Analysing The Air Traffic Trend Based On Passenger And Landing Count In United States

1st Chandana Haluvarthi Prabhudeva
MSc in Data Analytics
National College of Ireland
Dublin
X22167099@student.ncirl.ie

2nd Pratheek Gogate
MSc in Data Analytics
National College of Ireland
Dublin
x22159789@student.ncirl.ie

***Abstract*—This study seeks a better understanding of the increase in air travel and the number of aircraft landings in relation to the operating airline, passenger count, and landing count, among other fields, for the years 2005 and after. ETL, modelling and visualization techniques are used to check the growing airline industry; the two datasets are driven from the United States travellers' details taken from the US government which is an open source.**

***Index Terms*—Landing count,Passenger Count**

I. Introduction

People nowadays choose airways for travelling for business and pleasure as it is the fastest means to reach the other end of the world, this, in turn, made air transportation the most significant worldwide. This has increased the number of flights, passenger counts, and aircraft landings, which has introduced a few difficulties to the airlines. This project focuses on comprehending and examining two of the difficulties airlines currently face: Airport traffic and Safety and Maintenance. In order to maintain the resources such as runways, gates and ground personnel, the airlines need to know the landing count to allocate the areas and people for the tasks. Sometimes runway traffic is at its peak, it creates a tight schedule, and it's very difficult to manage resource allocation, people are not pleased if their flights get cancelled due to the poor resource and manpower. Passengers are not cooperative if the flight is overloaded and doesn't comply with safety regulations. Planning and optimizing the airways' operations are essential as it determines the pricing strategies. To understand these growing issues, we have taken the dataset related to the passenger details and flight details with the passenger count and landing count from the US government portal. Two datasets that are related to passenger details and flight details are used in this project. The passenger details dataset is in unstructured JSON format, and the flight details dataset is in CSV format. Both datasets are initially loaded into the NoSQL MongoDB. Then they are reloaded into Postgres SQL after cleaning the dataset, which is used for Python libraries to visualize the highest passenger count for the different airlines, the landing count, and a few other factors. A linear regression algorithm is used to check the landing count prediction based on the factors available.

II. Related Works

Paper [1] shows a simple method to find and measure the impact of advertised departure delays on arrival delays. A propagated departure delay can be defined as an event in which the arrival delay of an incoming flight exceeds the ground buffer of the next flight. This paper focuses on individual flights and their immediately preceding flight and does not count or focus on the cause of delay propagation through the entire sequence of previous flights. Results from the studies in this paper show that several important factors cause the impact of advertised departure delays on arrival delays.

Paper [2] tells how to use the state-of-the-art SLAM algorithm, ORB-SLAM2, for 6DoF position estimation of aircraft while doing manual landing when modern navigation systems are unavailable. The proposed algorithm is used to help the pilot with additional information about the state of the aircraft during landing, helping improve safety. ORB-SLAM2's performance is measured using a simulated runway scene generated with Unreal Engine and an onboard stereo camera. The results showed that ORB-SLAM2 can issue warning messages to the pilot when the aircraft's descent rate or glideslope is incorrect, thereby improving safety during landing. Additionally, the algorithm is lightweight, real-time, and self-contained, making it vest suitable for using the same in other UAVs with limited capabilities.

Paper [3] shows an Airport Carrying Capacity Forecast (AECC) methodology for planning and building appropriate airport development modes. The method used in this system was a dynamics model based on the Driving Force-Pressure-State-Response (DPSR) framework and selected 17 key variables from different dimensions. AECC predictions are determined by system dynamics model simulations and AGA-PP (Accelerated Genetic Algorithm Projection Pursuit) model calculations. It is necessary to analyze the appropriate development mode of airports in different cases. A case study of the Pearl River Delta Airport suggested that the AECC forecasting method based on the SD model and the AGA-PP model can provide decision support to relevant departments of the airport.

In this article [4], they investigated passenger flow at an airport terminal and built a model to predict its distribution

based on flight arrangements. The study found peak passenger count in different areas of the terminals at 30-minute intervals and varied at 60-80 minute intervals. RD values are used to explain this peak-shifting feature, with typical overall peak RD values of 0.6 to 0.8. The RD had its peak of 0.2 during the COVID-19 pandemic, which clearly reflected a reduced passenger count. This study provided useful information about airport terminal passenger flows for the design and functions of airport terminals.

Paper [5] used a hybrid evaluation method to evaluate five passenger throughput (APT) forecast models at 203 different airports across China. The model is fitted using historical data and validated using data from 2015 to 2019. The study showed that airports with higher APTs fitted the models and generally performed better. Also found that Complexity was not directly related to accuracy. Time series models, causal models, market share methods, and analogy-based methods could effectively predict the APT in 88 percent of the airports studied if the parameters were set correctly. However, insufficient historical data and external forces such as extension, displacement, and earthquakes can affect accuracy.

Paper [6] describes a probabilistic prediction model that uses DASHlink data to numerically forecast the vertical velocity of an aircraft during the time of landing. This model used a Bayesian neural network approach to quantify uncertainty and support risk-conscious decision-making. The methodology consists of five steps: clustering, touchdown point identification, data smoothing, input variable dimensionality reduction, and Bayesian recurrent neural network training. The model was validated against flight test datasets and observed a satisfactory performance.

The study in paper [7] was performed to assess pilot performance during the landing phase using flight high-speed access recorded data, using three landing parameters as indicators for pilot evaluation; a performance evaluation model for landing maneuvers was developed based on risk assessment principles. Based on this model, they constructed an in-flight landing performance evaluation system and demonstrated that it can accurately evaluate the pilot's landing performance. The study concluded that this method could be a practical tool for airlines to manage landing risk and improve airline pilots' training and design to avoid future accidents.

Paper [8] describes the identification of an indirect link between flight data and the risk of runway overrun accidents. The authors proposed a data-driven approach that used data analytics techniques and machine learning tools to classify flights as safe or vulnerable. An augmented tree classifier was trained to classify flights accurately, and the contributing factors were extracted. The analysis showed the weightage of certain factors and led to new insights into possible approaches to flight safety. This study highlighted the importance of understanding in-flight data collection to reduce flight runway accidents.

Paper [9] presents a simple approach to predict delays in the US air transportation system using passenger-centric data from Twitter. The study showed that including this data improves prediction accuracy with flight-centric data. Researchers have developed more efficient and accurate predictive models by analyzing the importance of different traits. Passenger-centric data gives a better understanding of the air transportation system as a whole and is more accessible than flight-centric data. Further research could refine this approach and extend it to other regions as well where flight-centric data are also available.

III. METHODOLOGY

A. DATASET DESCRIPTION

Below 2 data-sets were be used in the paper: These datasets are taken from US Government data website.

1) Dataset 1: – Air Traffic Passenger Statistics

This includes very important attributes like, Operating time, Active Airline, Active Airline under IATA Code, advertised airline, advertised airline IATA code, Geographic Summary, Geographic Area, and Type of Business. Code, Fare Type Code, Terminal, and Boarding Area basically contribute to the number of passengers. This data is read from the API in JSON format using the Python library. This has 12 columns and 50730 rows of data which one can study and better understand using visualization and understanding certain patterns.

2) Dataset 2: - Air Traffic Landings Statistics

This dataset includes aircraft landing information, including total landing weight, area and geographic summary. In addition, it contains information about the aircraft, such as model, manufacturer, body type and version. This one too includes information on uptime, activities and airline marketing. It contains approximately 57381 data records. The data is read in Python from a file in CSV format. With this data, we can get a lot of useful information regarding flight landings based on different aircraft attributes and geographic information.

B. DETAILED DESCRIPTION OF DATA PROCESSING ALGORITHMS

The two datasets that are chosen for the study in this analysis are taken from the US government's open-source website. The passenger details dataset is taken from the Application Program Interface (API) and stored in a Python data frame. While the Flight details dataset is taken as the CSV file(comma-separated file) and is stored in the Python panda library.

1) Extract: Data is read and then stored in a Python data frame and transferred into the NoSQL database MongoDB using a flexible schema to store it. There is no need to create a separate schema as we do in the other databases. It stores the data in document format, which makes it more flexible and dynamic. The data in MongoDB is stored in Binary JSON(BSON) format, a binary representation of the data. It can be retrieved back in JSON format. It can store either structured or unstructured data. We can connect to MongoDB with the help of py libraries with the local host and port 27017. We create a collection for each dataset.

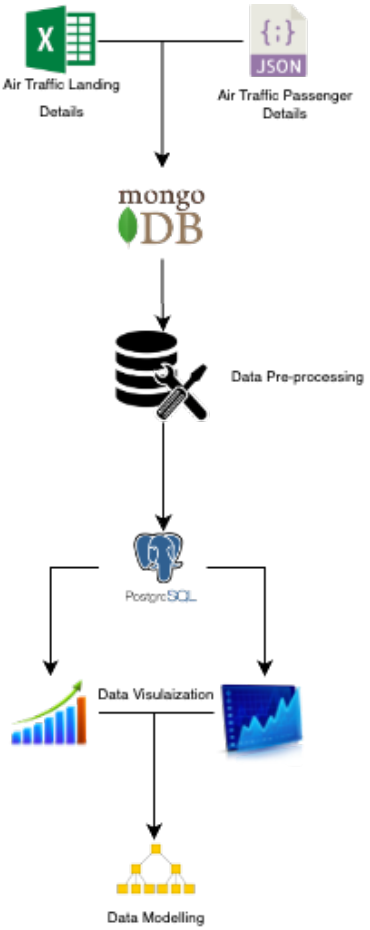

Fig. 1. Data Flow

2) Transform: The data is retrieved from MongoDB in Json format and stored in the Python data frame in a structured format. This data is transformed and changed according to the analysis that we need to do. Data – Preprocessing is important as it can improve the reliability of the data and remove missing, incorrect data and unimportant data.

The data - preprocessing steps done for our data are as below:

- The null values in the data add more complexity and disrupt the performance as they are missing values. They do not add any value to the analysis. They make the analysis more complex as there is nothing to fit in that place, so these values are replaced with NaN values for the string, and we replace them with zero for the numeric data. If there are duplicate data in the fields, they are removed.
- During the analysis, only a few fields are required to get the complete analysis few fields that are provided by the data sources are unnecessary, such fields are removed in the initial steps to reduce the unnecessary processing steps. In our dataset we have removed OperatingAirlineIATACode, PublishedAirlineIATACode and PublishedAirline as they are just codes than the real names of the airlines.
- The duplicate records in the dataset have to be removed as they lead to misleading results. NaN values have to be removed as they are not useful.
- We have split the date field into month and year so that the visualization can be done in a better way. We have aggregated the fields if necessary to improve the data and its usage of it. To combine the data into a single field by aggregation.

3) Load: The data that has been converted will be stored in PostgreSQL, which is often called Postgres, which is an object-relational database management system. It is a flexible relational database which is open-source. Datasets are stored in two different tables in a structured format.

IV. RESULTS AND EVALUATION

A. The flight details from the Postgres is used to visulaize

Fig. 2. Trend of total landed weight over time

The line graph plotted in Figure 2 for TotalWeightlanded per year essentially helps us understand flight usage trends. So, in the chart, we can observe a sudden increase in 2006, and there was steady growth until 2019 when it was at its peak. Covid 19 has slowed down progress significantly, as we can see a strong downtrend from 2019, although starting to improve in 2021 but not peaking like in 2019.

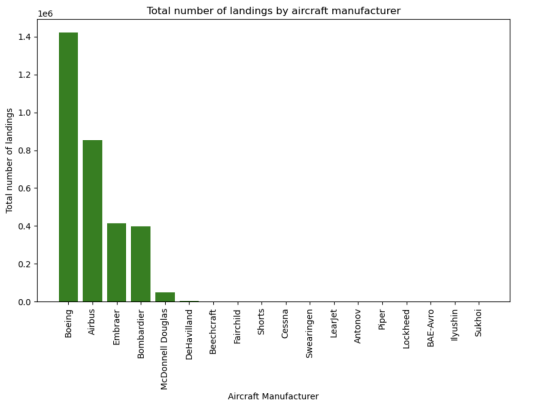

Fig. 3. Total number of landings by aircraft manufacturer

The bar graph drawn in Figure 3 shows aircraft manufacturers by the number of landings, essentially helping us to identify the main manufacturers. So looking at the numbers, we can see that Boeing is the boss of the plane industry, and Airbus, Embraer and Bombardier are close competitors. These four dominate the industry, with all the others not contributing much.

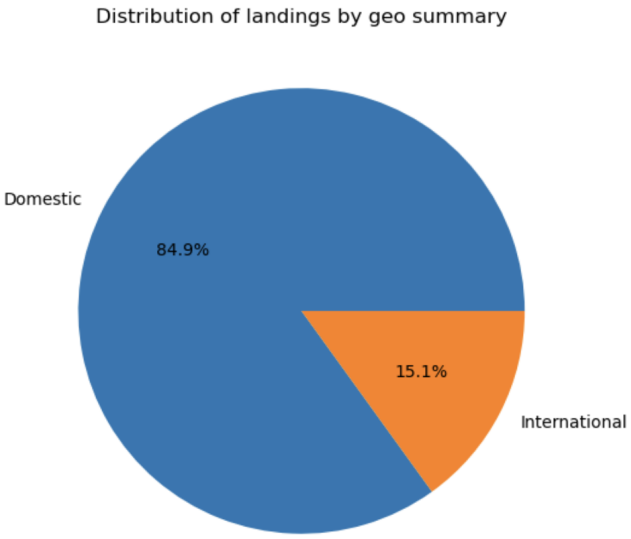

Fig. 4. Distribution of landings by geo summary

The pie chart above shows that the pie chart is based on whether the landed flight is international or domestic. So from Figure 4, we can easily say that Francisco Airport handles more domestic flights than international flights, with about 85 perect of domestic flights.

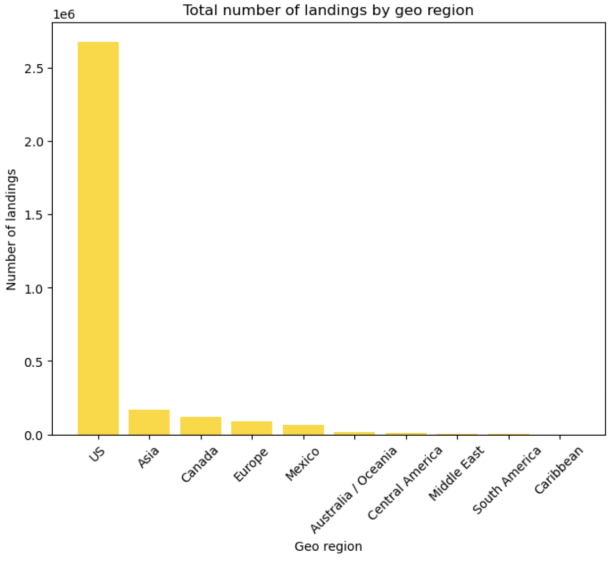

Fig. 5. Total number of landings by Geo Region

We've drawn a bar graph in Figure 5 above, where the number of landings is based on different regions, we already know that Francisco is more of a domestic airport than an international airport, and hence the US is the one with the highest number of flights, Asia and Canada being next in the list and all others followed them, with the Caribbean being the region with the lowest number of flights landed.

B. The passenger details from the Postgres is used to visulaize

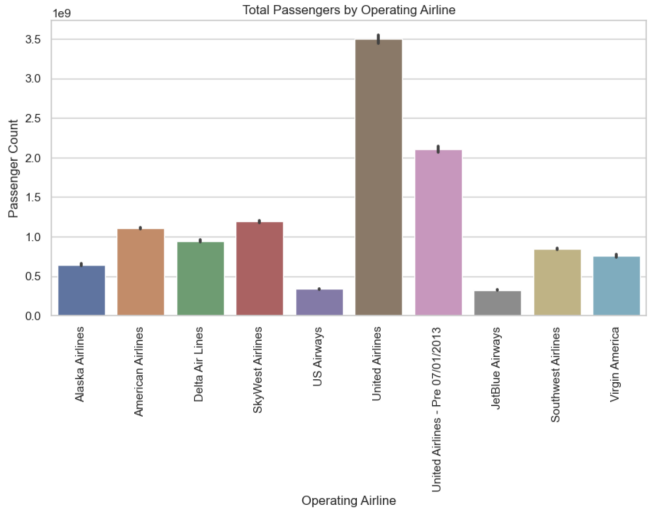

Fig. 6. Total Passengers by Operating Airline

The above barplot is to find out the best Airlines taking passenger count as the performance metric. So in figure 6, we can observe that United Airlines is the best considering it has the highest number passenger count by fair margin among its other competitors, and Skywest Airlines, American Airlines, and Delta Airlines are in the next places respectively. Here we have plotted for Top 10 Airlines based on above mentioned condition.

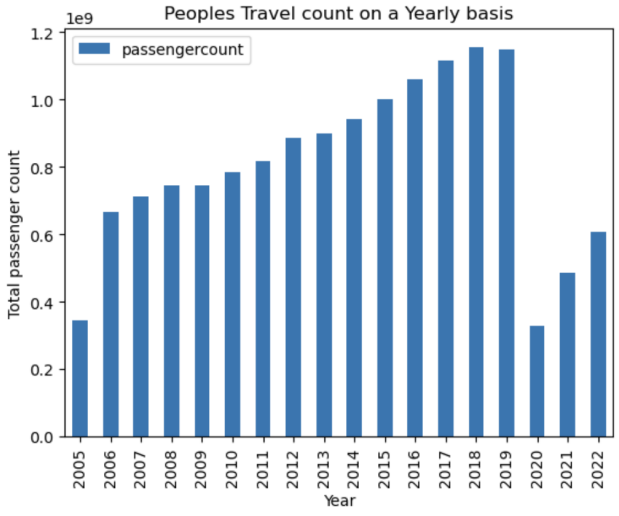

Fig. 7. People's Travel count on a Yearly basis

The objective of this Barplot was to find people's trends when it comes to travelling in that particular airport. So we plotted Passenger counts against the year in Figure 7, and we could observe that during 2005 flight travel was not so trendy, but after that, it suddenly picked pace in 2006, and from there onwards, there was a study growth in the trend till 2019. Then as we all know, the 2019 end was horrendous due to COVID-19, and it had a huge impact on passenger travel we could see a sudden collapse in 2020, and now, after the travel restrictions were lifted again, there is a constant upward trend.

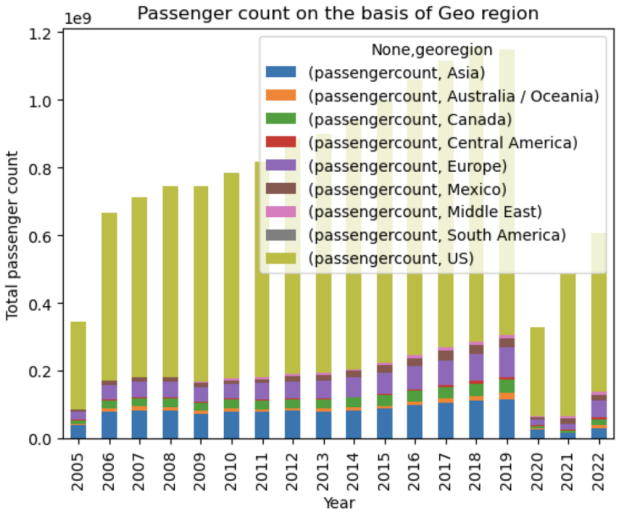

Fig. 8. Passenger count on the basis of Geo region

The AIM of this group bar chart was to show the passenger count on the basis of different regions each year. So when we plotted this in Figure 8, There was no surprise for the first place as it was the US for all the years with the highest number of passengers and its because of the fact that the airport which we were analyzing was from the US and through our analysis we already know that more domestic flights are running here than international. So Asia was second on the list, followed by Europe and other regions, with Australia being the lowest.

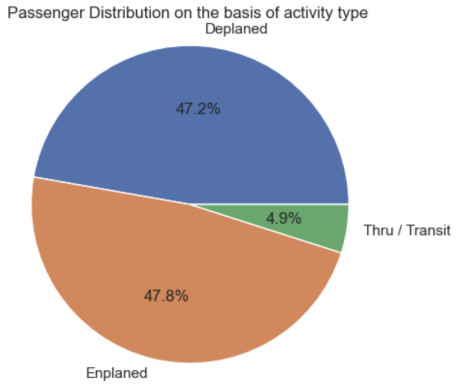

Fig. 9. Passenger Distribution on the basis of activity type

This Pie chart was drawn to visualize and take insights regarding activity types based on passenger count so looking at Figure 9, we can see that around 5 percent of passengers are entering the airport because of transit which basically means they will be waiting for their connecting flights and remaining 95 percent is all most equally distributed with around 47 percent between passengers Deplanning and Enplanning which is nothing but people going out of the airport after the journey and people coming to the airport to start a journey.

C. The combined data of passenger and flight from the Postgres is used to visulaize

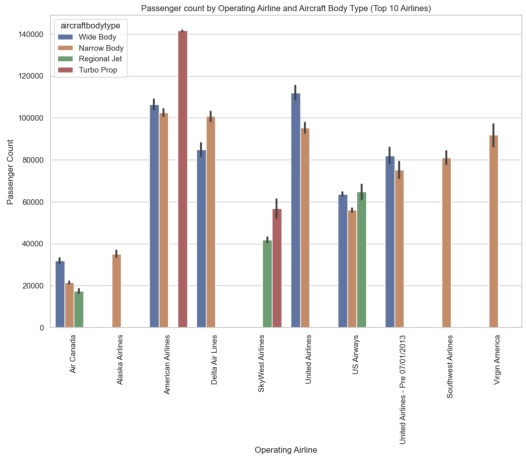

Fig. 10. Passenger count by Operating Airline and Aircraft Body Type (Top 10 Airlines)

The top 10 Airlines have been plotted with body types based on the Number of Passengers, In Figure 10, we can see a Barplot which actually tells which Airline the maximum number of people have travelled based on the type of flight. So from the figure, we can say that in American Airlines the maximum number of people have travelled and in that people, preferred Turbo Prop body type the most and Regional Jet type should have been more addressed. Looking at the graph, we can also say that Most of the Airlines use Narrow bodies as we can see them in all the Airlines in the graph, and Regional Jet is the least-used body type among airlines.

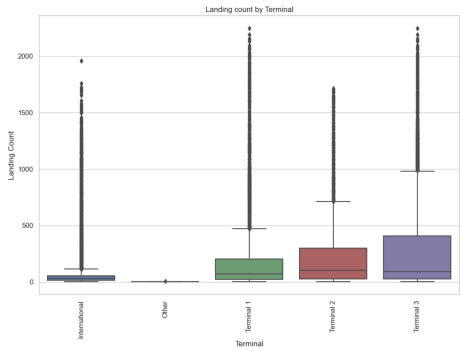

Fig. 11. Landing count by Terminal

In Figure 11, we have plotted the Landing count based on Terminals based on our combined dataset, and we can clearly say that Terminal 3 has the highest number of flights and Terminal 2 and Terminal 1 in the next places. We could also observe that most are skewed toward the graph's upper part, indicating that the values fall in the third quartile. Also, we could see comparatively fewer landings in the International terminal than in other terminals, indicating that more domestic flights are being landed at that airport.

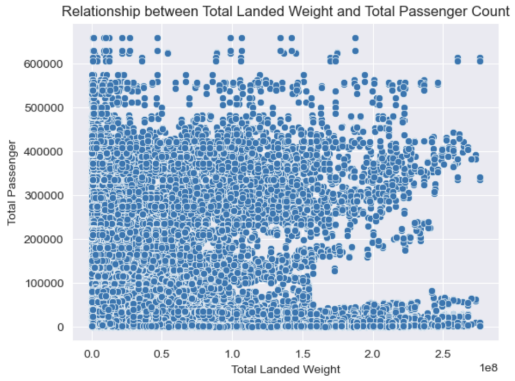

Fig. 12. Relationship between Total Landed Weight and Total Passenger Count

A scatter plot is usually plotted to understand the relationship between 2 columns. Hence, here we plotted the Total landed weight against the passenger count in Figure 12 domain knowledge. We expect the Total weight should depend on the Total passenger count assuming everybody carries luggage. Still, when we plotted the graph, no such relationship was found as the values are scattered throughout the graph, indicating no linear relationship.

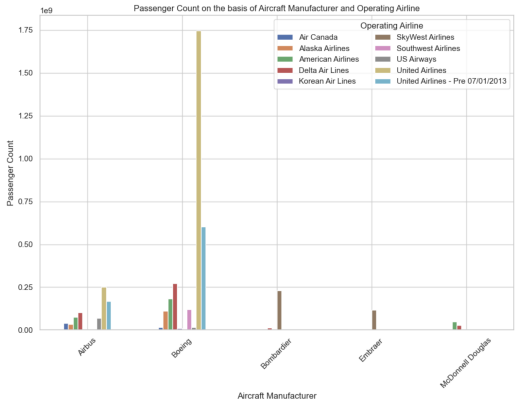

Fig. 13. Passenger Count on the basis of Aircraft Manufacturer and Operating Airline

The Objective of the above Graph, Figure 12, is to find from which Aircraft Manufacturer more passengers are travelling per airline. So looking at the graph, we can observe that Boeing and Airbus are the ones where maximum people use it and if look at them carefully we could say United Airlines by Flights from Boeing and Airbus only. Whereas Skywest Airlines usually buys from Bombardier and Embraer, American Airlines buys from Boeing and Airbus, and all the other airlines have very few passengers.

V. MODEL BUILDING

```
In [32]: # Import necessary libraries
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score

# Load data into a Pandas DataFrame
data = Combined_fetched_data

# Define the features and target variable
X = data[['passengercount', 'year', 'month', 'totallandedweight']]
y = data['landingcount']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Instantiate a decision tree classifier and fit the model
clf = DecisionTreeClassifier()
clf.fit(X_train, y_train)

# Make predictions on the test set
y_pred = clf.predict(X_test)

# Evaluate the accuracy of the model
accuracy = accuracy_score(y_test, y_pred)
print('Accuracy:', accuracy)

Accuracy: 0.9488973574061348
```

Fig. 14. Landing Prediction

```
In [36]: # Import necessary libraries
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score

# Load data into a Pandas DataFrame
data = Combined_fetched_data

# Define the features and target variable
X = data[['landingcount', 'year', 'month', 'totallandedweight']]
y = data['passengercount']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Instantiate a decision tree classifier and fit the model
clf = DecisionTreeClassifier()
clf.fit(X_train, y_train)

# Make predictions on the test set
y_pred = clf.predict(X_test)

# Evaluate the accuracy of the model
accuracy = accuracy_score(y_test, y_pred)
print('Accuracy:', accuracy)

Accuracy: 0.0042481414380120822
```

Fig. 15. Passenger Count Prediction

A decision tree classifier is a machine-learning algorithm which recursively partitions the data into subsets based on the input feature and assigns each subset to a class label. The passenger count and landing count predictions are calculated in this analysis, the landing count is predicted with an accuracy of 94 percent, and we have got to know that the passenger count data is not suitable for this type of prediction and requires additional data to do so.

VI. CONCLUSION

The analysis of this project shows that the passenger count has increased over the years but has decreased during the pandemic. Still, the passenger count is highly dependent on the airlines and their manufacturer. The landing count has also increased over the years in the major regions and depended on aircraft type and model; the visualizations helped back this result. The study also helps to predict the landing count using Decision Tree with an accuracy of 94 percent, which can be used for airline improvement and airport management.

REFERENCES

[1] Airline delay propagation: A simple method for measuring its extent and determinants Author links open overlay panel Jan K. Brueckner a, Achim I. Czerny b, Alberto A. Gaggero c in 2022 https://www.sciencedirect.com/science/article/pii/S0191261522000741

[2] Evaluation of ORB-SLAM based Stereo Vision for the Aircraft Landing Status Detection Chao-Chung Peng; Rong He; Chin-Sheng Chuang in 2022 https://ieeexplore.ieee.org/abstract/document/9969111

[3] A System Dynamics Prediction Model of Airport Environmental Carrying Capacity: Airport Development Mode Planning and Case Study by Qiuping Peng,Lili WanORCID,Tianci Zhang,Zhan Wang *ORCID andYong Tian in december 14 2021. https://www.mdpi.com/2226-4310/8/12/397

[4] A prediction model to forecast passenger flow based on flight arrangement in airport terminals Author links open overlay panelLin Lin a 1, Xiaochen Liu a 1, Xiaohua Liu a, Tao Zhang a, Yang Cao b in 17 jan 2022 https://www.sciencedirect.com/science/article/pii/S2666123322000423

[5] Evaluating Prediction Models for Airport Passenger Throughput Using a Hybrid Method by Bin Chen 1,2ORCID,Xing Zhao 1,* andJin Wu 1, on 8 dec 2022 https://www.mdpi.com/2076-3417/13/4/2384

[6] Bayesian Deep Learning for Aircraft Hard Landing Safety Assessment Yingxiao Kong; Xiaoge Zhang; Sankaran Mahadevan on 7 april 2022 https://ieeexplore.ieee.org/abstract/document/9751234

[7] A Method of Applying Flight Data to Evaluate Landing Operation Performance Lei Wang,Jingyi Zhang,Chuanting Dong,Hui Sun and Yong Ren in 2018 october https://www.tandfonline.com/doi/abs/10.1080/00140139.2018.1502806

[8] A data-driven model for safety risk identification from flight data analysis Author links open overlay panelMickael Rey a, Daniel Aloise b c, François Soumis b c, Romanic Pieugueu in september 2021 https://www.sciencedirect.com/science/article/pii/S2666691X21000439

[9] Predicting and Analyzing US Air Traffic Delays using Passenger-centric Data-sources Philippe Monmousseau, Daniel Delahaye, Aude Marzuoli, Eric Féron on 13 jul 2019 https://hal-enac.archives-ouvertes.fr/hal-02178441/document

[10] https://catalog.data.gov/dataset/air-traffic-passenger-statistics

[11] https://catalog.data.gov/dataset/air-traffic-landings-statistics