Airplane Crash Analysis

1908 – 2009

GADANSKI Aleksandar, GIZATULLIN Timur, SHAKIRZIANOV Iusuf, SHATOKHIN Anton, STEFANOVIC Zlata

A paper presented for the GE2324 Art and Science of Data course.

Department of Computer Science
City University of Hong Kong
Hong Kong
March 2024

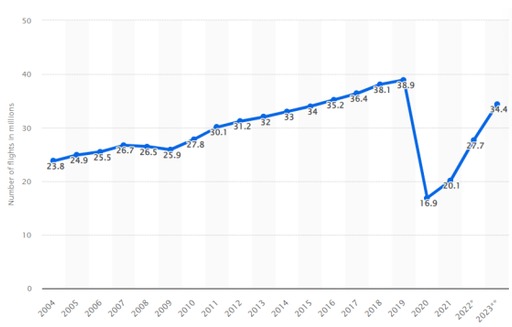
Contents

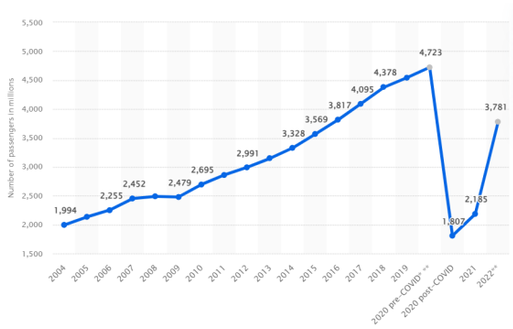
Abstract

Aviation has become a common part of humans' lives since the first airplane was created. In 2019, before the COVID-19 pandemic started, more than 106,000 flights [1] and more than 12 million passengers [2] were operated daily. Airplanes are treated as the safest transport type, but air crashes still happen, and in the worst cases, they can lead to hundreds of deaths. In this paper, we will analyze the data about airplane crashes while providing some main insights and tendencies, such as the average number of deaths per crash, the most dangerous airlines, and the most popular causes of crashes.

1 Introduction

In the past decades, airplanes have become much more popular because of their time efficiency and safety. That is why it can be clearly seen that the number of flights per year, as well as the number of passengers, had a positive tendency up to 2020 when the COVID-19 pandemic occurred, and the number of flights dropped significantly.





One of the key reasons for their safety is that airplanes have become much more digitalized and computerized. It is believed that airplanes are the safest transport type as the prevailing part of work is done by computers. Interactions between an airplane and a human tend to be zero as it is believed that it will prevent any human-factor mistakes and reduce the number of crashes.

However, fatalities still happen. The reasons for these fatalities may range from metal corrosion to terrorist attacks. But definitely, reasons depend on many aspects. One of the most important of them is time. For instance, it is pretty obvious that during wartime, there will be many more crashes, mostly in the military forces, and their causes will be related to the war. That is why it is important to consider many different aspects and get the most relevant information.

1.1 Research Target

The aim of this research is to analyze the number of crashes and deaths per year caused by aircraft accidents, indicate the tendency of these numbers, and figure out the causes of peaks that occurred. Another goal is to analyze airlines and aircraft models and compare them with each other. And the last part is to find the most popular cases of air crashes.

2 General analysis of flight data

The first interesting insights about airplane crashes are the places where they happened, the number of crashes, and the number of deaths by year as a result of air crashes.

2.1 Map visualization

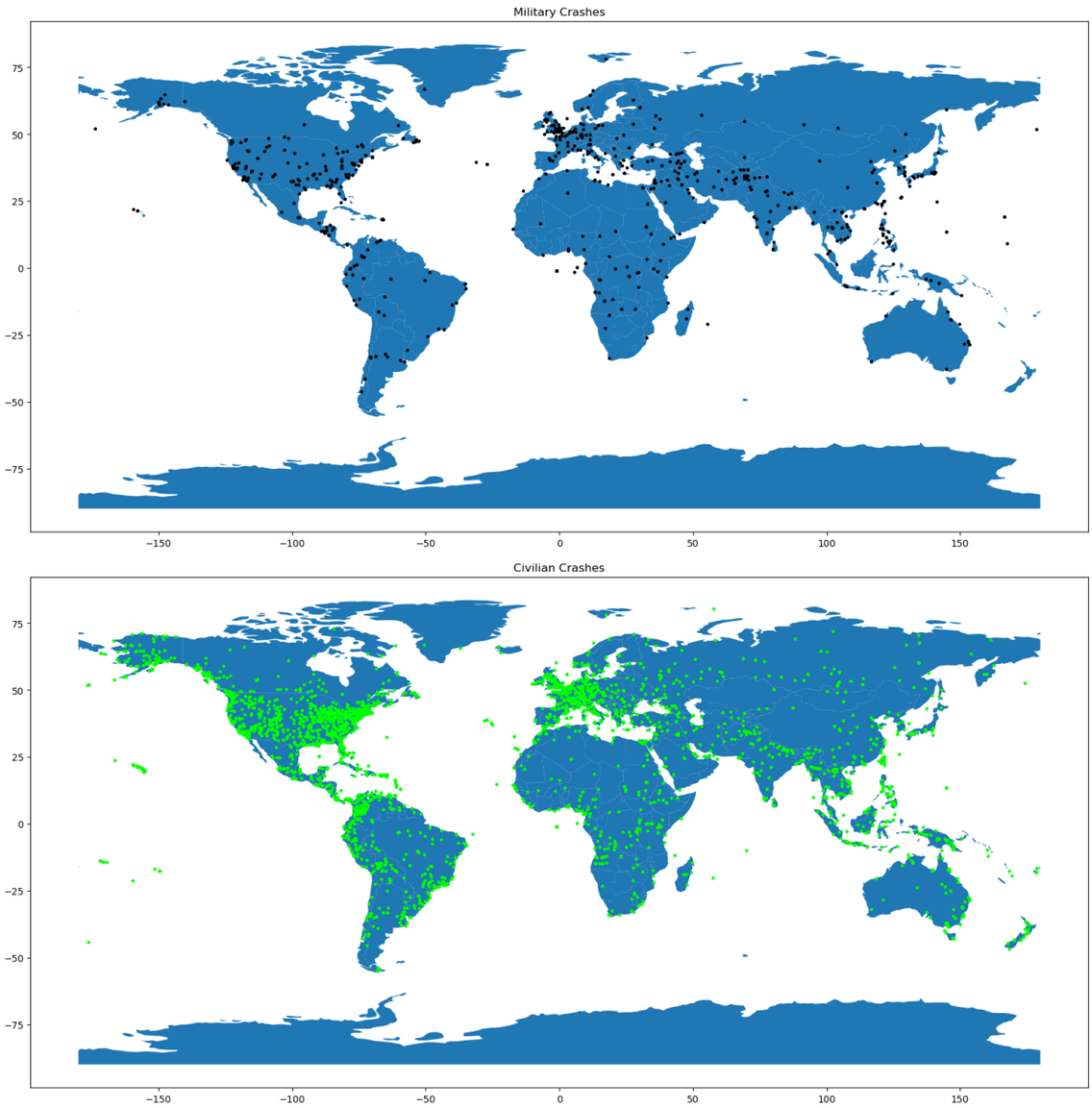
In this part, we will look into the geographical distribution of crashes and how they are spread across the globe.

2.1.1 Methodologies

Geopy.geocoders' Nominatim was used to locate the crashes on the map. One of the biggest problems was spelling inconsistency in the "Location" column of the database. Also, as some locations were not precise, we had to remove all the excessive words such as "near the" or "off the" or names of the countries that no longer exist, e.g., Yugoslavia or Czechoslovakia. In the end, we have locations for around 90 percent of the data records.

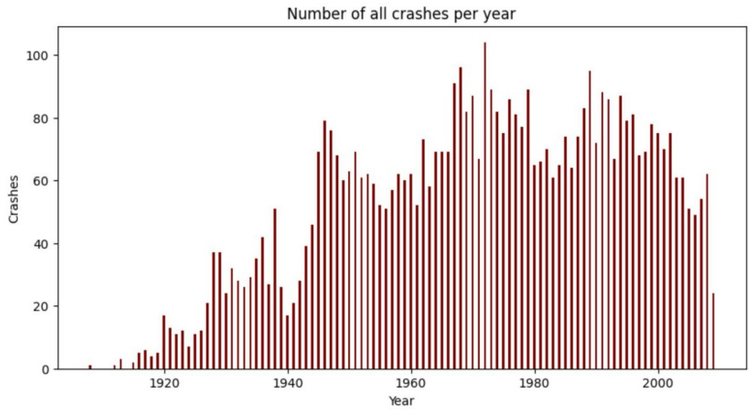
2.1.2 Results

Most of the military plane crashes happened in locations where wars occurred, like Vietnam, Europe, Korea, the Middle East, and Afghanistan. In contrast, civil crashes are more uniformly spread over six continents, more dependent on the number of flights in the area than anything else. Also, it is noticeable that there were a lot of crashes in the USA, even though there were no wars on their territory. The reason is the significant number of military training and aircraft tests they have annually.



2.2 Analysis of the number of crashes and the number of deaths per year

As can be seen from the graphs, the number of crashes had an increasing tendency up to 1972, where it reached its peak. After that, numbers fluctuated while significantly increasing or decreasing occasionally and finally turned to a decreasing tendency, showing that the safety of airplanes increases relative to digitalization and the implantation of modern technologies.



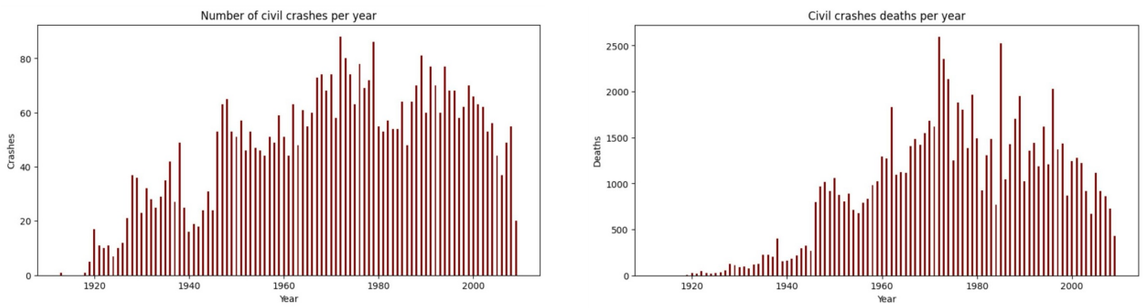
However, these graphs contain both military and civil plane crashes, which is why it may be unfair to analyze this graph and these stats without separating military and civil cases.

2.3 Separating Military and Civil cases

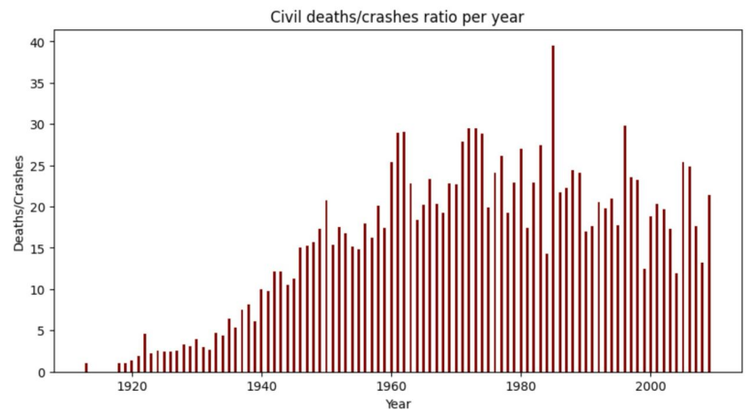
The process of distinguishing military and civil crashes was done by extracting the words such as "Military" from the "Operator" column data. If a word of such kind was present, that case was considered to be military, otherwise it was considered civil.

2.4 Results of separation and their analysis. Civil cases

The results for civil crashes are shown in the graph below.



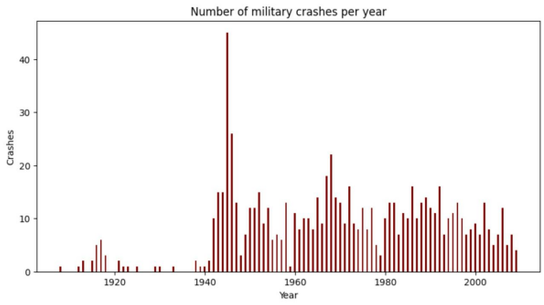
The total number of deaths reached its peak in 1972 (2595 deaths), and similar numbers occurred in 1985 (2528). Even though the number of accidents in 1985 was not extraordinary, that year had the highest deaths per accidents ratio in the whole history of aviation. As can be seen from the graph below, accidents that year averaged 39.5 deaths per crash, which is more than 32% higher than the second result (29.86 deaths/crash).

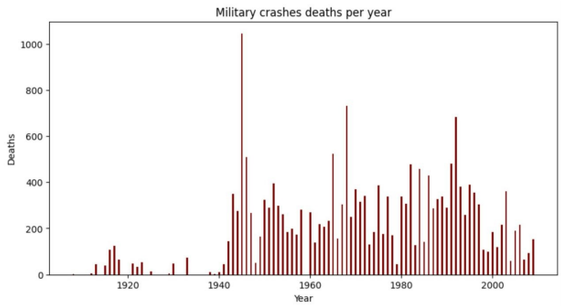


Although the number of crashes significantly decreased in the early 2000s, from the deaths per crash ratio chart, we still can observe that its value fluctuated around 15 in that period. The main reason for this is that the average capacity of airplanes increased. For instance, in the 1990s, Boeing released new upgraded versions of their airicrafts with an increased capacity, such as Boeing 737-800 and 737-900. Also, a lot of new massive planes were produced: Boeing 777 with more than 300 seats available, Airbus A330 with around 250 seats, and other models.

2.5 Results of separation and their analysis. Military cases

The graphs representing the number of crashes and deaths per year in military crashes are shown below.





It is easy to guess that it reached its peak in 1945 during World War II. However, we can also notice two more standing out from the crowd years: 1968 and 1992, where the number of deaths was over 600.

To analyze what were the factors for those crashes, we decided to find the most popular words among places of crashes because our supposition was that the increased amount of crashes was related to some military conflict, and the easiest way to check it is to find the most popular locations of those accidents.

As for 1968, the most popular word occurring 10 times was 'Vietnam' when the second most popular word was 'South', occurring only 6 times, while other words were presented not more than 2 times. Therefore, it is easy to conclude that the increased amount of crashes and deaths in 1968 was because of the Vietnam War.

The same method didn't work for the 1992 case as the most popular geographical words had an appearance of 2 times only. Also, it is noticeable that the number of crashes that year was almost average, which was why we decided to analyze the number of deaths in those accidents. It turned out that in 1992, there were two military crashes, which are the 3rd and 4th accidents with the highest number of deaths (158 and 157) in the whole history.

One of them was an accident in Libya where "Mikoyan-Gurevich MiG-23UB" collided with a Boeing 727-2L5 of the Libyan Arab Airlines Flight 1103 on 22 December 1992. That crash was the deadliest aviation disaster to occur in Lybia at that time.

Another one was the disaster of a Nigerian Air Force Lockheed C-130H Hercules on 26 September 1992. Three engines out of four failed during the take-off, leading to 158 deaths.

2.6 Conclusion

We can conclude that civil crashes are widely spread across 6 out of 7 continents. Also, the number of accidents and deaths in civil crashes had a decreasing tendency in the early 2000s. However, there were no significant changes in the deaths per crashes ratio, and we can suppose that it relates to the fact that the average capacity of aircrafts is growing fast, which is why the average number of deaths is maintaining almost the same value.

As for military cases, we have shown that most plane crashes happened in war locations, such as Vietnam, Europe, Afghanistan, and others. Also, we have shown that peaks in the number of accidents and deaths were caused by major military conflicts, such as World War II and the Vietnam War.

3 Analysis of airlines and aircraft models

In this part, we will analyze and compare different civil airlines and military forces so as to find the airlines and airforces with the most number of crashes and deaths. Also, we will analyze and compare those stats for aircraft models engaged in civil and military accidents.

3.1 Extracting data

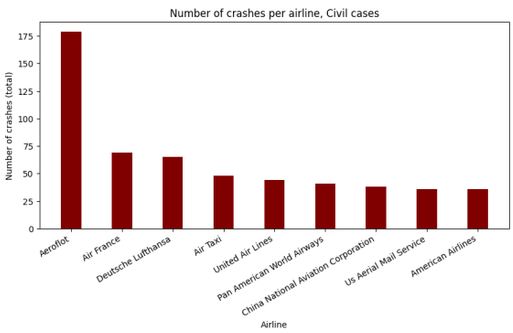
First of all, we used the same method of separation between military and civil cases as for the general data analysis and then we extracted information about the operator from the "Operator" column of the data table.

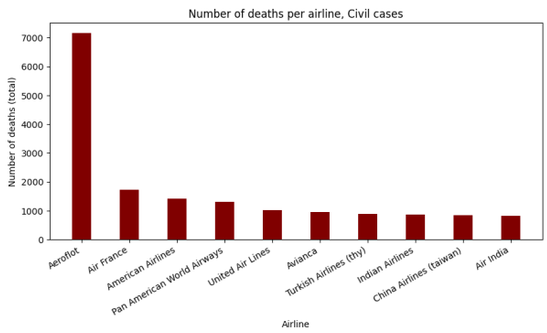
However, we noticed that a lot of military forces were presented with different names, distinguished by the addition of a space or a dash, etc. As the dataset with military forces was not that big, it was easy for us to find the most popular countries and then find all the names of forces for those countries. Then, we summarized all the cases to a unique name (e.g., "United Kingdom Air Force") and got the results for the most represented countries.

As for aircraft models, we simply extracted them from the "Type" column that represents the model of the airplane used in that flight. However, we didn't want to separate planes by their modifications as the airplanes of the same models are almost the same, and the main difference between modifications is usually in some small details, such as the cabin design or its capacity. For the majority of aircraft, their names consist of some words followed by a dash and the number of the model. That is why we separated all the strings representing aircraft names by dashes and took the first two items (or one if there was only one). After that, we deleted all the last letters representing the modification of the model. Definitely, we checked the correctness of our data. The only problem was with one name – "Britten-Norman BN-2" because it was represented as "Britten-Norman" because of the dash in the name of the model. For this case, we used an additional check to separate the models of this manufacturer.

3.2 Airlines analysis and comparison. Civil airlines

As can be seen from the graphs, "Aeroflot" is the leader in the number of crashes and deaths. However, that does not necessarily mean that "Aeroflot" is the most dangerous company because in the USSR "Aeroflot" was the only operating company, which is why it has many more flights compared to other companies.





It is noticeable that "American Airlines" is represented on the first graph at the last position, having only 36 crashes. However, on the second graph, it is the top-3 company with 1421 deaths, which means that the average number of deaths per crash for this airline is ≈ 39.47.

At the same time, we can see contrary statistics for the "Deutsche Lufthansa" and "Air Taxi" companies, which have 65 and 48 crashes and 396 and 182 deaths correspondingly. From this, we can get that "Deutsche Lufthansa" has ≈ 6.09 deaths per crash, and "Air Taxi" is averaging ≈ 3.79 deaths per crash. Both these values are significantly lower than the one we calculated for "American Airlines".

The reason for that is the lower capacity of planes used by "Air Taxi", as they mainly used small jets and operated short-distance flights. In the "Deutsche Lufthansa" case, after looking deeply into the data, we noticed that such a small average number of deaths per crash is caused by a small number of passengers aboard.

3.3 Airlines analysis and comparison. Military forces

Among all countries, the U.S. Military Forces were engaged in the highest number of accidents and caused the highest number of deaths.

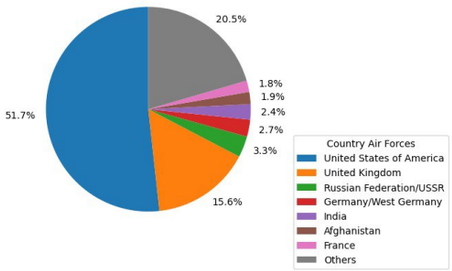


Figure: Crashes per Country Air Force

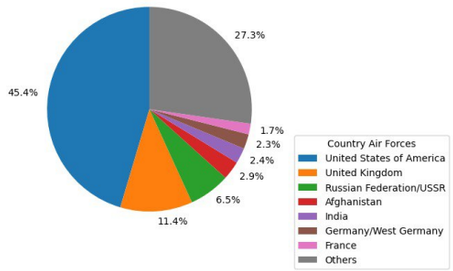
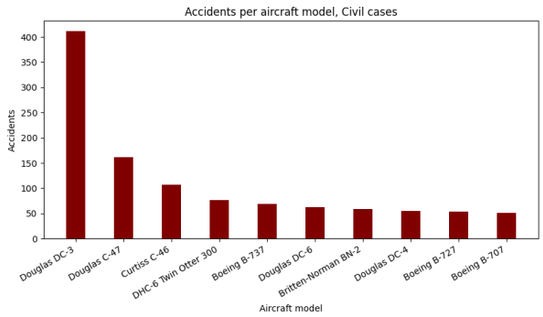


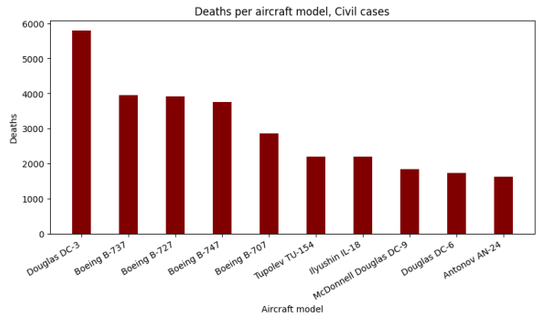
Figure: Deaths per Country Air Force

Among other countries, there is only the UK that has a substantial amount of both crashes and deaths.

3.4 Aircraft models comparison. Civil cases

The two graphs below represent the number of crashes and deaths per aircraft model used for civil flights.





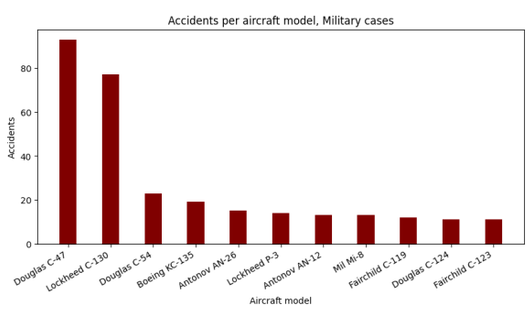
As can be observed, there is a "Douglas DC-3" Aircraft that was engaged in the highest number of accidents. Because of this, it also leads to the number of deaths in total. However, it is noticeable that in the deaths statistic, we can also see that 3 Boeing models are close to the same number of deaths. Even though these models had far fewer accidents, the reason for the increased number of deaths is the capacity of those models. "Douglas DC-3" can take only 21 to 32 passengers, whereas most of the Boeing models have a capacity of more than 100 passengers.

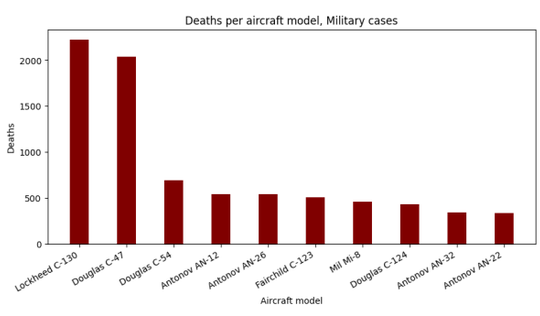
As we can see, "Douglas C-47" is not presented on the second graph because the total number of deaths for this airplane is 1444. After getting this number, we can calculate the average number of deaths in accidents with this aircraft model. Its value is ≈8.969, which is less than 9, and the capacity of "Douglas C-47" varies from 21 to 30.

For the "Boeing B-747", the total number of accidents is as small as 32, which is why it is not presented on the first graph. However, the average number of deaths is ≈117.06, and the capacity of different modifications can be from 300 to 650.

3.5 Aircraft models comparison. Military cases

The two graphs below represent the number of crashes and deaths per aircraft model used for military flights.





Two models that were involved in the highest number of crashes and led to the highest number of total deaths are "Douglas C-47" and "Lockheed C-130".

As we have shown, American Air Forces had the highest number of accidents among all others, and these two aircrafts were the most popular to be used in American Air Forces. That is why it is natural that these two models have the highest numbers of crashes and deaths.

3.6 Number of passengers onboard – number of fatalities correlation calculation

It is natural to suppose that the number of deaths correlates to the plane's capacity, and it can be observed from the evidence described above. That is why, in this part, we decided to calculate the correlation between the number of passengers onboard (unfortunately, we do not have data about the capacity of each plane) and the number of deaths in accidents.

3.6.1 Methodologies of calculating correlation

As always, we cleaned all the incorrect data where we had NaN values instead of numbers in cells. After that, we said that column X is "Aboard" and column Y is "Fatalities". To calculate the correlation, we first calculated the following numbers:

$$\sum X = \sum_{i=1}^{n} X_i$$

$$\sum X^2 = \sum_{i=1}^{n} X_i^2$$

$$\sum Y = \sum_{i=1}^{n} Y_i$$

$$\sum Y^2 = \sum_{i=1}^{n} Y_i^2$$

$$\sum Y^2 = \sum_{i=1}^{n} Y_i^2$$

$$\sum XY = \sum_{i=1}^{n} X_i \cdot Y_i$$

Where n is the total number of data cases we had (5246).

After calculating these values, we used the following formula to find the correlation:

$$r = \frac{n \cdot \sum XY - \sum X \cdot \sum Y}{\sqrt{[n \cdot \sum X^2 - (\sum X)^2] \cdot [n \cdot \sum Y^2 - (\sum Y)^2]}}$$

3.6.2 Results of computations

After all the computations, the result was $r \approx 0.757$, which means that the number of passengers onboard and the number of fatalities have a strong positive correlation, proving our supposition.

3.7 Conclusion

In this part, we analyzed airlines by their total numbers of crashes and deaths. Also, we analyzed the trends in crashes by aircraft models and made a supposition that the number of deaths correlates with the number of people aboard (and therefore correlates with the capacity of the airplane). The supposition was finally proved.

4 The causes of crashes

To identify the causes, we implemented the K-means method and Topic modeling techniques such as Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF).

4.1 Pre-processing for causes

The Airplane Crashes Since 1908 dataset [3] contained a column labeled "Summary", which was encompassed of short descriptions of each crash. The problem we encountered was that the summary did not explicitly state the crash cause. We analysed and extracted the information from it in order to classify the accidents into specific groups - main crash causes. To cleanse the data we lowered all the characters in the summary, removed punctuation, and eliminated words unrelated to the cause. These words included "plane", "flight", "miles", different geographical locations, numbers etc.

4.2 Extracting features

We used the Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer to extract the features of the text. TF-IDF (Term Frequency-Inverse Document Frequency)[8] is a measure of the relevance of a word in a document. The common applications include natural language processing (NLP) and information retrieval tasks. To extract features from the flight summaries we needed to numerically represent the words in the documents. By using TF-IDF, we could quantify the relevance of each term in the documents and represent it as a vector.

We can split the TF-IDF method into two parts: term frequency (TF) and inverse document frequency (IDF):

- Term Frequency (TF): The term frequency measures how frequently is the occurrence of a term relative to the total number of terms in the document. A higher TF indicates that the term is more important in that particular document – in our case, the short flight summary. We can write this as:

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{1}$$

  Where $f_{t,d}$ is the count of the term t in the document d.

- Inverse Document Frequency (IDF): The inverse document frequency aims to measure the importance of a term related to its rarity across the entire corpus – in our case, the whole flight summary dataset. It is calculated by dividing the total number of documents in the corpus by the number of documents that contain the term. IDF assigns higher weights to terms that appear in a smaller number of documents, indicating their significance in distinguishing documents from each other. We can compute it as:

$$idf(t,D) = log\frac{N}{|\{d : d \in D \wedge t \in d\}|} \tag{2}$$

  Where $N = |D|$ is the total number of documents and $|\{d : d \in D \wedge t \in d\}|$ is the number of documents d where the term t appears. By taking the logarithm of the IDF, the scaling emphasizes the differences in rarity between terms.

  The TF-IDF score for a term in a specific document is calculated by multiplying the term frequency (TF) with the inverse document frequency (IDF). A higher TF-IDF score indicates that the term is both frequent within the document and relatively rare across the corpus, making it potentially more informative and significant.

4.3 K-Means Clustering

K-means clustering is a method of vector quantization and its objective is to divide a set of data into clusters where each data point is assigned to the nearest cluster — cluster whose mean is closest to the data point. After computing the TF-IDF values for each summary, we transformed them into numerical representation – a sparse matrix, which we then converted into a 2D dense array that had the column parameter of the original IDs of the crashes.

4.3.1 Defining the number of clusters:

To find out the optimal sub-number of centroids, we used the silhouette score evaluation[4] and inertia evaluation on k values from 5 to 100. The silhouette score measures the compactness and separation of the clusters. Firstly, it calculates the average distance between data points in the same cluster, also known as the "Intra-cluster distance" or "a" value. We compute it as:

$$a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i, j) \tag{3}$$

Where i is a data point in the cluster C_I, and $d(i, j)$ is the distance between data points. We can interpret this as how close the points in each of the clusters are. Secondly, we calculate the "inter-cluster distance" or "b", which shows how close data points are to clusters they are not a part of.

$$b(i) = min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i, j) \tag{4}$$

After performing these steps, we calculate the Silhouette Score using the following formula:

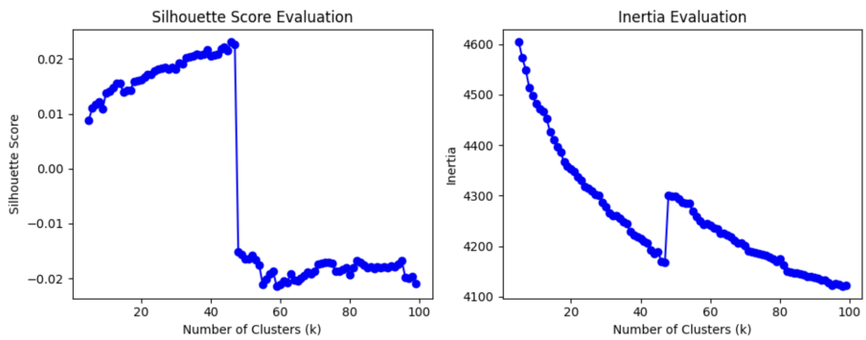
$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}, if |C_I| > 1 \tag{5}$$

The closer this score is to 1, the better matched the data points are.

The inertia evaluation is another metric for assessing the quality of clustering by summing up the squared Euclidean distance between each data point and the center of its cluster.

$$\sum ||i - C_I||^2 \tag{6}$$

After computing these scores for the k values from 5 to 100 we get the following graphs:



The silhouette score has a gradual, almost stagnant growth until the value of 45, where it drops. By testing different values and manually determining where the clusters separate best, we chose the k to be the value 10.

4.3.2 The implementation

We then implemented the K-means method with the TF-IDF features as input. Given a set of observations $(x_1, x_2, ..., x_n)$, where each observation is a d-dimensional real vector, k-means clustering aims to divide the observations into k sets(clusters). Let's denote these sets with $S = S_1, S_2, ..., S_k$. The aim is to find:

$$\begin{gathered} \arg\min \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2 = \\ \arg\min \sum_{i=1}^{k} |S_i \mathrm{Var} S_i| \end{gathered} \tag{7}$$

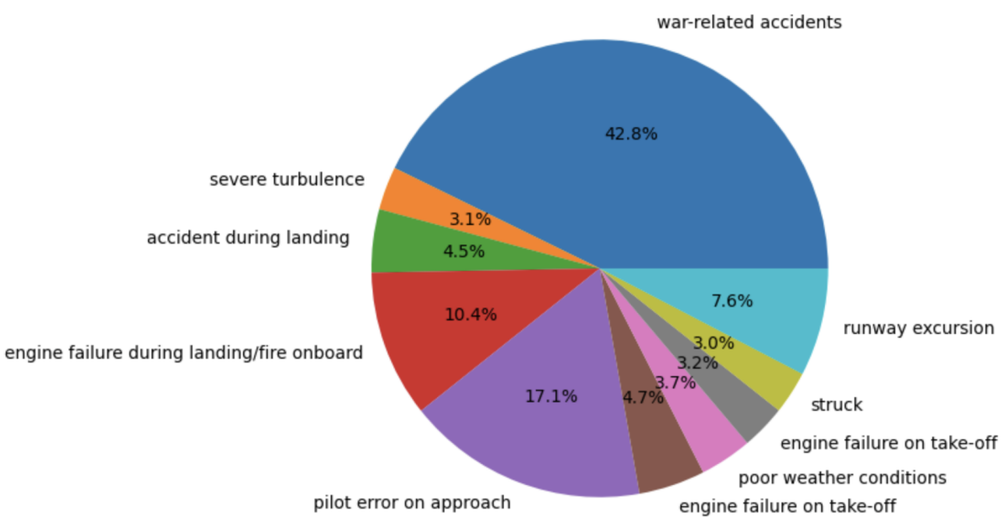
Where μ_i is the mean of points in S_i.

This is done by randomly choosing the first k cluster centers and assigning each of the data points a cluster they belong to. After classifying them into clusters, for each of them we calculate new centers as the mean of the current data points in the cluster. We repeat the process until the clusters do not change.

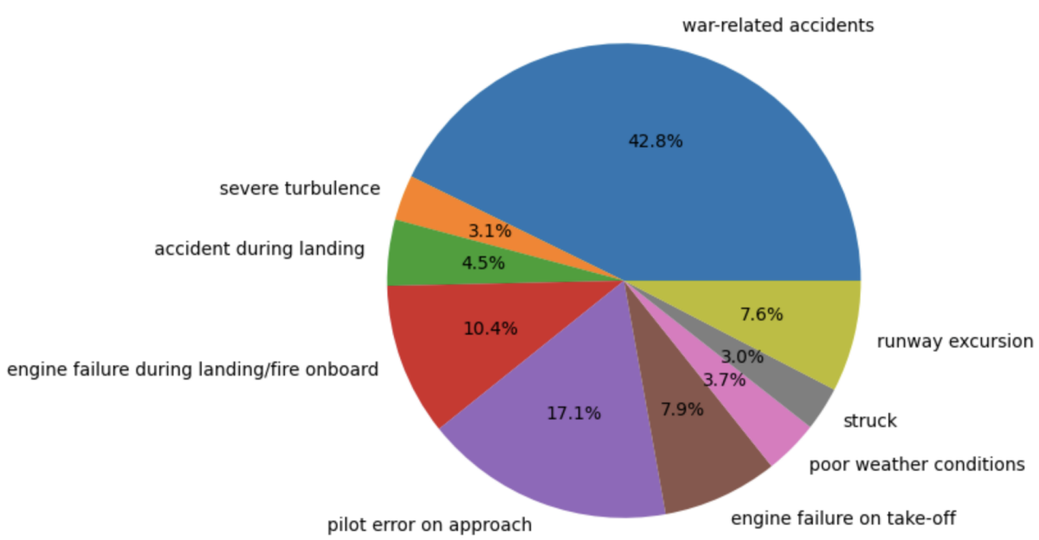
4.3.3 K-means Results

For each cluster, we extracted the top 5 keywords with the highest TF-IDF scores and then manually concluded the cause of the air crash for each of the clusters. We then tested our observations based on the initial IDs and full descriptions. The concluded causes are as follows:

- shot, air, fire, pilot, killed → war-related
- approach, pilot, terrain, altitude, ground → pilot error on approach
- engine, failure, landing, fire, emergency → engine failure during landing/fire onboard
- runway, short, approach, landing, overran → runway excursion
- taking, shortly, minutes, engine, airport → engine failure on take-off
- attempting, land, landing, runway, struck → accident during landing
- conditions, weather, vfr, adverse, continued → poor weather conditions
- takeoff, engine, shortly, failure, runway → engine failure on take-off
- turbulence, thunderstorm, severe, wing, area → severe turbulence
- mountain, struck, flew, poor, weather → struck a mountain



Since we have overlapping labels for poor weather conditions, we made them into a single label:



The dominant causes are war-related incidents, followed by pilot error on approach. Human error and system failures make up around 47.4%, while weather conditions and turbulence make up only around 6.8%.

Additionally, we can note that 6 out of our 10 clusters contain the words ”landing”, ”takeoff”, ”land”, ”approach” and other similar terms. These clusters sum up to 47.4% out of the 57.2% of non-war-related accidents. This suggests that the vast majority of crashes happen during these stages of flight.

To make conclusions about postwar-related causes of air crashes, we divided the data into two sections – before and after 1950.

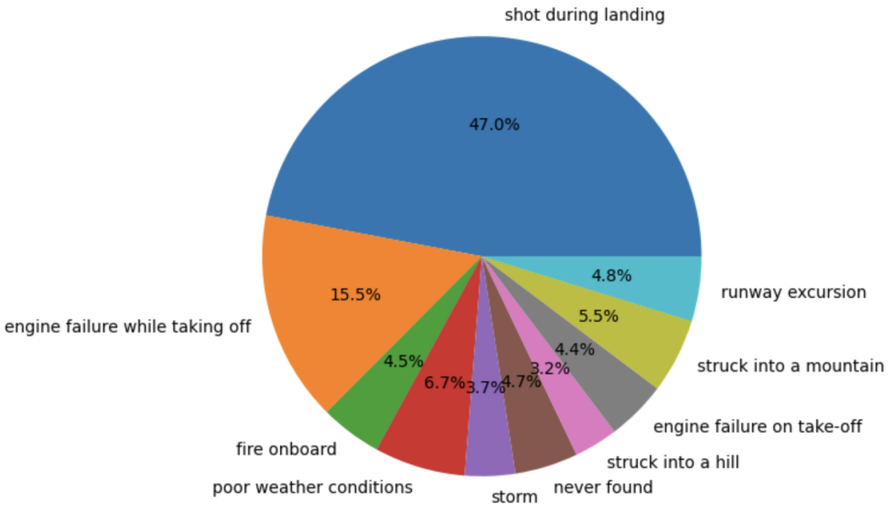


Figure: Causes of incidents before 1950

The main cause of accidents is war-related: the aircrafts being shot with 47% of the crashes.

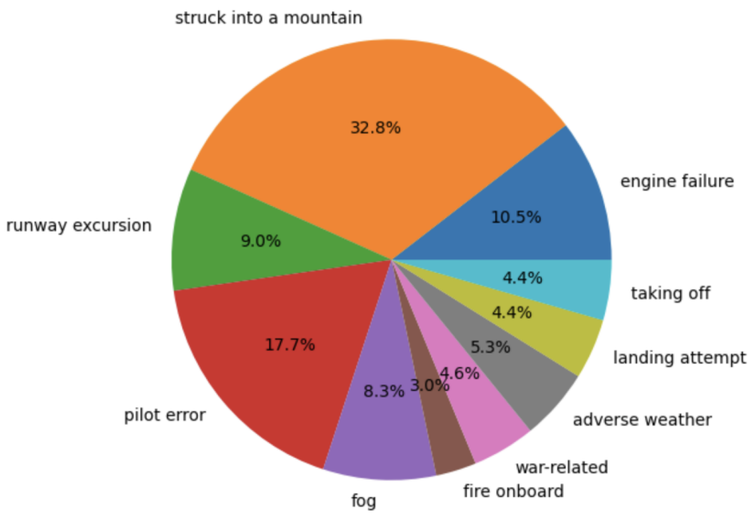


Figure: Causes of incidents after 1950

The main cause of crashes after 1950 is ”struck into a mountain”, which is inconclusive if it is a human or system error. The second biggest cause is pilot error, followed by runway excursion.

4.4 Topic modeling

Topic modeling is a popular technique in natural language processing (NLP) and machine learning that aims to discover hidden thematic patterns or topics within a collection of documents. Two commonly used topic modeling algorithms are Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF).

4.4.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA)[5] is a generative probabilistic model of a corpus, often used for tasks of unsupervised machine learning. It assumes the generating process for the corpus has the characteristics that follow:

- Each document is a bag of words – it disregards word order and grammar.
- Each document contains multiple possible topics.
- Every topic is a mixture of words.

The generative process that LDA assumes consists of choosing two variables N and θ such that $N \sim Poisson(\xi)$ and $\theta \sim Dir(\alpha)$, where N is the number of words in the corpus. Afterward, it chooses a topic $z_n \sim Multinomial(\theta)$ and then selects a word w_n from $p(w_n|z_n, \beta)$ a multinomial probability conditioned on the topic z_n.

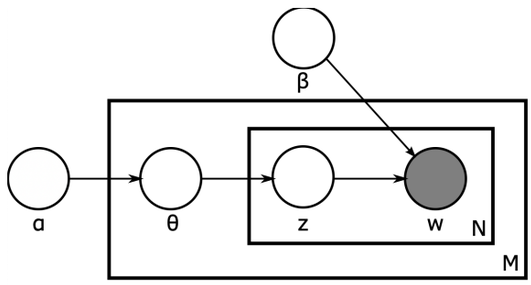


Figure: Plate notation of the LDA model

The LDA model initially assigns each of the words in every document a random topic from the assumed number of topics – k.

For each word in a document, it computes $p(z_n|d)$, the probability of a document d belonging to a topic z_n based on the proportion of the words in the document assigned to the topic, excluding the current word. Next, the proportion of assignments to topic z_n over the documents that contain the word w is calculated. We denote it as $p(w_i|z_n)$. If it is concluded that the word is likely to be in a topic, all the words in the said document become more likely to be in it and vice versa. We update the probability accordingly:

$$p(w_i \wedge z_n) = p(z_n|d) * p(w|z_n) \tag{8}$$

To implement this we used the LatentDirichletAllocation model from sklearn.decomposition. The model uses an iterative inference alhorithm called "variotinal inference". It randomly initializes the topic-word and document-topic distributions and iteratively updates the vectors for them to maximize the likelihood of the the data being observed until convergence.

After estimating the model parameters, each document is represented as a distribution over topics. From this data, we can infer the most important topics in the documents.

4.4.2 Non-Negative Matrix Factorization

Non-Negative Matrix Factorization (NMF) is another popular technique for topic modeling. It aims at factorizing a Document-Term matrix V as two lower-ranked non-negative matrices: Document-Topic matrix W and Topic-Term matrix H. In our case, documents were the lines of text from the "Summary" column.

NMF is a non-exact matrix factorization technique. This means that you cannot multiply W and H to get back the original document-term matrix V.

The matrices W and H are initialized randomly. And the algorithm is run iteratively until we find a W and H that minimize the cost function.

The cost function is the Frobenius norm of the matrix $V - W \cdot H$, as shown below:

$$minimize||V - WH||$$

The Frobenius norm of a matrix A with m rows and n columns is given by the following equation:

$$||A||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}$$

Where a_{ij} is the element of the matrix A in i-th row and j-th column.

We used the NMF model from the sklearn.decomposition module. The model initially sets document-topic and topic-word matrices to random non-negative values. It then uses

projected gradient descent to update the matrices while maintaining non-negative values and minimizes the reconstruction error.

Once we minimaze the cost function we are left with the topic-word and document-topic matrices which represent the importance of each word in each topic and the presence of a topic in each document. From these values we extracted the most important topic in each document and the keywords from it.

4.4.3 Coherence score

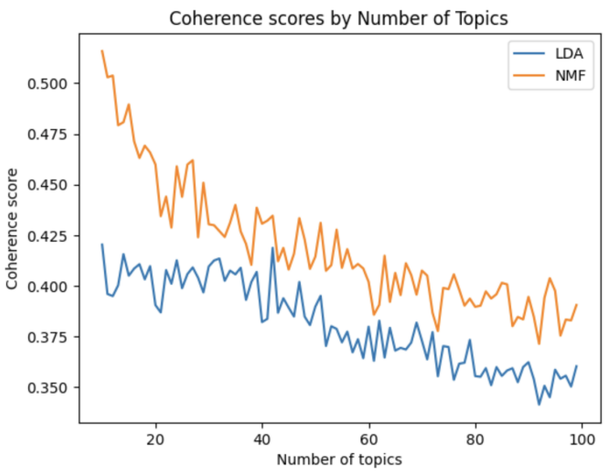
Coherence score is a metric used to evaluate the quality of topics generated by topic modeling algorithms such as LDA and NMF.[6] It measures the degree of semantic coherence or interpretability of the discovered topics. A higher coherence score indicates that the topics are more coherent and easier to interpret. It is computed as:

$$C_{xy}(f) = \frac{|P_{xy}(f)|^2}{P_{xx}(f)P_{yy}(f)} \tag{9}$$

Where P_{xx} is the m-by-m matrix of power spectral densities and cross power spectral densities of the input and P_{xy} is the m-dimensional vector of cross power spectral densities between the inputs and y, P_{yy} is the power spectral density of the output.

We used the CoherenceModel from the gensim.models module and the c_v coherence measure - mean of the pairwise similarity scores.

We calculated the coherence score for both methods for values $[10, 100]$:



LDA peaks at 10 topics, while NMF has the highest coherence score at 11 topics. The LDA score has a steep drop with each added new topic, while NMF stays more stagnant.

We decided to perform both LDA for 10 topics and NMF for 11.

4.4.4 Topic Modeling Results

The top words in the nine different topics generated by LDA and their counts are:

- disappeared, test, loaded, river, gravity – 209
- engine, failure, fuel, wing, mountains – 824
- fuel, landing, attempting, emergency, ditched – 159
- mountain, taking, shot, flew, shortly – 1312
- approach, pilot, weather, conditions, poor – 81
- mail, cause, undetermined, taking, unknown – 54
- takeoff, missing, engine, went, british – 56
- jungle, shot, helicopter, enemy, taking – 721
- landing, runway, attempting, land, engine – 415

- ocean, engine, power, lines, overloaded – 995

The coherence score for this run was ≈ 0.4204, which indicates a moderate level of coherence.

The count refers to the number of documents where the particular topic was the most significant one. We can see that in the majority of the flight summaries there was a war theme. The next leading topic is related to engine power or overloaded aircraft followed by a topic of engine failure.
The top words in the nine different topics generated by NMF and their counts are:

- engine, failure, lost, experiencing, right – 235
- mountain, flew, struck, fog, hit – 632
- takeoff, shortly, aborted, stalled, overloaded – 257
- conditions, vfr, adverse, weather, continued – 320
- approach, final, crew, pilot, descent – 555
- attempting, land, fog, burned, airport – 117
- taking, shortly, minutes, airport, lost – 1290
- runway, short, overran, fog, hit – 222
- poor, weather, conditions, visibility, conditions – 232
- landing, emergency, attempt, make, gear – 776
- struck, ground, high, trees, wing – 188

The coherence score that was calculated is 0.5208, which suggests a moderate level of coherence.

The most common topic is related to an aircraft taking off, followed by the topic of an aircraft landing. This suggests that most of the air crashes happened during these two periods of flight.

4.5 Conclusion on the Causes of Crashes

The most difficult aspect of determining the cause of an airplane accident is that most crashes have several of them. [7] By implementing the K-means and topic modeling algorithms and manually analyzing the keywords we identified various contributing factors:

- war-related
- pilot errors
- engine failure
- fire onboard
- runway excursions
- adverse weather
- severe turbulence
- collisions with terrain, usually mountains

The analysis revealed that the dominant causes of the crashes were war incidents, accounting for approximately 43% of all crashes, followed by a pilot error that contributes to around 10%. A study by Oster that analyzed air crashes from 1990 to 2006 suggests that human error was the sequence-initiating cause in 40% of those accidents.[7]

Furthermore, we observed that a significant amount of the accidents occurred during the critical stages of flight – landing and takeoff. Our results highlight the importance of safety measures and regulations during these periods.

The conclusions drawn from our analysis are based on a limited dataset [3], further research may be necessary to validate our findings.

5 Summary

5.1 Results obtained

Overall, in our project, we analyzed places and the number of crashes and deaths that occur annually and found the tendency of these numbers. We also analyzed why peaks of those numbers occur. As for civil cases, we understood that those increases are mainly connected with the rising capacity of aircrafts, and for military cases, they are usually caused by large-scale wars.

Additionally, we analyzed different airlines and aircraft models and their performance throughout this period. We made a supposition that the number of deaths in an accident correlates with the number of passengers aboard, which was proved by finding the correlation coefficient.

Another important feature of our project is that we analyzed the most popular causes of crashes. We divided the results into two periods: before 1950 and after 1950. As for the first period, we got 47% for the cause related to war: shot during landing. The second most popular cause was engine failure. Regarding the period after 1950, the results were completely different and much more diverse: 32.8% got the "struck into a mountain" reason. 4 other outstanding causes were: "pilot error", "engine failure", "runway excursion", and "fog".

The coherence score for those results was over 0.5, which shows that the accuracy is moderate.

5.2 Possible future improvements

There are 3 foreseeing improvements to our analysis:

1. Compare airlines and find the safest and the most dangerous airlines. For this, we should assume that different airlines have various numbers of flights: for example, "Aeroflot" was the only company in the USSR, which is why this company had many more flights than a lot of other companies. The best way to properly compare different airlines is to calculate the number of crashes per fixed number of flights: for instance, 100,000 flights.

2. The same method can be implemented for aircraft models, as, for example, "Douglas DC-3" was widely used in the USA, and this model was used in many more flights than other models.

3. Additionally, more accurate research on civil crash stats can be performed to find out the reasons for the peaks and overall tendency. For example, one can calculate the average capacity of airplanes involved in crashes and analyze those numbers.

4. Lastly, a more comprehensive analysis could be done for the airline crashes. By labeling a part of the data with the corresponding causes, we could make a training set for different deep learning methods. Another possible approach would be utilizing Large Language Models (LLMs) to extract the causes from flight summaries. In both cases, a proper confirmation from labeled data would be needed to confirm the accuracy.

References

[1] IATA. (June 1, 2023). *Number of flights performed by the global airline industry from 2004 to 2022, with a forecasts for 2023 (in millions).* https://www.statista.com/statistics/564769/airline-industry-number-of-flights/

[2] IATA. (June 30, 2022). *Number of scheduled passengers boarded by the global airline industry from 2004 to 2022 (in millions).* https://www.statista.com/statistics/564717/airline-industry-passenger-traffic-globally/

[3] Grandi, S. (2016). *Airplane Crashes Since 1908* [Data set]. Kaggle. Retrieved from https://www.kaggle.com/datasets/saurograndi/airplane-crashes-since-1908

[4] Shahapure, K. R., & Nicholas, C. (2020, October). Cluster quality analysis using silhouette score. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 747-748). IEEE.

[5] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(4-5), 993-1022.

[6] Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). *Optimizing semantic coherence in topic models.* In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 262-272).

[7] Oster, C. V., Strong, J. S., & Zorn, K. (2010). *Why airplanes crash: Causes of accidents worldwide.* Oxford University Press.

[8] Bafna, P., Pramod, D., & Vaidya, A. (2016). *Document clustering: TF-IDF approach.* In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)* (pp. 61-66). IEEE.

Contribution List

- GADANSKI Aleksandar (agadjansk2): 20% (data preprocessing; analysis of airlines and aircraft models; map visualization)
- GIZATULLIN Timur (tgizatull2): 20% (data insights for general data analysis and analysis of airline and aircraft models parts; report)
- SHAKIRZIANOV Iusuf (ishakirzi2): 20% (presentation; coordination)
- SHATOKHIN Anton (ashatokhi2): 20% (general data analysis)
- STEFANOVIC Zlata (zstefanov2): 20% (causes of crashes analysis; report)