This dataset contains the full history of airplane crashes involving civil, commercial and military transport throughout the world, from 1908 till 2009. The data was originally collected by “Socrata” a software company based in Seattle, Washington. According to Wikipedia Socrata was originally launched as “the worlds easiest database” and they marketed themselves as a online software as a service provider.

The dataset was hosted on https://opendata.socrata.com/Government/Airplane-Crashes-and-Fatalities-Since-1908/q2te-8cvq by the company. Unfortunately the original source of the dataset is no longer available which we suspect has something to do with fact that “Socrata” was acquired by “Tyler Technologies” in 2018. Our source for this dataset is https://data.world/data-society/airplane-crashes.

Since the original source for the dataset is no longer available it is difficult for us to understand the rational behind collecting and compiling it. It is also difficult to fully understand and interrogate the methods used to collect the dataset of this scale. However we can speculate that this dataset was collected as a means for the company to invite people from a diverse range of backgrounds to come and explore this dataset as a sample within their online database service which in turn promotes their service and gives people a chance to experience it first hand.

We can also further speculate that the data set was compiled and hosted as an attempt to promote the Open Data Network initiative launched by the company in July, 2014 which was designed to foster collaboration between governments and the private sector. Hence this could very well have been put together as a result of this collaborative effort to share different forms of data including crime data, transit data, incident data and expenditure data.

The data was easily accessible in Excel. Furthermore, it was also very easy to access it in Python as a Pandas Data Frame. As far as importing the data is concerned no issues arose but a slight hiccup that happened was that the Dates were being recognized in ‘Object’ format as opposed to ‘datetime64’ format hence the data had to be reimported while explicitly parsing the ‘Date’ column as dates. This was necessary so that the data might be ordered with respect to dates.

The dataset contains 13 dimensions listed below including their types and what they represent.

1. Date	datetime64[ns]	The date of crash
2. Time	object	Local time, in 24 hr. in the format hh:mm
3. Location	object	Location of the crash
4. Operator	object	Airline or operator of the aircraft
5. Flight #	object	The flight number as assigned by the operator
6. Route	object	Complete or partial route flown prior to the accident
7. Type	object	Aircraft type/model
8. Registration	object	International Civil Aviation Organization registration of the aircraft
9. cn/In	object	Serial number of the plane
10. Aboard	float64	Total number of people aboard
11. Fatalities	float64	Total number of deaths
12. Ground	float64	Total number of fatalities on ground as a result of the crash
13. Summary	object	Brief description of the accident and cause if known

- Based on this dataset we can observe that a total of 5268 crashes have been recorded from 17-09-1908 to 08-06-2009.

- We can also observe that the average survival rate from 17-09-1908 to 08-06-2009 is 16.5%.

- We can observe here that the highest number of crashes (a total of 15 for each) have happened in Moscow, Russia and Sao Paulo, Brazil.

- We can also observe that Aeroflot was the operator for at least 9 of the crashes (out of 15) that happened in Moscow, Russia.

- It can be observed that the highest number of crashes within the dataset have been flown by "Aeroflot" accounting for 179 crashes followed by "Military - U.S Air Force" accounting for 176 crashes.

- We can observe that out of 179 crashes suffered by Aeroflot 9 of those have been in Moscow, Russia followed by 4 near Moscow, Russia.

- The Military - U.S Airforce has suffered a total of 13 out of 176 crashes while flying on the training route.

- The highest number of crashes have occurred on the training route accounting for a total of 81 crashes followed by the sightseeing route which accounts for a total of 29 crashes.

- The most common type of aircraft to crash has been the Douglas DC-3 accounting for a total of 334 crashes.

Looking at the first 20 rows of the dataset we can observe that a significant number of values are missing within each of the columns hence we analyze each column further to assess the state of missing values.

Our analysis shows that around 79.7% of the total values are missing within the Flight# column which is then followed by 42.1% within the Time column. However on a more positive note the columns we are particularly interested in exploring didn't have as many missing values as demonstrated in the table below.

1. Route	32.4% Missing Values
2. Date	0% Missing Values
3. Location	0.4% Missing Values
4. Operator	0.3% Missing Values
5. Type	0.5% Missing Values
6. Aboard	0.4% Missing Values
7. Fatalities	0.2% Missing Values
8.	

This data is accurate for researchers trying to understand patterns or correlations between crashes and the various factors that might have contributed to the final outcome. This analysis can then be useful to help prevent future accidents by determining the most common factors leadings to airplane crashes. The data may also be used to simulate different scenarios and estimate chances of survival while taking various factors into account such as the type of plane or the number of people aboard or the crash track record of the operator.

It is unclear whether this dataset accounts for onboard crew members or not.