.scheduledTime"]
    final_dep.to_csv("/home/akshata/Flight_schedules/depart_dir/final_dep.csv")

    #for arrival
    #converting string to datetime
    final_arr['departure.scheduledTime']=pd.to_datetime(final_arr['departure.scheduledTime'
    ])
    final_arr['departure.estimatedTime']=pd.to_datetime(final_arr['departure.scheduledTime']
    )
    final_arr['arrival.scheduledTime']=pd.to_datetime(final_arr['arrival.scheduledTime'])
    final_arr['arrival.estimatedTime']=pd.to_datetime(final_arr['arrival.scheduledTime'])
    #handling missing data
    final_arr['departure.delay']=final_dep['departure.delay'].fillna(0)
    final_arr['arrival.delay']=final_dep['arrival.delay'].fillna(0)

    #derive columns
    final_arr["flight_duration"]=final_arr["arrival.scheduledTime"]-
    final_arr["departure.scheduledTime"]
    final_arr.to_csv("/home/akshata/Flight_schedules/arrival_dir/final_arr.csv")
```

5. **Concatenating both the Departure as well as Arrival data and storing the Final Flight data-**

```
def final_data():

    arr=pd.read_csv("/home/akshata/Flight_schedules/arrival_dir/final_arr.csv")
    dep=pd.read_csv("/home/akshata/Flight_schedules/depart_dir/final_dep.csv")
    final=pd.concat([arr,dep])
    final.to_csv("/home/akshata/Flight_schedules/flights.csv")
```

6. **Get and Merge the airport data with flights data–**

```
def Airport():

    #get data
    data_df=pd.read_csv("/home/akshata/Flight_schedules/flights.csv")
    airport=pd.read_csv("/home/akshata/Flight_schedules/airport_data.csv")
    #airport=pd.Dataframe(“https://raw.githubusercontent.com/jpatokal/openflights/master/d
    ata/airports.dat“)
    #merge data
    join_df2=pd.merge(data_df[data_df['type']=='arrival'],airport,left_on='departure.iataCode
    , 'right_on='Airport', how='inner')
    join_df1=pd.merge(data_df[data_df['type']=='departure'],airport,left_on='arrival.iataCode
    ,'right_on='Airport', how='inner')
    final_merge=pd.concat([join_df1,join_df2]).drop_duplicates().reset_index(drop=True)

    final_merge.to_csv("/home/akshata/Flight_schedules/flightairport_data.csv")
```

7. **Clean the Merged data and add a derived column as follows -**

```
def clean_process_Merge():

    data=pd.read_csv("/home/akshata/Flight_schedules/flightairport_data.csv")
    data_df=data.drop(columns=['Unnamed: 0.3', 'Unnamed: 0.2', 'Unnamed: 0.1', 'Unnamed:
    0.4', ‘Unnamed: 0’]
    data_df['Flight_type']=np.where(final_merge['Country']=='India', ‘Domestic’,
    ‘International’)

    data_df.to_csv("/home/akshata/Flight_schedules/flight_final_data.csv")
```

8. **A DAG file is created by calling all the functions that are created -**

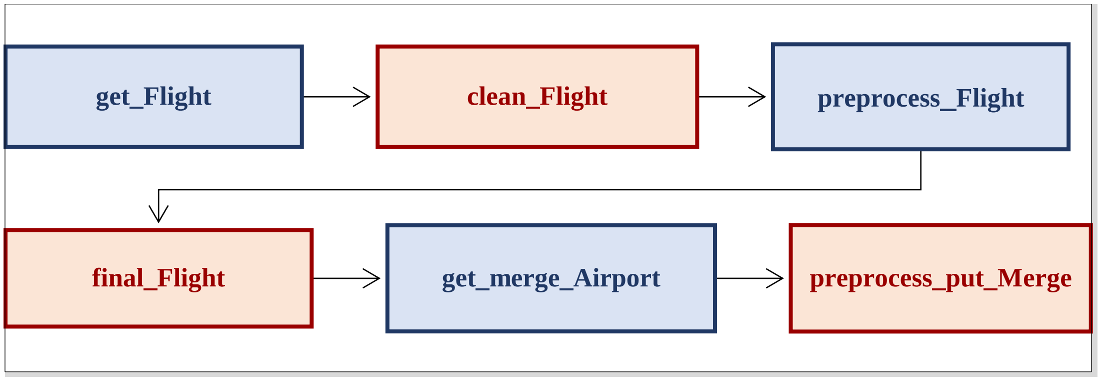
```
default_args = {'owner': 'akshata', 'start_date': dt.datetime(2024, 10, 10), 'retries': 1,
'retry_delay': dt.timedelta(minutes=5)}

with DAG('flight_dag', default_args=default_args, schedule_interval=timedelta(minutes=5))
as dag:

    get_Flight=PythonOperator(task_id=’getdata”, python_callable=getData)

    clean_Flight=PythonOperator(task_id='clean', python_callable=clean_data)
    preprocess_Flight=PythonOperator(task_id='preprocess' ,
    python_callable=preprocess_data)
    final_Flight=PythonOperator(task_id='final', python_callable=final_data)
    get_merge_Airport=PythonOperator(task_id='GetMerge' , python_callable=Airport)
    preprocess_put_merge=PythonOperator(task_id='PreprocessPut',
    python_callable=clean_preprocess_Merge)

    get_Flight >> clean_Flight >> preprocess_Flight >> final_Flight >> get_merge_Airport
    >> preprocess_put_merge
```



VISUALIZATION AND ANALYSIS –

The below dashboard gives insights of the following –

1. The total flights on October 20-21 are 1088, from which the total flights to departure from the airport are 605 and total flights that will arrive at the airport are 483. The total number of domestic flights is 893 while international flights are 195.
2. The status of flight distribution shows that 42% flights are active, that is on time, 34.74% is scheduled, 0.92% are cancelled or delayed and 0.83% are delayed. For flights that are arriving 21.51% are landed at the airport.
3. A map highlights the countries visited most frequently from this airport. The top countries include Saudi Arabia (11.28%), UAE (21.54%), and India itself with some regional breakdowns.
4. The top 10 Airlines by number of flights includes IndiGo at top with more than 200 flights which is followed by Air India and Vistara.
5. From October 20 12pm to October 21 12am is the peak time for flights that depart from the airport.

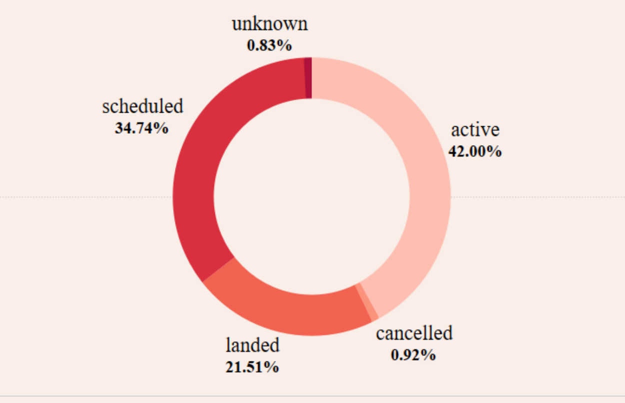
Insights on Flights at Chhatrapati Shivaji Maharaj International Airport

Total Flights

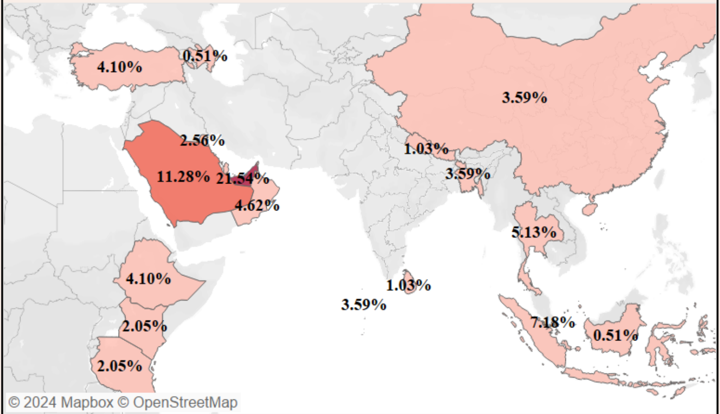
1,088

Flight type	Type arrival	departure
Domestic	384	509
International	99	96

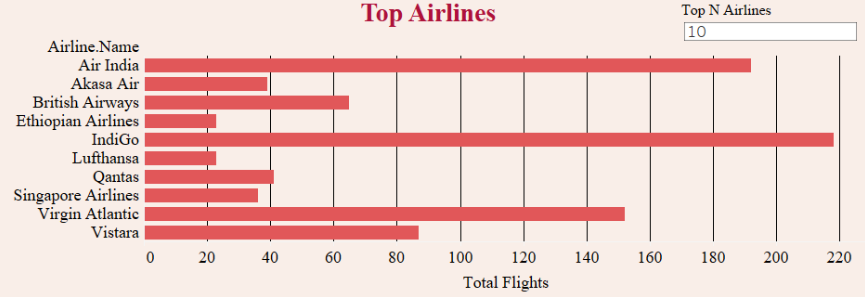
Flight Status Distribution



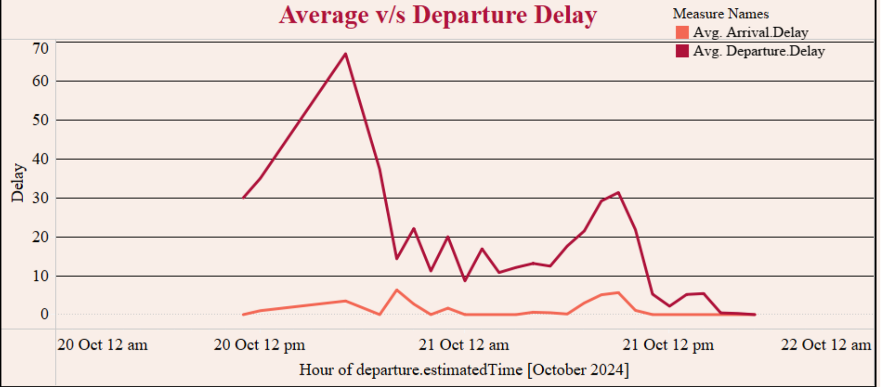
Most Visted Country

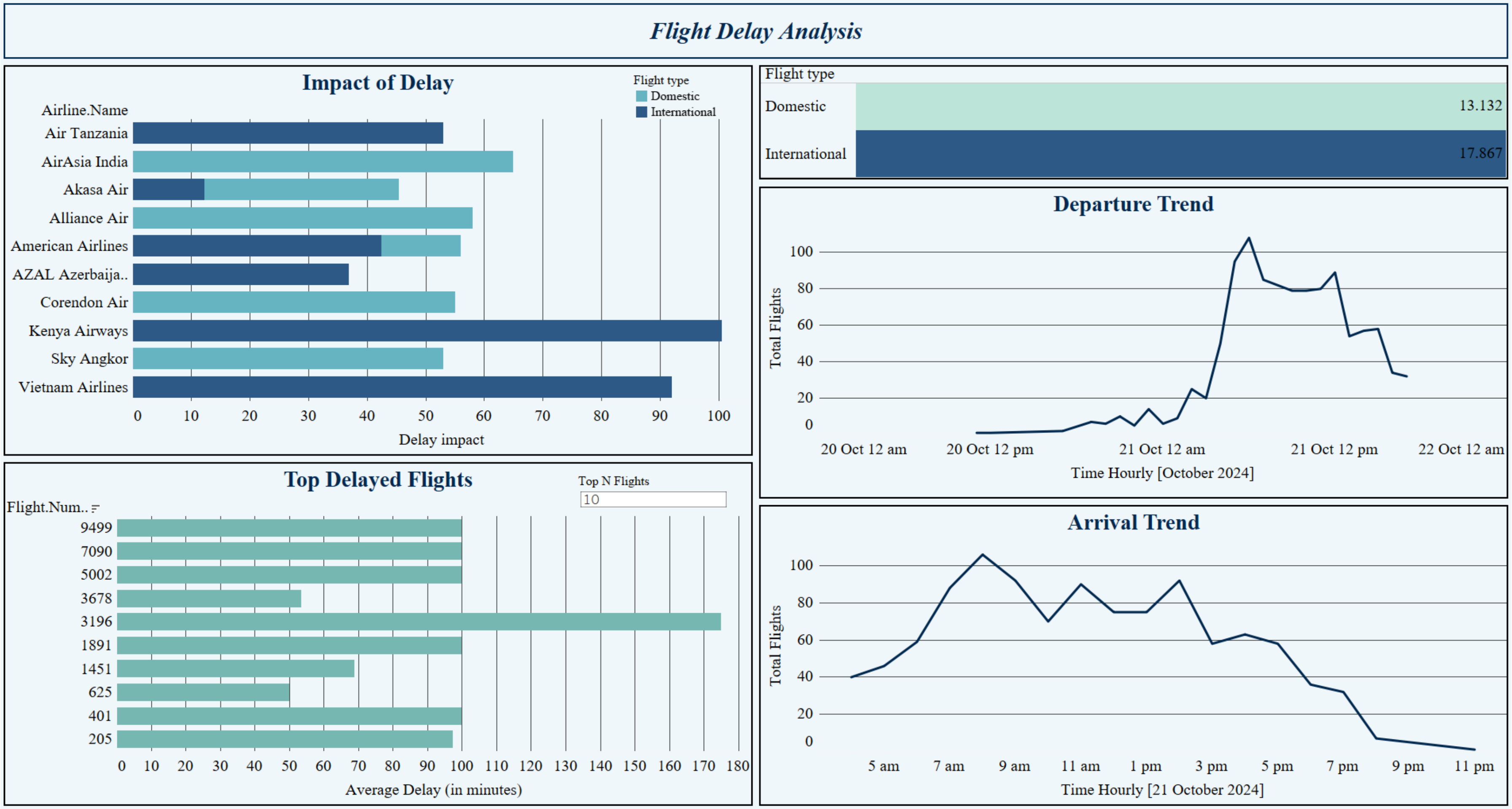


Top Airlines



Average v/s Departure Delay




Flight Delay Analysis
Impact of Delay
Flight type
Domestic
International
Airline.Name
Air Tanzania
AirAsia India
Akasa Air
Alliance Air
American Airlines
AZAL Azerbaija..
Corendon Air
Kenya Airways
Sky Angkor
Vietnam Airlines
0
10
20
30
40
50
60
70
80
90
100
Delay impact
Flight type
Domestic
13.132
International
17.867
Departure Trend
Total Flights
100
80
60
40
20
0
20 Oct 12 am
20 Oct 12 pm
21 Oct 12 am
21 Oct 12 pm
22 Oct 12 am
Time Hourly [October 2024]
Top Delayed Flights
Top N Flights
10
Flight.Num..
9499
7090
5002
3678
3196
1891
1451
625
401
205
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
Average Delay (in minutes)
Arrival Trend
Total Flights
100
80
60
40
20
0
5 am
7 am
9 am
11 am
1 pm
3 pm
5 pm
7 pm
9 pm
11 pm
Time Hourly [21 October 2024]

The above dashboard gives insights of the following –

1. Airlines like Air Tanzania and Coren don Air have the longest delays, with impacts reaching close to 100 minutes for some international flights.
2. There is a peak in departure delays during the late hours of October 20th and early hours of October 21st in Departure Trend chart.
3. In the Arrival Trend chart, a peak around 9 AM on October 21st, followed by a steady decline throughout the day in arrival delays.
4. In top delayed flights, the flight with flight number 3198 shows the longest delay of 170 minutes followed by flight 9499 and 7090 with a delay of approximately 100 minutes and so on.