Airplane Crashes and Fatalities Since 1908

Rafael Leite
Klingler College of Arts and Sciences
Marquette University
Milwaukee, WI, USA
rafal.leite@marquette.edu

Abstract
Even though air travel is the safest mean of transportation in the United States, a lot of people still fear flying from one place to another. The purpose of this project is, by using methods such as exploratory analysis, K-means clustering, and topic modeling, to investigate the main causes of airplane crashes. The findings were not surprisingly unexpected. Human error, made by the pilot or crew member, along with engine failure and poor weather conditions lead the causes of plane accidents.

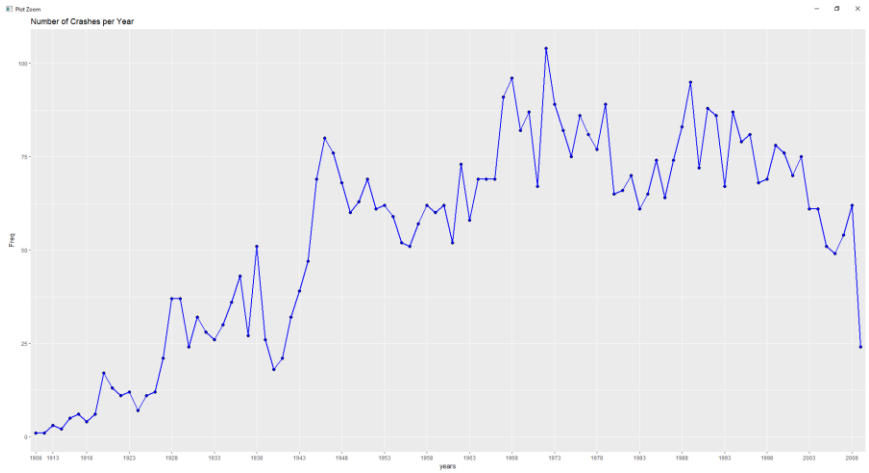
Introduction
Air travel has become more accessible to people in different parts of the world. Because of that, more people are using this type of transportation. This increase in number of passengers has been accompanied by an increase of safety measures. It is common knowledge that air travel is safer than other types of transportation, especially automobiles. In fact, commercial planes are the safer mean of transportation in the United States [1]; however, there are a lot of people that still get nervous when they need to use them. The chances of a plane crash are very slim, but when an accident occurs, it usually leaves no survivors, and this is what scares people the most. The purpose of this project is to investigate the main causes of plane accidents, as well as to answer questions like: Have airplanes become more secure? And are we more likely to survive a plane crash now than we were in the last century?

Data Set

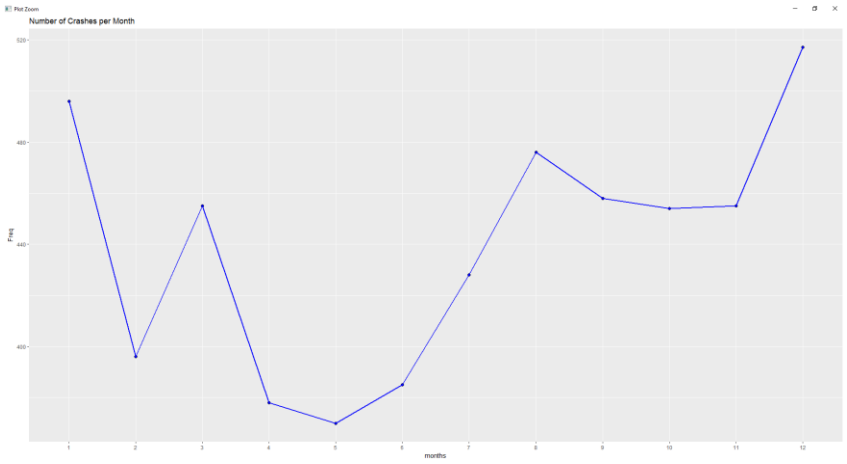
The data set used for this project is the "Airplane Crashes and Fatalities Since 1908" data set. It contains more than 5,000 records of plane accidents from 1908 to 2009. Its attributes are: Date, Time, Location, Operator, Flight #, Route, Type, Registration, cn/In, Aboard, Fatalities, Ground, and Summary.

Analysis/Evaluation

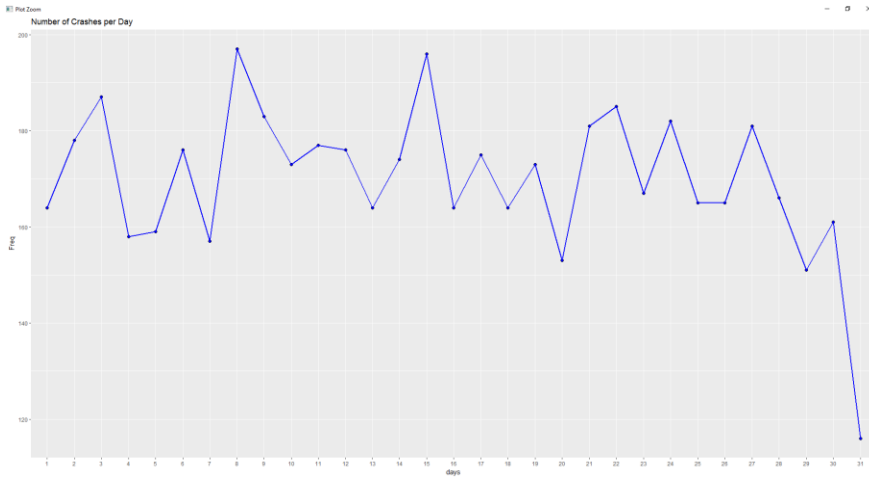
The link to the Github repository with the source code containing all the calculations and graphs can be seen below.
https://github.com/rzufil/COSC4931_DataScience
To have a better insight of the distribution of plane crashes throughout the years, exploratory analysis was conducted. RStudio was used to create the graphs below.

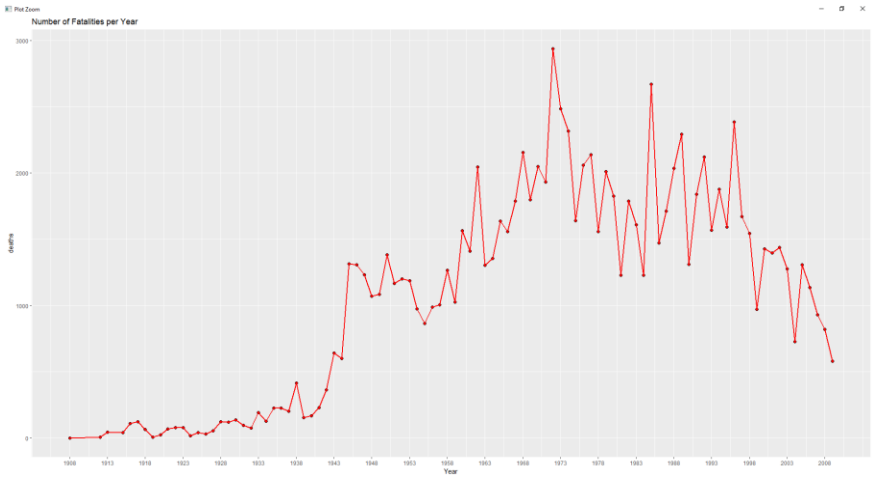


The first graph shows the frequency of plane crashes per year. A huge increase of plane crashes can be seen during the mid-40's. The frequency stabilizes during the 50's and 60's. It spikes in the late 60's and hits its all-time high in the early 70's. It then returns to the frequency before the late 60's spike and later proceeds to plummet after the turn of the millennium. This is a very big indicator that airplanes today are not experiencing as many accidents as they were in the last century, especially considering that air travel is much more accessible today.



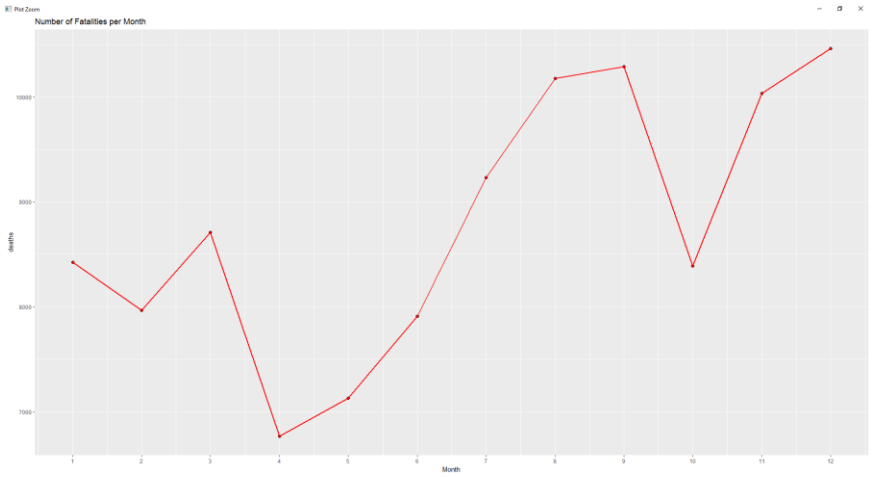
The frequency of accidents throughout the months of the year was also observed. The graph above shows a very high number of plane crashes during the last and the first month of the year. This finding was expected since there is an increase of travels during those months of the year due to holidays. May is the month with the lowest frequency of plane crashes, while December is the month with the highest frequency.

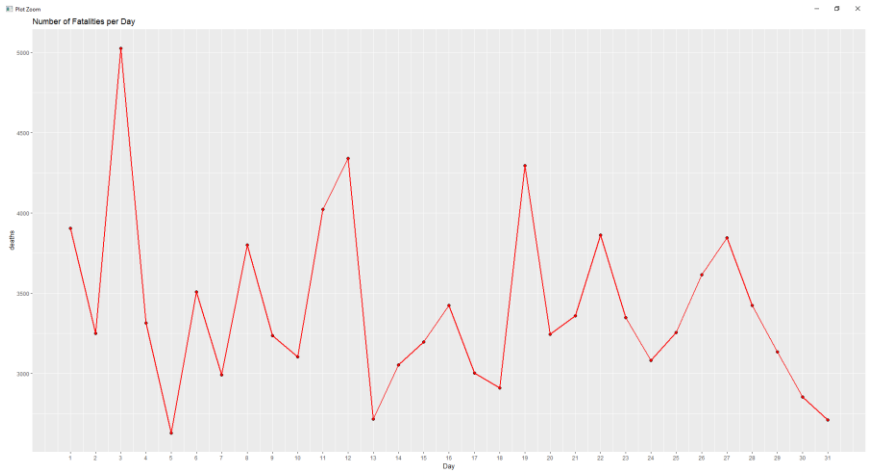


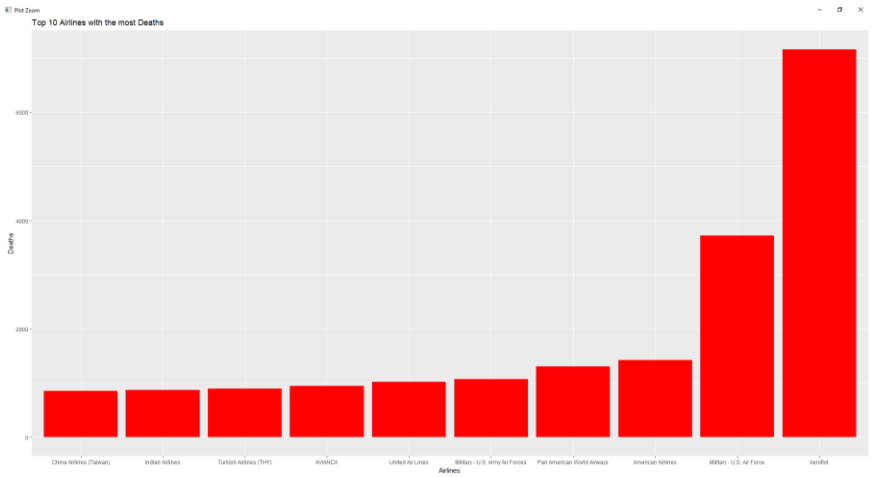


The third graph shows the frequency of plane crashes per day of the month. The graph does not show much variation throughout the month and is somewhat evenly distributed, except for the last day of the month, which exhibits a very low frequency of crashes compared to the other days of the month. However, this is also expected since not all months have 31 days.

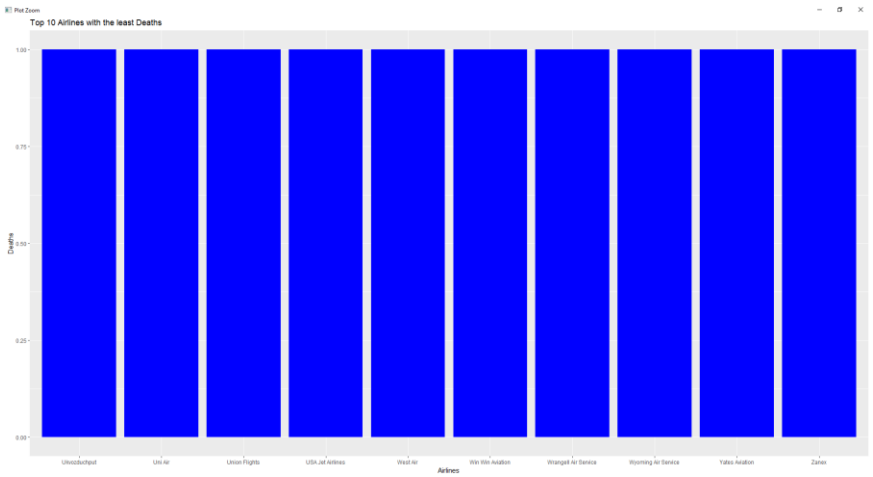
The number of deaths resulted from those plane crashes was also analyzed. The number of fatalities per year, month, and day somewhat followed the trend exhibited by the number of crashes per year, month, and day respectively. The graphs and results can be seen below.







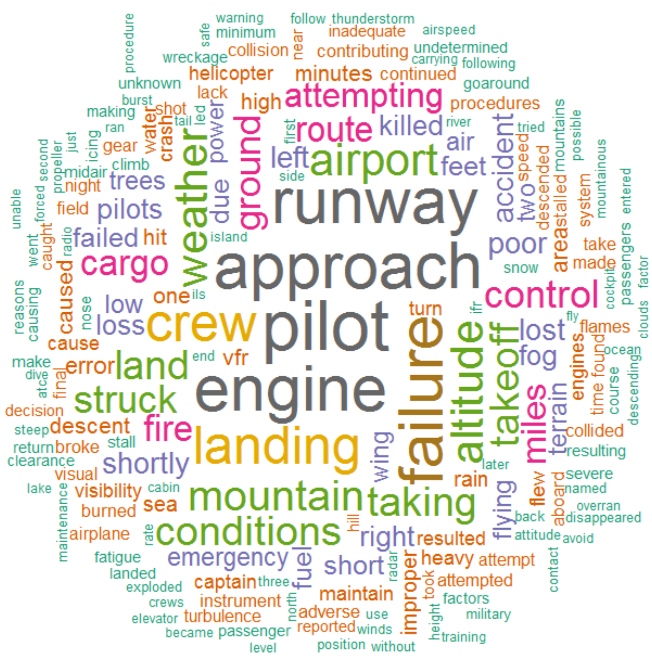
The results from the graphs hint that the chances of surviving a plane crash have not increased over the years. Otherwise the graphs would show a significantly higher decrease in the frequency of deaths in the past 20 years, something that was not observed. This indicates that although planes experience fewer crashes today, the number of deaths has not follower the trend. This makes sense since aircrafts today accommodates a lot more people than aircrafts in the last century.



Exploratory analysis was also conducted to investigate the top 10 airlines with the most accidents and the top 10 airlines with the least accidents. Graphs and results can be seen below.

The operators with the most fatalities caused by crashes during their flights are: Aeroflot, the U.S. Air Force, and American Airlines. Many airlines had only one occasion in which a person died during a flight that they were operating. However, in most cases those airlines appear only once in the data set. Therefore, we cannot conclude if they are indeed safe airlines.

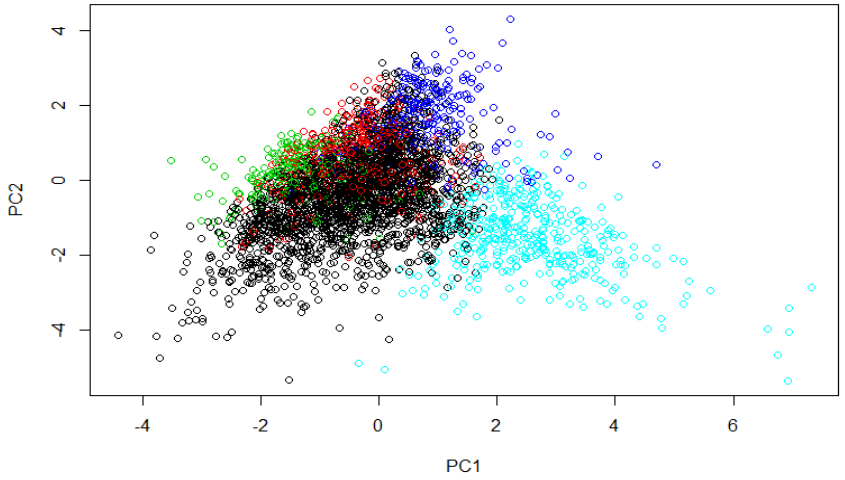
To have a better understanding of the main causes of plane crashes, text analysis was conducted. RStudio was also used for this analysis and to create the word cloud below.



	word	freq
pilot	pilot	1015
approach	approach	943
engine	engine	925
runway	runway	916
failure	failure	880
crew	crew	751
landing	landing	734
airport	airport	625
altitude	altitude	608
weather	weather	605

The list above shows the most frequent words in the “Summary” attribute of the data set. “Pilot” is the most cited word in the entire data set, with 1015 occurrences. Followed by “approach,” “engine,” “runway,” and “failure,” with 943, 925, 916, 880 occurrences respectively. The word cloud points that human error, made by the pilot or a crew member, is the most common cause of plane crashes. However, problems with the aircraft is also a common cause of accidents since engine was the third most mentioned word in the data set. It can also be inferred from the word cloud that most accidents happen when the aircraft is approaching the runway of the airport and not while mid-air.

To confirm the inferences made from the word cloud more tests needed to be done. To achieve that, the individual words were clustered into five groups by using K-means clustering. The result can be seen in the graph below.



Frequent words in cluster 1:
accident aircraft airport altitude approach attempting cargo conditions control crashed crew due engine failed failure feet fire flight flying fog ground killed land landing left loss lost low miles mountain pilot plane poor route runway short shortly struck takeoff taking weather
Frequent words in cluster 2:
aircraft cargo crashed engine flight mountain plane route runway struck takeoff
Frequent words in cluster 3:
aircraft crashed engine plane shortly takeoff taking
Frequent words in cluster 4:
aircraft airport attempting crashed land pilot plane runway
Frequent words in cluster 5:
accident aircraft altitude approach conditions crashed failure flight low mountain pilot plane poor route struck weather

From the first cluster of words we can interpret that accidents occur when there is an engine failure, which causes it to catch on fire while the airplane is landing or on the ground. The second cluster shows that in some instances the cause of the accident was the plane crashing into a mountain. From the third cluster, we can infer that accidents happen shortly after takeoff. The fourth cluster hints that crashes occur while the pilot is attempting to land on the runway. And the last cluster shows that some accidents happen due to poor weather conditions.

Topic modeling was also conducted in order to confirm the inferences made from the word cloud. Running LDA using Gibbs sampling resulted in the following top 10 terms in each of the five topics:

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	aircraft	plane	flight	crashed	engine
2	crew	crashed	pilot	approach	failure
3	altitude	runway	weather	airport	takeoff
4	struck	landing	mountain	land	taking
5	ground	cargo	conditions	miles	control
6	killed	fire	route	attempting	left
7	feet	short	poor	fog	shortly
8	route	feet	accident	plane	loss
9	approach	accident	flying	lost	failed
10	fire	poor	low	cargo	lost

Similar inferences can be done from the results of topic modeling. Therefore, both K-means clustering and topic modeling confirm the inferences made from the word cloud.

Conclusion

After all the evaluations, we can conclude that the critical moments during a flight is shortly after takeoff and during landing of the aircraft. And the main causes are human error, engine failure, and bad weather conditions. However, there is nothing to worry about while using air transportation since you are more likely to be involved in a car accident than die from a plane crash. Even though plane crashes are as lethally as they were in the past, they have become safer. The probability of a plane accident has decreased drastically in the past 20 years. Moreover, new technologies and safety policies will probably make air travel even safer in the future.

References

1. Isidore, Chris. May 13, 2015. What's the safest way to travel. CNN Money. http://money.cnn.com/2015/05/13/news/economy/train-plane-car-deaths/

2. Airplane Crashes and Fatalities Since 1908. Socrata OpenData. https://opendata.socrata.com/Government/Airplane-Crashes-and-Fatalities-Since-1908/q2te-8cvq