Draft.

What is the statistical likeliness to survive a plane crash

2.Problem Statement:

Ever since I was a kid airplanes and their ability to fly thousands of miles above the ground and thousands of miles across the globe fascinated me. I wanted to comprehend how humans had achieved such accomplishment and the intricates that came together to make this possible. How did we create such powerful machines, how is it possible that such heavy instrument that carries equally as heavy loads of cargo, passengers and fuel is able to surf over the clouds just like another bird. All these questions needed answer in my mind, but the question that I was intrigued the most and really needed and answer quickly was, what is the likeliness to survive a plane crash thousands of miles above the ground and at such high speed. In this project I will be demonstrating the likeliness to survive a plane crash based on the provided dataset by looking into the different causes that produce an accidents and the amount of losses reported. With the inspection of the dataset I will be able to answer other questions such as, what are the most common causes of an airplane accident, which airlines are categorized as most dangerous to fly with on regards of this specific dataset and the number of accidents registered, which locations are the most dangerous based on the number of accidents they produce, how many accidents happen yearly within the range of the dataset, which aircraft type has the most accidents recorded and which season are accidents more likely to occur.

3. Dataset used for this project:

The data set used for this project is Airplane_Crashes_and_Fatalities_Since_1908. A collection of aircraft accidents occured in between the years 1908 and 2009, this dataset contains airplane accidents but is not limited to airplanes, therefore it also includes recorded data of Zeppelin accidents and any other type of aircraft. This dataset contains 5269 entries of different accidents with 16 columns used to log the Date of the accident, Time of the accident, Location of the accident, Operator, Flight number, the Route, Type of aircraft, Registration number, Number of people aboard, Number of Fatalities, and a column with a summary explaining the apparent causes of the accidents. Some of the columns are missing values which has been accounted for through the study. The type of variables that the dataset contains are as followed:

·Int: Fatalities, People Aboard, Flight Number

·Categorical: Location, Route, Type, Registration, Summary

- Date – The date at which the accident occurred, written as month, day and year
- Time – The time at which the accident occurred
- Location – The approximate area in which the plane had the accident
- Operator – What company or group the aircraft was registered too
- Flight – The designated name of the flight path the aircraft was on e.g MH370
- Route – The flight path the aircraft was traveling on.
- Type – The model of aircraft involved in the accident

- Registration – The alpha-numeric that the aircraft is registered as

- Cn/ln – An abbreviation for construction number and line number, this denotes the number assigned to the aircraft production line and what number the specific aircraft rolled off the production line.

- Aboard – The number of people that were inside the aircraft at the time of the accident.

- Fatalities – Number of passengers that where killed as a result of the accident

- Ground – Number of fatalities that were on the ground and died as a result of the accident.

- Summary – A short written summary of what caused the accident to occur.

The most important variables going to be used in this study are people aboard, number of fatalities and summary. To accurately determine how likely it is for an individual to survive a plane crash I am going to focus on the summary to determine the different causes involved in a crash and look for the most common causes in order to establish a pattern. The number of people aboard and the number of fatalities are going to give me an accurate representation of how significant the accident was and how likely it is that a similar accident causes the same amount of loss and destruction. Other variables that are going to play an important role in this study are location and type, which in this case is referring to the type of aircraft. By looking at the amount of fatalities a location has linked to it, we can determine how dangerous a certain place can be and by looking at the type of aircraft involved we can determine if the accidents have some sort of pattern related with manufacture.

4. Tools used in the project:

In order to process, analyze, mold, study and produce the adequate data for this study I used an extended variety of tools that allowed me to work with the data in the most efficient way benefiting from the many different advantages each different tool could offer. In order for me to produce most of the output, graphs and tables python was the selected tool.

"Python is an interpreted, object-oriented, high-level programming language with dynamic semantics.." (CITATION APA) .The way I was able to use python was by writing different scripts that could iterate through the entire dataset in a fast manner to produce outputs seen in the Figures [1-14]. Python provided the flexibility necessary to apply basic math and compute various results used to answer the questions proposed for the study. It was also very useful when producing a variety of graphs in diverse colors and sizes that I thought necessary in order to properly display the data. Another similar tool I used to display data was R. *"R is a language and environment for statistical computing and graphics"*(CITATION APA). R was used to display the bar graphs seen in Figures [2,4,6], solely because python was not able to produce the desired layout. To achieve this I wrote a script where I was able to import the required libraries and display the data in a neat and suitable way. For text analysis and examination of sentiment and repetition I used Orange. *"Orange is a component-based data mining software."* (CITATION APA). With Orange I was able to analyze the Summary column of the dataset where the thought to be the causes of the accidents are explained. Since some of the cells are missing values, or the description is vague and does not contain much information the output obtained from Orange was very rough. The last tool used to analyze the dataset was Excel simply for visualisation purposes as I was able to load the CSV file and read it.

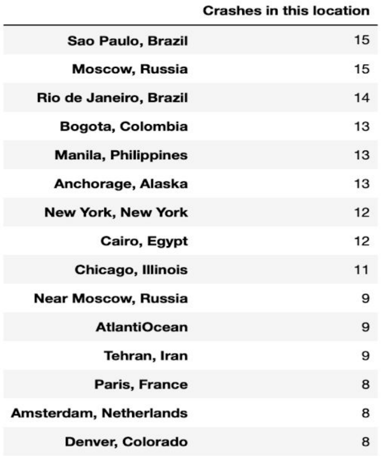
5. Data acquisition:

In order for the information to be extracted efficiently within a reasonable time manner I used a python script. With this script I was able to determine how many values the dataset was missing, how many entries the set contained, the number of how many people were reported aboard an aircraft for the entire dataset, how many people were reported as a fatality after an accident, also for the entire dataset, and deleted the columns that were not going to be useful for this study from the dataset. As these past values were all numbers, it was beneficial to use a script. To extract valuable data from the summary section of the set I used Orange for text mining functions; this allowed for a visual representation of what sentences were the most recurring for the 5962 entries of the set and therefore extracted and categorized the different types of accidents. I used this same technique to analyze the locations column and also get a visual of which location had the most occurrences in the dataset which therefore lead to be labeled as most dangerous location. But since I was interested in generating a list of the most dangerous locations I turned to the weight that each word carried in the set sorting it in descending order to produce the list. As the words obtained by the word cloud could be perceived as bague, I decided to read through the summary and the report that the word cloud generated in order for me to get a general idea of what the causes of an accident could be and generate my own script where this causes would be found in the summary so that they were easier to categorize, Ex: cause of accident engine failure, I made an array with the definition such as (‘engine failure’) and then adding possible causes of engine failure, fire or fail, resulting in this statement: *'engine failure': 'engine.*(fire|fail)'.*

6. Data preprocessing:(Explain the struggles you had while obtaining the data)

7. Data analysis and results: (need to structure the page)

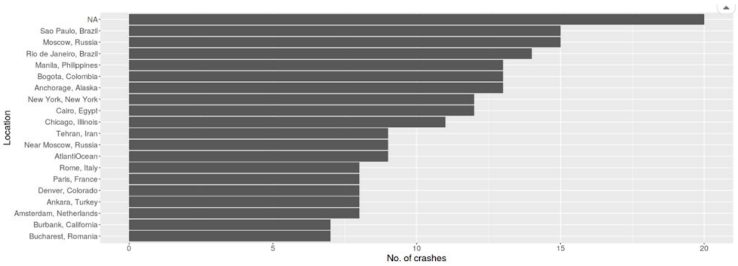
Figure 1.



	Crashes in this location
Sao Paulo, Brazil	15
Moscow, Russia	15
Rio de Janeiro, Brazil	14
Bogota, Colombia	13
Manila, Philippines	13
Anchorage, Alaska	13
New York, New York	12
Cairo, Egypt	12
Chicago, Illinois	11
Near Moscow, Russia	9
AtlantiOcean	9
Tehran, Iran	9
Paris, France	8
Amsterdam, Netherlands	8
Denver, Colorado	8

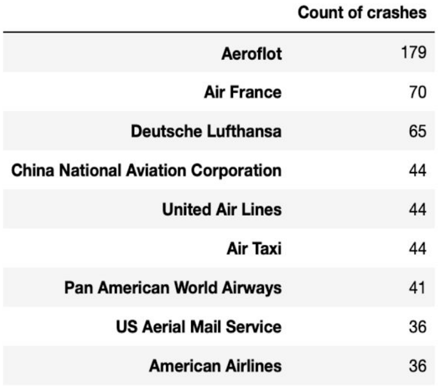
Note. This table shows the locations with the most amount of crashes reported listed in descending order with a minimal threshold of 8 crashes per location.

Figure 2.



Note. This graph shows the locations with most accidents reported. N/A means that the location of the accident could not be reported or the crash did not occur in land.

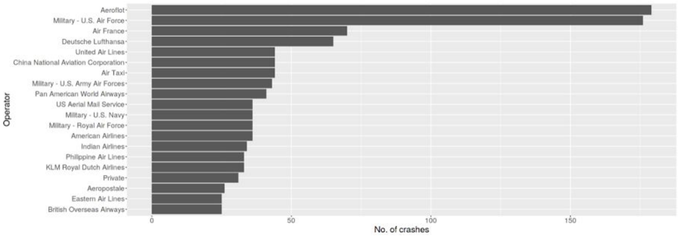
Figure 3.



	Count of crashes
Aeroflot	179
Air France	70
Deutsche Lufthansa	65
China National Aviation Corporation	44
United Air Lines	44
Air Taxi	44
Pan American World Airways	41
US Aerial Mail Service	36
American Airlines	36

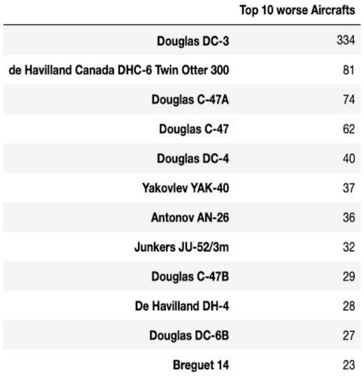
Note. This table shows the airlines with the most amount of crashes recorded in descending order with a minimal threshold of 36 crashes.

Figure 4.



Note. This graph shows the airlines with most accidents recorded including non-passenger planes such as the US Military for reference.

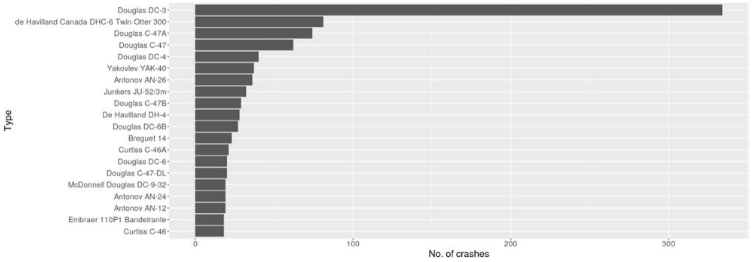
Figure 5.



	Top 10 worse Aircrafts
Douglas DC-3	334
de Havilland Canada DHC-6 Twin Otter 300	81
Douglas C-47A	74
Douglas C-47	62
Douglas DC-4	40
Yakovlev YAK-40	37
Antonov AN-26	36
Junkers JU-52/3m	32
Douglas C-47B	29
De Havilland DH-4	28
Douglas DC-6B	27
Breguet 14	23

Note. This table shows the different Aircraft types with the most amount of crashes recorded listed in descending order with a minimal threshold of 23 crashes.

Figure 6.



Note. This graph shows the Aircraft type with most accidents. With the threshold extended from 23 to 15.

Figure 7.

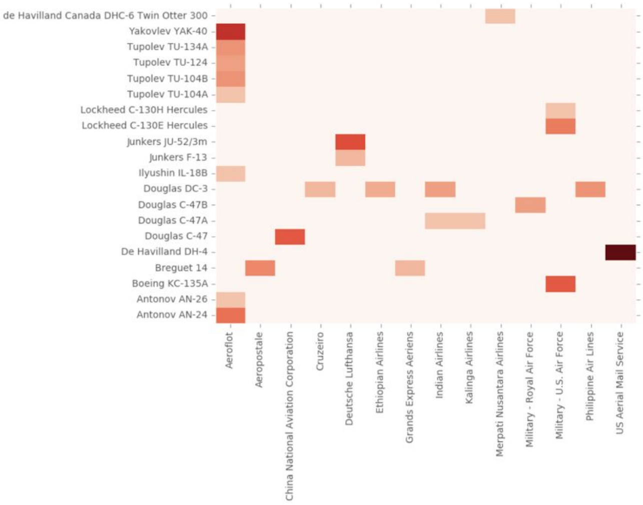
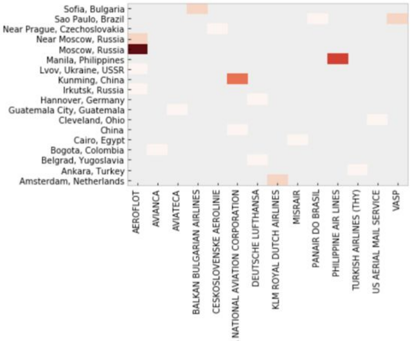
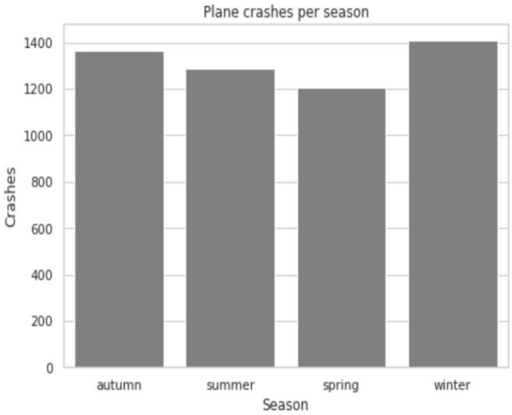


Figure 8.



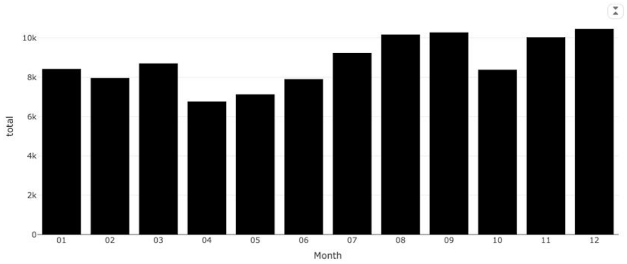
Note. Figure 7 & 8 are heat maps that show which airlines have crashed which type of aircraft and the locations of the accidents. The darker the color the more incidents recorded by the airline with the certain type of plane at a certain location respectively.

Figure 9.



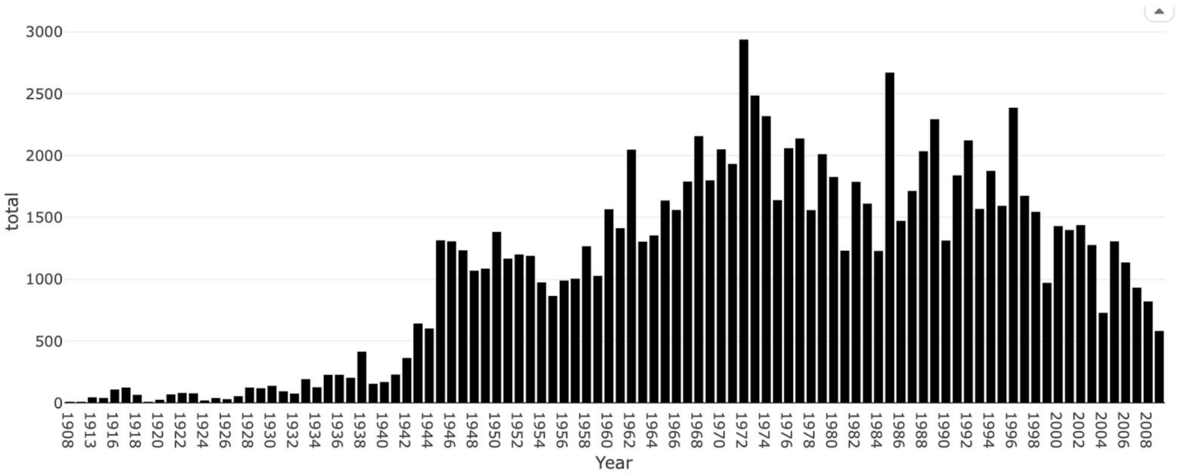
Note. This graph shows the seasons where accidents were reported.
·Spring: Months 3 to 5
·Summer: Months 6-8
·Autumn : Months 9-11
·Winter : Months 12-2

Figure 10.



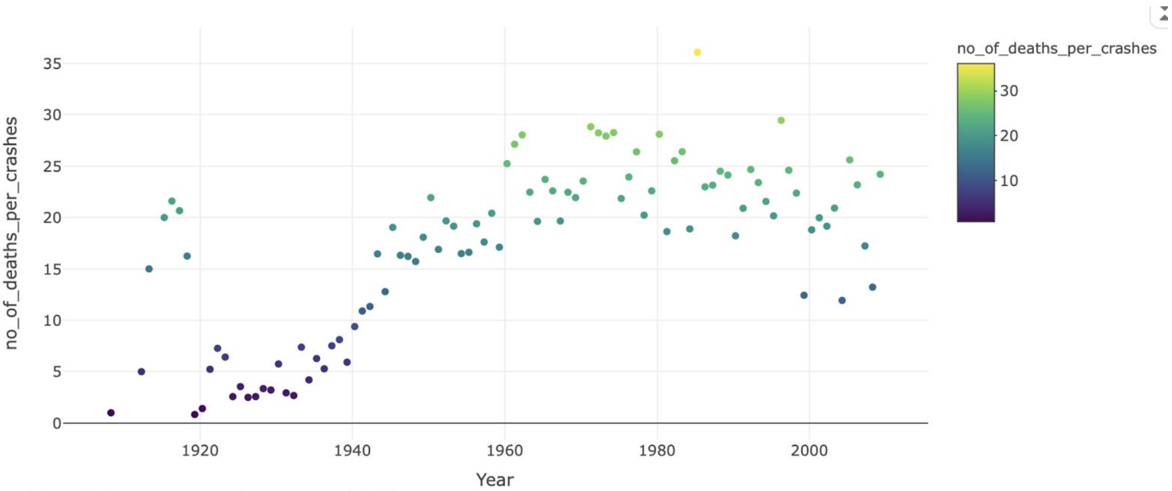
Note. This graph shows the months where different accidents have been reported.

Figure 11.



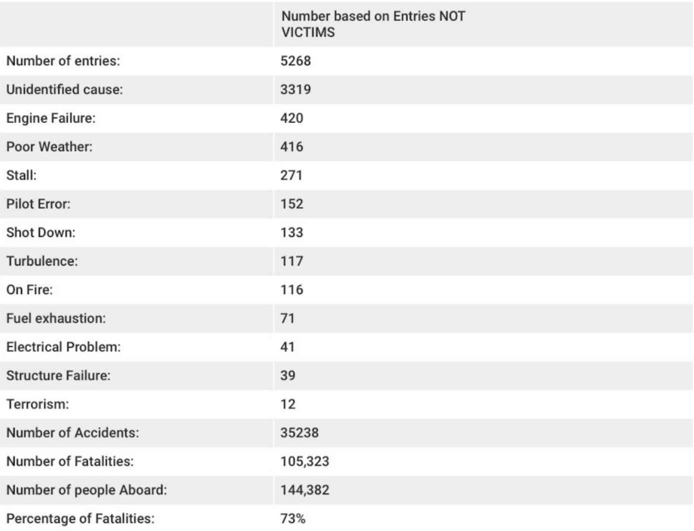
Note. This graph shows the amount of accidents recorded yearly from the year 1908 to 2008.

Figure 12.



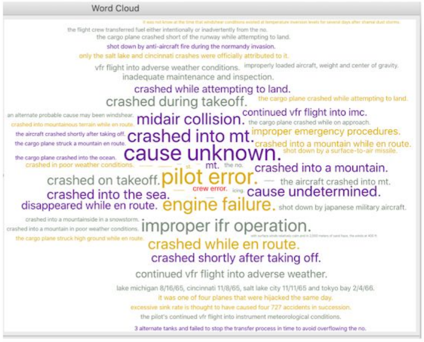
Note. This graph shows the amount of victims recorded yearly from the year 1908 to 2008.

Figure 13.



	Number based on Entries NOT VICTIMS
Number of entries:	5268
Unidentified cause:	3319
Engine Failure:	420
Poor Weather:	416
Stall:	271
Pilot Error:	152
Shot Down:	133
Turbulence:	117
On Fire:	116
Fuel exhaustion:	71
Electrical Problem:	41
Structure Failure:	39
Terrorism:	12
Number of Accidents:	35238
Number of Fatalities:	105,323
Number of people Aboard:	144,382
Percentage of Fatalities:	73%

Figure 14.


Word Cloud
it was not know at the time that windshear conditions existed at temperature inversion levels for several days after shamal dust storms.
the flight crew transferred fuel either intentionally or inadvertently from the no.
the cargo plane crashed short of the runway while attempting to land.
shot down by anti-aircraft fire during the normandy invasion.
only the salt lake and cincinnati crashes were officially attributed to it.
vfr flight into adverse weather conditions.
improperly loaded aircraft, weight and center of gravity.
inadequate maintenance and inspection.
crashed while attempting to land.
crashed during takeoff.
the cargo plane crashed while attempting to land.
an alternate probable cause may been windshear.
midair collision.
continued vfr flight into imc.
the cargo plane crashed while on approach.
crashed into mountainous terrain while en route.
the aircraft crashed shortly after taking off.
crashed into mt.
improper emergency procedures.
the cargo plane struck a mountain en route.
crashed into a mountain while en route.
cause unknown.
shot down by a surface-to-air missile.
the cargo plane crashed into the ocean.
crashed in poor weather conditions.
st.
mt.
the no.
crashed into a mountain.
crashed on takeoff.
pilot error.
the aircraft crashed into mt.
crew error.
icing.
cause undetermined.
crashed into the sea.
disappeared while en route.
engine failure.
shot down by japanese military aircraft.
crashed into a mountainside in a snowstorm.
crashed into a mountain in poor weather conditions.
improper ifr operation.
the cargo plane struck high ground while en route.
with surface winds relatively calm and in 2,000 meters of sand haze, the winds at 400 ft
crashed while en route.
crashed shortly after taking off.
continued vfr flight into adverse weather.
lake michigan 8/16/65, cincinnati 11/8/65, salt lake city 11/11/65 and tokyo bay 2/4/66.
it was one of four planes that were hijacked the same day.
excessive sink rate is thought to have caused four 727 accidents in succession.
the pilot's continued vfr flight into instrument meteorological conditions.
3 alternate tanks and failed to stop the transfer process in time to avoid overflowing the no.

Note. Figure 13 & 14 show the different reasons thought to be the causes of accidents reported.

8. Your timeline for completion:

9. Team workload and roles:

10. Citations: