Analysis of FAA Air Incidents Data

RAJANI MOHAN JANIPALLI

IN3061/INM430 Principles of Data Science

City University of London

***Abstract*-This paper presents the data analysis of general and aviation air incidents reported by FAA, for a period of over three decades. The data consists of various parameters of the incidents, that enable the analysis of the incidents from different points of view. The analysis aids to answer research questions, which try to find different unsafe aspects of flights based on these different points of view and also try to indicate the parameter(s) from the given data, that could be a possible factor due to which a particular aspect may be considered as unsafe. The analysis was done in an Jupyter notebook.**

I. INTRODUCTION

Travel is an essential activity of life and research have shown the benefits of travelling for human life [1]. Air transportation is one of the modes of travel and is in fact the fastest growing mode of transportation [2]. The demand for air travel has seen a steep rise in the past decade [3]. Even after a pause during the peak of COVID-19 pandemic, the airline industry managed to pick up quickly [4].

With a consistently increasing demand for air travel, aviation safety is one of the most important sections that airline industry looks at for its smooth progress. Analysing aviation safety and improving it further is an important step towards this [5].

Data analysis is a key tool in implementing these procedures. From operations to logistics to business analytics, data analysis is already playing a vital role in the aviation industry [6], [7]. It is also predicted to play a vital role in making the aviation safety better in future [8]. This paper inspects various points of air incidents from the given data of air incidents that have been reported in the USA and tries to investigate the factors behind the critical values of those points. This analysis would be useful to travellers as well as to various stakeholders of the aviation industry, to understand how the trend of aviation incidents has changed over the past three decades and also show them areas aviation safety that need attention and improvement. As a result, the aviation industry would be strengthened in its path to growth.

II. DATA

Data set is a record for general aviation and commercial air incidents from the year 1978 to 2015, taken from Kaggle [9] and the metadata is from FAA's AIDS website [10]. The data consists of 100,000 observations and 27 features. Each observation is a report of a discrete incident. Many of the features can be grouped into different categories.

Firstly, the features which cannot be grouped are as follows:

- AIDS Report Number – Consists of unique report number for each incident.
- Local Event Date – date of the incident.
- Event Type – type of the event, that is, incident or accident.
- Aircraft Damage – extent of damage that has happened to the aircraft.
- Flight phase – in which stage or part of the flight the incident occurred.
- Operator – airline which operated the aircraft in the incident.
- Aircraft Registration Nbr – registration number of the aircraft in the incident.

Then, the features which can be grouped by categories are as follows:

- Location category – Event City, Event State, Event Airport.
- Aircraft category – Aircraft Make, Aircraft Model, Aircraft Series.
- Flight category – Primary Flight Type, Flight Conduct Code, Flight Plan Filed Code.
- Threat category – Total Fatalities, Total Injuries.
- Engine category – Aircraft Engine Make, Aircraft Engine Model, Engine Group Code, Nbr of Engines.
- Pilot category – PIC Certificate Type, PIC Flight Time Total Hrs, PIC Flight Time Total Make-Models.

The variables in flight category are in aviation terms like personal, commercial etc. PIC in pilot category stands for Pilot In Command [11].

III. ANALYTICAL QUESTIONS

Looking at the data, many questions arised regarding the factors that might have caused the incidents. So, the research questions were put in two categories, one category was on the overall trend of the data and the other was based on different bases into which the variables could be categorized.

One of the motives of this paper was to understand how the trend of air incidents varied over the past thirty years. This would be addressed by the category of research question based on overall trend of the data, which is the first research question and is as follows:

1. What was the trend of number of accidents over the years? From the available data what could be a possible factor for that trend?

Another motive of this paper is to show the areas of aviation safety that need attention and improvement. This would be addressed by the second category of research questions based on different bases [12] and they are as follows:

2. Which was the most dangerous location and what could be the possible factors from the given data due to which that location was the most dangerous?

3. Which was the most dangerous aircraft and what could be the possible factors from the given data due to which that aircraft is the most dangerous?

4. Which was the most dangerous operator and what could be the possible factors from the given data due to which that operator is the most dangerous?

5. Which was the most dangerous flight type and what could be the possible factors from the given data due to which that flight type is the most dangerous?

6. Which was the most dangerous engine and what could be the possible factors from the given data due to which that engine was the most dangerous?

IV. ANALYSIS

The first step taken as a part of data prepossessing for analysis, was checking the information of the data which was loaded as a pandas data frame, to see the count of non-missing values and the data type for all the columns. One of the things that were observed was that the local event date column was of object data type and so it was converted into datetime data type, a step of data preparation. After that, the missing values in all the columns was checked. It was observed that there were many missing values most in non-numeric columns and not much in numeric columns. It was also realized that the variable names or any information about the last two columns were missing, so after exploring those columns they were dropped, another step of data preparation.

The first column consisting of report numbers for the incidents cannot be used for analysis as each observation in that column is a unique value. So, it was dropped.

A new column named year was created, which will have only the year but not the date and the month, a step of data derivation. This was done as the year column would be more practical to plot and do visual analysis, than the event date column. So, the local event date column was dropped.

Exploring the event type column showed that all the observations were incidents and it is not useful for analysis. Hence, it was dropped.

After exploring the total fatalities and the total injuries columns, based the paper by “A” a new column named severe was created, which would differentiate the observations based on the values of these two columns. Out of all the observations, only 773 observations word of the severe category.

One of the key steps of data preparation is the handling of missing values. As noted earlier, there were a lot of missing values in that data and most of them were missing in the non-numeric categorical columns. Since, the data was originally provided by FAA end is based on reports of the actual incidents importing numeric or non-numeric missing values using descriptive statistics would be in an inappropriate. Common methods like label encoding or mapping that are used to impute non numeric categorical missing values [13] we’re not used here, as there were large number of unique values for almost all the categorical columns and this would lead to inconsistent results on imputation.

Another key step of data preparation is handling outliers. For the columns total fatalities and total injuries, the visual analysis of the data through scatterplot showed some data points as outliers, but they were practically possible values and hence were not considered as outliers.

From the scatter plot of PIC flight time total hours, some outliers were visible which were not practically possible figures. Two times the mean of this column was greater than its standard deviation. But from the scatterplot it could be seen that a major chunk of the data was above even this value. So instead of going with the rule of thumb, visual inspection showed that most of the data was under the 40,000 figure and hence all the observations above it were removed as outliers.

Similar was the case for PIC flight time total make model column and hence, same method was followed, taking 20000 as the barrier for differentiating outliers.

After the handling of outliers, the number of observations under severe category was checked and it was realized that the observations of severe category dropped from 773 to 289, which is a huge fall.

So, it was felt that it’s better not to drop observations any further at an overall level of the data as this is leading to drop in the number of observations for columns which are very useful for the analysis. Thereafter, observations were dropped only in subsets created for analysis to answer the research questions, based on the necessity to drop them.

To answer the first research question, a count plot of the year column was plotted, followed by a line plot of the PIC flight time total hours and then a count plot of flight phase. After that, all the plots were compared. Thereafter, subsets of the data were created on different bases, which is another step of data derivation.

A copy of data set was created to be used for doing analysis to answer the remaining research questions.

For analysis to answer the second research question a subset of the copy data was created with location as the basis and event airport was selected as the key column that would determine Lee answer. As the analysis was to be based on the key column, dropping the observations with missing values in it was to be done. But this may lead to further reduction in the observations of the severe column, which was necessary for this analysis. So, instead of dropping missing values, further subsets of data were created into severe and not severe categories and then the analysis was done in these two categories. Determination of the frequency of values in other columns was used to support findings of the analysis.

Similar was the case for analysis to answer the third research question and hence the same procedure was followed.

For analysis to answer the fourth, fifth and sixth research questions, subsets of copy data were created with operator, flight type and engine as the bases, respectively. Unlike the above two cases, the subset data for these three cases didn't contain the severe column and hence the observations with missing values in the key column world dropped before the analysis. Thereafter, methods similar to above two cases were used.

V. FINDINGS AND REFLECTIONS

1. Overall trend of air incidents

Analysing the count plot of years showed that there was a fall in the number of air incidents over the years, although there were some intermittent surges, especially between 1986 and 1991. A comparison with flight time of the pilots indicated that there was a rise in the number of incidents whenever there was a fall in the flight time of the pilots. The count plot of flight phase displayed that level off touchdown was the phase in which most incidents happened, which is a part of landing procedure and may depend on pilot experience. Checking the pilot flight time for this particular phase revealed that 75 percentile of pilot flight time for this case was less than the mean pilot flight time in overall.

A similar analysis for the period between 1978 and 1980, which was the peak of air incidents in the given, presented results which were similar to the above observations.

For similar analysis in the period of 1986 to 1991, which was the period of highest surge in air incidents after the peak period, had a different result with roll out fixed wing as the flight phase in most incidents. But again, this phase is also a part of landing procedure [14]. The pilot flight time for this period was very similar to the above observations.

With all these, it can be concluded that the pilot flight time could be one of the key factors behind the overall trend of number of air incidents, although it should be noted that this might be due to the influence of the observations for the periods of the peak of air incidents and the highest surge of air incidents.

2. Location basis

The location-based analysis revealed that Denver International was the airport with highest incidents for the not severe category and Reno Stead was the airport with highest incidents for the severe category. For the not severe category, analysis of other columns indicated results somewhat similar to that of the overall trend above, but from an airport point of view this may also imply the need of attention towards infrastructure and maintenance of the airport. For the severe category, the analysis of other columns did not produce any clear results.

3. Aircraft basis

The aircraft-based analysis revealed that 172 and 182 were the models with highest incidents for the not severe and severe categories respectively. Interestingly, both these models were of the Cessna make. From the analysis of other columns for both not severe category and severe category, it can be concluded that the aircraft make is clearly the one of the factors for the models being unsafe and although not as clear as the aircraft make, the flight type could be another factor.

4. Operator basis

The operator-based analysis revealed that Delta Airlines Inc. was the operator with highest incidents. From the analysis of other columns for this operator, it can be concluded that the choice of aircraft make could be one of the possible factors behind this.

5. Flight type basis

The flight type based analysis revealed that personal was clearly the flight type with highest incidents. From the analysis of other columns for this flight type, it can be concluded that the pilot flight time could be one of the factors behind this.

6. Engine basis

The engine-based analysis revealed that 0235L2C, an LYC make, was the engine with highest incidents. From the analysis of other columns of this engine, it can be concluded that engine make could be one of the factors behind this.

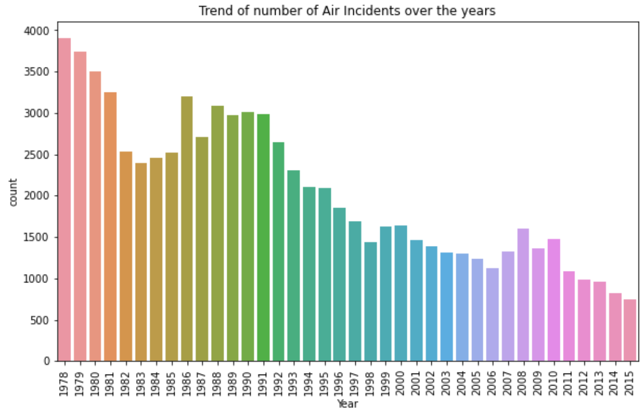

Figure 1: Trend of number of Air Incidents

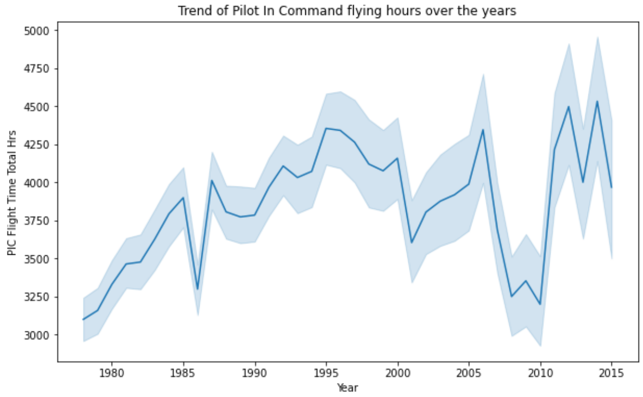

Figure 2: Trend if flying hours of Pilot in Command

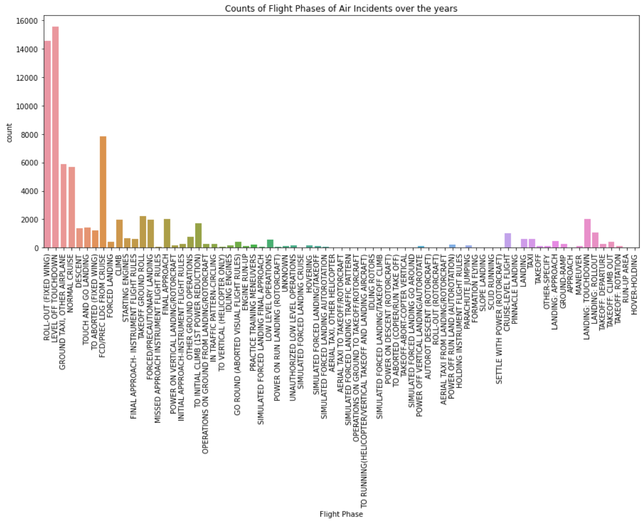

Figure 3: Counts of Flight Phases of Air Incidents

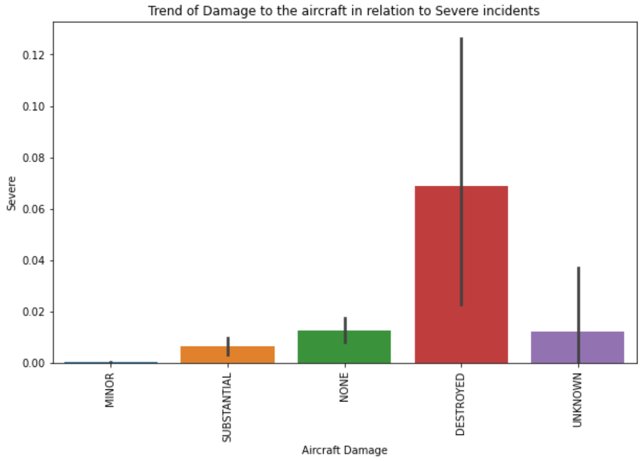

Figure 4: Trend of Damage to Aircraft with severity of incident

TABLE 1

Table of Word Counts of different sections		
Section No.	***Section Title***	***No. of Words***
	Abstract	102
I	Introduction	255
II	Data	240
III	Analytical Questions	300
IV	Analysis	946
V	Findings and Reflections	595

REFERENCES

[1] C.-C. Chen and J. Petrick, “Health and Wellness Benefits of Travel Experiences A Literature Review,” *J. Travel Res.*, vol. 52, pp. 709–719, Nov. 2013, doi: 10.1177/0047287513496477.

[2] “The Five Fastest-Growing Transportation Industries, and the Stories Behind Their Rise,” *ZipRecruiter*, Mar. 02, 2019. https://www.ziprecruiter.com/blog/transportation-industry-spotlight/ (accessed Dec. 21, 2021).

[3] “Future of Aviation.” https://www.icao.int/Meetings/FutureOfAviation/Pages/default.aspx (accessed Dec. 21, 2021).

[4] ICAO, “Effects of Novel Coronavirus (COVID-19) on Civil Aviation: Economic Impact Analysis.” [Online]. Available: https://www.icao.int/sustainability/Documents/Covid-19/ICAO_coronavirus_Econ_Impact.pdf

[5] C. V. Oster, J. S. Strong, and C. K. Zorn, “Analyzing aviation safety: Problems, challenges, opportunities,” *Res. Transp. Econ.*, vol. 43, no. 1, pp. 148–164, Jul. 2013, doi: 10.1016/j.retrec.2012.12.001.

[6] A. E. E. Eltoukhy, Z. X. Wang, F. T. S. Chan, and X. Fu, “Data analytics in managing aircraft routing and maintenance staffing with price competition by a Stackelberg-Nash game model,” *Transp. Res. Part E Logist. Transp. Rev.*, vol. 122, pp. 143–168, Feb. 2019, doi: 10.1016/j.tre.2018.12.002.

[7] IATA, “DATA SCIENCE HYPE OR RIPE FOR AVIATION?,” Jun. 2019. https://www.iata.org/contentassets/d72edc56c3814aac8ec508fdf8555a52/data-science-hype-or-ripe-for-aviation-white-paper.pdf

[8] “How Big Data and AI can be deployed for better aviation safety | SafeClouds.eu Project | Results in brief | H2020 | CORDIS | European Commission.” https://cordis.europa.eu/article/id/300588-how-big-data-and-ai-can-be-deployed-for-better-aviation-safety (accessed Dec. 21, 2021).

[9] “Aviation Accidents and Incidents (NTSB, FAA, WAAS).” https://kaggle.com/prathamsharma123/aviation-accidents-and-incidents-ntsb-faa-waas (accessed Dec. 22, 2021).

[10] “Data & Information.” https://www.asias.faa.gov/apex/f?p=100:2:::NO::: (accessed Dec. 22, 2021).

[11] "Pilot in command | Civil Aviation Safety Authority." https://vfrg.casa.gov.au/general/pilot-responsibilities/pilot-in-command/ (accessed Dec. 21, 2021).

[12] H. Lee *et al.*, "Critical Parameter Identification for Safety Events in Commercial Aviation Using Machine Learning," *Aerospace*, vol. 7, no. 6, Art. no. 6, Jun. 2020, doi: 10.3390/aerospace7060073.

[13] "scikit-learn : Data Preprocessing I - Missing/categorical data - 2020." https://www.bogotobogo.com/python/scikit-learn/scikit_machine_learning_Data_Preprocessing-Missing-Data-Categorical-Data.php (accessed Dec. 22, 2021).

[14] ECCAIRS 4.2.8, "ECCAIRS 4.2.8 Data Definition Standard Event phases." [Online]. Available: https://skybrary.aero/sites/default/files/bookshelf/1814.pdf