DS200-Project II

Airplane Crashes (1908 - 2024)

Team: Nichol Flowers Tonderai Madamba

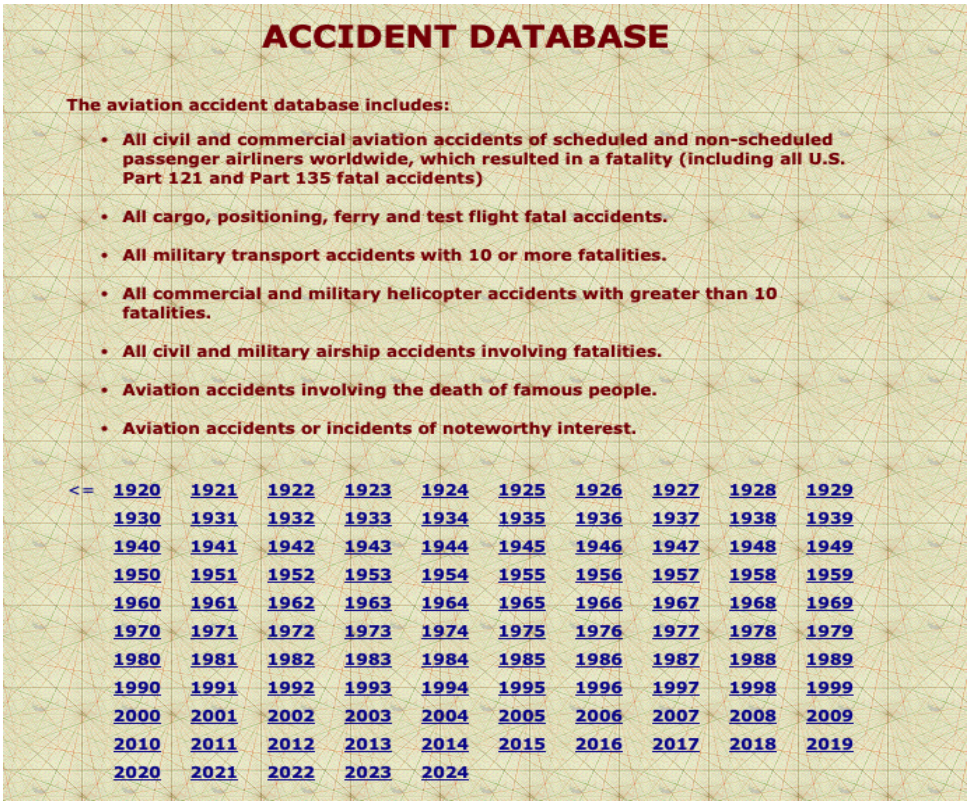
Github repository:

https://github.com/UC-Berkeley-I-School/Project2_Nichol_Tonderai

Data Source:

We scraped our data from https://www.planecrashinfo.com/database.htm created by a private citizen who collected this information for general interest.

Dataset:



ACCIDENT DATABASE

The aviation accident database includes:

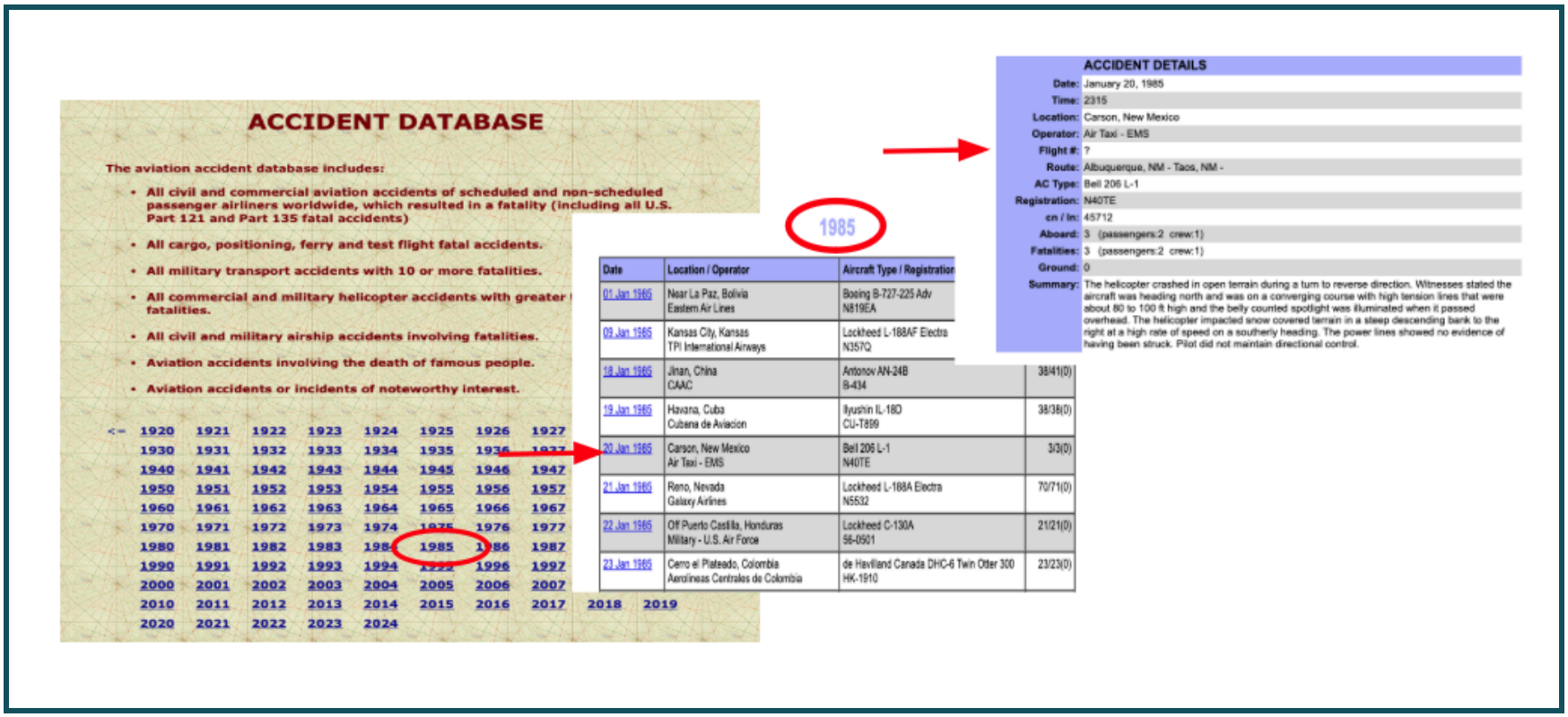
- All civil and commercial aviation accidents of scheduled and non-scheduled passenger airliners worldwide, which resulted in a fatality (including all U.S. Part 121 and Part 135 fatal accidents)
- All cargo, positioning, ferry and test flight fatal accidents.
- All military transport accidents with 10 or more fatalities.
- All commercial and military helicopter accidents with greater than 10 fatalities.
- All civil and military airship accidents involving fatalities.
- Aviation accidents involving the death of famous people.
- Aviation accidents or incidents of noteworthy interest.

<= 1920 1921 1922 1923 1924 1925 1926 1927 1928 1929
1930 1931 1932 1933 1934 1935 1936 1937 1938 1939
1940 1941 1942 1943 1944 1945 1946 1947 1948 1949
1950 1951 1952 1953 1954 1955 1956 1957 1958 1959
1960 1961 1962 1963 1964 1965 1966 1967 1968 1969
1970 1971 1972 1973 1974 1975 1976 1977 1978 1979
1980 1981 1982 1983 1984 1985 1986 1987 1988 1989
1990 1991 1992 1993 1994 1995 1996 1997 1998 1999
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009
2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
2020 2021 2022 2023 2024

Database Structure:

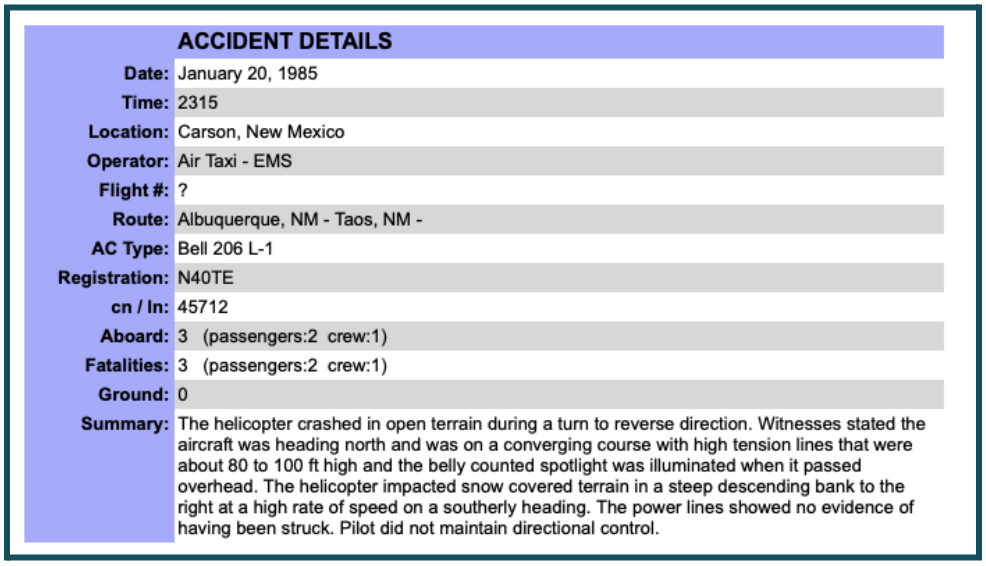
The accident database has three hierarchical layers. At the very top of the hierarchy, is the Accident Database where a user can click the year of interest. This takes them to the second page, which contains a table of the crashes recorded for that specific year. Upon clicking a specific date of interest, the Accident Details page is displayed with the details specific to that crash.

Example:

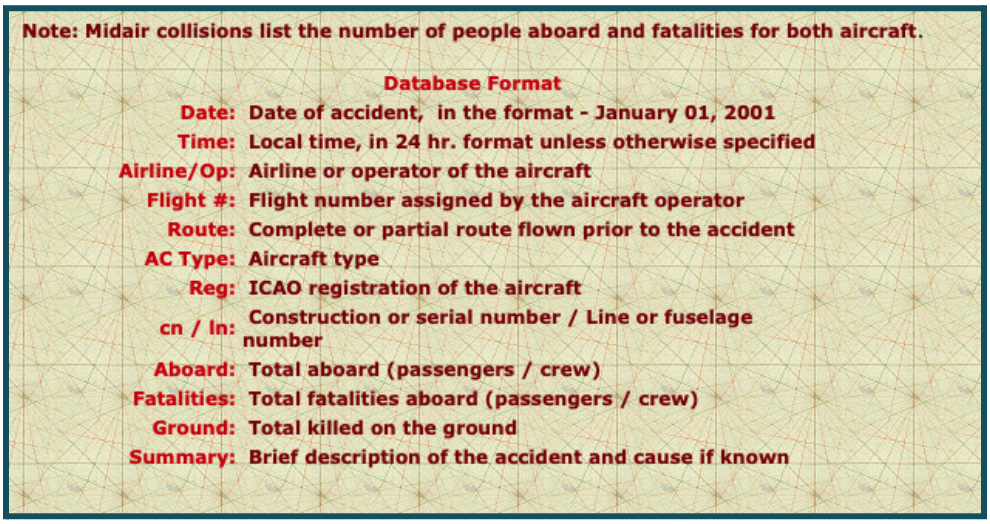


Data Contents:

The Accident Details page has 11 sub fields, which are described in the Database Format table.



ACCIDENT DETAILS	
Date:	January 20, 1985
Time:	2315
Location:	Carson, New Mexico
Operator:	Air Taxi - EMS
Flight #:	?
Route:	Albuquerque, NM - Taos, NM -
AC Type:	Bell 206 L-1
Registration:	N40TE
cn / ln:	45712
Aboard:	3 (passengers:2 crew:1)
Fatalities:	3 (passengers:2 crew:1)
Ground:	0
Summary:	The helicopter crashed in open terrain during a turn to reverse direction. Witnesses stated the aircraft was heading north and was on a converging course with high tension lines that were about 80 to 100 ft high and the belly counted spotlight was illuminated when it passed overhead. The helicopter impacted snow covered terrain in a steep descending bank to the right at a high rate of speed on a southerly heading. The power lines showed no evidence of having been struck. Pilot did not maintain directional control.



Note: Midair collisions list the number of people aboard and fatalities for both aircraft.

Database Format

Date:	**Date of accident, in the format - January 01, 2001**
Time:	**Local time, in 24 hr. format unless otherwise specified**
Airline/Op:	**Airline or operator of the aircraft**
Flight #:	**Flight number assigned by the aircraft operator**
Route:	**Complete or partial route flown prior to the accident**
AC Type:	**Aircraft type**
Reg:	**ICAO registration of the aircraft**
cn / ln:	**Construction or serial number / Line or fuselage number**
Aboard:	**Total aboard (passengers / crew)**
Fatalities:	**Total fatalities aboard (passengers / crew)**
Ground:	**Total killed on the ground**
Summary:	**Brief description of the accident and cause if known**

Ingesting the data:

Beautifulsoup was used to scrape the 3-layer database.

Context

Since its invention by the Wright brothers, flying has transformed the way we explore the world. Over the years, aircraft technology and aviation practices have evolved drastically, enhancing both safety and the comfort of air travel. Today, aviation is considered one of the safest modes of transportation, thanks to the high standards set by regulatory boards and rigorous pilot training. Despite these advancements, though rare, airplane crashes still occur and have led to countless fatalities, instilling fear worldwide.

For this reason, we would like to explore airplane crash trends related to location, airline operator, fatalities and causes.

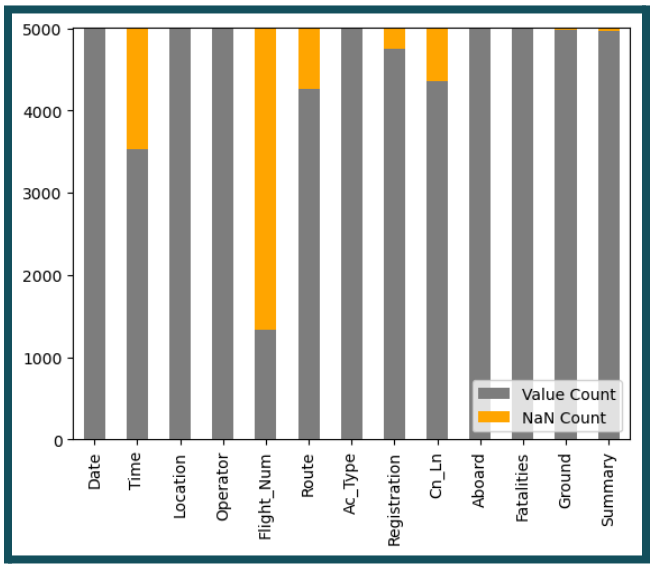
The Focus

1. Where are airplane crashes most common?
2. How has the number of plane crashes (and fatalities) changed over time?
3. What operators have had the most crashes?
4. What operators have had the most fatalities?
5. Are there common themes in the crash causes for different periods of time?

Initial Exploration

After ingesting the data, the first step was to count the null numbers (missing data entries) in each column the results shown in the following bar graph.

Bar graph displaying nulls:



Lucky for us, our columns of interest (Date, Location, Operator, Aboard, Fatalities, Ground, Summary) are not really impacted by nulls.

Verification:

- Confirmation of date range:1920 - 2024. *Data actually starts in 1908.*
- Randomly select years and examine the data details. *9/11 totals for casualties on the ground were duplicated.*
- Check for other duplicate entries. None found.
- Variable Analysis: Fine except the Location field. *No standard naming*

Unfortunately, the location field is unstructured and allows a wide range of inputs. Some locations included country names while others did not. Some locations were listed in their local language (instead of English) while others were displayed as their historical colonial names.

Example:

```
36                Mount Phou-Lassy, French Indo-China
112                      Near Eaeka, French Cameroons
335                   Off Nuka Hiva, French Polynesia
487                      Near Gao, French West Africa
703                  Bangui, French Equatorial Africa
1085                 Bangui, French Equatorial Africa
2451                    Off Dakar, French West Africa
2716    Gustavia, St. Barthélémy, French West Indies
3128                     Off Moorea, French Polynesia
3533             Mont Blanc, French Alps, Switzerland
3951                          Obock, French Somaliland
4160                        Raiatea, French Polynesia
Name: Location, dtype: object
```

Since the location is critical to our study, we resolved inconsistencies by standardizing the data using GeoPy and other sources. We assigned state and country codes, such as “TX” (Texas) and “USA ' (United States), to ensure uniformity.

GeoPy Integration (to resolve location data issue)

Using the crash in Carson, New Mexico on January 20, 1985 as a reference, GeoPy will format the location respectively shown in the following images:

“GeoPy helps Python developers locate the coordinates of addresses, cities, countries, and landmarks across the globe using third-party geocoders and other data sources.” As a result GeoPy fixed most of these inconsistencies as a result we added additional fields/columns to our data.

Before GeoPy:

```
Mount Phou-Lassy, French Indo-China
Pacific Ocean, 325 miles east of Wake Island
Off Irish coast
Black Sea, Gulf of Karkinitsky
Near Eaeka, French Cameroons
Ocean, 800 miles east of Newfoundland
Off Freetown, Sierre Leone
Off  Bimini
Atlantic Ocean, 110 miles West of Ireland
Off Malaya
Off the Panama coast
Near Hong Kong International Airport
Near Gao, French West Africa
Near Sofia, Bugaria
Near Kariba, Rhodesia (Zimbabwe)
Near Lidköping, Västergötland, Swden
Atlantic Ocean off Florida
Almelund, Minnisota
```

After GeoPy:

	Location	Country_location	Code	Country_Geopy
36	Mount Phou-Lassy, French Indo-China	{'place_id': 237996895, 'licence': 'Data © Ope...	LAO	Luang Prabang, Laos
112	Near Eaeka, French Cameroons	{'place_id': 66116545, 'licence': 'Data © Open...	CMR	Cameroon
335	Off Nuka Hiva, French Polynesia	{'place_id': 32894214, 'licence': 'Data © Open...	PYF	French Polynesia, France
487	Near Gao, French West Africa	{'place_id': 288997898, 'licence': 'Data © Ope...	MLI	Gao, Mali
703	Bangui, French Equatorial Africa	{'place_id': 68581124, 'licence': 'Data © Open...	CAF	Bangui, Central African Republic
1085	Bangui, French Equatorial Africa	{'place_id': 68581124, 'licence': 'Data © Open...	CAF	Bangui, Central African Republic
2451	Off Dakar, French West Africa	{'place_id': 290570699, 'licence': 'Data © Ope...	SEN	Dakar Region, Senegal
2716	Gustavia, St. Barthélémy, French West Indies	{'place_id': 291212938, 'licence': 'Data © Ope...	GLP	Guadeloupe, France
3128	Off Moorea, French Polynesia	{'place_id': 32894214, 'licence': 'Data © Open...	PYF	French Polynesia, France
3533	Mont Blanc, French Alps, Switzerland	{'place_id': 93764562, 'licence': 'Data © Open...	CHE	Switzerland
3951	Obock, French Somaliland	{'place_id': 69700534, 'licence': 'Data © Open...	DJI	Obock, Djibouti
4160	Raiatea, French Polynesia	{'place_id': 32839627, 'licence': 'Data © Open...	PYF	French Polynesia, France

Initial Analysis

We kicked off our analysis by aggregating crash totals and summing the Passengers Aboard, Passenger Fatalities and Ground Fatalities columns.

country	Total_Crashes	Total_Aboard	Fatalities_Plane	Fatalities_Ground
United States	1087	26713	15663	3179.0
Russia	260	10963	8563	90.0
Brazil	179	4576	3205	65.0
Canada	160	3211	2192	15.0
Colombia	157	3524	3072	38.0
United Kingdom	144	3593	2614	121.0
France	138	3787	3077	32.0
Germany	129	3408	2847	66.0
India	108	3695	2890	195.0
China	103	4173	2665	55.0
Indonesia	103	3892	2722	74.0
Mexico	87	1668	1269	63.0
Italy	87	2438	1859	26.0
Philippines	74	2675	1321	32.0
Australia	71	859	741	0.0
Tajikistan	67	3589	3039	0.0
Spain	67	3374	2482	2.0
Venezuela	56	1450	1238	71.0
Peru	54	1972	1530	6.0
Democratic Republic of the Congo	52	1266	922	330.0

state	Total_Crashes	Total_Aboard	Fatalities_Plane	Fatalities_Ground
California	124	3769	1673	86.0
Alaska	107	1678	1026	0.0
New York	59	3098	1623	2778.0
Texas	50	942	635	1.0
Ohio	44	355	229	0.0
New Jersey	38	735	472	13.0
Florida	37	1255	607	1.0
Illinois	35	1089	665	14.0
Pennsylvania	32	745	446	2.0
Colorado	30	736	308	1.0
Washington	26	744	419	26.0
Missouri	26	565	229	0.0
Hawaii	24	1017	236	0.0
New Mexico	22	442	157	0.0
North Carolina	21	500	374	1.0
Wyoming	21	239	200	0.0
Indiana	21	322	302	22.0
Maryland	19	338	315	0.0
Virginia	19	668	570	125.0
Nevada	18	427	362	10.0

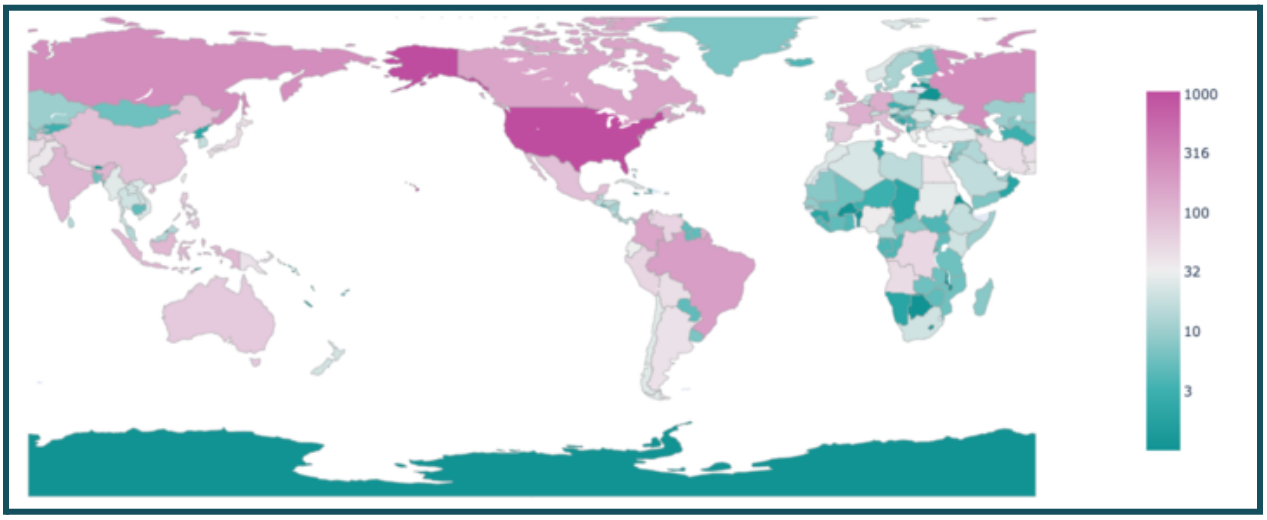
Take away:
We see a correlation between the number of crashes and fatalities, as expected. However, the number of crashes does not necessarily dictate the number of fatalities.

It's important to note that crashes in this database can be commercial planes, private planes, cargo, or military. For this reason, you can have high crash totals without a high number of fatalities.

Looking at the US, California and Alaska have the most crashes by far. Although Alaska has the 2nd highest crash total , its fatalities and total passengers aboard are not quite as high. New York has a high number of fatalities, despite its lower number of crashes (59), when compared to Alaska (107). What stands out is the 2,778 number of fatalities on the ground in New York; specifically due to the tragic events on 9/11/2001.

Mapping the Data:

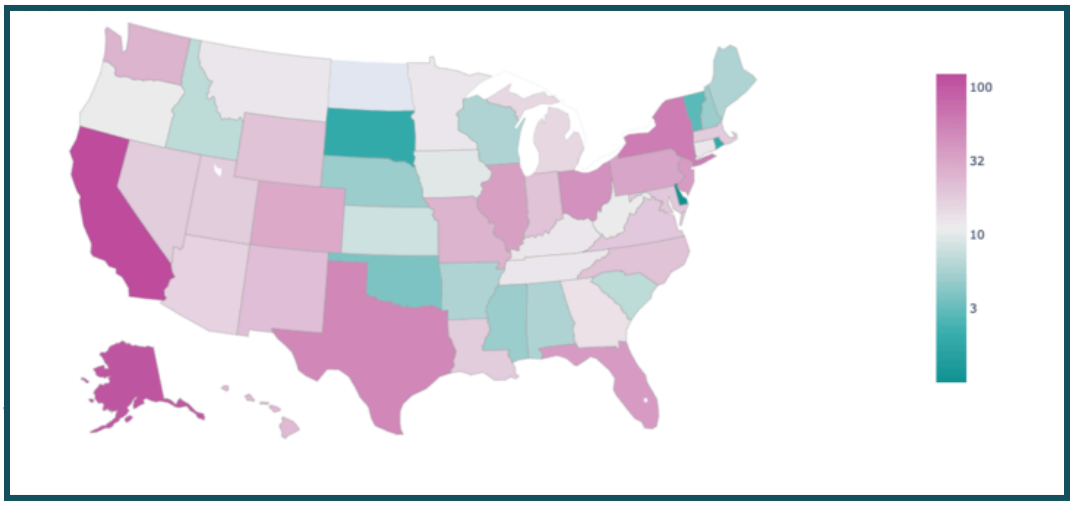
Global Crashes:



Take Away: Most of the crashes are on the left and around the Pacific Ocean. Pink means the total crashes is greater than 30.

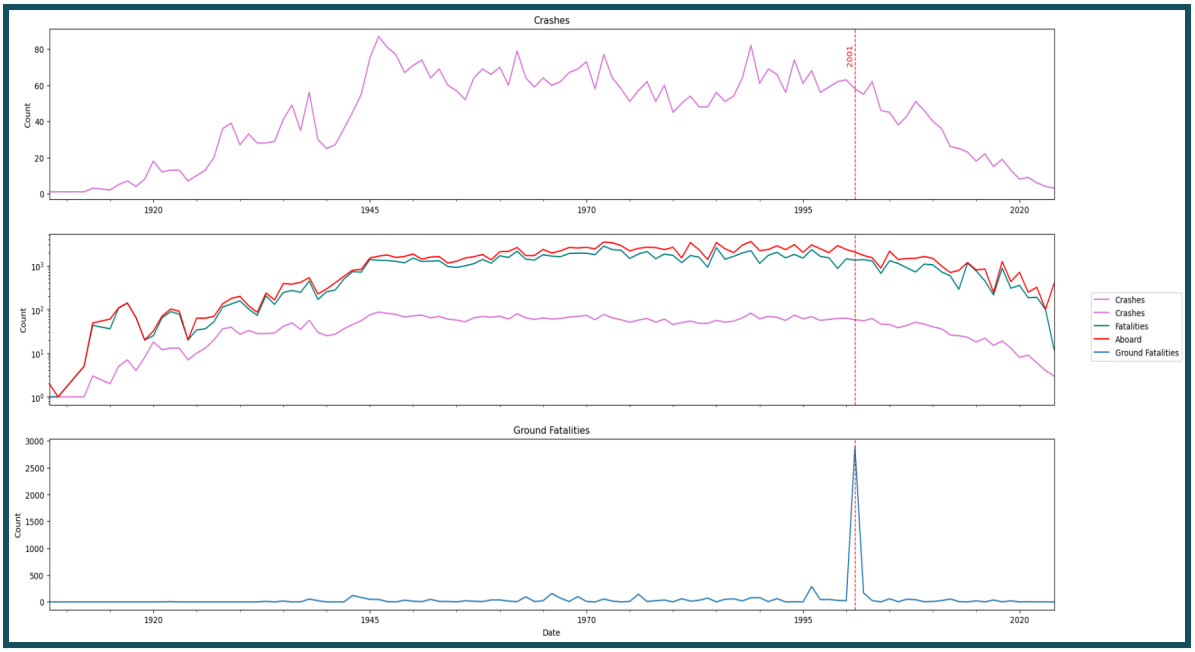
\

USA Crashes:



The crashes are not evenly distributed. California and Alaska have the most. There seems to be fewer crashes in central parts of the USA, and most of the crashes are happening on the coastal edges of the country.

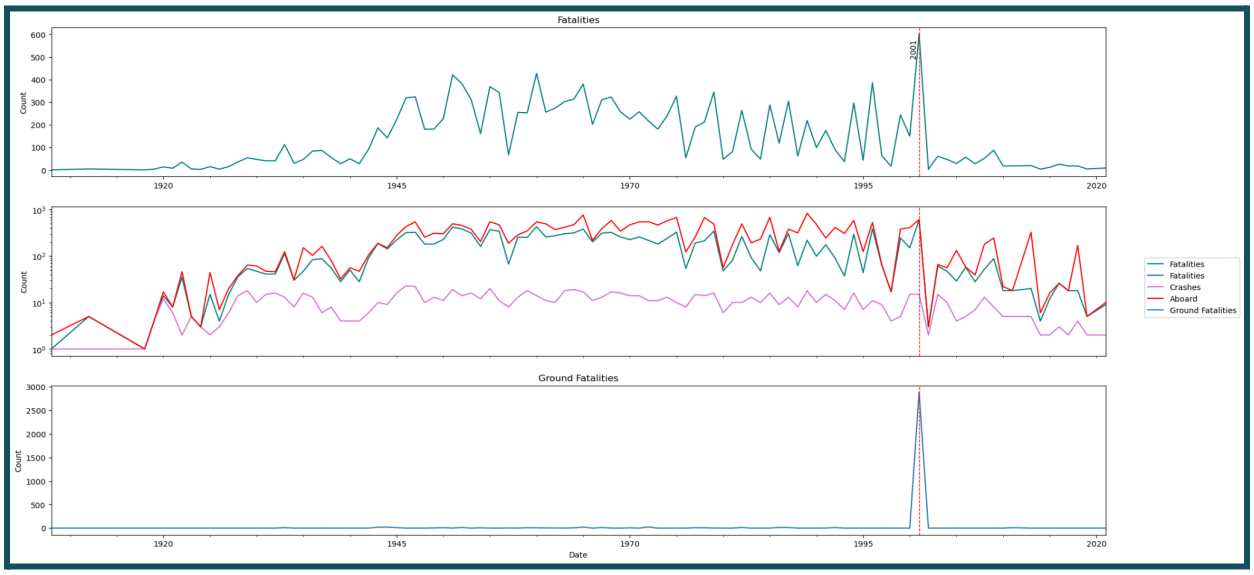
Plotting crashes, fatalities, and aboard (Globally):



Take Away: The first plot on the graph above shows the total number of crashes per year since 1908. For some context, [four engine airliners] were introduced right after 1945, meaning flying was a regular occurrence as a day-to-day form of transportation. We used a log scale in the second plot because different magnitudes. A correlation between crashes, fatalities and passengers aboard across time is shown. There isn't much of a difference between Crashes and Fatalities with time, also shown is that the death rate or chances of you surviving after a

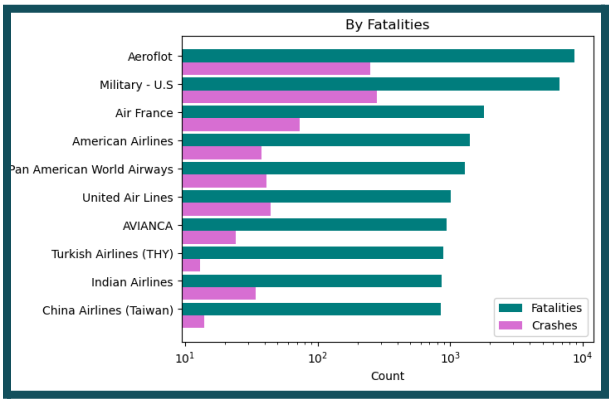
crash occurs or to simply put it is the marginal difference between Aboard and Fatalities across time. From 1945 to about 2001 there are no significant changes that can be observed, across all the categories. 9/11 happened in 2001 hence a huge spike with the fatalities on the ground in 2001. For some context that's when the TSA (Transportation Security Administration) was formed. From the plot it's evident that this had an impact spanning 2 decades and counting.

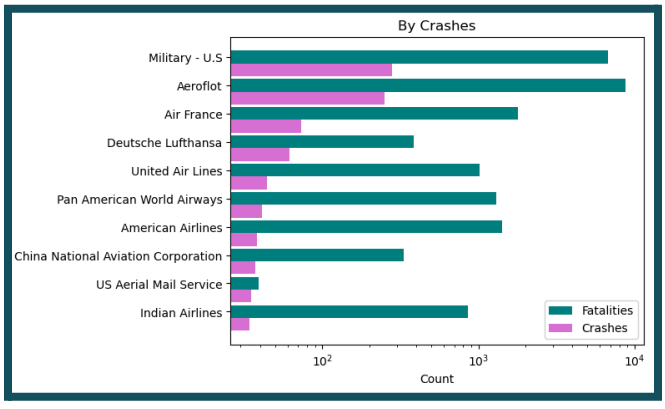
Plotting crashes, fatalities, and aboard (USA):



Take Away: The USA figure 2 shows the same correlation with the global plot (the middle figure with all the plots is in log scale); these two figures cannot be easily distinguished from each other, meaning that the USA data in the global data has a strong influence on it. The number of fatalities drastically decreased after 2001, similar to the global trend.

Plotting crashes, fatalities (By Operators)

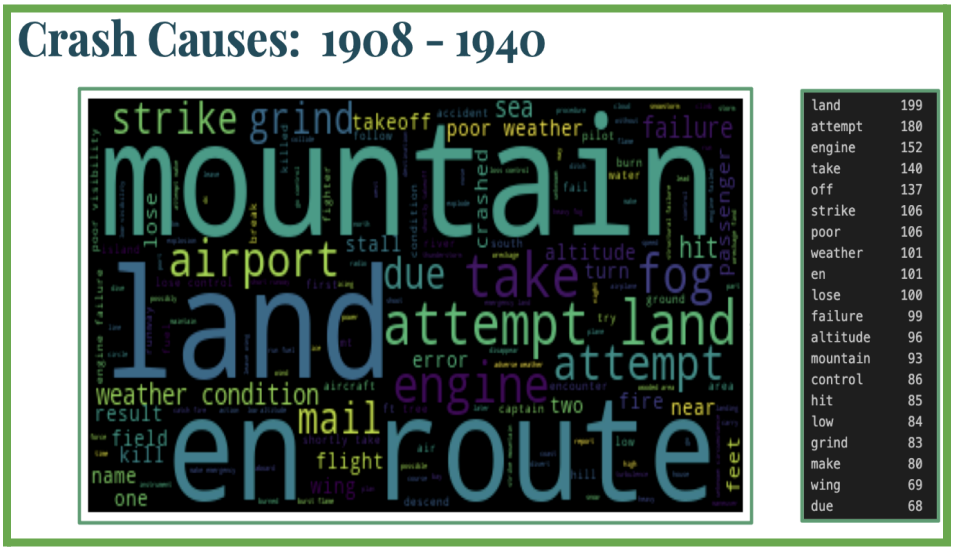




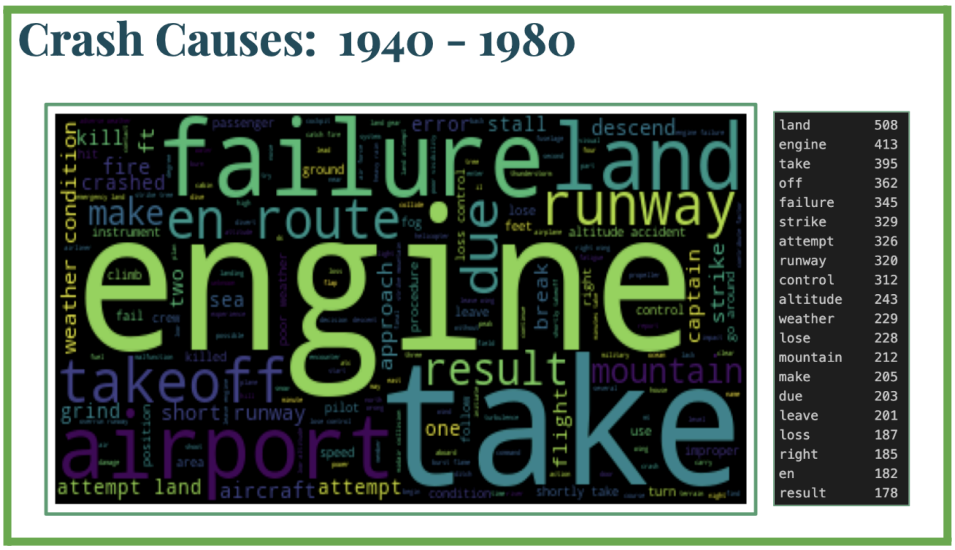
Take Away: We combined all U.S Military crashes into one group. The U.S military as a single operator has the second-highest fatalities and the highest crashes. Pen Am World Airways, Aeroflot, Avianca, are U.S, Russian, Colombian owned respectively. In the Top 10 by fatalities, we see 40% operators from the USA alone, which is the most of any single country. On the Top 10 by number of crashes by operator this number increases to 50%.

Crashes Cause:
Another data column we used was the summary of the crash report, see page 3. As you can see this is a free text field. We grouped all the summary data into 3 bins spanning ~40 years. We used the NLP tool kit for normalization: {remove stopwords, and lemmatizer}, this was done to remove irrelevant words and group the ones that have a similar verb. We also added additional filters to remove words like crash since we are dealing with crash data so this would be common everywhere.

1908-1940:
The verb 'land' appears most frequently, which might suggest overuse. However, terms like "engine", "take off", "failure" and "weather" also occur frequently indicating weather related issues and maintenance issues are quite prevalent.



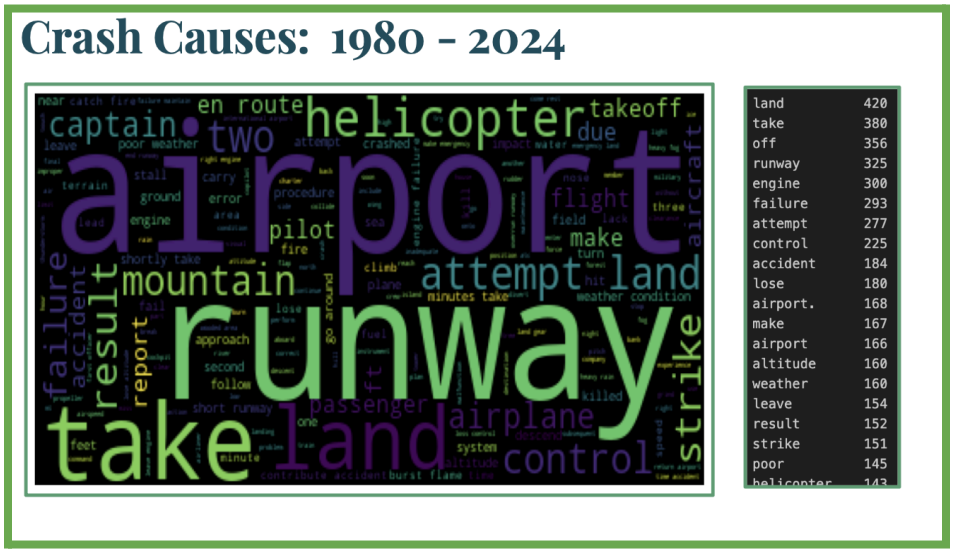
1940 - 1980: Similarly, words like “mountain” and “strike” are very interesting suggesting that these could be areas that are influencing crashes, this is consistent too. “Weather” is also an interesting one, and “runway” too. The only difference is that the magnitude is amplified by more than double in most cases. From the time plot we can see that flying as a mode of transportation increased after 1945.



.

1980-present:

Covid happened at the end of 2019-2020, and this data goes to 2024. It's surprising that the count of words is a little less compared to the previous period. Engine and run way, take off are very common words much like the previous periods and weather is still an issue.



Conclusion:
In conclusion, our comprehensive analysis of airplane crash data has shed light on several key insights regarding the frequency, distribution, and contributing factors of aviation incidents.
Firstly, our findings underscore that the U.S emerges as a focal point for airplane crashes, particularly along coastal regions such as California, Texas, Florida and Alaska. This concentration of incidents prompts a deeper examination into the underlying causes and regulatory frameworks governing air travel within the country.

Furthermore, our exploration reveals a significant influence of historical events, such as the tragic events of September 11, 2001, on the patterns of crashes and fatalities. The subsequent establishment of the Transportation Security Administration (TSA) reflects a pivotal moment in aviation safety, leading to noticeable improvements in mitigating risks associated with air travel in terms of crashes and fatalities.

Moreover, our analysis of operator-specific data highlights the diverse landscape of stakeholders involved in aviation incidents. While the U.S. military emerges as a prominent operator with a significant number of crashes, a closer examination reveals a broader global representation among the top operators by fatalities and crashes.

The temporal (time series) analysis of crash causes across distinct time periods highlights the evolving trends and challenges in aviation safety. From the prominence of “engine” “failures” to the persistent influence of “weather” conditions and operational factors like “take-off” and “runway” issues, our study underscores the multifaceted nature of crash causation.

In addressing our initial inquiries, our investigation reveals that airplane crashes are most common in the U.S, with notable variations in distribution and severity across different regions and operators. Over time, advancements in regulatory oversight and technological innovations have contributed to enhancing aviation safety, yet persistent challenges persist, necessitating continued vigilance and proactive measures to mitigate risks and ensure passenger well-being.

In summary, our analysis serves as a valuable resource for understanding the dynamics of airplane crashes, informing stakeholders across the aviation industry and regulatory bodies in their ongoing efforts to improve safety standards and practices.