Evaluation

PREDICTIVE ANALYTICS

Team Number - 13

1. Tanay Garg - 500108691
2. Yash Varshney - 500110167

Predicting Flight Delays

To predict flight delays using a dataset with multiple features like flight info, weather, and air traffic conditions.

Project Goal:

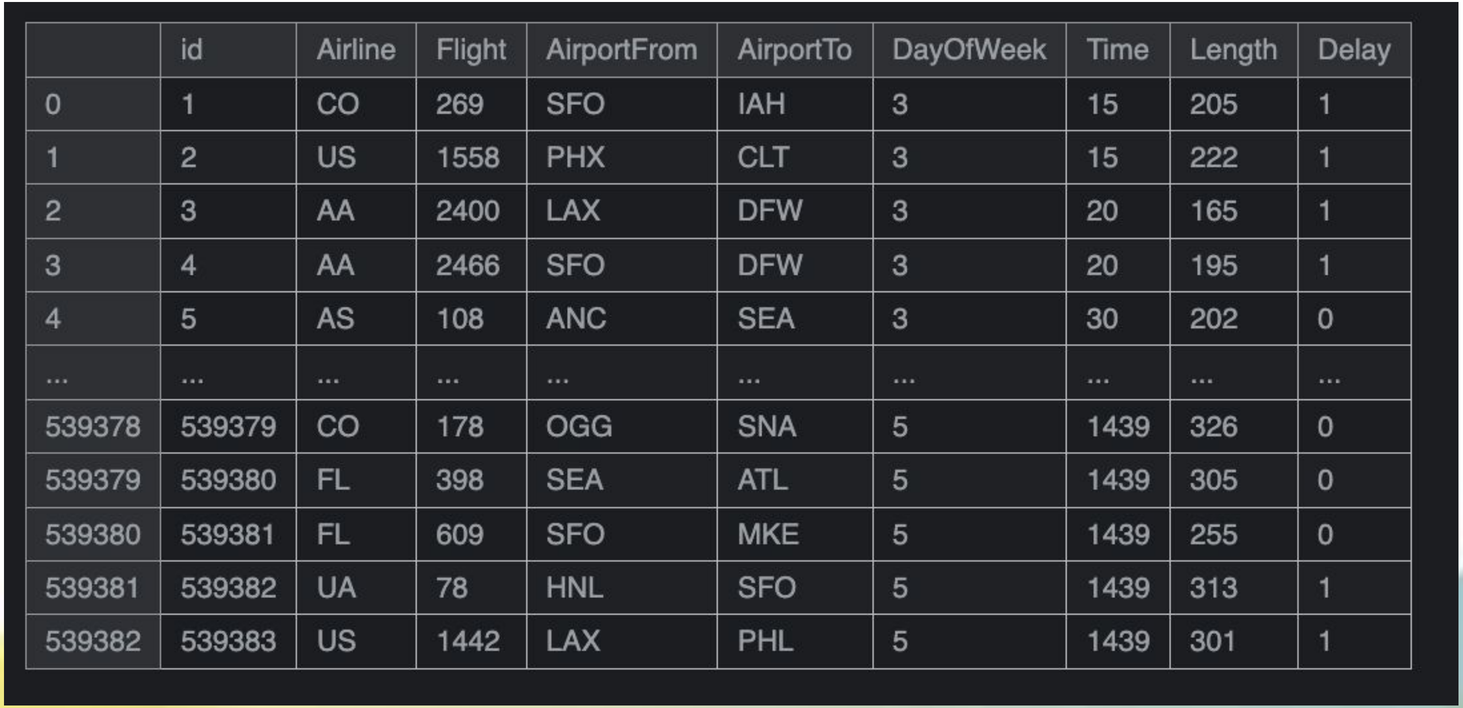
To predict flight delays using a dataset with multiple features like flight info, weather, and air traffic conditions.

Why This is Important:

Delays affect passengers, airlines, and airport operations. Predicting delays helps improve planning and efficiency.

Dataset

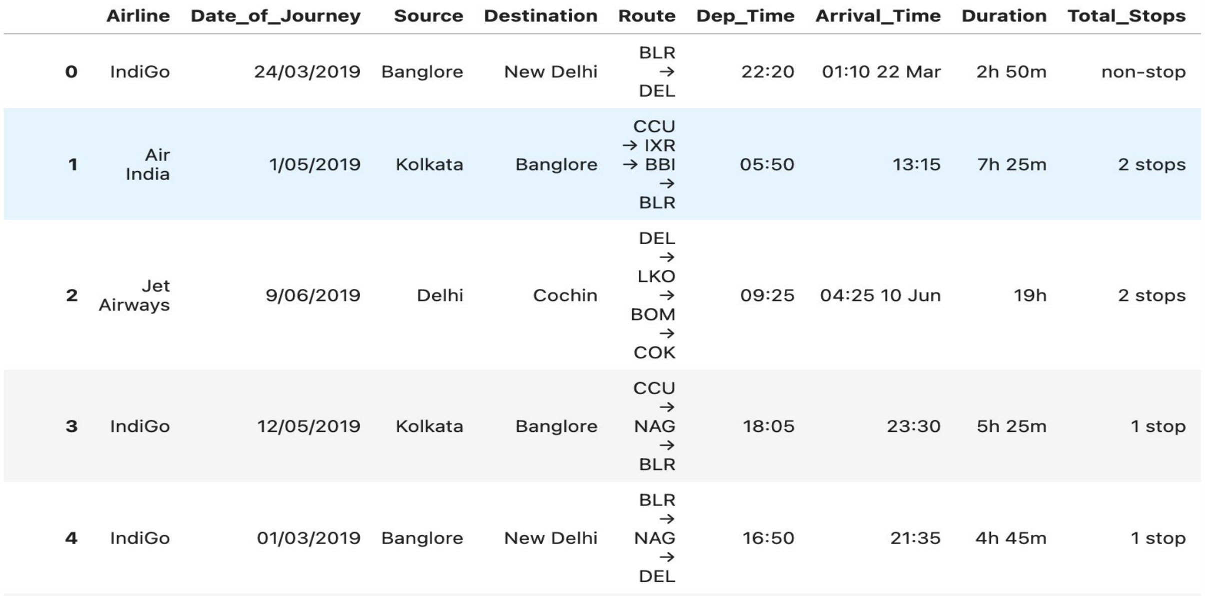
Combining 2 Datasets, One from Kaggle and another from Github (Indian_Flight_Dataset)



	id	Airline	Flight	AirportFrom	AirportTo	DayOfWeek	Time	Length	Delay
0	1	CO	269	SFO	IAH	3	15	205	1
1	2	US	1558	PHX	CLT	3	15	222	1
2	3	AA	2400	LAX	DFW	3	20	165	1
3	4	AA	2466	SFO	DFW	3	20	195	1
4	5	AS	108	ANC	SEA	3	30	202	0
...	...	...	...	...	...	...	...	...	...
539378	539379	CO	178	OGG	SNA	5	1439	326	0
539379	539380	FL	398	SEA	ATL	5	1439	305	0
539380	539381	FL	609	SFO	MKE	5	1439	255	0
539381	539382	UA	78	HNL	SFO	5	1439	313	1
539382	539383	US	1442	LAX	PHL	5	1439	301	1

Dataset (5.45 lakhs data entries)

Indian_Flight_Dataset (Github)



	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops
2	Jet Airways	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop
4	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop

Parameters

- **Temporal Features:**

  Day of Week, Month, Season.
- **Weather Information:**

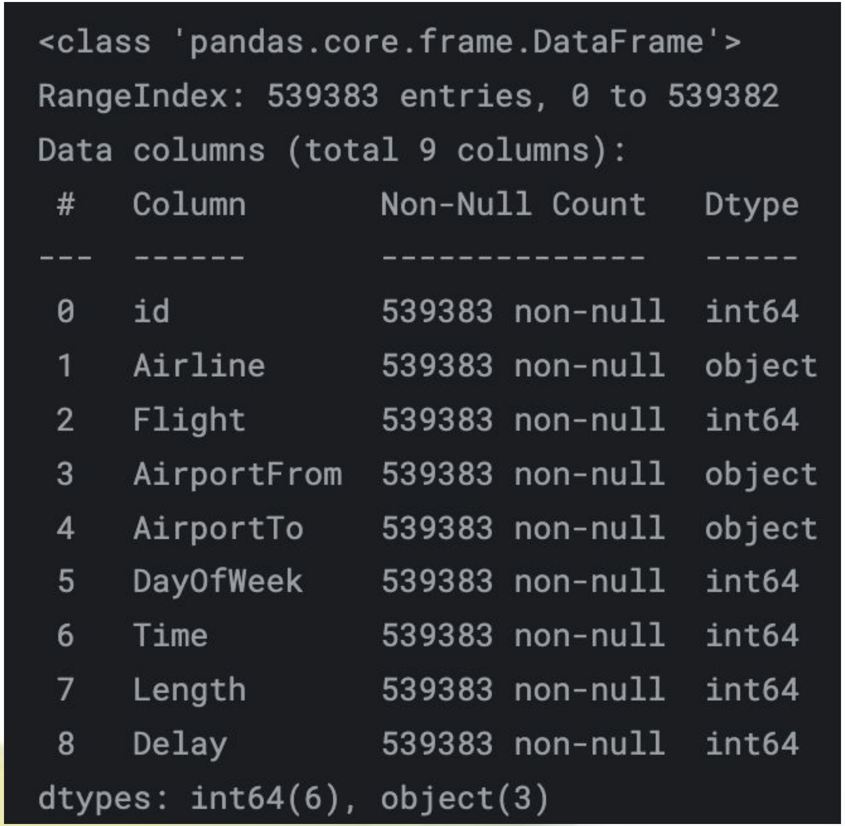
  Departure and Arrival Airport Weather.
- **Air Traffic Control Factors:**

  Air Traffic Volume, Runway Availability.
- **Operational Factors:**

  Aircraft Type, Crew Information.
- **Historical Delay Data:**

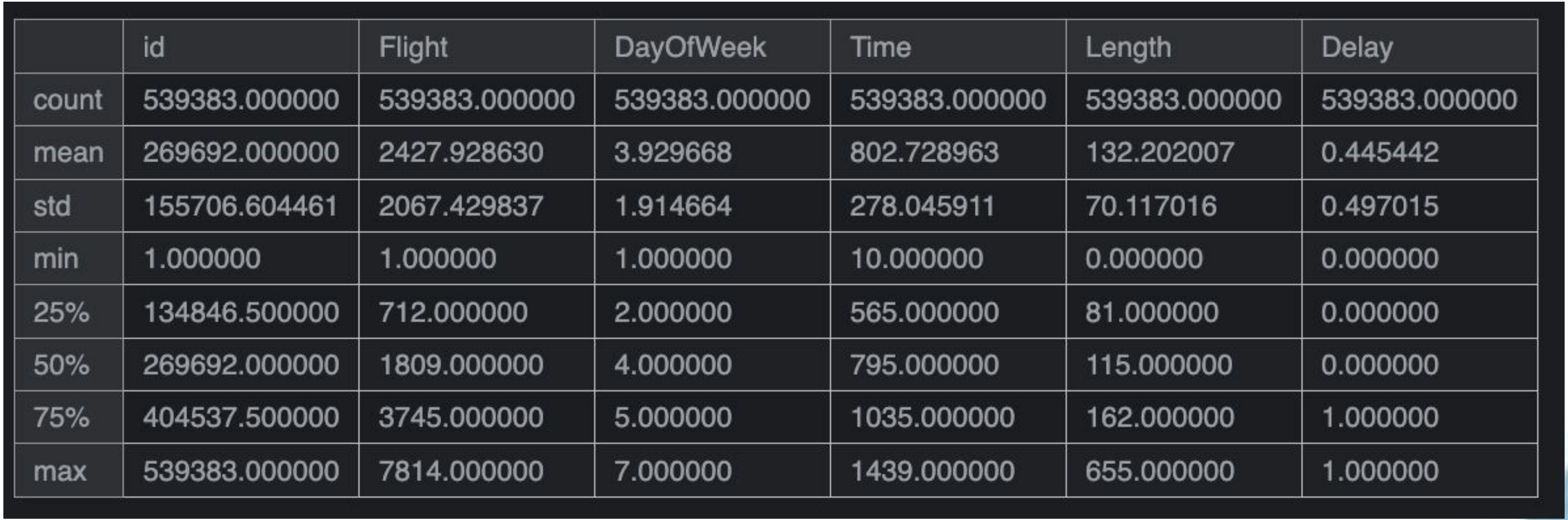
  Previous Delays, Delay Reasons.

Data Preprocessing



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 539383 entries, 0 to 539382
Data columns (total 9 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   id           539383 non-null  int64
 1   Airline      539383 non-null  object
 2   Flight       539383 non-null  int64
 3   AirportFrom  539383 non-null  object
 4   AirportTo    539383 non-null  object
 5   DayOfWeek    539383 non-null  int64
 6   Time         539383 non-null  int64
 7   Length       539383 non-null  int64
 8   Delay        539383 non-null  int64
dtypes: int64(6), object(3)
```

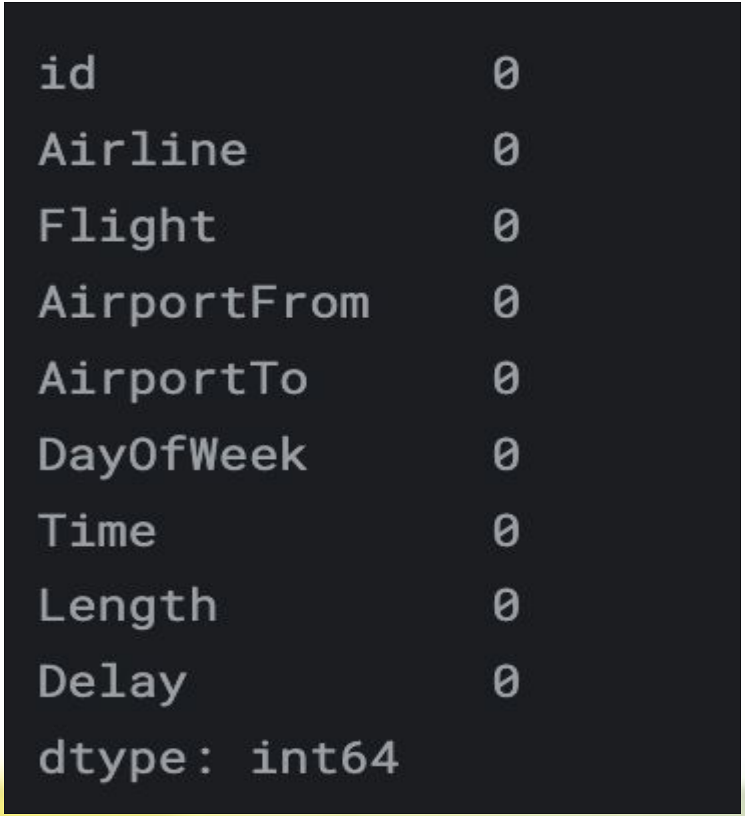
df.info()



	id	Flight	DayOfWeek	Time	Length	Delay
count	539383.000000	539383.000000	539383.000000	539383.000000	539383.000000	539383.000000
mean	269692.000000	2427.928630	3.929668	802.728963	132.202007	0.445442
std	155706.604461	2067.429837	1.914664	278.045911	70.117016	0.497015
min	1.000000	1.000000	1.000000	10.000000	0.000000	0.000000
25%	134846.500000	712.000000	2.000000	565.000000	81.000000	0.000000
50%	269692.000000	1809.000000	4.000000	795.000000	115.000000	0.000000
75%	404537.500000	3745.000000	5.000000	1035.000000	162.000000	1.000000
max	539383.000000	7814.000000	7.000000	1439.000000	655.000000	1.000000

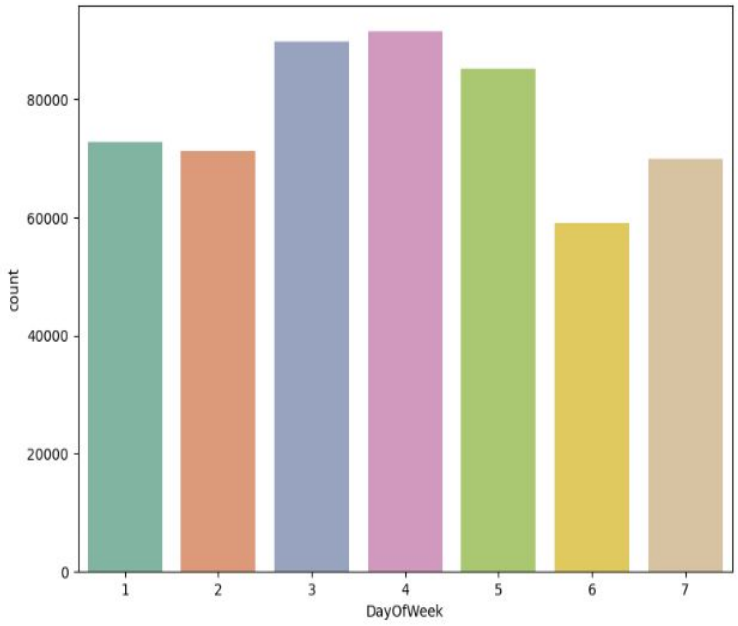
df.describe()

No Null Values Found in Dataset

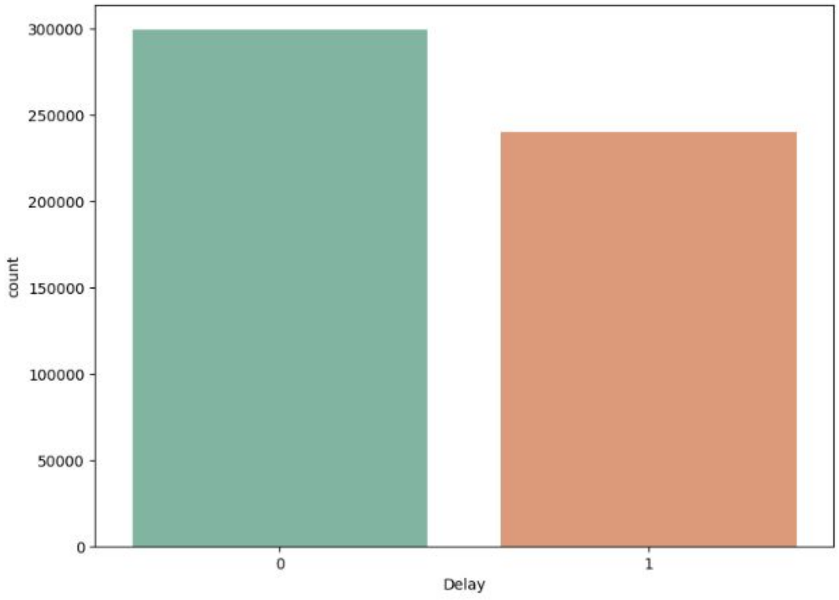


```
id             0
Airline        0
Flight         0
AirportFrom    0
AirportTo      0
DayOfWeek      0
Time           0
Length         0
Delay          0
dtype: int64
```

df.isnull().sum()

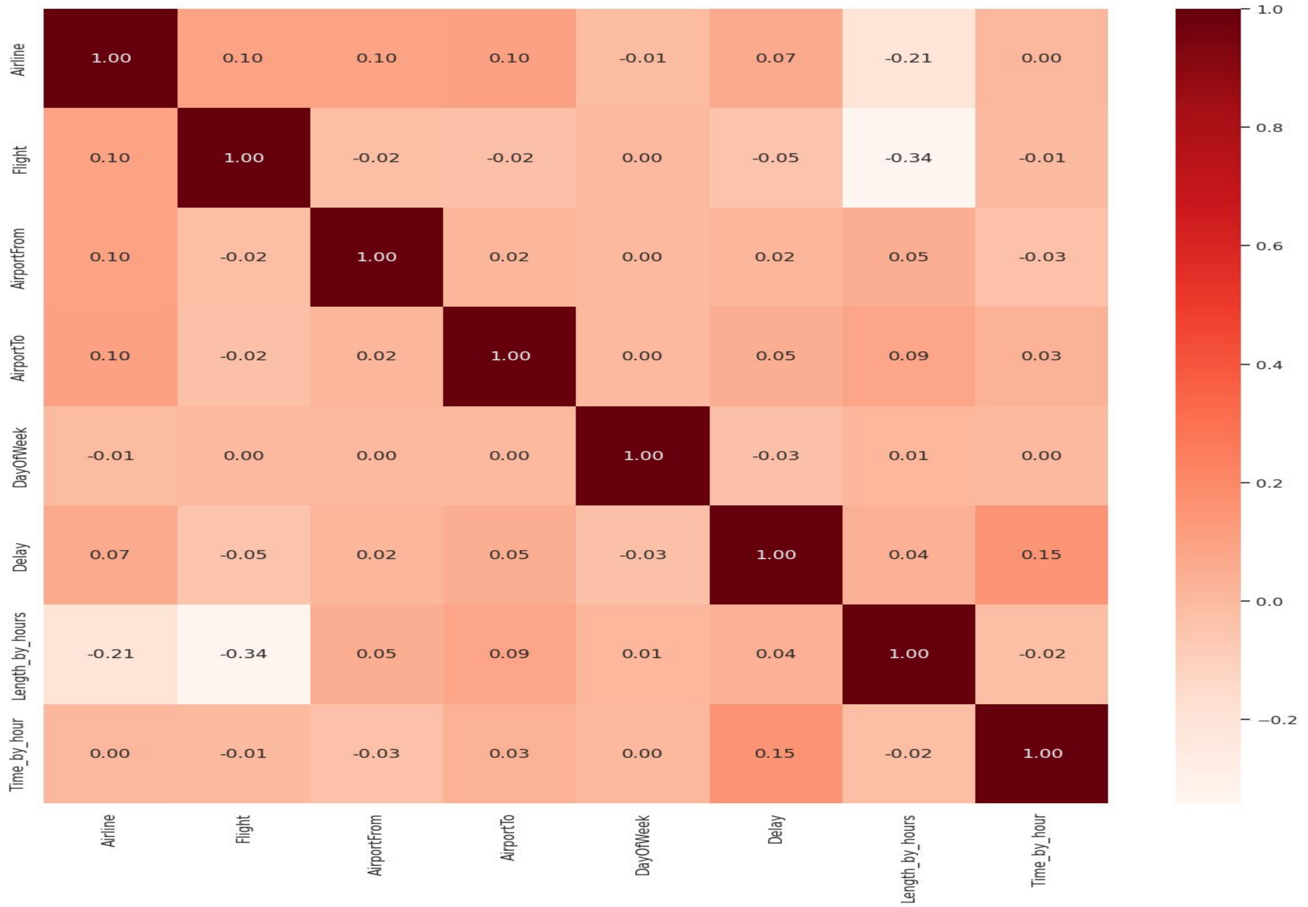


Flight count for days of week



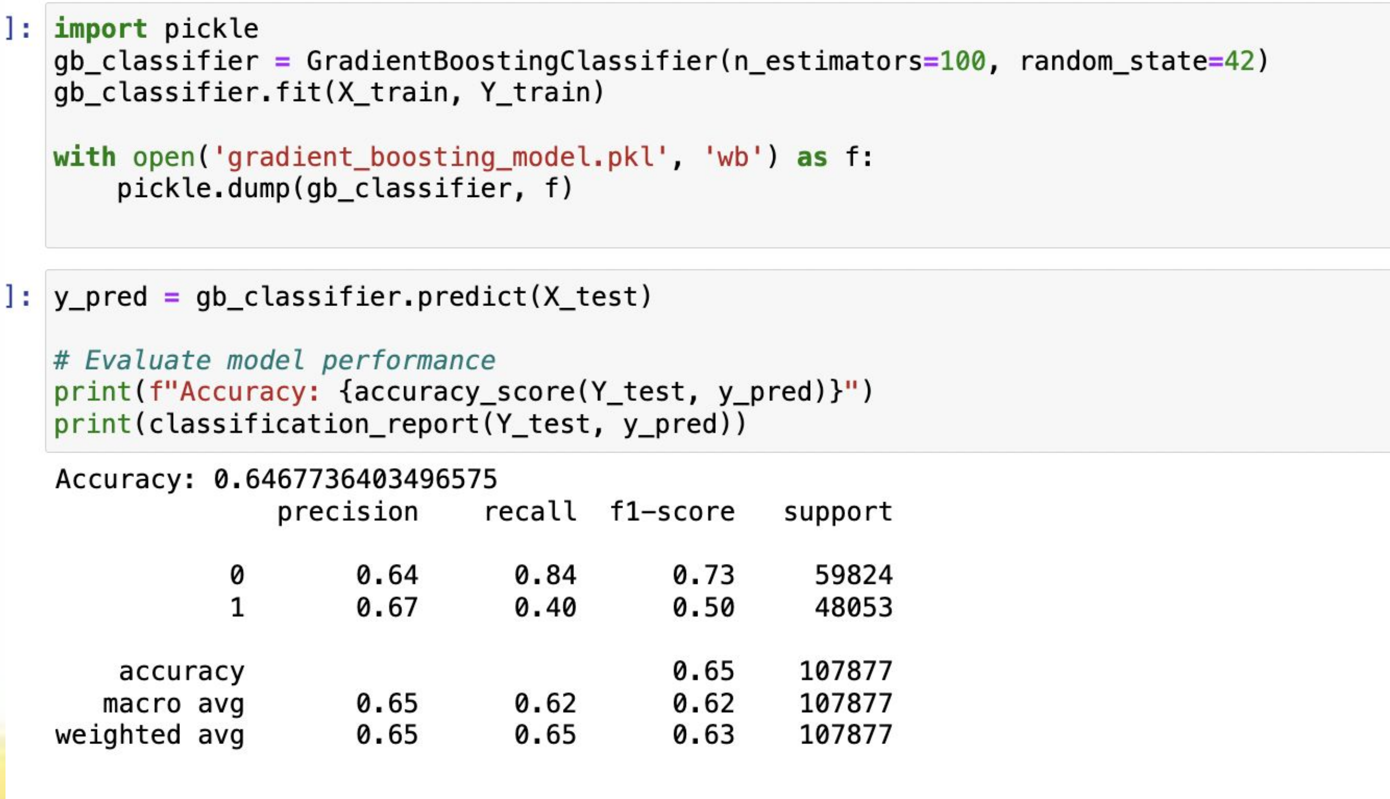
Delayed v/s Non Delayed Flight

Correlation Matrix



	Airline	Flight	AirportFrom	AirportTo	DayOfWeek	Delay	Length_by_hours	Time_by_hour
Airline	1.00	0.10	0.10	0.10	-0.01	0.07	-0.21	0.00
Flight	0.10	1.00	-0.02	-0.02	0.00	-0.05	-0.34	-0.01
AirportFrom	0.10	-0.02	1.00	0.02	0.00	0.02	0.05	-0.03
AirportTo	0.10	-0.02	0.02	1.00	0.00	0.05	0.09	0.03
DayOfWeek	-0.01	0.00	0.00	0.00	1.00	-0.03	0.01	0.00
Delay	0.07	-0.05	0.02	0.05	-0.03	1.00	0.04	0.15
Length_by_hours	-0.21	-0.34	0.05	0.09	0.01	0.04	1.00	-0.02
Time_by_hour	0.00	-0.01	-0.03	0.03	0.00	0.15	-0.02	1.00

Training Model Using Gradient Booster



```
]: import pickle
   gb_classifier = GradientBoostingClassifier(n_estimators=100, random_state=42)
   gb_classifier.fit(X_train, Y_train)

   with open('gradient_boosting_model.pkl', 'wb') as f:
       pickle.dump(gb_classifier, f)
```

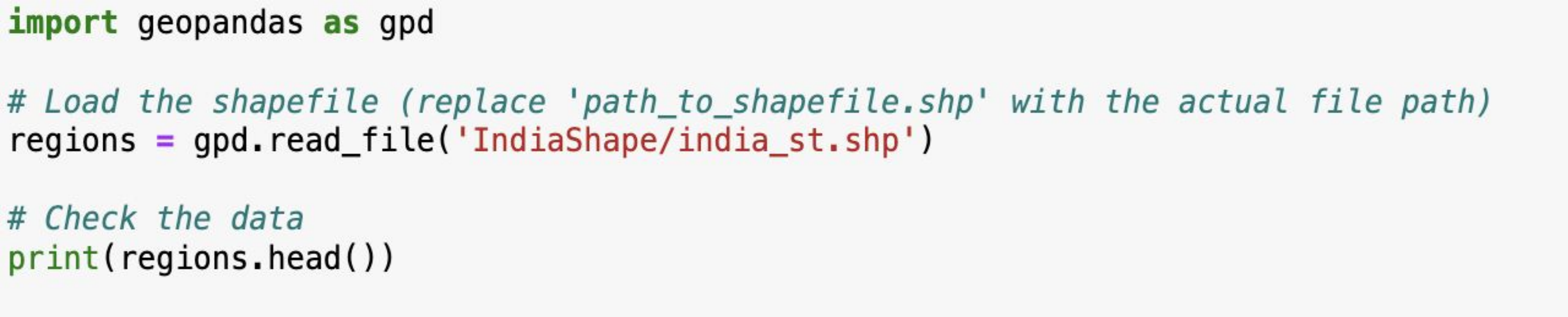
```
]: y_pred = gb_classifier.predict(X_test)

   # Evaluate model performance
   print(f"Accuracy: {accuracy_score(Y_test, y_pred)}")
   print(classification_report(Y_test, y_pred))
```

```
Accuracy: 0.6467736403496575
              precision    recall  f1-score   support

           0       0.64      0.84      0.73     59824
           1       0.67      0.40      0.50     48053

    accuracy                           0.65    107877
   macro avg       0.65      0.62      0.62    107877
weighted avg       0.65      0.65      0.63    107877
```

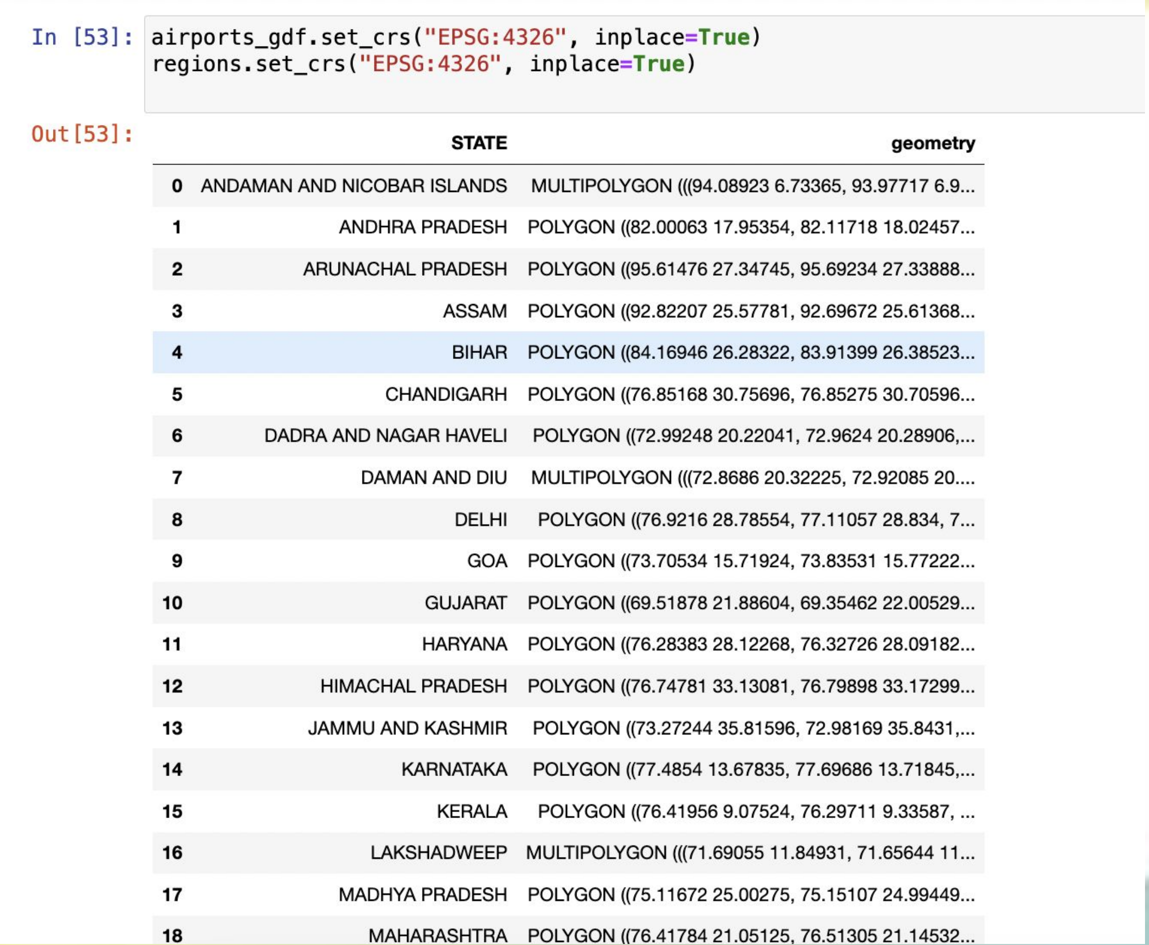
What Is Geopandas?

- **Definition**: GeoPandas is an open-source Python library that simplifies working with **geospatial data** (data with location-based attributes).
- **Purpose**: It extends the capabilities of Pandas, enabling **easy handling** and **analysis** of **spatial data** in a similar way to handling regular tabular data.
- **Key Features**:
  - Integrates **geometry data types** (like points, lines, and polygons).
  - Supports **spatial operations** (e.g., overlay, spatial joins).
  - Works well with **Shapely**, **Fiona**, and **Pyproj** libraries for geospatial data processing.
  - Visualizes spatial data easily using **Matplotlib**.
- **Use Cases**:
  - **Mapping** and **spatial analysis**. Analyzing and **visualizing geographic patterns**.Widely used in fields like **urban planning**, **transportation**, and **environmental science**.

Shapefile Of India



```
import geopandas as gpd

# Load the shapefile (replace 'path_to_shapefile.shp' with the actual file path)
regions = gpd.read_file('IndiaShape/india_st.shp')

# Check the data
print(regions.head())
```

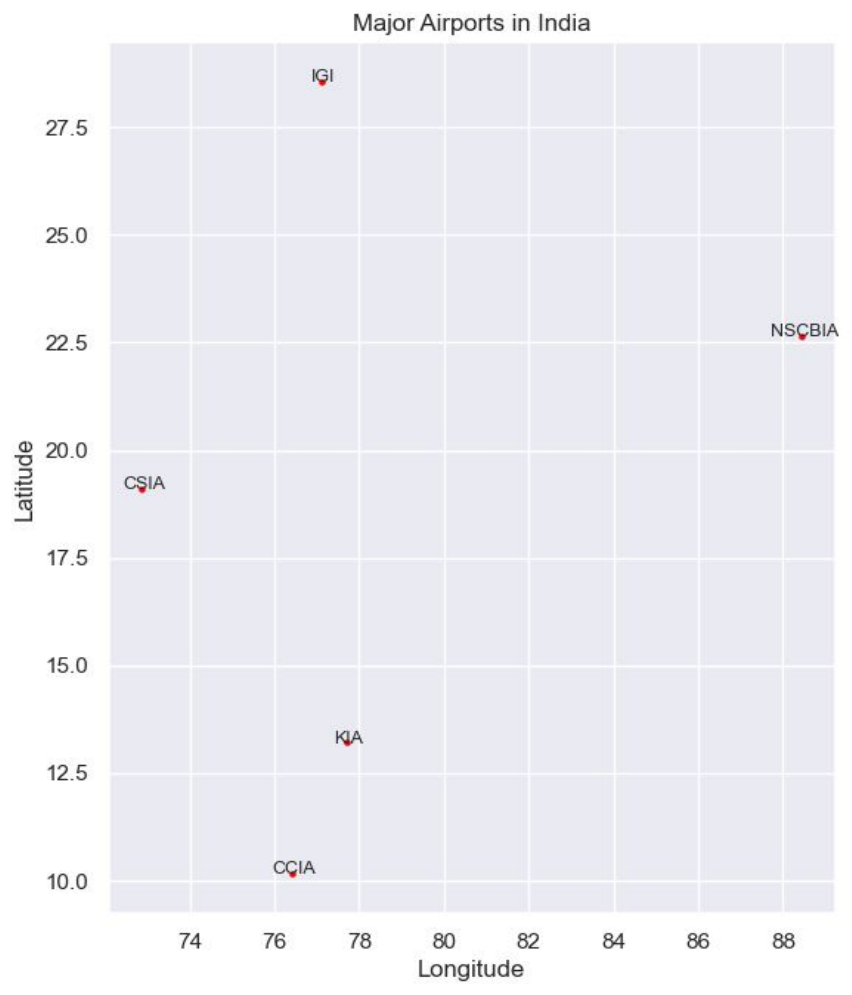


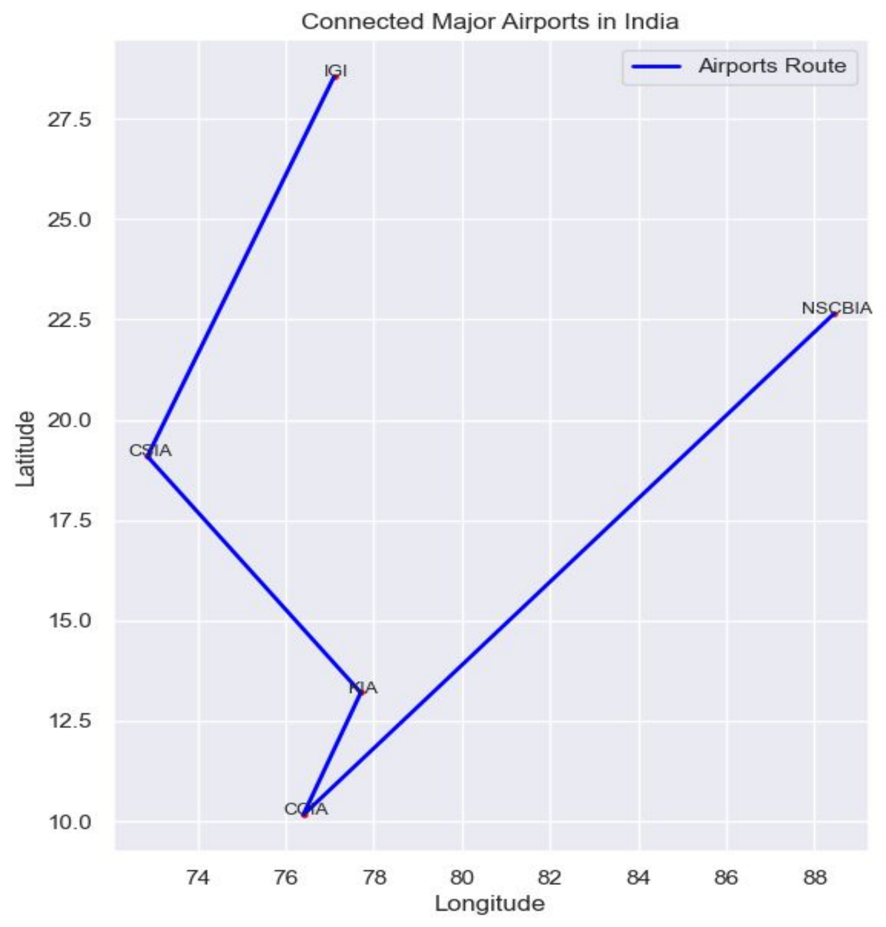
In [53]:
```python
airports_gdf.set_crs("EPSG:4326", inplace=True)
regions.set_crs("EPSG:4326", inplace=True)
```

Out[53]:

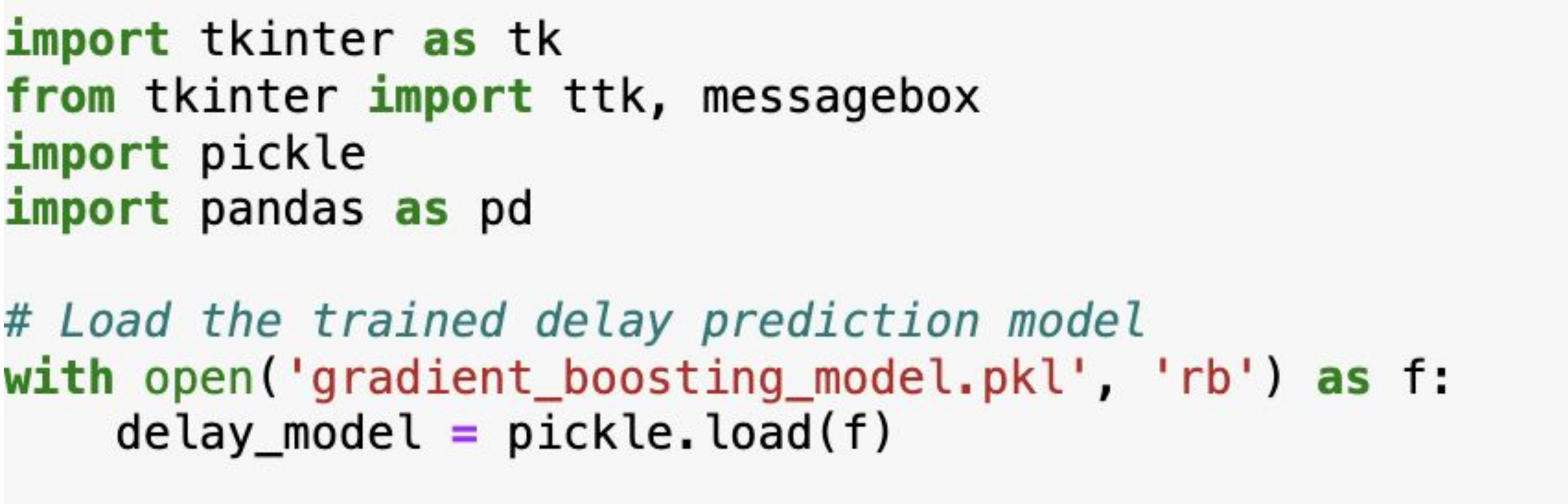
	STATE	geometry
0	ANDAMAN AND NICOBAR ISLANDS	MULTIPOLYGON (((94.08923 6.73365, 93.97717 6.9...
1	ANDHRA PRADESH	POLYGON ((82.00063 17.95354, 82.11718 18.02457...
2	ARUNACHAL PRADESH	POLYGON ((95.61476 27.34745, 95.69234 27.33888...
3	ASSAM	POLYGON ((92.82207 25.57781, 92.69672 25.61368...
4	BIHAR	POLYGON ((84.16946 26.28322, 83.91399 26.38523...
5	CHANDIGARH	POLYGON ((76.85168 30.75696, 76.85275 30.70596...
6	DADRA AND NAGAR HAVELI	POLYGON ((72.99248 20.22041, 72.9624 20.28906,...
7	DAMAN AND DIU	MULTIPOLYGON (((72.8686 20.32225, 72.92085 20....
8	DELHI	POLYGON ((76.9216 28.78554, 77.11057 28.834, 7...
9	GOA	POLYGON ((73.70534 15.71924, 73.83531 15.77222...
10	GUJARAT	POLYGON ((69.51878 21.88604, 69.35462 22.00529...
11	HARYANA	POLYGON ((76.28383 28.12268, 76.32726 28.09182...
12	HIMACHAL PRADESH	POLYGON ((76.74781 33.13081, 76.79898 33.17299...
13	JAMMU AND KASHMIR	POLYGON ((73.27244 35.81596, 72.98169 35.8431,...
14	KARNATAKA	POLYGON ((77.4854 13.67835, 77.69686 13.71845,...
15	KERALA	POLYGON ((76.41956 9.07524, 76.29711 9.33587, ...
16	LAKSHADWEEP	MULTIPOLYGON (((71.69055 11.84931, 71.65644 11...
17	MADHYA PRADESH	POLYGON ((75.11672 25.00275, 75.15107 24.99449...
18	MAHARASHTRA	POLYGON ((76.41784 21.05125, 76.51305 21.14532...

Airport Mapped By GeoPandas



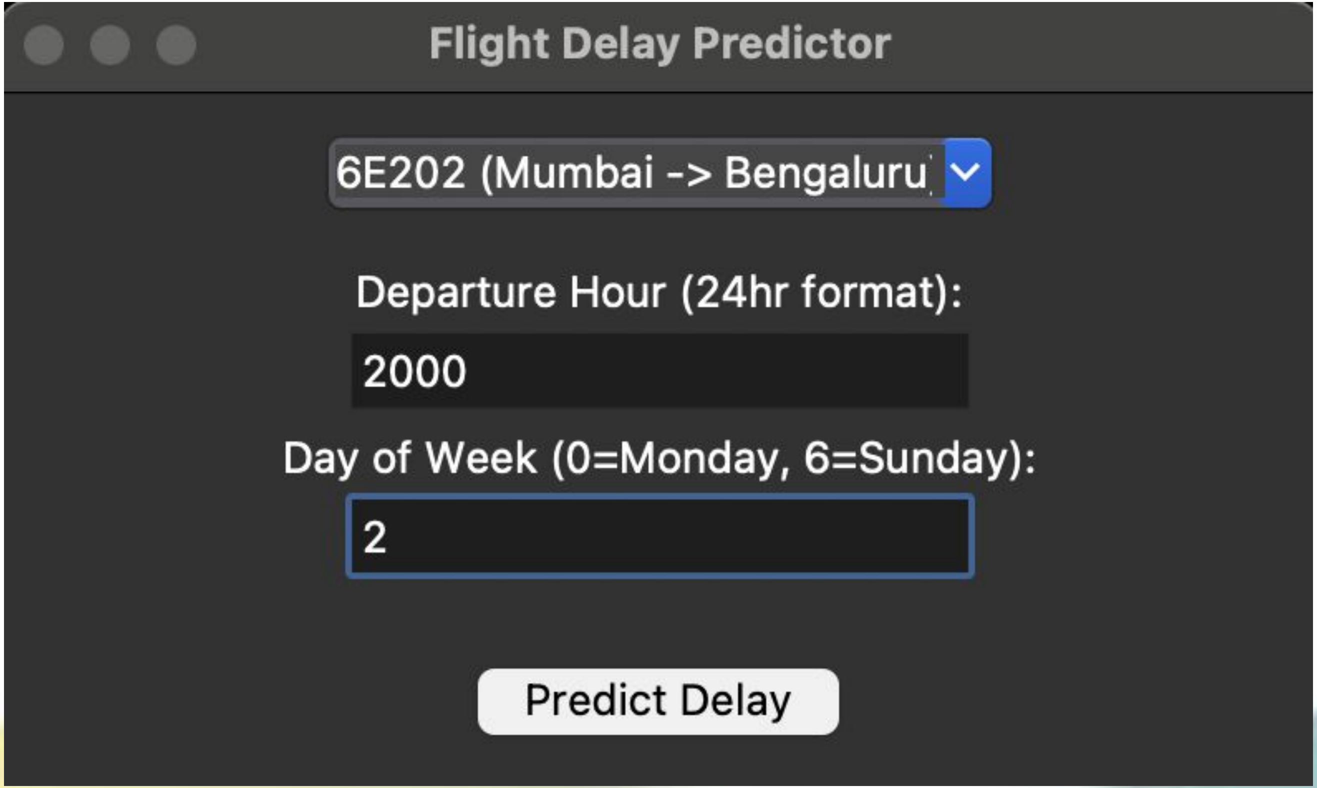

Connected Major Airports in India
Airports Route
IGI
NSCBIA
CSIA
KIA
CCIA
Latitude
27.5
25.0
22.5
20.0
17.5
15.0
12.5
10.0
Longitude
74
76
78
80
82
84
86
88

Loaded Model in Tkinter



```
import tkinter as tk
from tkinter import ttk, messagebox
import pickle
import pandas as pd

# Load the trained delay prediction model
with open('gradient_boosting_model.pkl', 'rb') as f:
    delay_model = pickle.load(f)
```

User InterFace



Flight Delay Predictor

6E202 (Mumbai -> Bengaluru

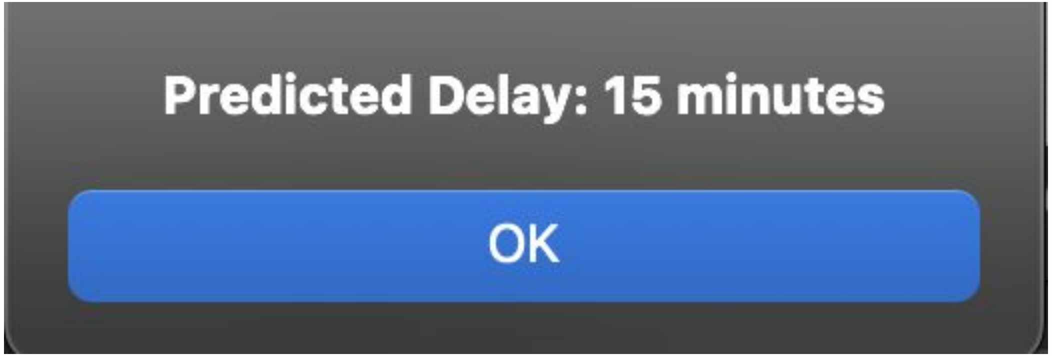
Departure Hour (24hr format):

2000

Day of Week (0=Monday, 6=Sunday):

2

OUTPUT



Predicted Delay: 15 minutes

References

- Indian Flight Dataset : https://github.com/OludolapoAnalyst/Indian_Flight_Data
- GeoPandas Documentation : https://geopandas.org/en/stable/docs.html
- India Shapefile : https://www.indiaremotesensing.com/2017/01/download-india-shapefile-with-official.html