Trusted AI – Aviation (Skyshield)

Saurav Subash Prasad, Vishwas Shivakumar, Akhilesh Gowda Mandya Ramesh,
Sai Surya Pulagam, Anirudh Emani

INTRODUCTION

The data we will be exploring is the FAA's AIDS (Accident and Incident Data System) dataset, which contains information about air accidents and incidents that have occurred since 1975. Our primary focus will be on cause factors (found in the "a" files) and NTSB investigation remarks (found in the "e" files). By examining this information, we hope to gain a deeper understanding of the cause-and-effect relationship between various root causes and the resulting incidents.

After thoroughly cleaning this dataset and filtering out the information we need, we drew a knowledge graph that represents the five most common primary cause factors, examples of incidents having these cause factors, and the contributing cause factor for each. We also investigated how these incidents have changed over the past few decades. We looked at various geographic and anthropological data points as you will see in section 4.

Table 1. Some examples of air incidents

Date	Location	Cause	Investigation Remarks
01/01/2022	New York City	Pilot Error	The pilot failed to properly execute a landing approach.
03/15/2022	Los Angeles	Equipment Failure	The engine on the plane failed mid-flight.
06/20/2022	Chicago	Weather	The plane encountered severe turbulence due to a thunderstorm.
09/10/2022	Atlanta	Human Factors	The crew was fatigued and made errors during pre-flight checks.

1 INSIGHT NEEDS

Through our conversations with Lindsey Michie from The University of Notre Dame, we have established the following insight needs:

1. To determine the **correlation** between various cause factors, and to determine the **cause-and-effect** relationship between the different cause factors and incidents.
2. To identify **clusters or patterns** in the cause factor data.

Due to insufficient time, we were unable to determine the correlation between different cause factors. However, we were able to establish cause-and-effect relationships between cause factors and incident types and have visually represented them in the knowledge graph.

After running extensive analyses on the AIDS dataset, we've also identified multiple trends and patterns in the data, such as the changing average age of pilots, the geographic distribution of air accidents and the decreasing number of air accidents in general.

1.1 Stakeholder Analysis

For this project, we identify the following stakeholders:

1. Sponsor – Lindsey Michie, University of Notre Dame
2. Regulatory Authorities – Agencies like FAA (United States) or DGCA (India) can benefit from studies like these as they learn about significant cause factors behind civil aviation safety incidents.
3. Civil Aviation Industry – While air travel is already perhaps the safest means of transportation, learning to mitigate and proactively avoid incidents with high fatality rates, it can become even more safer than it is today.

2 DATA ACQUISITION

The data that we used is the FAA's Accident and Incident Data System (AIDS). It is a database that contains information on all aviation accidents and incidents in the United States since 1975. The data is collected from a variety of sources, including pilots, air traffic controllers, and mechanics, and is used to identify trends and patterns in aviation accidents and incidents. This information is then used by the FAA and other organizations to develop policies and procedures to improve aviation safety. The acquisition of this data can be a complex process, but it provides valuable insights into how accidents and incidents occur and can be prevented in the future.

2.1 Description of Data

These are the most common cause factor codes in the dataset. We will focus our attention on these, and also on the cause factors that have the highest average fatality rate.

Cause Factor	Description
GC	"Ground Control" – Improper aircraft operation on the ground.
LO	"Level Off" – The aircraft failed to or wasn't properly levelled off after departure.
HO	The aircraft failed to avoid an obstruction or foreign object.
GN	Landing Gear not extended.
AS	"Airspeed/Stall" – The aircraft failed to maintain sufficient airspeed.

3 ANALYSIS METHODS

In analysing the FAA's Accident and Incident Data System (AIDS) dataset, we employed a range of different analysis methods to gain a comprehensive understanding of the causes and effects of aviation accidents and incidents.

3.1 Exploratory Data Analysis

The dataset was originally presented in various files, including 10-14-2010_attention.doc, afilelayout.txt, aidcodes.doc, aircraftseries.doc, and airport.doc, each describing different aspects of the AIDS data. We consolidated these files into one, removing null values and performing exploratory data analysis to achieve our research objectives. For the word cloud, we conducted tokenization, stop-word removal, and identified the top frequency words.

To make sense of the word cloud, we converted all remark codes into their corresponding actual words. Given the dataset's considerable size, it required a significant amount of time and effort to process. We attempted several word cloud platforms, such as wordcloud.com, but due to the dataset's large size, we were unable to render the word cloud successfully. Even limiting the words to the top 50 most frequent words was insufficient, and we ultimately had to resort to using Python to generate the word cloud.

3.2 Trends & Patterns

We investigated the possible effects of various cause factors, and identified how different factors may lead to different outcomes in terms of accidents and incidents. To better understand the trends and patterns of accidents and incidents over time, we created geospatial and temporal visualizations, such as maps and line graphs that allow us to see how accidents and incidents have changed over the years and which areas are more prone to crashes. This will also allow us to identify any shifts or changes in aviation safety and maintenance practices that may have contributed to changes in accident and incident.

Finally, we used natural language processing (NLP) and clustering algorithms to analyse the remarks section of the dataset, identifying common themes and phrases that may provide additional insights into the causes and effects of accidents and incidents. By combining these different analysis methods, we can gain a comprehensive understanding of the complex factors that contribute to aviation safety and develop more effective strategies for preventing accidents and incidents in the future.

4 Visualizations

4.1 Knowledge Graph

We began working on a knowledge graph that highlights five most common primary cause factors behind civil aviation incidents. Moving outwards from the center of the graph, we can see the cause factors, the incidents that are linked to the cause factor codes, and the contributing factors/reasons behind these incidents.

Figure 1. A knowledge graph representing the five most common cause factors.

For example, overloading a freighter aircraft and not properly securing the cargo pallets inside the hold resulted in the plane spinning out of control and crashing after takeoff as it was unable to maintain the necessary airspeed.

4.2 Map Visualizations

To create a map visualization of aviation accidents, we imported relevant data from the FAA's AIDS dataset that includes the date, location, type, and cause factor codes. After importing, a suitable map type, like a world or US map, can be selected to add data using filters and color-coding to identify patterns or trends.

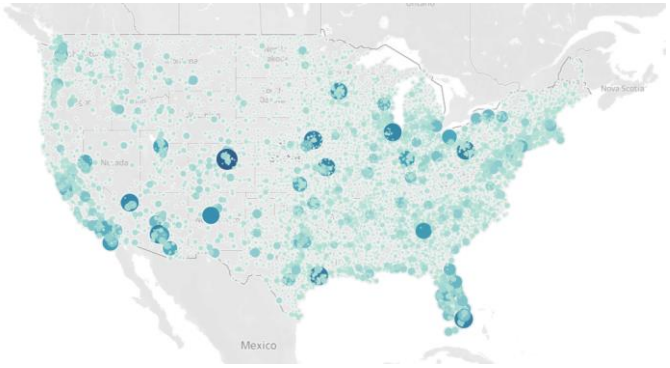

Figure 2. Frequency of air accidents in the Lower 48

Geocoding is generally used to convert location data into longitude and latitude coordinates, but since the AIDS dataset doesn't have them, a third-party geocoding service is used. The created map visualization can help in identifying the spatial patterns of accidents and incidents, leading to insights on factors affecting aviation safety and better strategies for the future.

Figure 2 shows how major United States civil aviation hubs and densely populated regions have more air incidents. In 1971, the busiest airport in the US was O'Hare International Airport in Chicago. Atlanta's Hartsfield-Jackson International Airport became the busiest airport in the US in 1998, and it has held that title ever since.

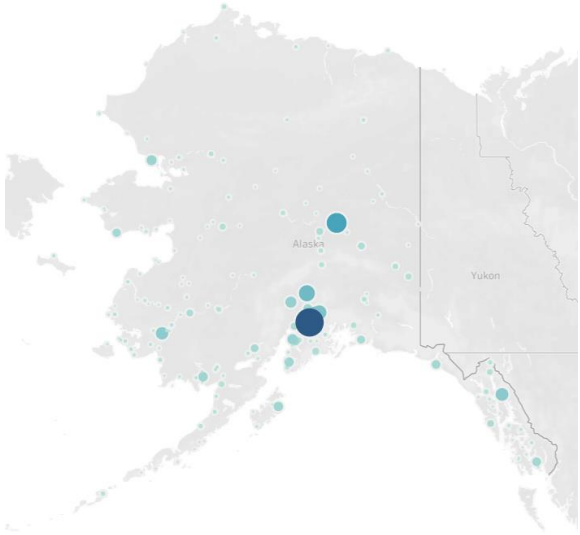

Figure 3. Frequency of air accidents in Alaska

Since 1998, aviation regulations and policies have become increasingly strict, resulting in a safer air travel experience and a decrease in the number of aviation accidents. These regulations cover a wide range of areas, including aircraft design and maintenance, pilot training and qualifications, air traffic control procedures, and airport security measures. Additionally, advancements in technology have played a significant role in improving safety, with the development of systems such as collision avoidance and weather forecasting helping to prevent accidents. The combination of stricter regulations and technological advancements has contributed to a decrease in the number of aviation accidents and has helped to make air travel one of the safest modes of transportation available today.

While Hartsfield-Jackson Atlanta International Airport is the busiest airport in the United States, it is not the airport with the highest number of aviation accidents or incidents. The airport with the highest number of reported aviation accidents and incidents in recent years is Ted Stevens Anchorage International Airport, located in Anchorage, Alaska. (See Figure 3) The majority of these incidents are related to smaller aircraft and cargo planes, which are common in the region due to Alaska's rugged terrain and remote communities.

4.3 Tableau Dashboard

We started off by identifying the key metrics and KPIs that are important to our stakeholder and their stakeholders. These might include metrics such as accident and incident rates over time, the top primary cause factors for accidents and incidents, and the types of aircraft and equipment involved in accidents and incidents.

Once we have identified the key metrics, we can use Tableau to create a series of interactive visualizations that allow users to explore the data in more detail. For example, created a line chart showing accident and incident rates over time, with filters that allow users to drill down into specific types of accidents or incidents or filter by location or other variables.
We also created a bar chart showing the top primary cause factors for accidents and incidents, with filters that allow users to drill down into specific cause factor codes or filter by other variables such as aircraft type or location.

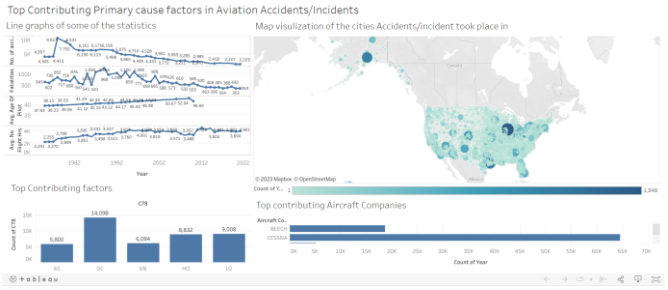

Figure 4. A screenshot of our Tableau Dashboard

To make the dashboard more interactive, we can add features such as dropdown menus or slider bars that allow users to adjust the time or other variables in real-time. We can also use colour coding and other visual cues to highlight important trends or outliers in the data.

Ultimately, the goal of the dashboard is to provide our stakeholder and their stakeholders with a comprehensive, interactive view of the data, allowing them to explore the factors that contribute to aviation accidents and incidents and develop more effective strategies for preventing them in the future.

The dataset under consideration records the average age of pilots until 2012, which has resulted in a sudden decline in the average age of pilots. However, certain columns in the FAA dataset are incomplete or deprecated since 2010, with no explicit mention of the deprecated columns and no updates from the FAA regarding the same.

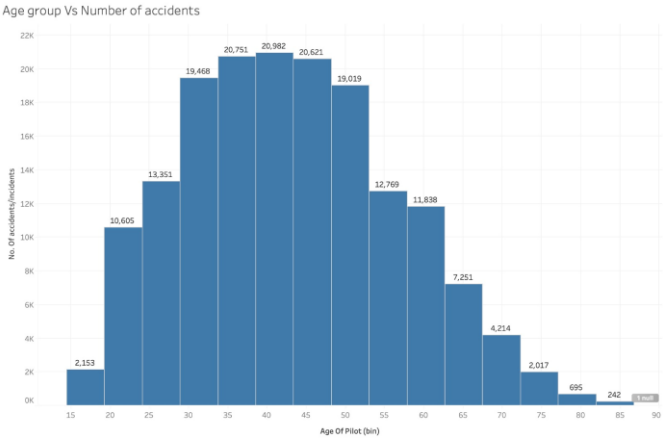

Figure 5. Histogram depicting the number of incidents by pilot age

The histogram analysis of the average age of pilots against the number of accidents/incidents has revealed certain data points that deviate from the norm, such as bins with ages below the minimum legal age of 20. The inclusion of such data points is questionable, and requires further investigation to confirm their status as outliers or not. Furthermore, it is noteworthy that pilots in the age bracket of 35-50 have experienced the highest frequency of accidents, which aligns with the average age of pilots falling within this range.

The inclusion of numerous informative visualizations into a dashboard is a critical and valuable tool for stakeholders seeking to obtain insights and make data-driven decisions. The dashboard provides a thorough picture of key performance indicators and essential metrics by integrating the necessary data elements into a consolidated area.

The dashboard allows stakeholders to engage with the data by filtering, sorting, and altering the visuals to obtain deeper insights into the data's trends and patterns. It is critical to be able to personalize the dashboard to match specific needs and preferences, as this allows stakeholders to obtain vital information quickly and efficiently.

4.4 Word Tree

We utilized Natural Language Processing (NLP) techniques in Python to analyze remarks data on aviation accidents and incidents, specifically those reported by the Federal Aviation Administration (FAA) in the United States. Our approach involved leveraging algorithms and tools to extract, analyze, and derive insights from textual data, identifying patterns and relationships. We preprocessed the text data using NLP libraries such as NLTK and spaCy, which included removing stop words, stemming, and lemmatizing words, and extracting key phrases or topics.

Our analysis of the FAA's accident and incident data yielded a wealth of information about aviation safety, including the root causes of such events. By using NLP techniques to process the textual remarks that accompanied each incident report, we identified the most frequent and important factors that contributed to such events. To gain a better understanding of their impact on aviation safety, we utilized visualizations such as word clouds and word trees to highlight these factors.

Figure 6. A word cloud highlighting primary keywords from NTSB's remarks.

Our findings revealed that the most common causes of crashes in the FAA's data were the inability to avoid obstacles, improper landings, brake failures, parts failing, forgetting to extend the landing gear, failure to maintain proper height, and poor pre-flight checking. These insights provide valuable information on the key factors contributing to aviation accidents and incidents, which will enable the aviation department to develop strategies to improve aviation safety. Additionally, the visualizations we created helped to emphasize the most frequent or important words and phrases in the remarks data.

By using NLP techniques in our analysis, we gained a deeper understanding of the root causes of aviation accidents and incidents, which can aid in developing more effective safety measures to prevent them.

5 Interpretation of Results

The map visualization of aviation accidents can help identify any spatial patterns in the accidents and incidents over time. For example, we are able to identify regions or airports where accidents are more common – such as around Anchorage, or areas where accidents have decreased over time – such as over Chicago and Atlanta. This can help stakeholders in the aviation industry to identify potential areas for improvement in safety procedures and protocols or develop special rules for areas with bad weather conditions like Alaska.

The dashboard created in Tableau provides a comprehensive overview of the accident and incident data, allowing stakeholders to easily compare and analyse different variables and trends. This can help identify potential correlations between cause factors and accident types and provide insights into the most effective strategies for improving aviation safety.

For example, Texas, California, and Florida have a very large number of private aircraft – in the tens of thousands. This would explain the larger number of incidents and the more distributed nature of these incidents compared to states like Georgia and Illinois where the numbers are concentrated around hub airports at Atlanta and Chicago.

The NLP analysis of the remarks about accidents and incidents can help identify any common themes or factors that contribute to accidents, such as pilot error or mechanical failures. This information can be used to develop targeted interventions and training programs to address these issues and improve overall safety in the aviation industry.

References

[1] *RPubs - Air-crash Data Visualization*. (n.d.). https://www.rpubs.com/adwsb/aviaccidents

[2] Fox, T., Howell, M. A., Senatore, M., & Varghese, S. (n.d.). *Visualizing the FAA Aviation Accident Database*. http://cluster.ischool.drexel.edu/~cchen/courses/INFO633/10-11/g1.pdf

[3] *General Aviation Accident Dashboard: 2012-2021*. (n.d.). https://www.ntsb.gov/safety/data/Pages/GeneralAviationDashboard.aspx

[4] *Federal Aviation Administration - Data Downloads*. (n.d.). https://av-info.faa.gov/dd_sublevel.asp?Folder=%5CAID

[5] Bird, Steven, et al. "Natural language processing with Python and NLTK." O'Reilly Media, Inc., 2009.

[6] Paulheim, Heiko. "Knowledge graph refinement: A survey of approaches and evaluation methods." Semantic Web 8.3 (2017): 489- 508

[7] Detection of Aircraft, Vehicles and Ships in Aerial and Satellite Imagery using Evolutionary Deep Learning URL:https://www.diva- portal.org/smash/get/diva2:1609134/FULLTEXT02.pdf

[8] *Clifford Law Offices.* An Analysis of 37 Years of Airplane Crash Data. https://www.cliffordlaw.com/aviation-accidents-and-incidents/

[9] *Bean, Larry. "*These States Top the Charts for Private Aviation." Robb Report. https://robbreport.com/motors/aviation/top-private-aviation-states-eg17-2753874/