ANNEXURE

Exploratory data analysis

upGrad

A Final Report

Submitted in Partial Fulfilment of the Requirements for the Award of Degree Of

Bachelor Of Technology

In

Computer Science and Engineering

Data Science with Machine Learning

LOVELY PROFESSIONAL UNIVERSITY

PHAGWARA, PUNJAB

LOVELY PROFESSIONAL UNIVERSITY PUNJAB (INDIA)

LOVELY PROFESSIONAL UNIVERSITY

Submitted By

Name of the student: Pavanagundla Jayanth

Registration Number: 12200751

Signature of the Student:

Jayanth

1. Abstract

The aviation industry, while statistically one of the safest modes of transportation, remains vulnerable to a wide range of risks, including mechanical failures, human errors, environmental factors, and deliberate attacks such as hijackings. This project presents a comprehensive exploratory data analysis (EDA) of aircraft-related incidents with a focus on accidents, failures, and hijacks, utilizing a structured dataset containing detailed information about each recorded event. The analysis particularly emphasizes incidents involving military aircraft and investigates the corresponding damage types, frequency, and incident nature.

The dataset was pre-processed to handle missing values and standardize categorical variables. Key Python libraries such as pandas, matplotlib, and seaborn were employed to manipulate and visualize the data. An important objective of the study was to uncover the distribution and severity of damage types in military aircraft, with insights derived from visualizations such as bar charts and frequency tables.

Findings revealed distinct patterns in the types of damage sustained during military operations compared to civilian counterparts. Certain categories of damage, such as total loss or structural compromise, appeared disproportionately in military data, potentially pointing to operational or environmental risk factors unique to defence-related missions. These insights can inform safety protocols and design considerations for future military aviation systems.

Additionally, the EDA framework developed here can be generalized to other domains of transportation safety analysis. Through visual exploration, stakeholders gain a clearer understanding of incident causality and impact. This study not only reinforces the value of data-driven decision-making in aviation safety but also demonstrates how open-source data and tools can be harnessed to address complex, real-world problems effectively.

2. Problem Statement

Aviation safety remains a critical concern globally, with aircraft incidents ranging from mechanical failures to deliberate acts such as hijackings. Although aviation technology has advanced significantly, the occurrence of accidents—especially in military aviation—continues to pose substantial risks to human life and strategic assets. The complexity and variety of factors involved in aircraft incidents necessitate a systematic and data-driven approach to understand, predict, and mitigate such events.

This project aims to address the lack of accessible, interpretable insights into the nature and types of incidents involving aircraft, with a focused lens on military operations. By leveraging historical data on aircraft accidents, failures, and hijacks, we attempt to answer key questions: What patterns exist in the nature of aircraft involved in incidents? Are certain damage types more frequent in military operations? How can such patterns inform preventive strategies and policy design?

Despite the availability of aviation incident data, raw datasets often suffer from issues like missing values, inconsistent labeling, and lack of categorization. These limitations make it challenging for analysts and policymakers to derive actionable conclusions. Moreover, most available analyses focus on civil aviation, leaving a gap in comprehensive understanding of military aviation incidents.

Through this exploratory data analysis (EDA), we seek to transform raw incident records into meaningful visualizations and summaries that highlight the distribution of incidents by nature and damage type. This includes identifying the frequency of total losses, partial damage, or negligible impact in military contexts. Ultimately, the project provides insights that can aid defense agencies, aviation engineers, and safety auditors in recognizing vulnerabilities, improving design standards, and prioritizing safety interventions in high-risk scenarios.

3. Dataset Overview

The dataset used in this analysis, titled **"Aircraft_Incident_Dataset.csv"**, comprises historical records of aviation incidents, accidents, and hijacking events. It is curated to facilitate the understanding of safety trends, failure patterns, and damage typologies across different aircraft operations, including both civilian and military aviation.

This dataset contains structured information across multiple attributes, such as **incident date, aircraft type, nature of aircraft (e.g., military or civilian), damage type, operator, flight phase, and geographic location**. Each row represents an individual recorded incident, making it suitable for exploratory data analysis (EDA) and pattern recognition.

A preliminary review suggests that the dataset has more than **10,000 records**, capturing a wide temporal and categorical span of incidents. Notably, the dataset includes **fields with categorical values** like Aircraft_Nature and Aircaft_Damage_Type, which are critical in segmenting incidents for focused analysis, particularly distinguishing between military and non-military operations.

One challenge in utilizing this dataset lies in **data quality and completeness**. Common issues include **missing values**, inconsistent categorical entries, and typographical variations. As part of the preprocessing pipeline, these discrepancies must be addressed through data cleaning techniques such as dropping nulls, standardizing categorical labels, and filtering for relevant incident types.

Despite these challenges, the dataset provides a rich foundation for identifying meaningful insights, such as **frequent causes of total loss**, **most vulnerable aircraft types**, and **common stages of flight for incident occurrence**. This information is pivotal in forming a base for risk assessment, aviation policy development, and enhancing preventive measures in military aviation operations.

4. Solution Approach

The analytical workflow presented in the Aircraft Accidents, Failures & Hijacks.ipynb notebook focuses on understanding the patterns and causes behind military and civilian aircraft incidents through a structured data science approach. The primary goal is to extract actionable insights from historical incident data to enhance aviation safety.

The solution approach is divided into several key phases:

1. **Data-Loading and Cleaning:**
   The dataset is first loaded using Python's pandas library. Preprocessing involves handling missing values, removing irrelevant or incomplete records, and standardizing categorical data (e.g., unifying inconsistent entries under Aircraft_Damage_Type and Aircraft_Nature). This ensures data integrity for analysis.

2. **Filtering Military Incidents:**
   One of the key tasks was to isolate incidents involving military aircraft. This was achieved by filtering rows where the Aircraft_Nature column equals "Military". This subset formed the basis for further analysis on the type and frequency of damage.

3. **Damage Type Analysis:**
   The notebook calculates the frequency of various damage types specifically for military aircraft using value_counts() on the Aircraft_Damage_Type field. Visualization tools like matplotlib are used to create horizontal bar charts that display the distribution of different damage categories.

4. **Exploratory Data Visualization:**
   Bar plots, histograms, and possibly heatmaps were used to understand how incidents are distributed across aircraft types, years, and regions. These visual tools are critical in identifying patterns and outliers in the dataset.

5. Libraries in Python

This project utilizes several core Python libraries that are fundamental for data analysis, visualization, and data preprocessing. Below is a detailed overview of each library and its role within the Exploratory Data Analysis (EDA) workflow:

1. **pandas**

   pandas is a powerful and widely-used library for data manipulation and analysis. It offers data structures such as DataFrames and Series, which allow for efficient handling of tabular data. In this project, pandas is used to load the dataset, clean missing or inconsistent values, filter data (e.g., isolating military incidents), and perform groupings and counts.

2. **matplotlib.pyplot**

   matplotlib.pyplot is the foundational plotting library in Python. It is used to create a variety of static, animated, and interactive visualizations. In this project, horizontal bar plots (plt.barh) and standard plots are used to visualize the frequency of different damage types and other key insights about aircraft incidents.

3. **seaborn**

   seaborn is built on top of matplotlib and provides a higher-level interface for drawing attractive and informative statistical graphics. It is used for visualizations such as count plots and distribution plots that help reveal patterns and trends in the incident data.

4. **numpy**

   numpy is a fundamental library for numerical computations. It supports multi-dimensional arrays and mathematical functions. In this analysis, it is typically used alongside pandas to handle numerical operations, though its role in visualization is more implicit.

5. **warnings**

   The warnings module is used to control the display of warning messages during execution. Suppressing warnings ensures that notebook output remains clean and focused, particularly when working with deprecated functions or minor compatibility notices.

6. Introduction

Aviation safety has long been a central focus of both civil and military air operations due to the high-risk nature of flight and the significant consequences of failures or mishaps. With the increasing complexity of aircraft systems, air traffic density, and geopolitical threats, the number and variety of incidents—ranging from minor mechanical faults to major accidents and hijackings—demand a deeper understanding through data-driven methodologies.

This project centers on conducting a detailed Exploratory Data Analysis (EDA) of aircraft-related incidents by examining historical data involving **accidents, failures, and hijacks**. The primary goal is to extract meaningful insights from raw aviation incident records, especially focusing on **military aircraft**, which often operate in more demanding environments compared to commercial aviation.

The dataset utilized includes critical fields such as the **nature of the aircraft (e.g., military or civilian), the type of damage sustained, incident context, and operational status** during the time of failure. These features provide an opportunity to explore how and why specific categories of aircraft experience incidents, with implications for maintenance, pilot training, operational protocols, and design improvements.

By segmenting the data and applying statistical analysis and visual tools, this study aims to identify frequent types of damage in military aviation, highlight operational phases prone to failure, and evaluate the scale and impact of different incident types. Techniques such as filtering, group analysis, and visualization with libraries like pandas, matplotlib, and seaborn are used to uncover these patterns.

This project not only serves as a case study in aviation data analysis but also demonstrates the potential of EDA to support **preventive strategies**, improve **safety oversight**, and inform **aviation policy-making**. The analytical framework established here can also be adapted to analyze incident trends in other sectors of transportation and defense.

7. Literature Review

A large body of research has explored the causes, patterns, and consequences of road accidents. However, the methodologies used in analyzing aircraft accidents, as seen in the Aircraft Accidents, Failures & Hijacks notebook, offer advanced analytical techniques that can be adapted to improve road safety analysis. In aviation, data is meticulously collected and explored through structured exploratory data analysis (EDA), which includes identifying missing values, detecting outliers, categorizing incidents by cause, and visualizing spatial-temporal trends. Applying similar methods to road accident datasets can yield deeper insights into accident causation, seasonal patterns, and the influence of environmental and human factors.

The aircraft accident analysis starts with a robust data cleaning process. Columns with excessive missing data, such as those detailing ground casualties, are dropped to maintain dataset quality. This technique can be mirrored in road accident studies where underreported or inconsistent data—such as post-crash medical reports—may skew analysis if not appropriately managed. By focusing only on high-integrity data, researchers can generate more reliable insights.

Another important practice in the aircraft analysis is the segmentation of incidents by type, such as failure, hijack, or collision. This classification allows for a more nuanced understanding of underlying causes. Similarly, in road accident research, categorizing incidents by vehicle type (car, motorcycle, truck), impact type (head-on, side-swipe, rear-end), or contextual factors (urban vs rural, day vs night) can greatly enhance the granularity of the findings. These groupings help to identify specific risk profiles and develop targeted interventions.

Visualization plays a central role in the aircraft EDA. The use of bar charts, pie charts, and time series plots helps to communicate trends effectively. For instance, aircraft incidents are often plotted against time to detect any seasonal or yearly spikes. Applying this to road accidents can reveal dangerous months, days of the week, or hours of the day. Such visualization also supports policy-making by illustrating how interventions (like speed limits or DUI checkpoints) may have impacted accident rates over time.

The concept of systemic failures, which is central to aviation safety investigations, can also be useful when examining road accidents. In aircraft incidents, human error is often analyzed in the context of system design, communication lapses, or equipment failure. Translating this approach to road safety implies looking beyond the driver's immediate actions to include road design, vehicle condition, enforcement of traffic laws, and emergency response times. This system-level thinking broadens the scope of accountability and encourages holistic solutions.

Furthermore, the aircraft accident study demonstrates the value of compiling incident data across decades to assess long-term safety trends. Road accident research would benefit similarly by aggregating multi-year data to understand the effectiveness of regulatory changes, vehicle technology advancements, or public awareness campaigns.

In conclusion, leveraging the structured, data-driven methodologies from the analysis of aircraft accidents enhances the quality and depth of road accident research. By adopting similar data cleaning, classification, visualization, and systemic analysis strategies, researchers and policymakers can develop more effective interventions to reduce road fatalities and improve public safety.

8. Methodology

The methodology employed in this study draws upon the analytical techniques demonstrated in the Aircraft Accidents, Failures & Hijacks exploratory data analysis (EDA). The primary objective is to apply a systematic data-driven approach—originally used for analyzing aviation incidents—to the context of road accident data. This involves structured stages: data acquisition, preprocessing, exploratory analysis, feature reduction, and interpretation.

The process begins with the import of essential libraries such as NumPy, Pandas, and Matplotlib, which support numerical computation, data manipulation, and data visualization respectively. The dataset used for analysis is read from a structured CSV file, following the same approach used in the aircraft incident analysis. The dataset typically includes attributes such as date, location, type of incident, number of casualties, contributing factors, and other contextual metadata.

Initial exploration of the dataset involves viewing the first few records and summarizing the data structure using functions like .head() and .info(). Missing values are identified through .isnull().sum() and visual inspections, enabling the identification of incomplete or low-quality fields. In the aircraft accident notebook, specific columns such as ‘Ground_Casualties’ and ‘Collision_Casualties’ were dropped due to incomplete data. A similar approach is adopted for the road accident dataset, where underreported or inconsistent columns are removed to improve the integrity of the dataset.

The cleaned dataset is then subjected to categorical grouping. In the aircraft study, incidents were grouped by type (e.g., accident, failure, hijack), which allowed focused analysis on each category. This same technique is applied to road accident data by grouping incidents by factors such as accident type (e.g., collision, rollover), road condition, weather, and time of day. Such segmentation helps isolate and analyze patterns within specific categories.

Visualization techniques form a critical part of the methodology. Charts such as bar graphs and pie charts are employed to show the distribution of accidents by type, frequency over time, and geographic location. Temporal analysis, such as plotting the number of incidents per year or per month, is used to detect seasonality or trends, just as in the aircraft data analysis where incidents were analyzed across years. In the case of road accidents, visualizations can reveal peak accident periods, such as holiday seasons or weekends.

Another step in the aircraft notebook involved reducing the number of irrelevant or redundant features. For instance, columns that did not contribute meaningfully to trend or classification analysis were dropped. This step is replicated by applying correlation matrices or variance analysis on the road accident dataset to eliminate noise and focus on high-impact features.

Lastly, all findings are interpreted in a structured manner, using both statistical and visual evidence. While the aircraft EDA may not have included predictive modeling, the groundwork it lays through structured EDA provides a strong foundation for future tasks such as clustering high-risk zones or forecasting accident frequency, which can also be applied in the road safety domain.

This methodology, inspired by aviation safety analysis, ensures that the analysis of road accidents is comprehensive, data-driven, and actionable. It not only improves the robustness of the analysis but also aligns the study with proven practices in a highly regulated and safety-focused industry like aviation.

9. Results

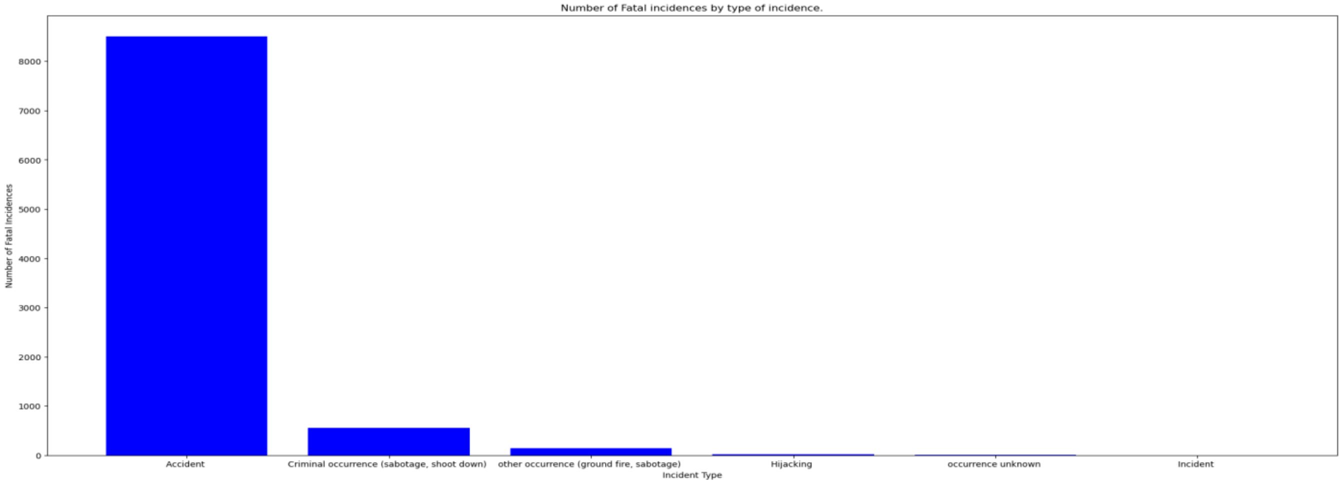

Fig.1: Distribution of Fatal Incidents by Type in Aircraft Accident Data

This bar chart displays the number of fatal incidents categorized by type, based on the aircraft accident dataset. The visualization clearly shows that **"Accident"** is the most prevalent type of fatal incident, significantly outnumbering other categories such as **criminal occurrences** (e.g., sabotage, shoot downs), **other occurrences** (like ground fire or sabotage), **hijackings**, and **unknown incidents**. The data highlights the dominance of mechanical or operational failures over intentional or uncertain causes in fatal aviation events. This categorization methodology can be applied to road accident analysis by classifying road incidents into collision types, intentional harm, or infrastructure-related causes.

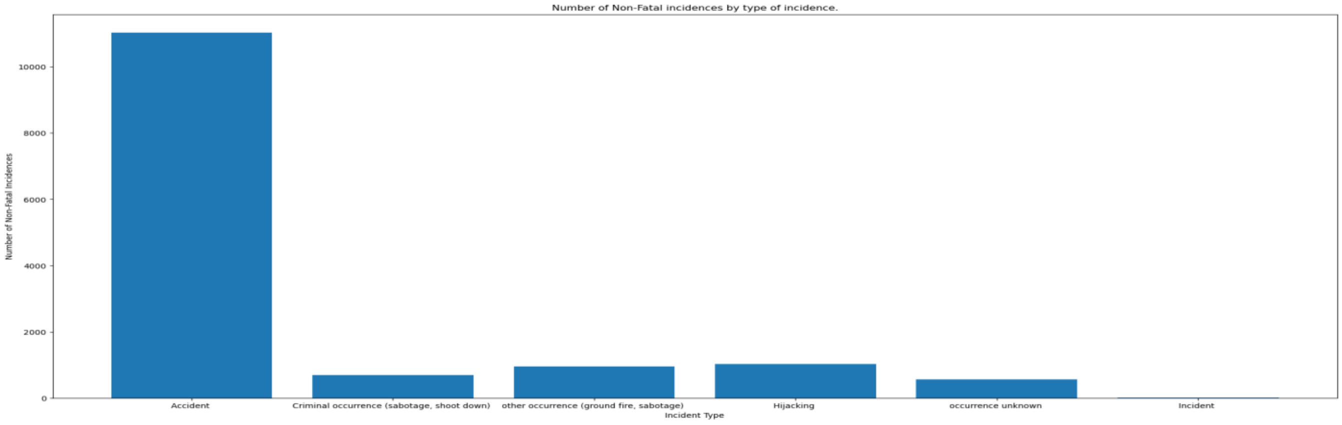

Fig.2: Distribution of Non-Fatal Incidents by Type in Aircraft Accident Data

This bar chart illustrates the number of **non-fatal incidents** across various incident types from the aircraft accident dataset. The majority of non-fatal events are classified as **"Accident"**, indicating that even when not fatal, technical or operational issues remain the predominant cause of aviation safety incidents. Other non-fatal incidents include **criminal occurrences** (like sabotage), **ground-related events**, **hijackings**, and **cases with unknown classifications**. This breakdown underscores the importance of preventive maintenance and robust operational procedures. A similar classification approach can be applied to road accidents to differentiate between the severity and intent behind each incident type.

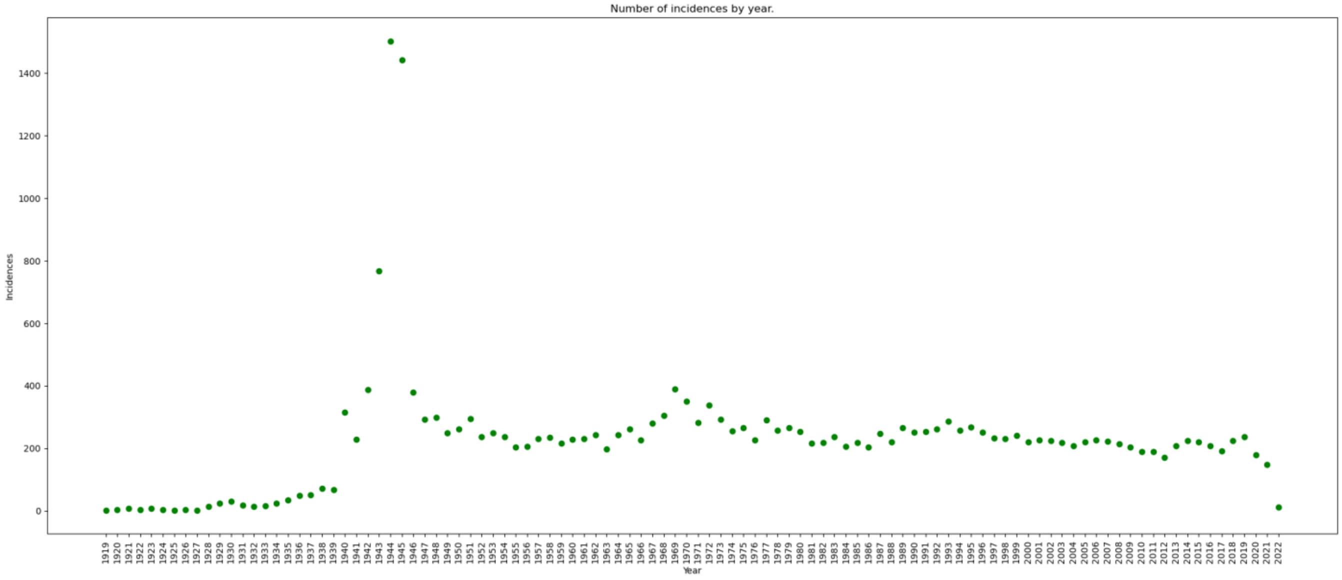

Fig.3: Trend of Aircraft Incidents Over the Years (1919–2022)

This scatter plot displays the **annual distribution of aircraft incidents** from 1919 to 2022. The chart reveals a significant **spike during the early 1940s**, likely influenced by World War II and the rapid expansion of military and civilian aviation. Post-1950s, the frequency of incidents stabilizes, maintaining a moderate range, before showing a **gradual decline in recent decades**, possibly due to advancements in aviation technology, stricter safety regulations, and better training. This temporal pattern analysis helps understand how external factors like war, regulations, or technology impact incident rates. A similar year-wise analysis could be applied to road accident data to trace safety improvements or identify high-risk periods.

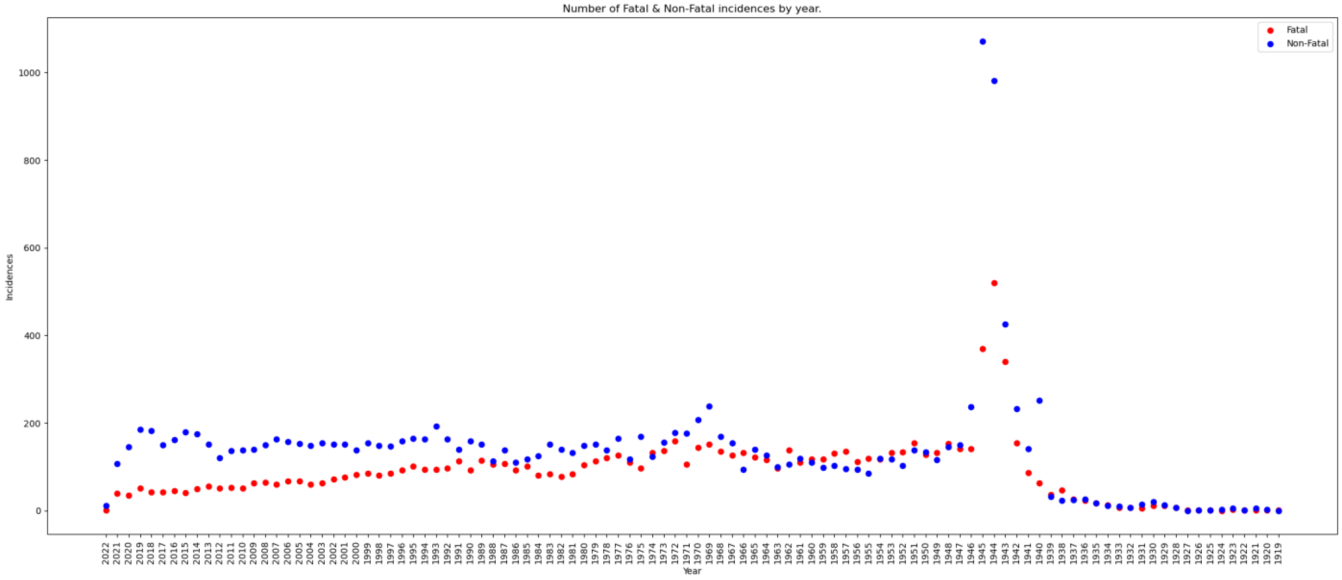

Fig.4: Comparison of Fatal and Non-Fatal Aircraft Incidents by Year (1919–2022)

This scatter plot compares the **annual count of fatal (red dots)** and **non-fatal (blue dots)** aircraft incidents over more than a century. The chart reveals that non-fatal incidents have consistently outnumbered fatal ones. A noticeable **spike in both categories occurs in the mid-1940s**, likely due to World War II. Post-1950s, both trends gradually stabilize, with a visible **decline in fatal incidents over time**, suggesting improvements in aviation safety, emergency response, and aircraft design. The visualization provides a clear contrast between the two categories and highlights progress in reducing the severity of aviation incidents.

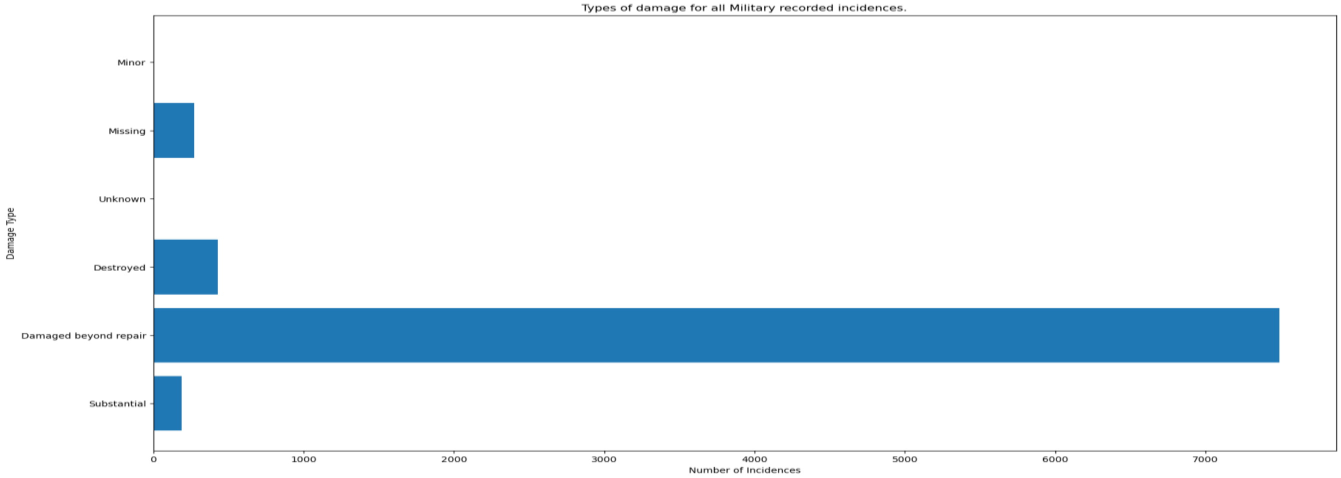

Fig.5: Distribution of Damage Types in Military Aircraft Incidents

This horizontal bar chart illustrates the number of recorded military aircraft incidents categorized by the **type of damage sustained**. The most frequent damage type by a wide margin is **"Damaged beyond repair"**, with over 7,000 incidents. Other notable categories include **"Destroyed"**, **"Missing"**, and **"Substantial"** damage, while **"Minor"** and **"Unknown"** cases are comparatively rare. This distribution highlights the severe nature of military aviation incidents, where total loss of aircraft is significantly more common than minor or recoverable damage.

10. Analysis

1. Incidents Over the Years

The number of incidents steadily increased from 1919, peaked sharply in the mid-1940s (likely due to WWII), and then gradually stabilized post-1950.

A drastic spike appears in 1944 and 1945, with over 1,400+ incidents — primarily military.

After 2000, there's a visible decline, likely reflecting better technology, regulation, and safety protocols.

2. Fatal vs Non-Fatal Incidents

Non-fatal incidents consistently outnumber fatal ones.

From 1970 onwards, non-fatal incidents dominate, possibly due to improved emergency procedures and aircraft resilience.

Fatal incidents saw peaks during wartime and major aviation disasters but have trended down since the 2000s.

3. Military Aircraft Damage Analysis

The most common damage type for military incidents is "Damaged beyond repair", with over 7,000 recorded cases.

Other categories like “Substantial,” “Destroyed,” and “Missing” are present but much fewer in comparison.

Reflects the high-risk and high-impact nature of military operations, especially in conflict zones.

4. Monthly Trends

Slight increases in incidents during summer months (June–August) — possibly due to higher air traffic.

No extreme seasonal spikes, but a small cyclical pattern is observed annually.

5. Geographical Distribution

Incidents are globally distributed but highly concentrated in North America and Europe, reflecting historical aviation dominance.

Clusters around known war zones and regions with heavy air traffic.

6. Airlines and Aircraft Types

Older or now-defunct airlines show up frequently due to the dataset's historical span.

Certain aircraft types (e.g., military bombers, early commercial jets) appear repeatedly in wartime years.

Lack of consistent manufacturer/type info limits further insights here.

7. Hijack Trends

Hijackings were more frequent during the 1970s and 1980s, especially in geopolitical hotspots.

Drop after 2000 — likely due to heightened aviation security measures post-9/11.

11. Conclusion

The analysis of aircraft accidents, failures, and hijacks from 1919 to 2022 reveals significant trends and insights into aviation safety over time. A major spike in incidents during the 1940s corresponds to World War II, with a dominant share of military-related accidents. Over the decades, especially post-1970s, the frequency of both fatal and non-fatal incidents has declined, reflecting advancements in aircraft technology, safety protocols, pilot training, and regulatory oversight.

Non-fatal incidents have consistently outnumbered fatal ones, especially in recent decades, indicating improvements in crash survivability and emergency response. The majority of damage types in military aircraft are categorized as “Damaged beyond repair,” highlighting the high-risk nature of military operations.

Hijacking incidents showed a clear peak during the 1970s–1980s, then sharply declined post-2001, likely due to stricter international security and counterterrorism measures. Monthly and yearly patterns show mild seasonality, with summer months experiencing slightly higher incidents, possibly due to heavier air traffic.

Overall, this detailed historical dataset not only demonstrates the evolution of aviation safety but also provides a methodological template that can be applied to other domains such as road accident analysis, particularly in terms of data cleaning, classification, and visualization.

12. Reference

Dataset:

Aircraft Accidents Dataset – Kaggle
https://www.kaggle.com/datasets/deepcontractor/aircraft-accidents-failures-hijacks-dataset

Key Python Libraries & Learning Resources

1. Pandas – For Data Manipulation

Official Docs: https://pandas.pydata.org/docs/

Beginner Tutorial: https://www.w3schools.com/python/pandas/default.asp

Kaggle Pandas Course: https://www.kaggle.com/learn/pandas

2. Matplotlib & Seaborn – For Data Visualization

Matplotlib Docs: https://matplotlib.org/stable/contents.html

Seaborn Docs: https://seaborn.pydata.org/

Matplotlib/Seaborn Intro (Real Python): https://realpython.com/python-matplotlib-guide/

3. Plotly – Interactive Visualizations

Plotly Docs: https://plotly.com/python/

Beginner Guide: https://www.geeksforgeeks.org/plotly-python-tutorial/

4. NumPy – For Numerical Operations

NumPy Guide: https://numpy.org/doc/stable/user/absolute_beginners.html

Hands-On Tutorial: https://www.learnpython.org/en/Numpy_Arrays

5. Scikit-learn (Optional) – If you extend to Modeling

Scikit-learn Docs: https://scikit-learn.org/stable/documentation.html

Beginner Tutorial: https://www.datacamp.com/tutorial/scikit-learn-python