What is the statistical likeliness to survive a plane crash

Iho López Tobi

Data Mining

April 24th 2020

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 available in *[Figure 1]*. 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.

Figure 1.

Data Variables	Description
Cn/In:	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.
Summary:	A short written summary of what caused the accident to occur.
Ground:	Number of fatalities that were on the ground and died as a result of the accident.
Fatalities:	Number of passengers that were killed as a result of the accident
Registration:	The alpha-numeric that the aircraft is registered as
Location:	The approximate area in which the plane had the accident
Date:	The date at which the accident occurred, written as month, day and year
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
Aboard:	The number of people that were inside the aircraft at the time of the accident.
Time:	The time at which the accident occurred
Operator:	What company or group the aircraft was registered too

Note. Table definition for the variables in the dataset

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.." (What is Python? Executive Summary) .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 and do proper data analysis was R. *"R is a language and environment for statistical computing and graphics"*(What is R?). 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."* (Read the Docs). 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:

Acquiring the dataset used for this study did not require an extensive process. I started by accessing the website *"data.world"* that contained the dataset. In order for me to visualize the data and have full access to it I had to create an account or sign up with Google if I did not have an account on the site already. I chose to sign up with google to save time, once I was signed up I had full access to the set and I could view every element of the set, perform projects on the website itself such as running different scripts with different programming languages and even download it to my machine as it has a Open Database licence. The dataset is originally from a website called *" opendata.socrata.com"* but unfortunately the set was no longer on the site, it could have been deleted by the author, or deleted due to the site being deprecated in July 2020. Once I obtained full access I simply downloaded the set to my machine and started working on in.

6. Data preprocessing:

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 and which ones, how many entries the set contained, the number of how many people were reported aboard an aircraft, how many people were reported as a fatality after an accident, 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 as I could just perform math easily. Once I obtained the number determining how many values the dataset was missing I decided if the column was a numeric value, as for example Fatalities or Aboard, to fill in with the number 0, with the help of a script, if the column was a text column like summary I decided to leave it blank and prompt the script to not look at those with

missing information. If the columns where Time, Date Cn/ln, registration or route, I simply delete the rows as I decided to not use them in the dataset as they did not provide any relevant data to this particular study. 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. As the words obtained by the word cloud could be perceived as bague and did not give a clear insight, 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. Then I generate my own script with my own definition of what I thought to be fatalities based on the summary so that they were easier to categorize, Ex: cause of accident engine failure, I made an array with various definitions, such as (‘engine failure’), and then added possible causes of an engine failure such as fire or fail, resulting in this statement: *'engine failure': 'engine.*(fire|fail)'. [Figure1]* . Initially I used the same initial approach of using Orange, to analyze the locations section of the dataset to also get a visual of which location had the most occurrences in the dataset to produce the list of most dangerous locations. But since the world cloud was not able to produce the level of accuracy I needed since it was not parsing the name of locations correctly and I could not rely on the weight of the words on relation to the dataset I decided to use a python script to generate the list of most dangerous locations with a minimum threshold of 8 accidents recorded per location. *[Figure 2]*.

Figure 1.

```
data = pd.read_csv('Clean_data.csv')

failures = {
    'pilot error': '(pilot|crew) (error|fatigue)',
    'engine failure': 'engine.*(fire|fail)',
    'structure failure': '(structural fail)|(fuel leak)
    |(langing gear)',
    'electrical problem': 'electrical',
    'poor weather': '((poor|bad).*(weather|visibility)|
    thunderstorm)',
    'stall': 'stall',
    'on fire': '(caught fire)|(caught on fire)',
    'turbulence': 'turbulence',
    'fuel exhaustion': '(out of fuel)|(fuel.*exhaust)',
    'terrorism': 'terrorist|terrorism',
    'shot down': 'shot down',
}
```

Note. Definitions used to look for different causes of accidents reported in the Summary section of the dataset.

Figure 2.

```
loc_list = Counter(data['Location'].dropna()).most_common(15)
locs = []
crashes = []
for loc in loc_list:
    locs.append(loc[0])
    crashes.append(loc[1])
print('Top 15 the most dangerous locations')
pd.DataFrame({'Crashes in this location' : crashes}, index=locs)
```

Note. Definitions used to look for the locations with most accidents with a minimum of 8 accidents reported.

Figure 3.

```
Data.isnull().sum() #cal
```

```
Date                0
Time             2219
Location           20
Operator           18
Flight #         4199
Route            1706
Type               27
Registration      335
cn/In            1228
Aboard             22
Fatalities         12
Ground             22
Summary           390
dtype: int64
```

Note. List of dataset variables and missing values

7. Data analysis and results:

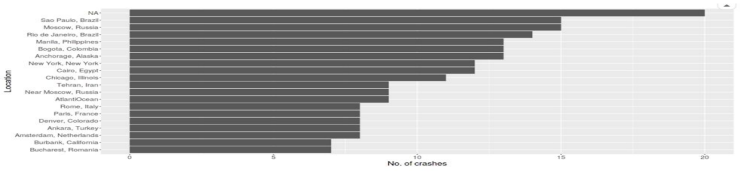
In this section I will be discussing the various findings I was able to obtain from the dataset and explain their significance and relevance to the study. One of the questions proposed was which locations were the most dangerous. I was able to produce a list of fifteen locations that recorded a high number of accidents with a minimum threshold of accidents at a location set at 8 accidents and the maximum at 15. [Figure 1-2] The reason why the minimum was set to 8 is because it was a consistent number for the amount of accidents recorded in a particular location, lower numbers did not make the data significant enough as the higher the number of accidents in a singular location the more likely it is to indicate that there is something in that particular location that makes it so prone to accidents. Based on the data Sao Paulo,Brazil and Moscow, Russia are the most dangerous locations due to the high number of accident associated to them, these two locations register over 400 deaths for Moscow and 300 for Sao Paulo. Geographically these two locations do not share many variables, the weather is drastically different from one another and the latitudes are very different as well. To answer why these cities are so dangerous we have to identify who is having trouble in said locations, and why, and most important what sort of trouble are the planes that crash in those locations having.

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.

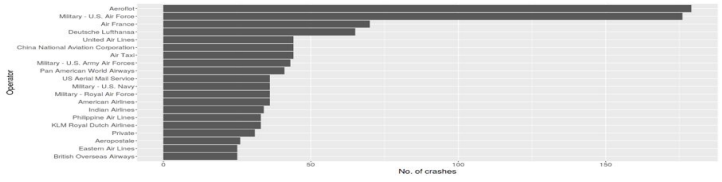
To identify the airline with the higher number of accidents I decided to look at the highest number of accidents register in the data set by a single airline. This returned the airline named Aeroflot, the Russian national airline with a total of 179 accidents producing over 7000 fatalities. Then I decided to look which airlines followed Aeroflot in terms of accidents and produced a top ten list with the airlines that had recorded the most amount of accidents. The airline that was considered to be the second most dangerous in the dataset was far off the mark from recording the all time high 179, it did not even record half of the max. Air France has produced well over a thousand fatalities but Aeroflot is seven times more deadly. The other airlines reported to be dangerous all were within a close range of Air France, which means that at some point in time accidents became very prone. Airlines identified this growing issue and implemented changes, but what happened with Aeroflot. With the piece of information stating that the Russian National Airline has produced the single highest number of incidents and fatalities, we can start to see a potentially developed relation between one of the most dangerous locations, Moscow and this airline. We still have to understand why these airlines are having so much trouble, in particular Aeroflot. What type of planes are the airlines using, to identify if the issues are due to manufacturer, crew, weather or a combination of all.

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.

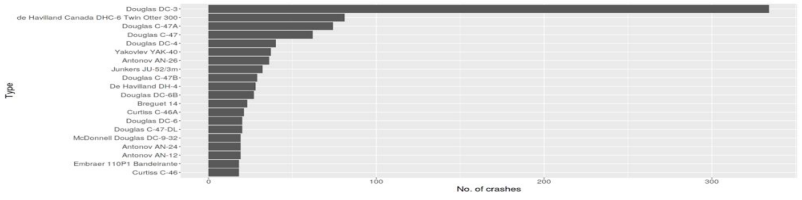
When it comes to aircrafts in the dataset and accidents reported, not all of these planes are transporting civilians or are used for regular transportation. It is common to see one type of airplane being used for the transportation of merchaderies, individuals and even used during times of war since the dataset records accidents from 1908 up until 2009. When an aircraft type is used for an extensive array of things, it becomes more prone to be exposed to accidents. Douglas DC-3 has reported 334 accidents and scored close to five thousand fatalities alone. No Aircraft has recorded even half of such letal statistics. We have to understand that this particular model of aircraft has served different purposes through its beginnings. The Douglas DC-3 was the primary type of airplane used during world war I & II to transport parachuters to different war zones, therefore it is reasonable that during times of war the aircrafts would be involved in some sort of fatal incident causing several lifes to perish. The DC-3 was also the favorite airplane type used by airlines to transport passengers due to it efficiency to carry cargo and passengers for long distance travel. Therefore we can see how the extensive use of this aircraft could have produced such high mortality rate.

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.

The following images are attached solely for reference purposes. The legend for these graphs is based on shades of red, the darker the shade of red the higher the amount of accidents recorded for that particular instance. Thanks to the heat map we can see which airlines used which type of aircraft. We notice that Aeroflot has a high intensity color for the aircraft type "*Yakovlev yak-40"* and all the other instances of aircrafts recorded by this airlane are solely Russian planes, no other company has been recorded using Russian Aircrafts. So therefore they could not have contributed to any of the accidents produced by the *"Douglas DC-3"*. Any of the accident that Aeroflot had could be link to safety issues with the way the Russian companies manufacture the planes or the way the crew had been trained rather than being a common problem with the DC-3 since the airline did not use such planes. None of the Airlines present in the chart have a extreme dark shade of red for the *"Douglas DC-3",* but the Airlines that do record incidents with this type of aircraft do so with a bright enough shade of red that indicates the popularity of the plane among different airlines and the amount of incidents this same type produced for different companies. Figure 8 represents the locations where the various companies had been reported to have accidents. As we can appreciate Aeroflot has very strong colors for accidents reported in or near Moscow, Russia and any other accident ever reported was also within the old URSS territory now formally known as Russia and nearby countries.

Figure 7.

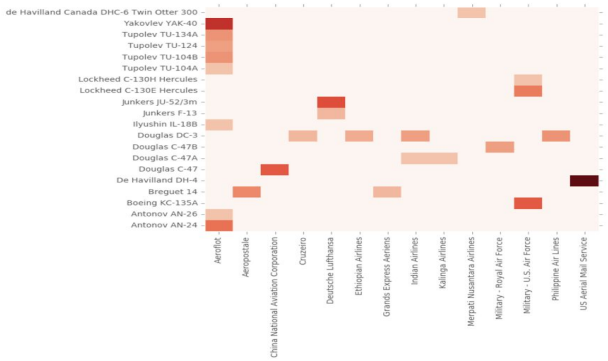
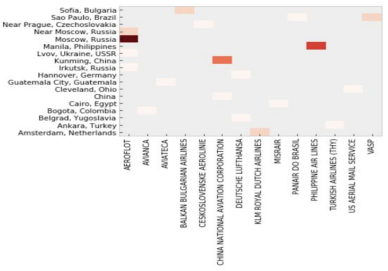


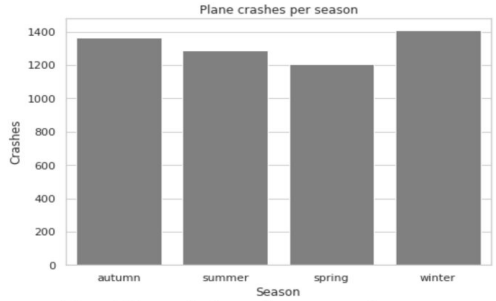
Figure 8.



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.

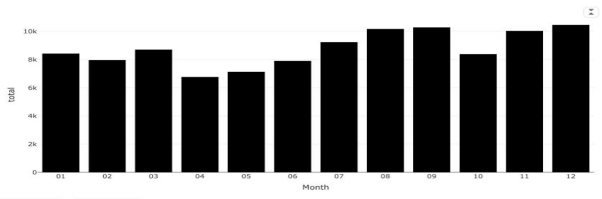
To properly understand the likeliness of surviving an airplane crash it is important to analyze the different external factors that are outside of human control, these factors can be predicted but not avoided and play a very important role in aircraft accidents as planes are exposed to these factors on a regular basis, seasons. Figure 9 represents the seasons where different number of accidents were reported to take place. As we can appreciate autumn and winter are the seasons were the higher number of accidents have been reported. If we think of the nature of these two seasons these are the ones containing the most drastic weather conditions ranging from heavy rains in autumn, to hailstones storm in winter that can seriously influence the performance of an aircraft and very much produce enough damage to cause an irregular performance mid flight that can end up in a serious catastrophe. Also an important observation from figure 9, is that summer does not fall far behind from autumn nor winter, we have to keep in mind that while in the northern hemisphere summer is known to be a warm and dry weather that is not the same for the southern hemisphere as summer is actually cold weather with snow storms and heavy precipitations. Figure 10 is a further breakdown of the months and the total number of fatalities that each month had recorded. With this amount of data gathered is time to put thing into perspective and understand how each piece plays a key role in terms of aircraft safety and transportation.

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

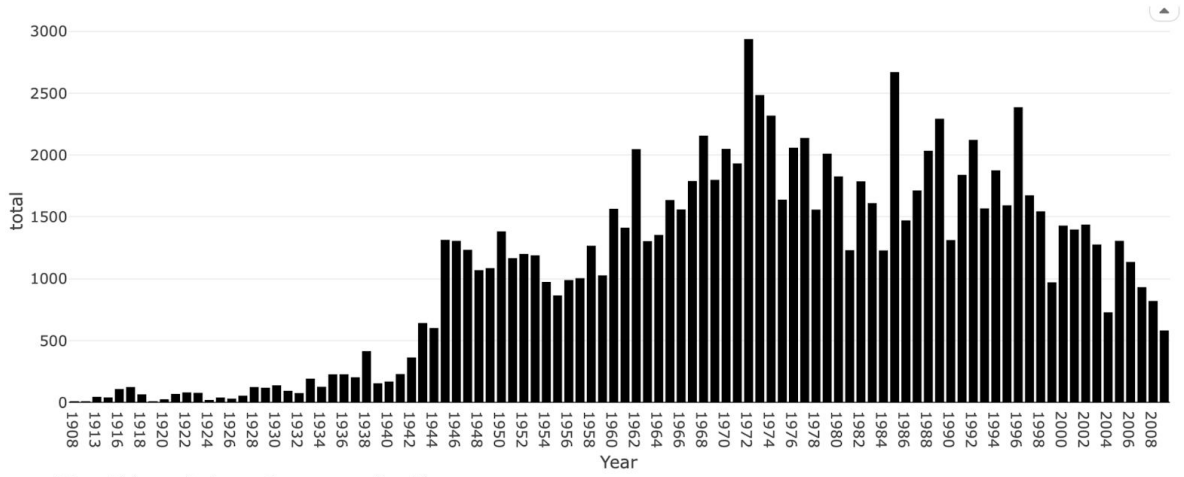
Figure 10.



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

Figure 11 is able to provide a complete breakdown through the years of how many accidents had been recorded yearly from the year 1908 to the year 2008. This graph is able to provide us with the visual of how aviation has evolved through the years and the quick adaptation of this. As it is appreciated there are a series of years were spikes in the number of accidents have been recorded these being year 1962, the most significant 1972, 1985, 1989 and 1996. To understand why these spikes have been recorded we have to understand what the demand for flying was at that time. In the year 1972 we can start to see a sudden burst in the demand of flying and going places that was not present before. People wanted to use a more efficient way of traveling. Flying was the way of going faster and allow to get further in less time. This demand caused for the manufacturers to rush production of their products, planes, as they had to fulfill the demand. Airlines where rushing the training of their crew members as they needed to put as many planes in the air as possible to be able to compete with other airlines that likely were doing the same in order to not lose clients. The under training of the crew would cause serious trouble as people would not be trained properly to take on extreme situations where specific training would be required that usually ended in a fatality. The often forgotten Air traffic controllers also played a key role in aircraft safety as they are in charge of supervising the traffic of planes in the air and in the airport to ensure no one is in danger. The demand experienced in 1972 did not account for the number of air traffic controllers needed, which lead to understaffing the control towers with people being overflowed with work by taking on too many planes at once and not being able to perform optimally resulting in the losing of planes, in air crashes, miss approaches to the runway, wrong runway landing and crashes in the tarmac as planes would land or take off in occupied runways. All these issues paved the way for the safety guidelines know in place that ensure our safety every time we fly.

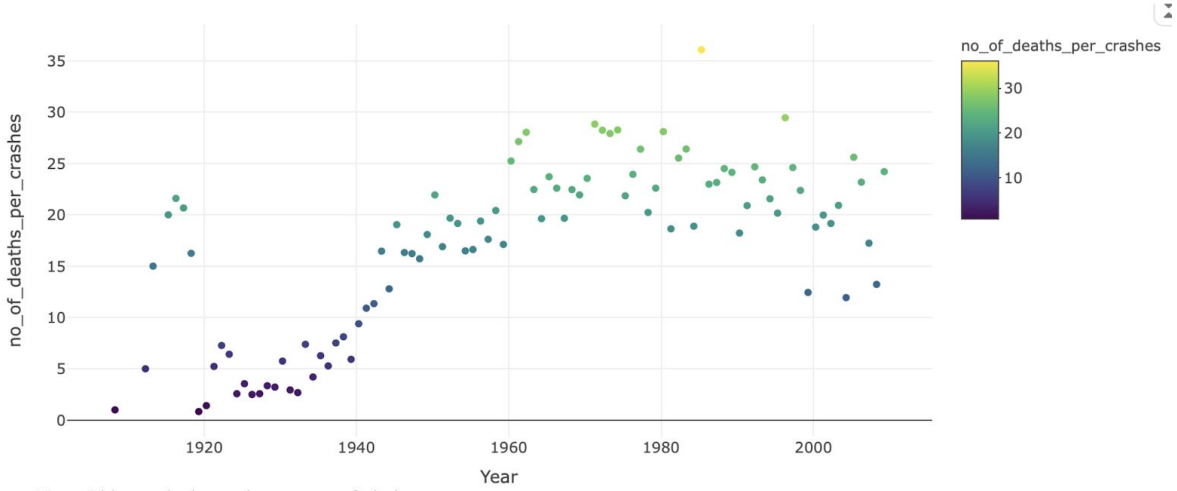
Figure 11.



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

Figure 12 is included for reference showing the amount of victims recorded through the years in a single crashes. The legend is represented as the brighter the color the higher the amount of deaths recorded.

Figure 12.



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

With all the data gathered through this study now is time to understand how each component plays an enormous role in the development of an aircraft accident. Figure 13 & 14 represent the most common causes of accidents found in the dataset that produced the 73% of fatalities. The highest cause recorded in the dataset and therefore most common cause of accident in regards of this particular dataset is engine failure. An engine failure is when the engines for a variety of reasons cannot produce enough trust and produce energy and stability to keep flying. It is very difficult to recover from an engine failure as the procedures to restart the engines take time to identify and resolve, in an environment where every second is valuable and critical an engine failure is the deadliest cause of accidents. Engines can fail for a variety of reasons, as for example when engines are unable to operate under extreme weather conditions that drive the engines to work overload and cause these to cease and fail. This is why Poor Weather is the following deadliest most dangerous non human related causes of accident. The ability of weather to change drastically and unexpectedly to very drastic levels is what makes it so dangerous for pilots.. If we recall the locations labeled to be the most dangerous in the dataset, São Paulo, Brazil and Moscow, Russia; we can understand that these locations do not share much weather similarities but instead the weather is very significant to each area respectively. Moscow,Russia is known to be in a very cold area with a lot of snow storms, hail storms, and below zero weather during winter time. These conditions are very harsh for the optimal operation of an aircraft. If there is a hail storm while flying the stones could be of a big enough diameter that can hit the engine fans damaging these preventing them from working correctly and resulting in engine failure from which it is very difficult to recover and ending up in a catastrophe (How Does Icing Affect Your Aircraft? (2018, February 13). Another cause issue is due to the weather ice sits on the wings of the planes incrementing the weight of the plane and not allowing for the friction of the air

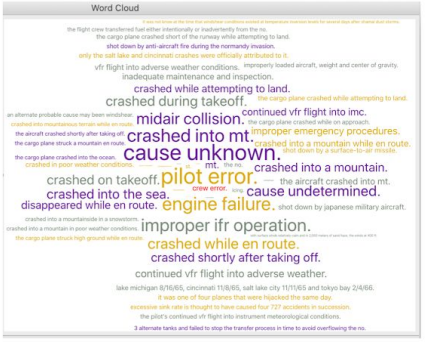
to distribute evenly through the wing, losing lift and finally causing the plane to plume down from the skies. As per São Paulo, it is a location characterized by tropical weather known for heavy rain and high winds. These conditions are critical as heavy rain can impair the visibility of a pilot, the Congonhas, Sao Paulo airport is located in the center of the city and is characterized by the extremely bad drainage of the runways which makes them really slippery, and the other characteristic of this airport is the mountain that pilots have to overcome in order to land, which under bad weather conditions this is really difficult to spot producing a high number of accidents(Congonhas Airport, Brazil.)(Schrader, R. (2020, March 18).The type of aircraft also plays a very important role in the resulting of airplane crashes. Different type of engines are going to be more prone to engine failures under harsh weather, or perhaps would be more prone to stalls if not enough trust could be produced due to the design of the engines. The DC-3 was the airplane type with most accidents reported of all times, these accidents most are attributed to the fifth cause of accidents in Figure 13, shot down. The DC-3 was the preferred airplane type used by the US military during world wars I & II therefore it was common for this plane to be caught in crossfire that would result in fatal accidents. The second most leading cause of accident for this plane type was engine failure and pilot error, but the reason why the DC-3's numbers are not concerning is because in relation to its popularity and the number of years this particular aircraft has been in service and the amount of airplanes produced, these numbers fall within reasonable parameters, in fact it would be very strange if this airplane type had recorded only half of the reported incidents it currently has. When it comes to worse airline we saw that Aeroflot was the worse airline of all times, we could start to see a relation between the type of airplanes the airline flew, only of Russian manufactured and the area where the accidents happened, Moscow. The Soviet era had had a very strong impact on nationalism which in this case meant that Russian

product was the only product to be used in Russia by Russian companies. With the collapse of the country's aircraft manufacturing industry, the company deemed to be involved in more frequent crashes. The more rare these aircrafts became so did the pieces required for proper routine maintenance, making it harder to service the planes that Aeroflot was using. The undermaintanace of the planes and the extremely harsh weather conditions of the area of operations, as Aeroflot was responsible for flying the entire length and breadth of the entire Soviet Union, carrying 100 million passengers in 1976, where the reasons Aeroflot was labeled one of the most deadliest companies of all times (Morris, H. (2018, February 12).

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.



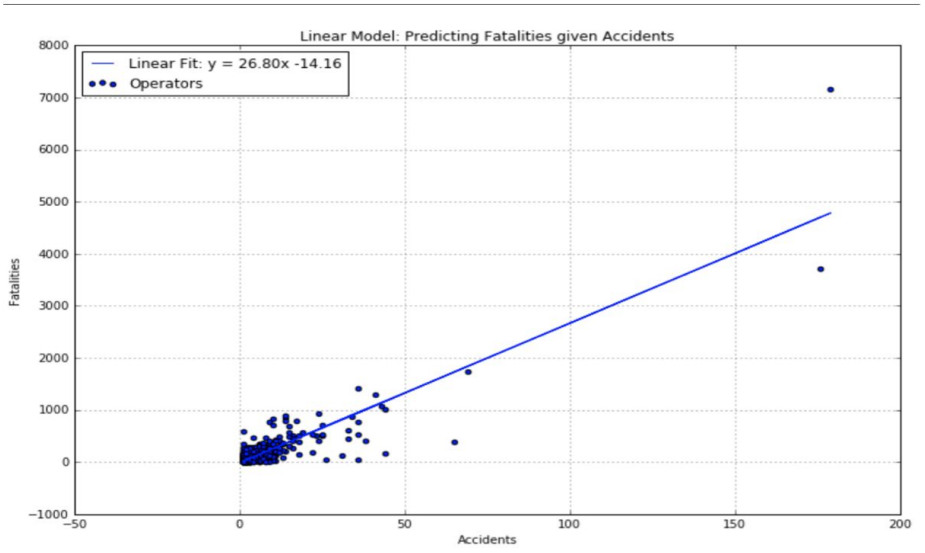
Note. Figure 13 & 14 show the different reasons thought to be the causes of accidents reported.

After understanding the different causes of airplane crashes and how the location, type of plane and even airline play a crucial aspect of safety how likely it is to survive a plane crash is the last question to answer. Figure 15,16 & 17 are going to be used to answer the question. Figure 15 is a linear model used to predict the amount of fatalities expected to be produced given an x amount of accidents and what the actual outcome was. We can see that at the very beginning of the linear fit the values jump all over the line, some are reported to be over,

some are reported to be under, but as we approach 50 there are a number of operators who have scored the number of fatalities to be over the line, which is a bad thing given that the linear regression model gives us the expected number of victims expected to be produced by that particular amount of accidents. At the very end of the line we can see how a company had recorded over 7000 victims for less than 200 accidents which is a very high number given that the regression model does not even get close to such values.

Figure 15.

Note. Figure 15 shows a linear regression model predicting how many fatalities an accident would produce.



With figure 16 we can see that the mean survival rate is of 16.5%, this means that if we were people in this particular dataset and we were to be involved in an airplane crash the chances of surviving the crash are of 16.5% . The minimum survivor in a single crash is 0 people, a very deadly crash and the maximum is 100 with a variance of 897.95 which mean that for this particular dataset the data points are very spread out from the mean and from one another(Roberts, D., & Roberts, F.)

Figure 16.

```
data_nobs = len(crashes["Survival Rate"])
data_mean = crashes["Survival Rate"].mean()
data_min = crashes["Survival Rate"].min()
data_max = crashes["Survival Rate"].max()
data_var = crashes["Survival Rate"].var()
data_skew = crashes["Survival Rate"].skew()
data_kurtosis = crashes["Survival Rate"].kurtosis()

print("Survival Rate Stats:")
print("Nobs: {}".format(round(data_nobs,2)))
print("Mean: {}".format(round(data_mean,2)))
print("Min: {}".format(round(data_min,2)))
print("Max: {}".format(round(data_max,2)))
print("Variance: {}".format(round(data_var,2)))
print("Skewness: {}".format(round(data_skew,2)))
print("Kurtosis: {}".format(round(data_kurtosis,2)))
```

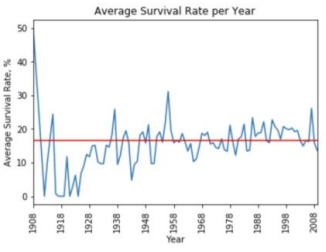
```
Survival Rate Stats:
Nobs: 5191
Mean: 16.57
Min: 0.0
Max: 100.0
Variance: 897.95
Skewness: 1.67
Kurtosis: 1.35
```

Note. Figure 16 shows the different survival rate statistics.

And lastly figure 17 is a visual representation of how the average survival rate had varied per year. What this means is that while the mean survival rate for this dataset was of 16.5% some years were reported to be above or below this mean value, the years reported below means that accidents were more deadly since more people were dying therefore the survival rate dropped, while those reported above the mean where better as more people were able to survive, the peek being 30% for year 1956.

Figure 17.

```
yearly_survival = crashes[["Date","Survival Rate"]].groupby(crashes["Date"].dt.year).agg(["mean"])
yearly_survival.plot(legend=None)
plt.ylabel("Average Survival Rate, %")
plt.xlabel("Year")
plt.title("Average Survival Rate per Year")
plt.xticks([x for x in range(1908,2009,10)], rotation='vertical')
plt.axhline(y=data_mean, color='r', linestyle='-')
plt.show()
```




```
train = pd.DataFrame(
np.random.rand(100, 3),
columns=['Fatalities','Aboard','Accidents'])
X = train[['Fatalities','Aboard']]
y = train[['Accidents']]
model = DecisionTreeRegressor().fit(X,y)
print("The first 5 predictions:",model.predict(X.head(5)))
```

```
The first 5 predictions: [0.82999536 0.20721209 0.95825625 0.37137628 0.79696111]
```

Note. Figure 17 shows the different average survival rate per year.

So given this particular dataset and after analyzing the different reasons that play into an airplane accident the chances of surviving an airplane crash for the people in this dataset was

of 16.5 percent meaning that the chances of surviving were really slim. And factors like weather, location, type of plane, airline and reason of accident played a really serious and significant role in determining if people were going to survive or not.

8. Your timeline for completion:

Week	Task	Day
Week 5,6,7,8,9	Data Analysis & Visualization	March 1 - April 19
Week 2,3,4	Data Cleaning/preprocessing	Feb 9 - March 1
Week 1	Data Acquisition	Feb 2-9

Week 1: Acquired the dataset from data.world

Week 2,3,4: Look for missing values in the dataset, filled in the missing values or eliminated irrelevant rows of information, made scripts for identifying the data types of the data and prepared information for analysis.

Week 5,6,7,8,9: Built the various scripts used for data visualization, built linear regression model, built script for text mining and used orange for text mining purposes, started working on the project report, finished all data visualization with meaningful information, finished linear model and decision tree, finish final project report.

9. Citations:

Airplane Crashes - dataset by data-society. (2016, December 4). Retrieved from https://data.world/data-society/airplane-crashes

Congonhas Airport, Brazil. (n.d.). Retrieved from http://www.orangesmile.com/extreme/en/scariest-runways/congonhas-airport.htm

How Does Icing Affect Your Aircraft? (2018, February 13). Retrieved from https://hartzellprop.com/icing-affect-aircraft/

Morris, H. (2018, February 12). Do Russian airlines still have a safety problem? Retrieved from https://www.telegraph.co.uk/travel/travel-truths/are-russia-airlines-safe/

Read the Docs. (n.d.). Retrieved from https://readthedocs.org/projects/orange3/

Roberts, D., & Roberts, F. (n.d.). Retrieved from https://mathbitsnotebook.com/Algebra1/StatisticsData/STSD.html

Schrader, R. (2020, March 18). Do Planes Fly in the Rain? Everything You Need to Know. Retrieved from https://www.skyscanner.com/tips-and-inspiration/do-planes-fly-in-the-rain

What is Python? Executive Summary. (n.d.). Retrieved from https://www.python.org/doc/essays/blurb/

What is R? (n.d.). Retrieved from https://www.r-project.org/about.html

Iho López Tobi