Visual Analysis of Aircraft Crashes and Fatalities in the History of Modern Aviation

Alessandro Alviani
School of Mathematics, Computer Science and Engineering
City, University of London

Abstract – In the last 50 years there has been a sharp decline in fatal aviation accidents. The year 2019 recorded 14 crashes over 36 million flights and 2020 eight accidents over 10.6 million. Although these figures make air travel the safest form of transport, there is still a general fear of flying and every year aircraft operators strive to make this industry even safer.
There is only one published paper that analyses the causes of aircraft crashes. Given the importance of the subject, the presented work proposes the study of a large collection of aircraft crashes with the aim to understand when, where, and how the reported accidents have occurred. The work is carried out applying principles of Data Science and state-of-the-art Visual Analytics (VA) techniques to the Aircraft Crashes (AC) dataset.
The data is mined with Python and visualized with Tableau, Matplotlib, Seaborn, WordCloud, Altair, Networkkx and pyLDAvis.

1. PROBLEM STATEMENT

This work studies an extensive collection of crashes (5008) occurred in modern aviation history (1908 – present), where the word "aviation" refers to the air travel of military and domestic aircraft [1]. The aim is to identify where, when and why the aircraft crashed, and ultimately answer the following questions:

1. What are the geographical areas where the most fatalities happened?
2. How do the number of fatalities change over the years?
3. What are the main reasons for crashes and fatalities?

The dataset is found on Kaggle and is originally scraped from PlanetCrashInfo.com. Despite the source being non-official, it is often referenced [2, 3, 4] and the data leads to findings in line with this paper's [5] and this source's [6]. However, being a private collection of crashes, this should be considered general interest information, and so the study's findings.
The dataset has 5008 rows and 17 columns which include information on dates, times, places, fatalities and a brief description of each accident. The data is relatively clean and contains small percentages of missing values in its key features.
Although the dataset requires extensive preparation and feature engineering, the presence of geographical data, time series, and text make this dataset suitable for visual analysis and to answer the study's questions.

2. STATE OF THE ART

The first paper reviewed is "Visual Analysis of Dynamic Networks with Geological Clustering" [7], presented in 2007 at the IEEE Symposium on Visual Analytics Science and Technology. This work compares the performance of National Football Teams (NFT) in the history of World Cups, which is essentially the study of performance in a network characterized by geological information. The authors present two complementing models with different levels of granularity. The NFT (nodes) are plotted circularly in space with their position and size determined by their measure of centrality (performance).

In one of the two models, to avoid over cluttering, the NFT are discretized into four groups: "winner", "strong", "medium", "weak".
Visualizations are interactive and generated with their custom-made software.
Although applied to a different domain, this paper inspires the presented work in five ways:

1. The National Football Teams can be thought as "countries" and their "level of performance" as "number of fatalities."
2. Similarly, the clustering applied on "level of performance" can be applied to the "number of fatalities".
3. Different levels of detail (they provide two) can offer different insights.
4. Interactivity can help extract information from networks that tend to grow very large and chaotic (i.e., the ability to zoom in and out and hover over elements).
5. Networks seem to be an effective way to link countries and causes of accidents.

The second paper, "Topic- and Time-Oriented Visual Text Analysis" [8], published in 2016 by the IEEE Computer Society, studies text. Due to its complex logic, intrinsic noise and high dimensionality, text can be a challenging data source to interpret for humans and machines. This paper presents the benefits of text analysis through interactive visualization that:

1. Define topics.
2. Represent them in time.

In topic modeling, the text is first processed, then clustered in similar topics and ultimately displayed in space. The authors suggest a four-stage iterative process where the user collects, mines, visualizes and interprets the text. The model's output is often inconclusive and human reasoning is required to define the topics or further tune the algorithm to generate more insightful ones. Once the topics are created they can be plotted on timelines and analyzed in a time-oriented way.
The authors propose compelling visualizations created with TIARA, TextPioneer and their custom-made software ThemeRiver.

Inspired by this paper, the presented work performs topic modeling on the AC dataset and uses pyLDAvis [9] to visualize the key-words. The obtained topics are subsequently plotted on a timeline (figure 10) to display how they contribute to the number of fatalities over time.

“Knowledge Generation Model for Visual Analytics” [10] is an insightful paper on interactions between human and computer, where computer is defined as the set of data, model and visualization. The authors suggest a model of knowledge generation that revolves around visualizations and human thinking.
This study’s workflow is inspired by this paper, as showed in diagram 1 in section 4.

3. PROPERTIES OF DATA

The dataset is originally composed of 17 columns and 5008 rows. Each row defines a plane crash, and the columns are its defining features. A summary of the dataset is presented in Table 1:

Table 1: Dataset summary.

Feature	Explanation	Format	Range
Date	Date of accident	yyyy:mm:dd	1908 - 2021
Time	Local time	hh:mm	24h
Location	Accident location	City, Country	Worldwide
Route	Accident route	City - City	Worldwide
AC Type	Aircraft type	String	311
Operator	Aircraft operator	String	2267
All Aboard	Total aboard (crew + passengers)	Integer	1 - 644
Passengers Aboard	Passengers aboard	Integer	0 - 614
Crew Aboard	Crew aboard	Integer	1 - 83
All fatalities	Fatalities (crew + passengers)	Integer	0 - 583
Passenger fatalities	Total Passenger fatalities	Integer	0 - 560
Crew fatalities	Total Crew fatalities	Integer	0 - 43
Summary	Brief description of the accident	Strings	0 – 2670 (mean: 220)

The features *“aircraft serial numbers”*, *“ground”*, *“registrations numbers”* and *“case id”* are left out from the analysis as non-relevant.

The initial exploratory visual analysis (EVA) is carried out in Tableau and Python and focus on the study of the outliers, feature distributions and missing values. Figure 1a presents an interactive bubble chart that provides information on each crash. Countries are sorted alphabetically: starting with “A” are placed in the middle and starting with ”Z” are distributed along the perimeter of the sphere (e.g., Australia is at the center, Zimbabwe on the border). In this chart the outliers stand out thanks to their size and color and, when hovered over, provide the user with a wealth of additional information on the crash (i.e., date, country, route, number of fatalities, crew fatalities and passenger fatalities). The bubble chart is coupled to the Raincloud (Micah Allen [11]) plot shown in Figure 1b., which combines a kernel density estimate plot (KDE) to a boxplot and the jittered raw data. This plot displays feature distribution, outliers and robust statistics (i.e., median and percentiles).
The two plots, working on different levels of granularity, complement each other and, together with the analysis of the outliers, provide additional wealth of information.

Figure 1a: “Bubble Chart” **Figure 1b:** “Rainbow Chart”

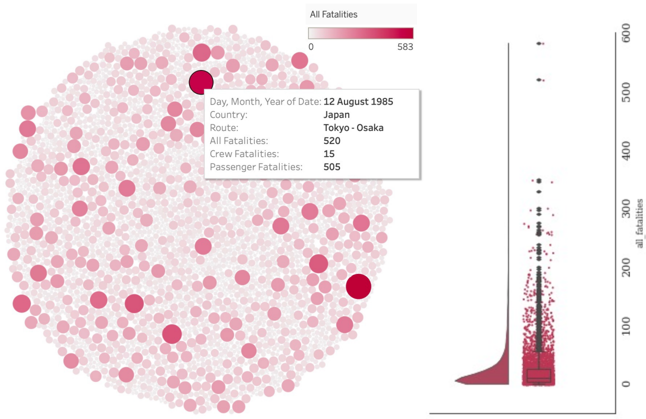

Figure 1: Outliers analysis.

The following visual analysis, on the missing values, is inspired by Sara Johansson Fernstad and Robert C Glen [12] who discuss the importance of extracting knowledge from this kind of data. In this study this is done by combining two visualization methods: histogram and matrix. The histogram shows the number (or percentage) of missing values for each feature but fails to provide information on distributions and relations. The nullity matrix (figure 2), implemented with the "Missingno" python library [13], shows how the missing values occur in time, how they relate to each other, and the quality of data completeness. The observations are ordered chronologically and a sparkline on the right summarizes the level of missingness over time. It is easy to observe that the features *"time"* and *"route"* have the most missing values, although not centered around any particular time. Data on passengers and crew fatalities is often missing, and whenever information on their presence aboard misses too. However, the feature *"all_fatalities"* suggests that data on total fatalities is always reported. This is explained with military aircraft’s fatalities.

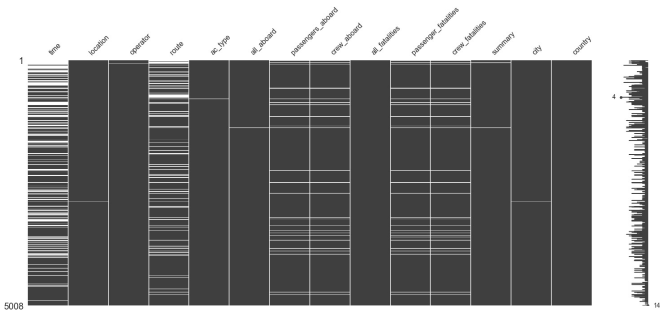

Figure 2: Nullity matrix.

4. ANALYSIS

A. Analysis Approach

Diagram 1 presents the workflow in the analysis of the Airplane Crashes (AC) dataset.

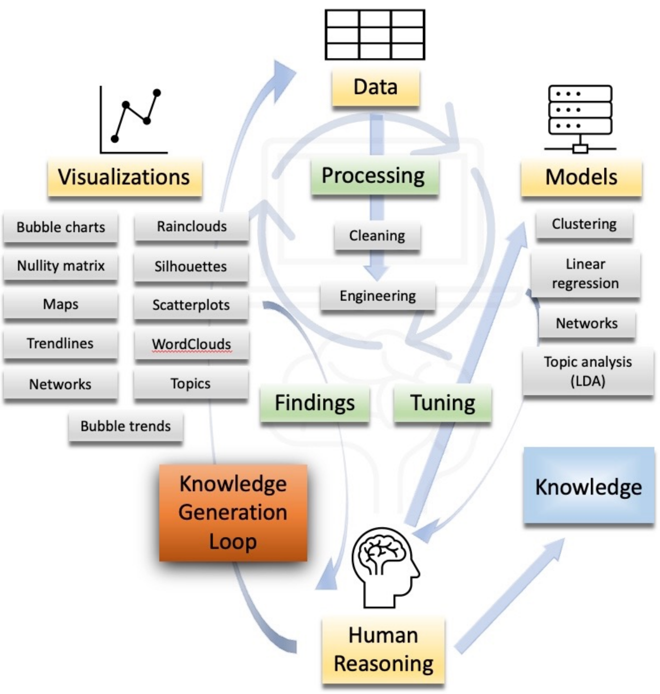

Diagram 1: Analysis workflow.

In this model, knowledge generation is a tight interaction between human and machine. Through models and visualizations, the user extracts information from the data, which is interpreted and flowed back into the system. This process allows an iterative refining of data and models, and results in a knowledge generation loop driven by human reasoning and supported by the machine.
The ratio of models and visualizations can vary depending on the dataset and task. In this study, as the analysis progresses, the models become more complex and require growing support from the visualizations.
The diagram presents sub-loops nested in bigger loops. These can be imagined as specific tasks that generate particular knowledge (findings). This knowledge is either flowed back into the system and refined to execute further tasks or presented as conclusive.
The whole process is driven by human reasoning, in an attempt to answer questions from the data.

In this workflow the dataset is first imported and then explored through models and visualizations. Initial models are as simple as descriptive summary statistics (i.e., mean, median, maximum, minimum and length) and provide a high-level understanding of boundaries and measures of centrality in the data. Then features are plotted to display outliers (figure 1a), distributions (figure 1b) and missing values (figure 2). The visual analysis suggests if the outliers are informative, where missing values occur, and which additional features need to be imputed and from which variables.
Features like *"time"* and *"location"* are cleaned and wrangled into a suitable format; lastly, new features (e.g., *"country"*) are derived.
The work advances with visualizations aimed at answering the research questions.
When the total fatalities are plotted on a map, it is a challenge to find a satisfying color palette. Silhouettes and scatterplots (figure 3) are used to inform on the best number of clusters and assign a distinguishing color palette based on fatalities and geolocation (figure 4).
Subsequently, when the fatalities are plotted on a timeline (figure 6a), the trendlines appears potentially misleading if not adjusted by the total number of flights per year and signal the need of finding this additional data.
Scatterplots (figure 5) and trendlines (Figure 6b) drive an attempt to impute the missing values in the new dataset sourced (data processing loop).
The same plots are later used to reject the imputation attempt and abandon the plan to calculate ratios on fatalities before 1970.
WordClouds (figure 7) are generated to clean the "*summary*" text, display the most crashed aircrafts and start the research on leading causes of crashes.
Interactive networks (figure 8) are used to display links between the causes of crashes and the countries and inform the LDA on the presence of similar topics.
Topic modeling, performed with latent Dirichlet allocation (LDA), allows to order similar crash causes in topics (figure 9), and finally to display them on a timeline (figure 10).

B. Analysis Process

After the exploratory visual analysis (EVA) and initial data preparation, the first research question is:

1. *What are the geographical areas where the most fatalities happened?*

This question is answered by plotting all the fatalities, grouped by country, on a map.
The challenge is given by extracting geographical coordinates from the feature *"location"*, which is provided in a problematic format (City, Country), moreover with several unknown cities and countries that need to be transformed into Tableau compatible IDs. This is resolved with some wrangling and each crash is successfully mapped to longitude and latitude.
After being aggregated by country, the total numbers of fatalities are plotted on a map. However, countries' fatalities range from tens to thousands and make finding an expressive range of colors difficult.

Logging the fatalities generates a much smaller range of colors but fails to display significant differences in numbers of fatalities.
A satisfactory color palette is found by clustering the countries per number of fatalities and considering their geography. The clustering algorithm is k-means; silhouettes and scatterplots are combined to identify the best number of clusters (Figure 3) as presented by T. Li et al.[14].

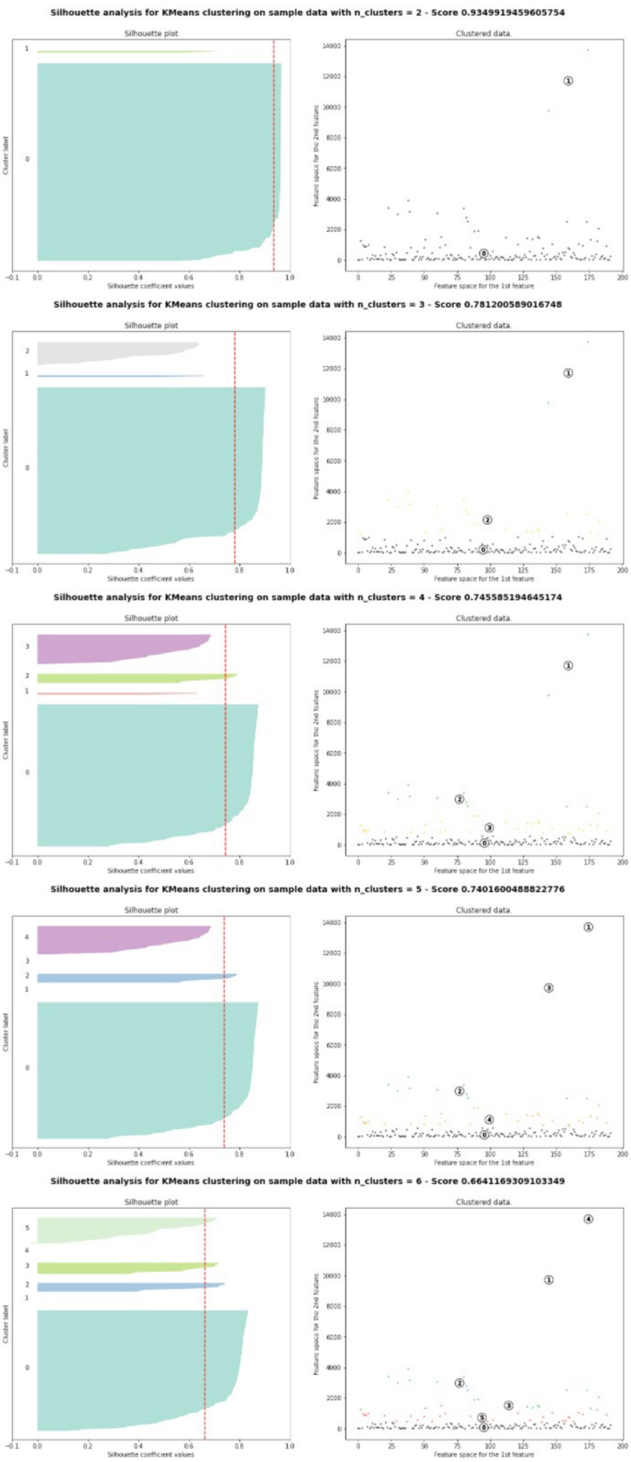

Figure 3: K-means silhouettes and scatterplots.

While the computational model suggests two as the best number of clusters (score: 0.93), the silhouettes and scatterplots offer a better way to interpret those clusters.
The silhouettes show the separation distance and the size of the clusters, the scatterplots how the points (countries) are assigned to each cluster
Looking at the scatterplots, the two outliers (top-right) are USA and Russia, which are geographically very distant. Along the x-axis there is the rest of the world, with much fewer fatalities.
Five clusters are chosen, due to geography, color separation and convenience in the number of clusters to base the following analysis on.
The number of fatalities is plotted on each country with size given by the cluster's number (Figure 4) .
The map shows a prevalence of fatalities in USA (13,968) and Russia (9,762). Canada, Brazil, and Southeast Asia have a medium level of fatalities; Southeast Africa, Scandinavian Countries, and Greenland present the least fatalities.

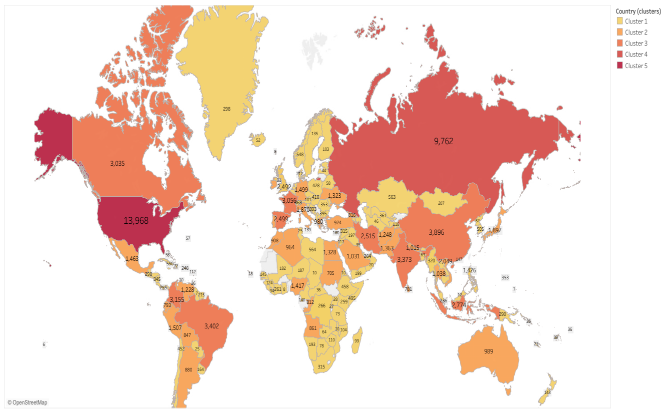

Figure 4: Number of aircraft fatalities per country.

2. *How do the number of fatalities change over the years?*

When the number of total fatalities is plotted on a timeline it becomes evident that this could be misleading, or only partially informative, if not adjusted by the number of total flights per year.
An additional dataset reporting the total number of flights per year is found but it only includes data since 1970, while the AC dataset starts in 1908.
Figure 5 shows an attempt to impute those missing values through the combined use of visual analysis and computational models.
The original dataset (fig. 5a) shows an exponential trend and an outlier in 2020, attributed to the reduced flights due to the pandemic. The curve is logged (fig. 5b) to achieve linearity (partly obtained) and use a linear regression model to impute the missing values. The ordinary least-squares (OLS) model is fitted with (5c) and without the outlier (5d) in 2020. The latter model is chosen due to the outlier slightly tilting the line to the wrong direction.
The results of this imputation are presented on the logged distribution (5e) and on the original curve (5f). The year 2020 is added back into the final data.
The result seems reasonable, but there is an area of high concern indicated by the red boxes.
This is further studied in Figure 6a, that shows the number of fatalities per year adjusted by the estimated and actual

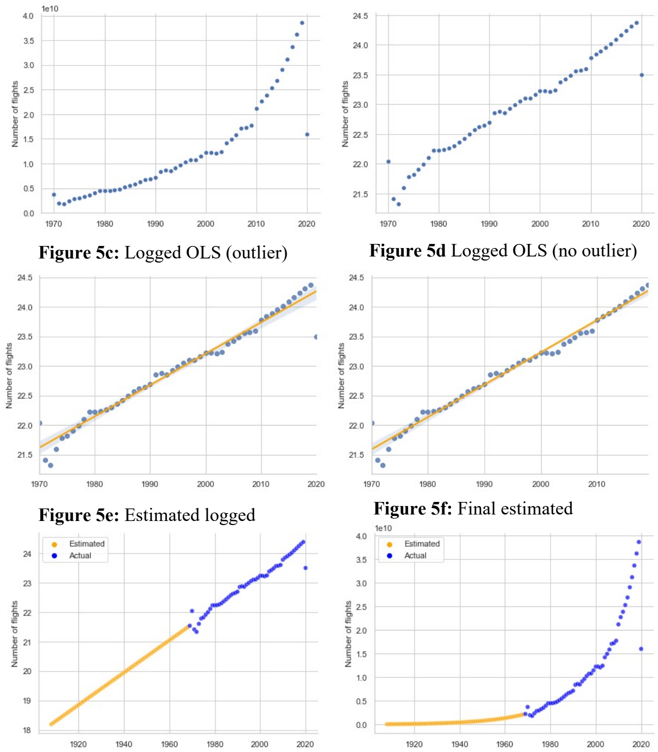

Figure 5a: Number of flights **Figure 5b:** Logged number of flights **Figure 5c:** Logged OLS (outlier) **Figure 5d** Logged OLS (no outlier) **Figure 5e:** Estimated logged **Figure 5f:** Final estimated

Figure 5: Missing values imputation (linear regression).

number of flights. The difference in trends before 1970 and afterward leaves severe doubts on the effectiveness of the previous linear regression, especially on early dates, before 1945. On the one hand it would seem reasonable to think that the number of fatalities per flight has been steadily going down due to the advancements in technology; on the other early days might have seen an initial peak of fatalities due to the increased number of flights and aircraft capacity.

It was impossible to find external data to validate either hypothesis and the imputation of dates before 1970 was abandoned as not robust enough, although fascinating.

Figure 6b shows the absolute number of fatalities per year and splits the aircraft in domestic and military, which is a feature engineered from *aircraft types* and *summaries*.

This plot shows a general uptrend of domestic fatalities until 1975 and downtrend afterward. However, the reason behind its spiky trend is only clear when this plot is coupled with the bubble chart presented in Plot 1a.

Looking at the spikes in fatalities in the year 1977 and 1985, these correspond to two major outliers: the Boeing crash on route Tenerife - Las Palmas (Spain), which caused 583 victims (highest death toll ever recorded), and the Boeing crash on route Tokyo - Osaka (Japan), which caused 520 fatalities (second biggest toll).

This correlation between high spikes in fatalities and outliers is found throughout the aviation history since World War 2 and leads to conclude that the high numbers of fatalities in those years are primarily the result of big tragedies, rather than a multitude of smaller fatalities.

Military aircraft fatalities follow a different trend: they peak in World War 2 and are on average around 15 times fewer than domestic aircraft fatalities. The two highest military death tolls are recorded in 2003, when 275 Islamic revolution militants crashed into a mountain in Iran and 2018, when an aircraft carrying 257 Algerian military forces caught fire after the take-off.

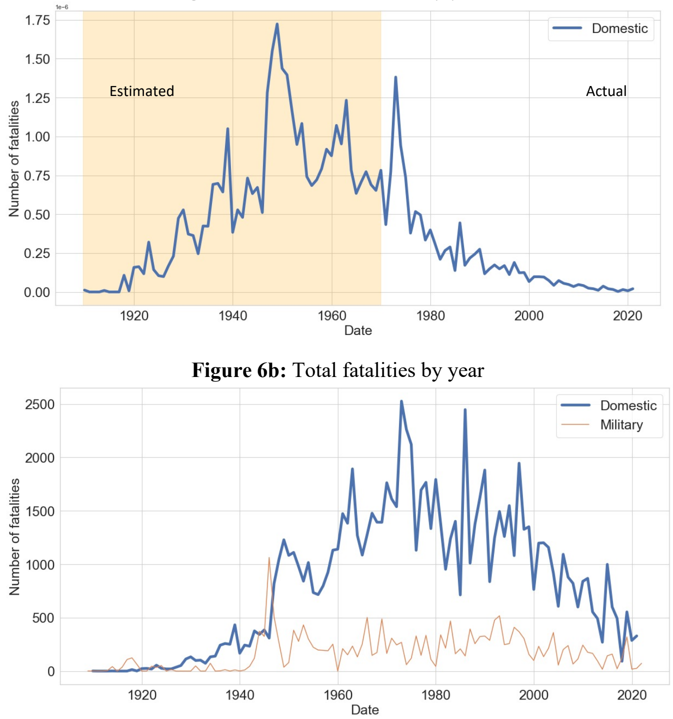

Figure 6a: Relative fatalities by year **Figure 6b:** Total fatalities by year

Figure 6: Fatalities per flight and total number of fatalities

3. *What are the main reasons for crashes and fatalities?*

The strategy to answer this question is to analyze the text in the crash summaries, identify keywords for fatality causes, link them to the countries, and finally study their frequency over time. WordClouds provide a high-level view on the words in a text corpus and it is the first visualization tool used to text analysis. Figure 7a presents a WordCloud of the types of aircraft involved in the most crashes, figure 7b the main causes of crashes.

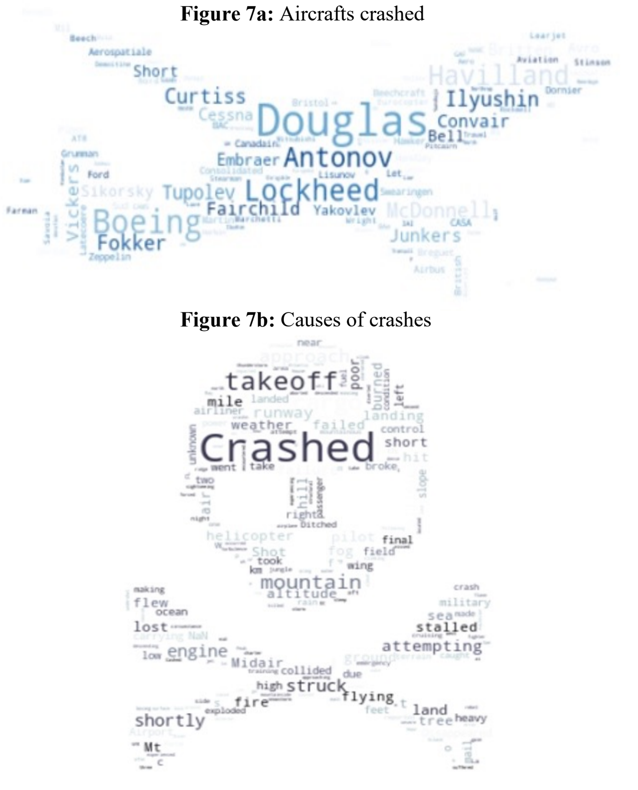

Figure 7a: Aircrafts crashed **Figure 7b:** Causes of crashes

Figure 7: WordClouds for aircrafts and causes of crashes.

Both WordClouds are generated after processing the text.
Although WordClouds help identify prevalent words in a text, they fail to provide much additional information.
To this purpose two interactive networks (Figure 8) are created (Python library Networkkx) [15]. In the first network (figure 8a) the countries are placed in space according to their measure of centrality (number of fatalities) and connected by edges to the five prevalent causes of crashes. Each edge's size is proportional to the number of fatalities for that specific cause.
In the center, countries close together share high fatalities related to similar causes, while fewer fatalities and different causes characterize outer countries.
This network provides a very granular view but needs to be interacted with and explored from close to provide valuable insight.
This limitation is resolved with the network in figure 8b which displays the countries in clusters according to the previous k-means model.
This network is less granular but provides more critical insights at a glance.
Each cluster is connected to ten crashes causes, twice as many as before and much easier to read (e.g., cluster 5 has a prevalence in fatalities due to *"failure"*, but no links to *"taking-off")*.
Each cluster is linked to 10 crash causes, giving 50 possible outcomes, but only 15 are presented, indicating that clusters mostly share the same causes for accidents.
What becomes noticeable from the clustered network is that the causes seem to be either connected (*"engine"*, *"failure"*) or similar to each other (*"approach", "landing", "runway"*).

Figure 8a: Interactive network of all countries and crash causes

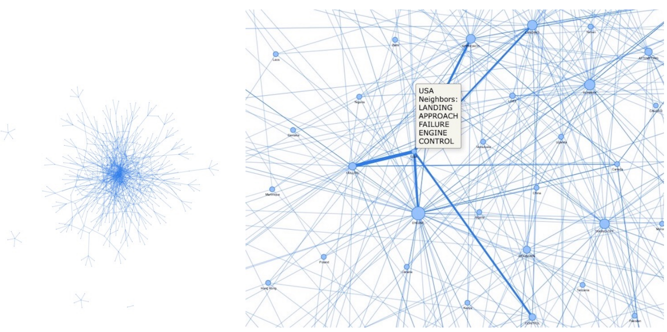

Figure 8b: Interactive network of clustered countries and crash causes

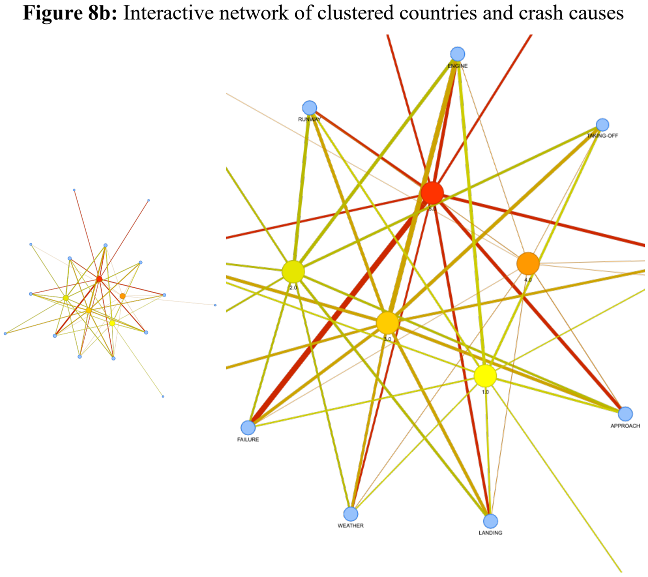

Figure 8: Interactive networks of countries and crash fatalities

Latent Dirichlet allocation (LDA), implemented with the Python package Gensim [16]), aims to resolve this.
LDA is a topic modeling algorithm, also known as the three-layer Bayesian probability model, that assigns each word to a topic based on conditional probabilities calculated on text, topic and feature words.
This model is used in association with WordClouds by Bashri and Kusumaningrum [17]. The model can be displayed as in figure (9), composed of two parts. In the first plot (9a) the words are clustered (topics) and visualized in space. The clusters positioning represents the semantic distance between topics, and the clusters' size the number of words contained.
The second plot (9b) displays the exact words and their number of occurrences in each cluster.
LDA requires the user to assign names to the clusters (topics) and this process needs generating several different models, identifying patterns and re-processing the text.

Figure 9a: LDA – Clusters of words (topics) in space

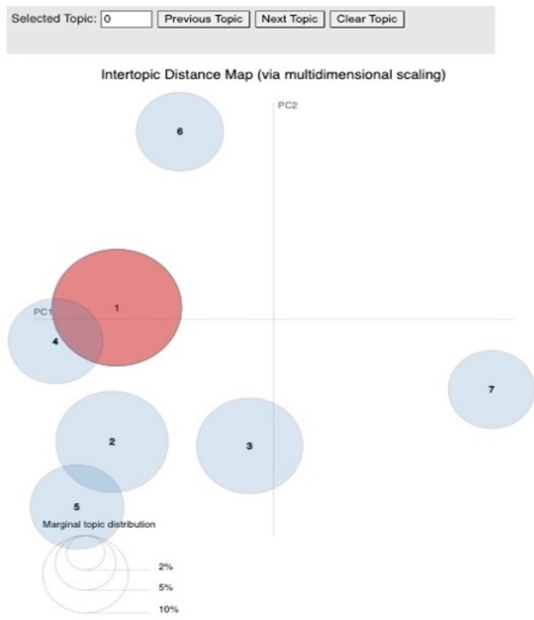

Figure 9b: LDA – Most relevant topics per topic

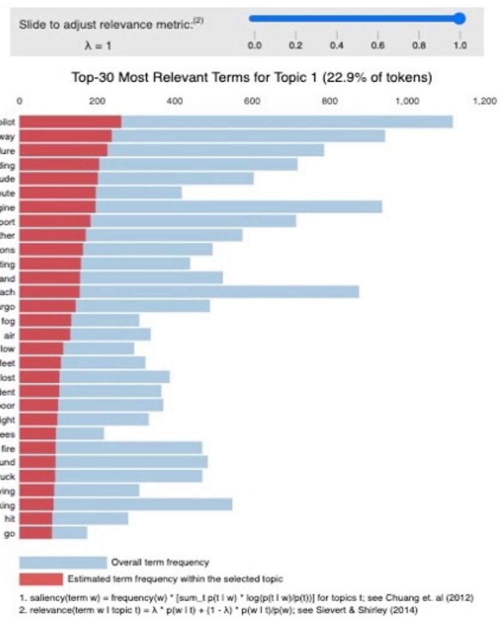

Figure 9: Latent Dirichlet allocation on aircraft crashes.

After several iterations keywords are identified and finally mapped to topics using the following logic:

{"hit": "struck",
"taken-down": "struck",
"shot": "struck",
"bomb**": "struck",
"attack**": "struck",
"explosive": 'struck'}

The topics generated and the corresponding crashes and fatalities are presented in table 2.

Table 2: Cause of crashes.

Topics	Crashes	Fatalities
Taking-off	711	19,973
Pilot (error)	661	15,347
Engine	594	12,178
Landing	498	12,622
Struck	423	9,377
Weather	419	9,393
Mountains	330	7,991

The above topics capture 78% of all the fatalities in the dataset.

C. Analysis Results

This work finds that in the modern history of aviation there are a total of 111,470 fatalities. Military aircraft fatalities (19,536) represent 17% of all fatalities (figure 6b) and are mostly reported in the USA (2,186) and Vietnam (1,399). The USA presents the highest total death toll (13,968), followed by Russia (9,762) (figure 4).
Domestic aircraft fatalities have a mean of 22.3 deaths per crash and are severely affected by accidents on big planes (e.g., Boeing), resulting in up to 583 fatalities per crash (figure 1).
The military aircraft Douglas (992) and the domestic Boeing (410) count the most crashes.
The leading causes of fatalities are listed in figure 7, detailed in Table 2 and displayed in time (figure 10).
Pilot errors, especially during the phases of landing and taking off, engine failures and adverse weather conditions are some of the prevalent causes.
Although the ratio of fatalities per flight has followed a steady downtrend since 1970 (figure 6a, 10b), some preventable fatalities still occur.
However, when looking at fatalities per number of flights, these appear to be minimal (< 1/1,000,0000).

Figure 10a: Main causes of total fatalities since 1970

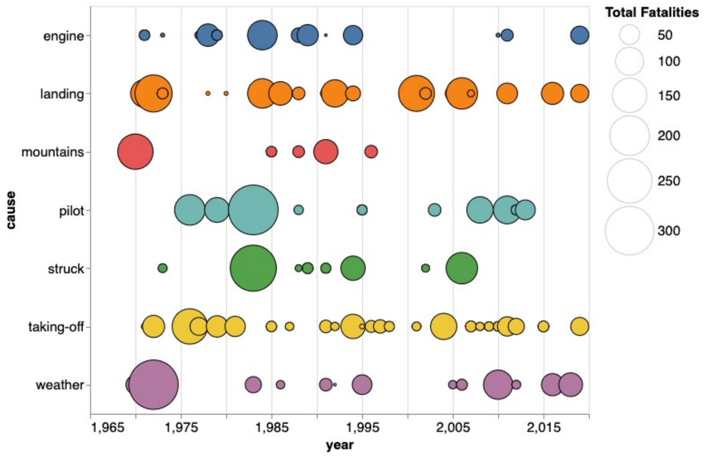

Figure 10b: Main causes of relative fatalities since 1970

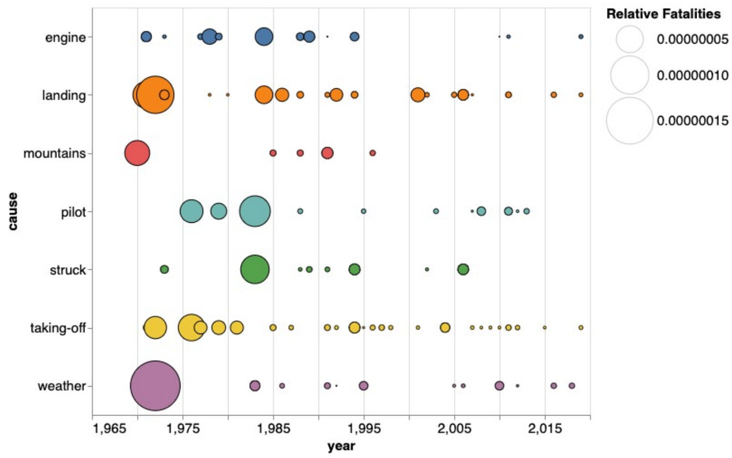

Figure 10: Total fatalities and adjusted fatalities

5. CRITICAL REFLECTION

The presented work was one of the first studies in aviation fatalities and follows M. Stephens' and W. Ukpere's paper [5], the first publication to study correlations between causes and locations of aircraft fatalities. As shown in their study, that uses data collected from different sources, finding historical data on this domain was complex. This limited the opportunity to compare and validate the findings.
Particularly, in this work the datasets came from non-official sources, exposing the analysis' findings to imperfections in the data. Limitations in the data also included the missing values for total flights before 1970 that constrained computing relative statistics on only the last 50 years.
Regarding methods and visualizations, the interactive "bubble chart" used for the analysis of the outliers (figure 1) offered valuable insight into the rest of the analysis and this study recommends its use over simpler forms of scatterplots.
Similarly, the nullity matrix (figure 2) provided important information on missing values that histograms alone could not. The combined use of these two plots is advisable in the analysis of missing values.
In geographical data processing, independent islands (e.g., Sealand), cities (e.g., Vatican City), sea crashes, and

military zones were assigned to the nearest country, generating approximation in geo-localization. More in-depth works may consider transforming those locations to longitude and latitude to achieve a more specific mapping.
When using silhouettes, elbow plots and scatterplots (figure 3), computational methods and visualizations often disagreed on the number of clusters. Silhouettes and scatterplots were more consistent than the elbow plots, but ultimately none of those visualizations was exhaustive and human reasoning was prevalent in making the final decision.
It is worth noting that shifting the analysis on the continents would have offered a natural clustering and certainly an interesting angle on the topic, as approached in M. Stephens' and W. Ukpere's paper [5].
This work used a continuous color palette to generate a progressive risk scale for each country (figure 4); a diverging color palette could have been used instead to increase the cluster's visual separation.
Scatterplots were a valuable tool to inform linear and polynomial regressions; this work provides a practical framework to implement them (figure 5).
The work on total and relative fatalities (figure 6 ,10) reminded of the importance of the metric used to represent a phenomenon, where total and relative metrics can tell a different story (figure 10a, 10b).
Networks (figure 8) were a valuable tool to explore links between countries and fatalities but, in domains with a high level of complexity, they either require interactive visualizations or important simplifications to be easily interpretable.
Although this study classifies the majority of all aircraft crashes (78%), further work on text analysis and topics modeling would have improved the number of classified crashes and the detail of topics generated.
Future studies may consider following that route to answer more specific questions on crashes and fatalities.

REFERENCES

[1] https://en.wikipedia.org/wiki/History_of_aviation
[2]https://en.wikipedia.org/wiki/List_of_aircraft_accidents_and_incidents_by_number_of_ground_fatalities
[3] https://aviation-safety.net/database/
[4] https://www.baaa-acro.com/
[5] Stephens, M. S., & Ukpere, W. I. (2014). An Empirical Analysis of the Causes of Air Crashes from a Transport Management Perspective. Mediterranean Journal of Social Sciences, 5(2), 699.
[6]https://www.statista.com/statistics/262867/fatal-civil-airliner-accidents-since-1945-by-country-and-region/
[7] A. Ahmed, X. Fu, S. Hong, Q. H. Nguyen and K. Xu, "Visual Analysis of Dynamic Networks with Geological Clustering," 2007 IEEE Symposium on Visual Analytics Science and Technology, 2007, pp. 221-222, doi: 10.1109/VAST.2007.4389027.
[8] W. Dou and S. Liu, "Topic- and Time-Oriented Visual Text Analysis," in IEEE Computer Graphics and Applications, vol. 36, no. 4, pp. 8-13, July-Aug. 2016, doi: 10.1109/MCG.2016.73.
[9] https://pyldavis.readthedocs.io/
[10] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis and D. A. Keim, "Knowledge Generation Model for Visual Analytics," in IEEE Transactions on Visualization and Computer Graphics, vol. 20, no. 12, pp. 1604-1613, 31 Dec. 2014, doi: 10.1109/TVCG.2014.2346481.
[11] Allen M, Poggiali D, Whitaker K et al. Raincloud plots: a multi-platform tool for robust data visualization [version 2; peer review: 2 approved]. Wellcome Open Res 2021, 4:63
[12] S. J. Fernstad and R. C. Glen, "Visual analysis of missing data — To see what isn't there," 2014 IEEE Conference on Visual Analytics Science and Technology (VAST), 2014, pp. 249-250, doi: 10.1109/VAST.2014.7042514.
[13] https://github.com/ResidentMario/missingno
[14] T. Li, P. Ye and S. Zheng, "State grid office system user clustering analysis based on K-means algorithm," 2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA), 2018, pp. 438-442, doi: 10.1109/ICBDA.2018.8367723.
[15] https://networkx.org/
[16] https://radimrehurek.com/gensim/
[17] M. F. A. Bashri and R. Kusumaningrum, "Sentiment analysis using Latent Dirichlet Allocation and topic polarity wordcloud visualization," 2017 5th International Conference on Information and Communication Technology (ICoIC7), 2017, pp. 1-5, doi: 10.1109/ICoICT.2017.8074651.

Word Count		
Abstract	152	- 48
Problem Statement	215	- 35
State of the Art	491	- 9
Property of Data	436	- 64
Analysis Approach	493	-7
Analysis Process	1442	- 58
Analysis Results	183	-17
Critical Reflection	482	-18
Figures	10	0
Diagram	1	0