**DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING (PYTHON PROJECT)**

**PROJECT REPORT**

(Project Semester January-April 2025)

***(Data Analysis and Visualization using Pandas, Matplotlib and Seaborn)***

Submitted by

Priyanka Kumari

Registration No - 12307499

Programme and Section – Btech(CSE) and K23FA

Course Code – INT 375

Under the Guidance of

**Ms. Sandeep Kaur (UID : 23614)**

**Discipline of CSE/IT**

**Lovely School of Computer Science and Engineering**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that **Priyanka Kumari** bearing Registration No. **12307499** has completed **INT375** project titled **“Data Analysis and Visualization using Pandas, Matplotlib and Seaborn”** under my guidance and supervision. To the best of my knowledge, the present work is the result of her original development, effort, and study.

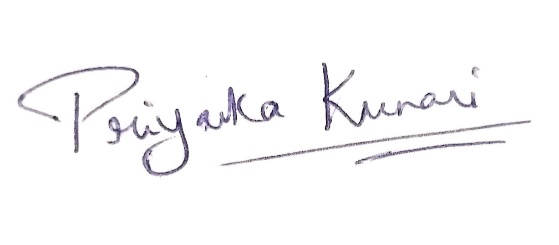
**Ms. Sandeep Kaur**  
**Assistant Professor**  
**School of Computer Science & Engineering**

**Lovely Professional University**  
**Phagwara, Punjab**

Date: **11-April-2025**

**DECLARATION**

I, **Aryan Gupta**, student of **Bachelors of Technology (B.Tech)** under CSE/IT Discipline at Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 11 – April - 2025 Signature –

Registration No.12307499

Name of the student – Priyanka Kumari

**ACKNOWLEDGEMENT**

I would like to express my sincere gratitude to **Ms. Sandeep Kaur Mam**, my project guide, for their invaluable support, guidance, and encouragement throughout the development of this project. Their expert insights and constructive feedback have been instrumental in shaping the project's outcome.

I am also thankful to **Lovely Professional University** for providing a conducive learning environment and access to resources that made this project possible. Additionally, I extend my appreciation to my professors and peers for their continuous motivation and insightful discussions, which greatly enhanced my understanding of the subject.

Lastly, I would like to acknowledge the unwavering support of my family and friends, whose encouragement has been a source of inspiration throughout this journey.

**Table of Contents**

1. Introduction

2. Source of dataset

3. Dataset Preprocessing

4. Analysis on dataset (for each objective)

1. General Description
2. Specific Requirements
3. Analysis results
4. Visualization

5. Conclusion

6. Future scope

7. References

**Introduction**

In the aviation industry, understanding the causes and patterns of flight delays is crucial for improving customer satisfaction and operational efficiency. This project, undertaken as a part of academic learning in the second year of B.Tech in Computer Science and Engineering, focuses on analyzing flight delay data using Python and Excel-based Exploratory Data Analysis (EDA) techniques. The dataset used for this project includes detailed information about various delay causes such as weather, security, carrier, NAS (National Aviation System), and late aircraft delays, along with flight and airport-specific data.

The goal of this analysis is to identify key patterns, correlations, and impactful delay causes across different airlines and airports. The project includes visualization techniques like pie charts, bar graphs, line charts, and heatmaps to make the findings more interpretable and insightful. By conducting this analysis, we aim to develop a foundational understanding of data handling, processing, and visualization using Python libraries such as Pandas, Matplotlib, and Seaborn.

This hands-on project not only enhances data analysis skills but also gives real-world insights into airline operations and the impact of delays, offering valuable learning and practical exposure to data-driven decision-making.

**Source of dataset**

The dataset used in this analysis is sourced from the U.S. Department of Transportation’s (DOT) Bureau of Transportation Statistics (BTS). This organization monitors the on-time performance of domestic flights operated by major U.S. airlines.

BTS provides monthly data through the Air Travel Consumer Report, which includes statistics on:

* On-time arrivals and departures
* Delays categorized by cause (e.g., weather, carrier, NAS)
* Cancellations and diversions

Since June2003, the BTS has made both summary statistics and detailed raw data available to the public. The information used in this project is publicly accessible and was published on the BTS website at the time of release.

Link of the dataset –

<https://transtats.bts.gov/OT_Delay/>

**Dataset Preprocessing**

Dataset preprocessing is a crucial step in any data analysis project. It ensures that the data is clean, consistent, and ready for further analysis and visualization. In this project, where the dataset contains air pollution data across various cities, several preprocessing techniques were applied to ensure the quality and usability of the dataset.

**1. Importing the Dataset**

The dataset was imported using Pandas, a Python library that makes it easy to work with tables (dataframes).

**2. Understanding the Columns**

The dataset included information like:

* Airline name (carrier\_name)
* Airport name (airport\_name)
* Number of arrival flights (arr\_flights)
* Delay details by cause (carrier, weather, NAS, etc.)
* Total arrival delay (arr\_delay)

#### 3. **Checking for Missing Values**

We checked if there were any empty (NaN) values in the dataset:

df.isnull().sum()

**Action Taken:** Most values were complete, and rows with critical missing data were either dropped or filled if needed.

**4. Renaming Columns**

To make it easier to understand, some column names were simplified

**5. Creating New Columns for Analysis**

Some new columns were created to support our analysis:

* Total Delay: Sum of all delay types for each row
* Average Delay per Flight

**6. Data Validation**

After preprocessing, the dataset was validated to ensure that all issues had been addressed. A summary of the dataset was reviewed using descriptive statistics and visualizations to confirm the correctness of preprocessing steps and the readiness of the dataset for further exploration.

**Analysis on dataset**

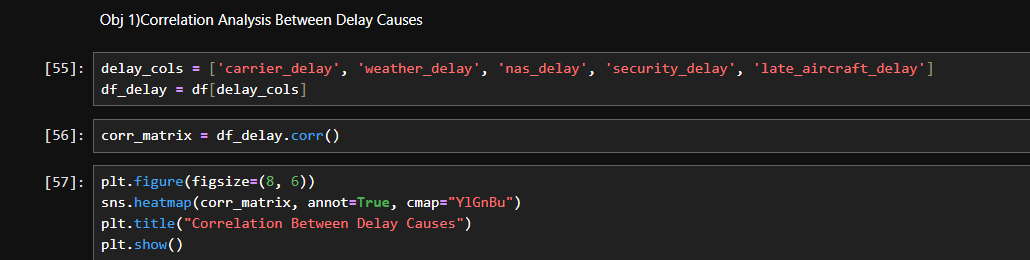
**1. Correlation Analysis Between Delay Causes**

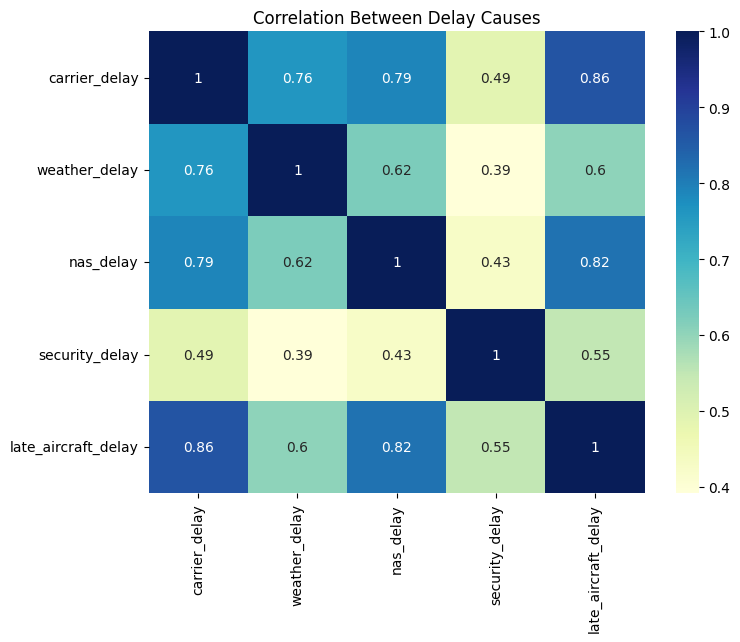
1. General Description**:** To find out if different delay causes ,like weather, NAS, carrier, etc. are related to each other.
2. Specific Requirements

* Selected the delay columns: carrier\_delay, weather\_delay, nas\_delay, security\_delay, late\_aircraft\_delay
* Used .corr() to calculate correlation
* Visualized the correlation matrix using a heatmap

Helps understand which delay causes commonly occur together and could be interrelated.

1. Analysis results

****

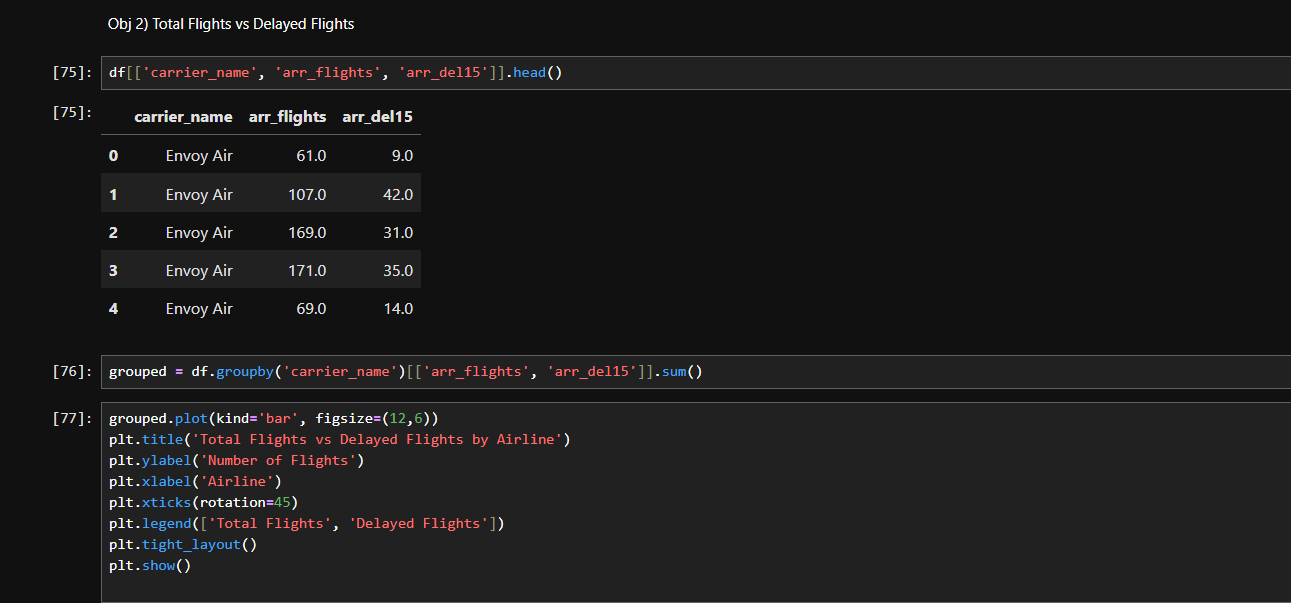
1. Visualization

**2. Total Flights vs Delayed Flights**

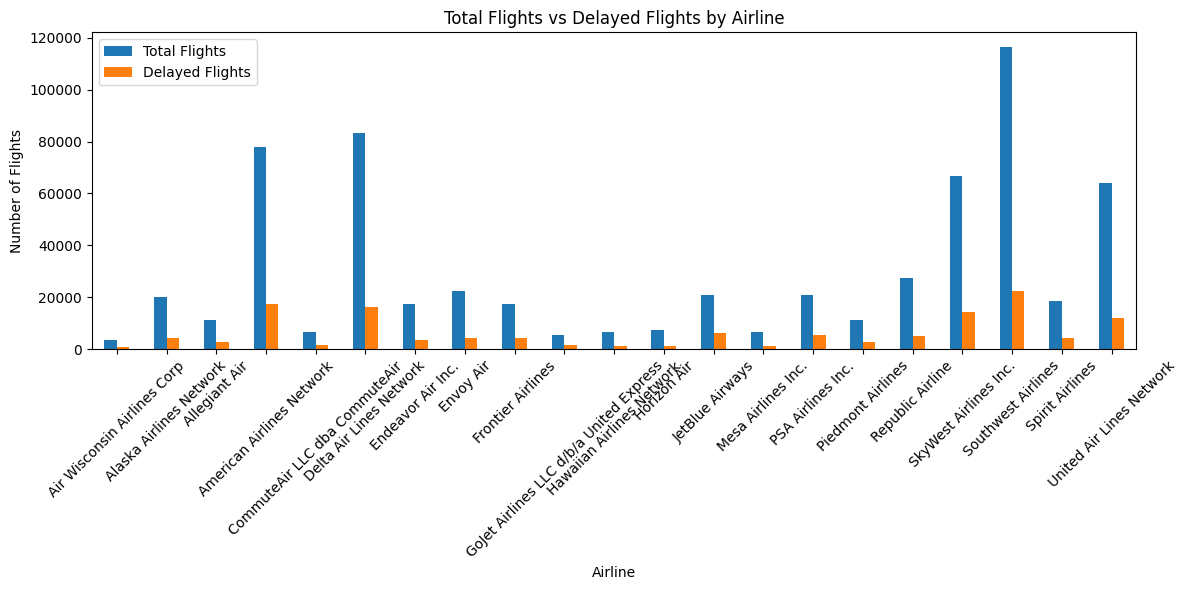
1. General Description**:** Compare how many flights were scheduled versus how many got delayed.
2. Specific Requirements:

* Plotted total flights ‘arr\_flights’ vs. delayed flights ‘arr\_del15’
* Used a bar chart to visualize the comparison for different airlinesShows which airlines experience more delays in proportion to their total flights.

1. Analysis results



1. Visualization

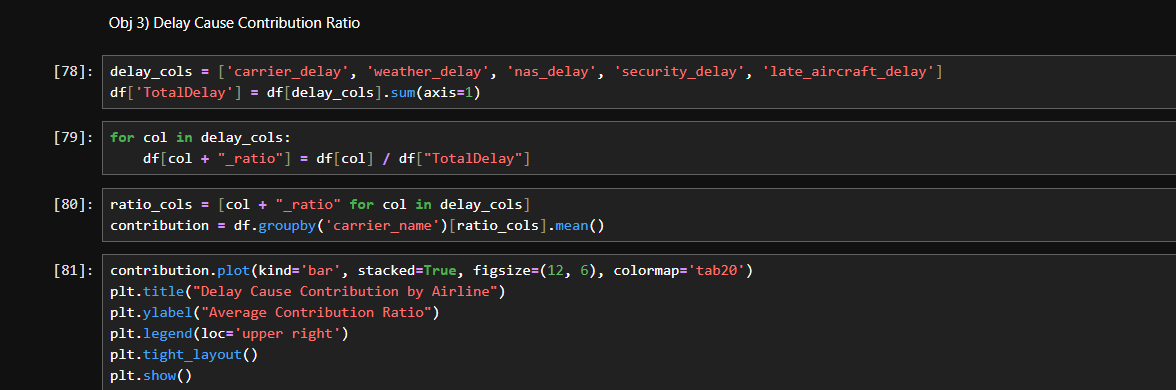


**3. Delay Cause Contribution Ratio**

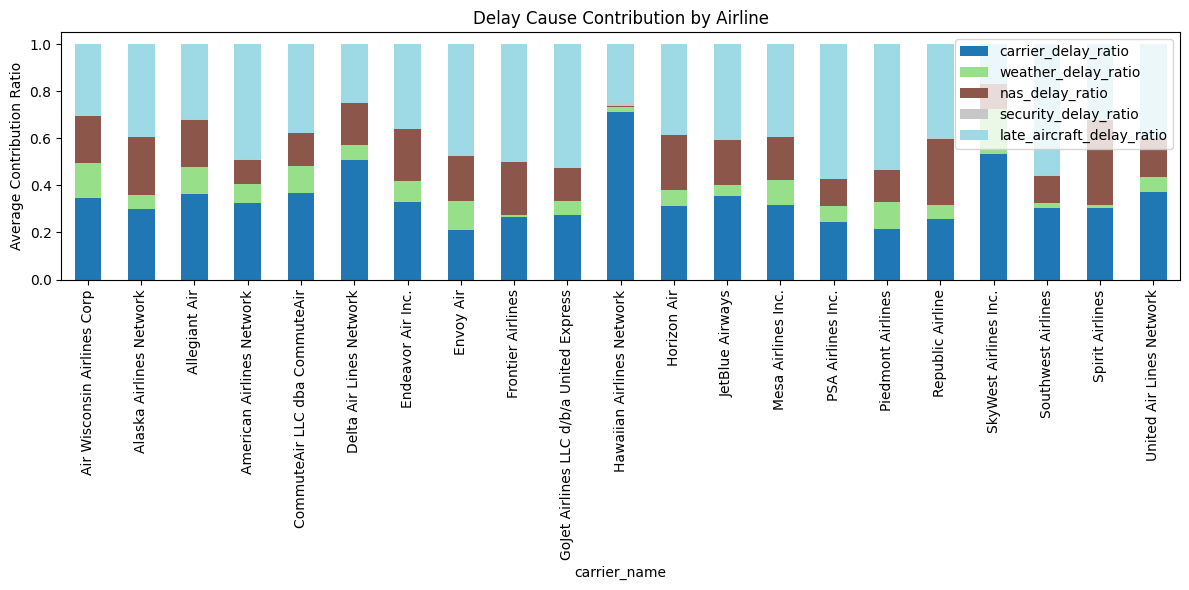
1. General Description**:** Find out how much each cause (weather, carrier, etc.) contributes to the total delay for each airline.
2. Specific Requirements:

* Calculated TotalDelay as the sum of all delay types
* Computed the percentage share of each delay type
* Used a stacked bar chart to visualizeHelps understand which causes affect which airlines the most.

1. Analysis results



1. Visualization



**4. Most and Least Reliable Airports**

1. General Description**:** Identify the most reliable (least delayed) and least reliable (most delayed) airports.
2. Specific Requirements:

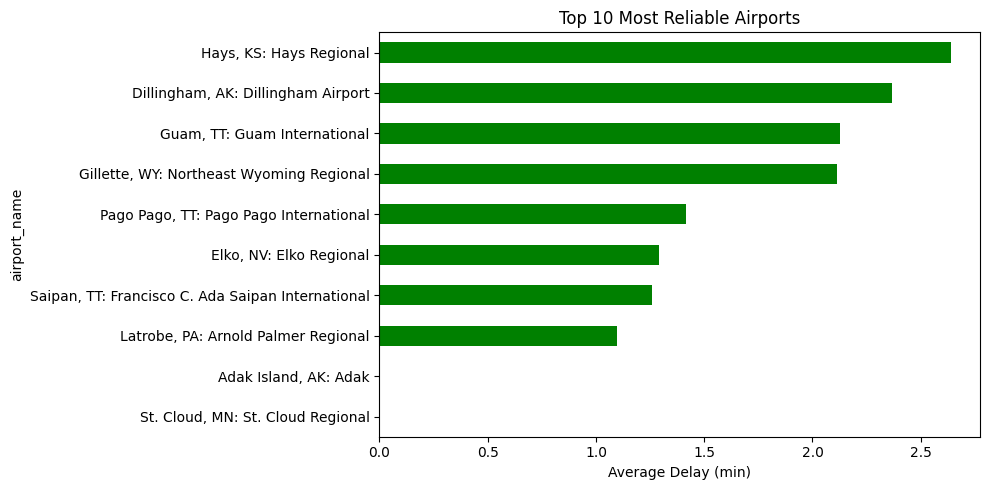
* Calculated avg\_delay = arr\_delay / arr\_flights
* Grouped by airport\_name
* Sorted and visualized top and bottom 10 using a horizontal bar chart

Airports with consistently low or high delays can be identified for performance reviews.

1. Analysis results



1. Visualization



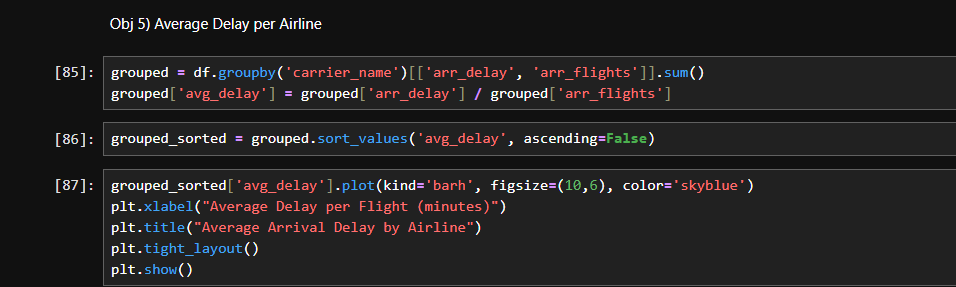
**5. Average Delay per Airline**

1. General Description**:** Find the average delay time for each airline to evaluate performance.
2. Specific Requirements:

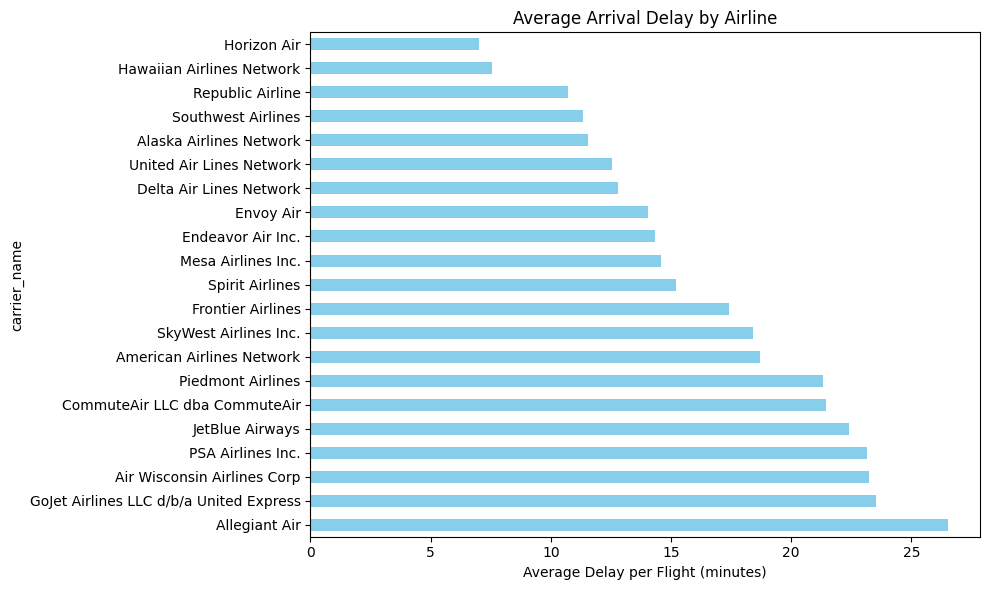
* Grouped by carrier\_name
* Divided total arr\_delay by total arr\_flights
* Visualized using a bar chart

Reveals which airlines are the most and least punctual.

1. Analysis results



1. Visualization



**6. Pie Chart of Delay Causes (Overall)**

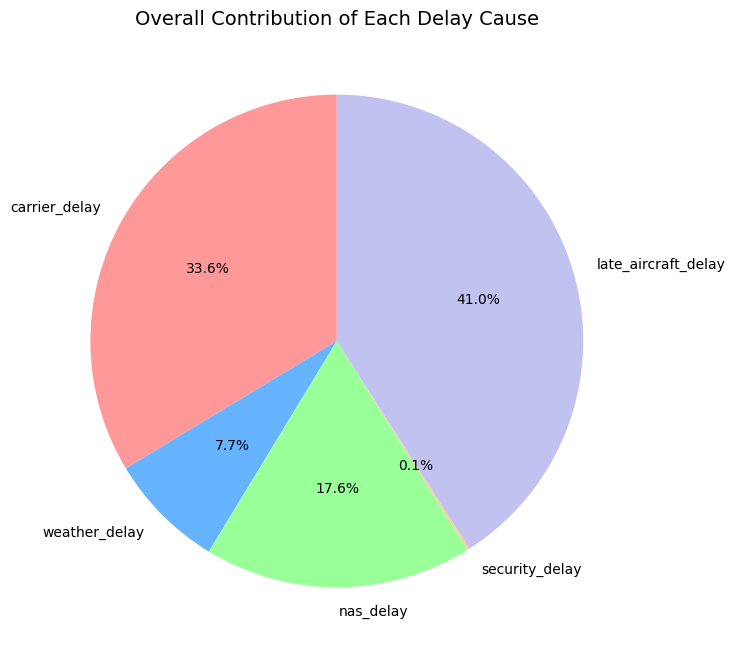
1. General Description**:** Understand what portion each delay cause contributes to overall delays across all airlines.
2. Specific Requirements:

* Summed all delay cause columns
* Used a pie chart to show distribution

Clearly shows the most common causes of delay (e.g., weather, late aircraft, etc.).

1. Analysis results



1. Visualization

**Conclusion**

The airline delay analysis project provided deep insights into the primary causes and trends associated with flight delays across various airports and airlines in the U.S. Through a combination of exploratory data analysis (EDA) techniques and visualizations in Python, we discovered which delay causes are most dominant, how airlines differ in performance, and which airports tend to be more or less reliable.

Some key takeaways include:

* LateAircraft and Weather are among the leading causes of delays.
* Certain airlines and airports consistently perform better in terms of on-time arrivals.
* The correlation analysis revealed dependencies between delay causes, which can help in proactive planning.
* Visual tools like bar charts, pie charts, and heatmaps made it easier to interpret complex patterns in delay data.

This analysis not only helped understand current trends but also lays a foundation for predictive modeling and operational improvements in the aviation sector.

**Future Scope**

Predictive Modeling for Delay Forecasting:  
Future work can include advanced machine learning models to predict delays based on current conditions (e.g., weather, time of day, flight load).

Real-Time Dashboard Integration:  
Develop real-time dashboards using tools like Power BI or Tableau for live monitoring of flight delays.

Geo-Spatial Mapping:  
Integrate geolocation data to map delays spatially for more actionable insights.

Passenger Impact Analysis:  
Analyze how delays affect passenger satisfaction, especially frequent flyers and connecting passengers.

Monthly/Seasonal Analysis Across Years:  
Extend the dataset across multiple years to detect seasonal trends and long-term delay patterns.

**References**

* Python libraries: pandas, matplotlib, numpy

Link of linkedIn –

[**https://www.linkedin.com/posts/priyanka-a61z\_dataanalysis-python-airlinedelays-activity-7318296441817346051-MaNk?utm\_source=share&utm\_medium=member\_desktop&rcm=ACoAAEb-7hkBqnsBYSzoUjySCr2gXKrYsLd35Bo**](https://www.linkedin.com/posts/priyanka-a61z_dataanalysis-python-airlinedelays-activity-7318296441817346051-MaNk?utm_source=share&utm_medium=member_desktop&rcm=ACoAAEb-7hkBqnsBYSzoUjySCr2gXKrYsLd35Bo)

Link of GitHub –

[**https://github.com/Priyanka-Kri6a/AirDelayAnalysis**](https://github.com/Priyanka-Kri6a/AirDelayAnalysis)