Project Title: Analysis of Historical Airplane Crash Data: Causes, Trends, and Implications

1. Introduction

Airplane crashes have been a major concern for the aviation industry since its inception. Despite significant advancements in aircraft technology and safety regulations, crashes continue to occur, resulting in the loss of many lives. In this project, I propose to conduct an exploratory data analysis of the comprehensive dataset on airplane crashes since 1908, to identify patterns and trends. Novel analysis was completed using machine learning algorithms, including Principal Component Analysis and K-Means to reveal better insights with some interesting visualizations. The methodology section explains the dataset, assumptions, and the data analysis. Finally, the questions that were presented in the proposal are answered in the conclusion section.

2. Methodology

2.1. Data Source

Airplane Crashes Since 1908 data set includes a full history of airplane crashes throughout the world, from 1908 to 2009. The data set has a total of 5268 observations with 13 features and some missing values described in Table 1. This data set was used to conduct the analysis described in the Analysis section.

Table 1. Data feature description and missing values

Feature Name	*Description*	*Data Type*	*Missing Values (#)*
Date	Date of the crash	object	0
Time	Time of the crash	datetime64	2219
Location	Location of the crash	object	20
Operator	Airline of the airplane	object	18
Flight #	Flight number	object	4199
Route	Crash route	object	1706
Type	Airplane type	object	27
Registration	Unique id	object	335
cn/In	Construction serial num.	object	1228
Aboard	Number of passengers	float64	22
Fatalities	Number of air fatalities	float64	12
Ground	Number of ground fatalities	float64	22
Summary	Text summary of the crash	float64	390

Unfortunately, the abovementioned crash data set doesn't include the total number of airplane departures worldwide. Furthermore, a secondary data set was used to estimate the total number of airplane departures annually. This data set was only used to investigate if there is a justifiable relation between the number of crashes and the total airplane departures worldwide.

2.2. Explanatory Data Analysis

A text analysis was conducted on the Summary and Operator columns of the dataset to determine whether the crashes were related to military or not. This analysis involved the use of a simple regular expression to identify relevant keywords and phrases that indicate military involvement. By examining the text data in this way, I may

have been able to classify each crash as either military or non-military with a high degree of accuracy. According to the available data as shown in Figure 1, there were a total of 5240 crashes recorded, out of which 781 were military crashes while 4469 were non-military crashes. This indicates that a significant portion of the crashes were non-military in nature. It is important to note that the causes and consequences of these crashes may vary greatly depending on a range of factors such as the location, type of aircraft, weather conditions, and human error. Further analysis of the data may provide valuable insights into the patterns and trends of crashes and help inform efforts to improve aviation safety. Non-military crashes will henceforth be referred to as crashes in this report and the analysis will be limited to those only.

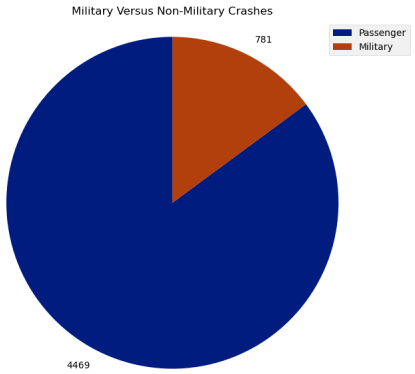


Figure 1. Total number of military plane crashes versus non-military crashes

According to Figure 2 that shows the crash breakdowns by month, the months of December and August have shown to have a higher frequency of airplane accidents compared to other months. One reason for this could be the fact that these months are peak travel seasons in many parts of the world, which leads to an increase in the number of flights and passengers. With more flights taking place, the probability of accidents happening also increases. Additionally, weather conditions during these months can also contribute to airplane accidents.

Figure 3 shows the crashes by day of the week. Thursday and Friday have a higher frequency of airplane accidents compared to other days of the week, there could be various reasons for this trend. One possible explanation is that Thursdays and Fridays are popular days for business travel, which can lead to an increase in the number of flights and passengers. Another possible explanation is that human factors, such as fatigue and stress, may play a role in the higher frequency of accidents on Thursdays and Fridays. Pilots and crew members who have been working throughout the week may be more fatigued and prone to errors, which can increase the risk of accidents.

Figure 4 summaries the crashes by the hour of the day. The hours of the day with the highest historical airplane accidents are between 7 AM and 6 PM. One possible explanation is that this time period covers the peak hours for air travel, which is when most flights depart and arrive at airports. With more flights taking place during this time, the probability of accidents happening also increases. However, no concrete explanation can be made using this data set.

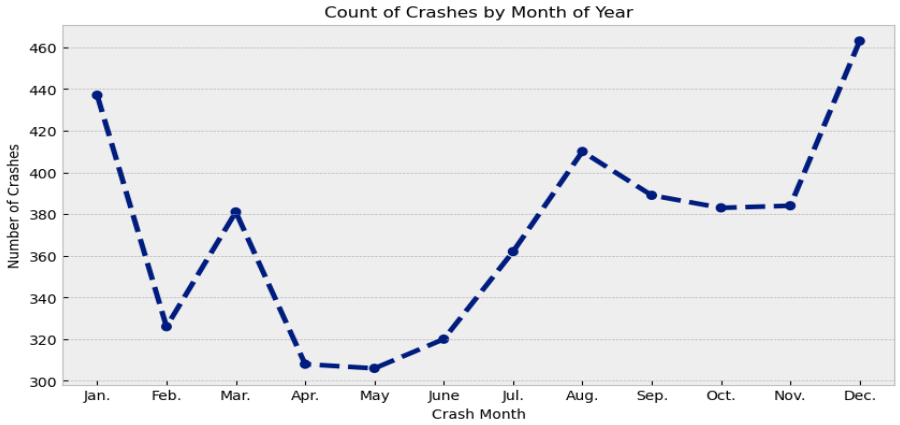


Figure 2. Total number of crashes by month across all years

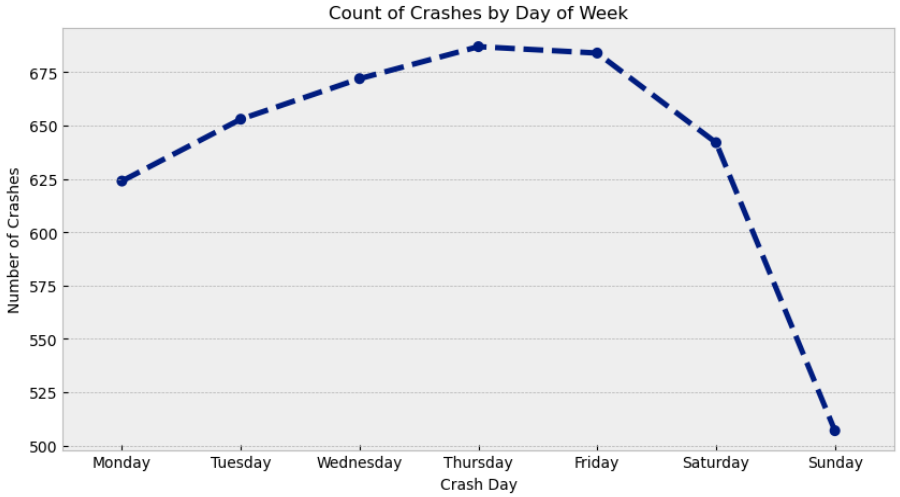


Figure 3. Total number of crashes by day across all years

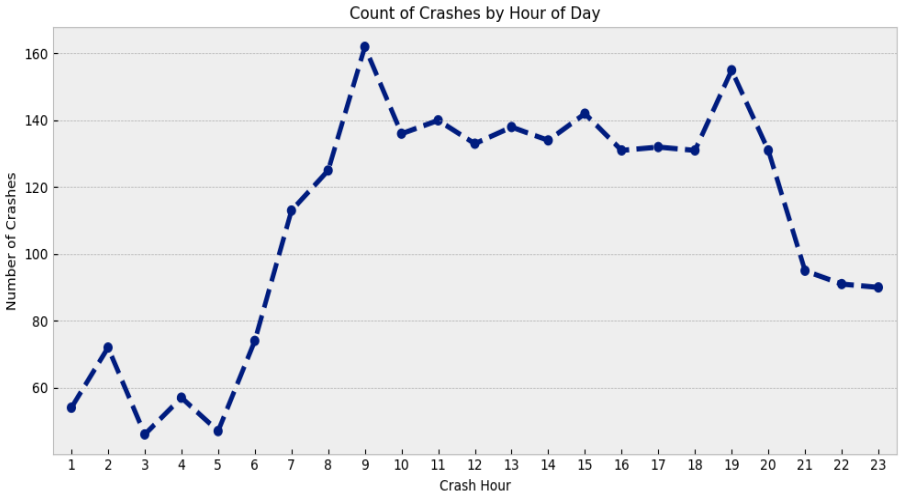


Figure 4. Total number of crashes by hour across all years

Figure 5 depicts the annual airplane crashes from 1908 to 2009. The data shows that there has been a general increase from 1908 to 1970 and a decline in the number of airplane crashes from 1970 to 2009, with occasional

spikes in certain years. This trend is likely due to advancements in technology and safety measures in the aviation industry, as well as increased regulation and oversight by governing bodies. Figure 6 shows the total number of air passengers per year versus the fatality. These two figures are unable to prove a reliable conclusion because the number of air passenger have gradually increased over the years according to a separate data set shown in Figure 7.

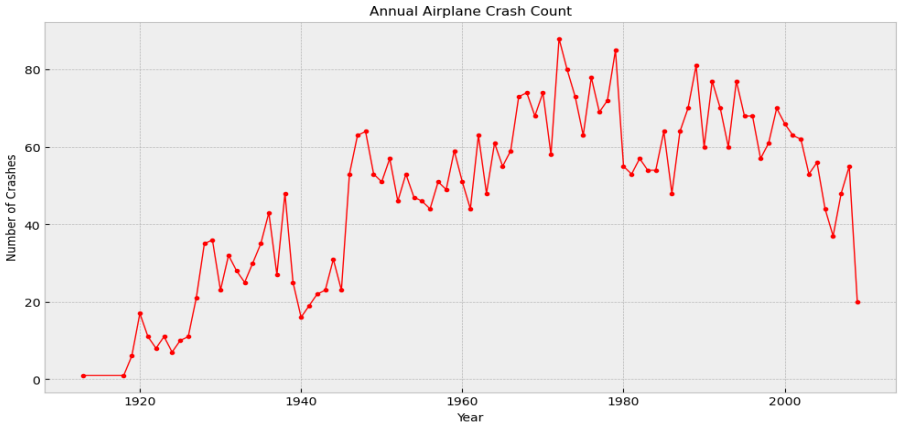


Figure 5. Annual number of airplane crashes

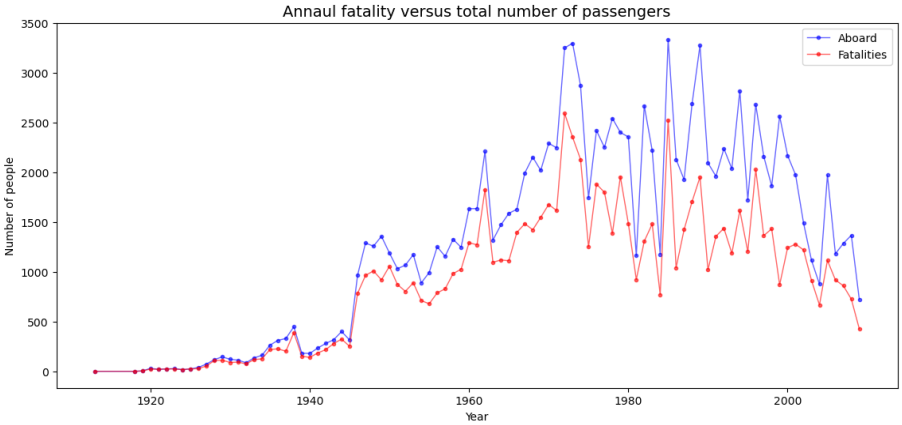


Figure 6. Annual fatality versus total number of air passengers per year

The ratio of fatalities per total number of air passengers can be a useful metric for assessing the safety of air travel. This metric is calculated by dividing the total number of fatalities in air accidents by the total number of air passengers for each year at Figure 8. In general, this ratio has decreased significantly over the years due to advancements in technology and safety measures in the aviation industry.

Figure 9 depicts the breakdown of airplane crashes by operator. According to the data, Aeroflot is responsible for the highest number of crashes among the operators. As previously mentioned, this may be due to factors such as the Soviet-era prioritization of production quotas over safety standards, inadequate pilot training, and insufficient maintenance of aircraft. However, it's worth noting that Aeroflot has taken significant steps to improve its safety record in recent years.

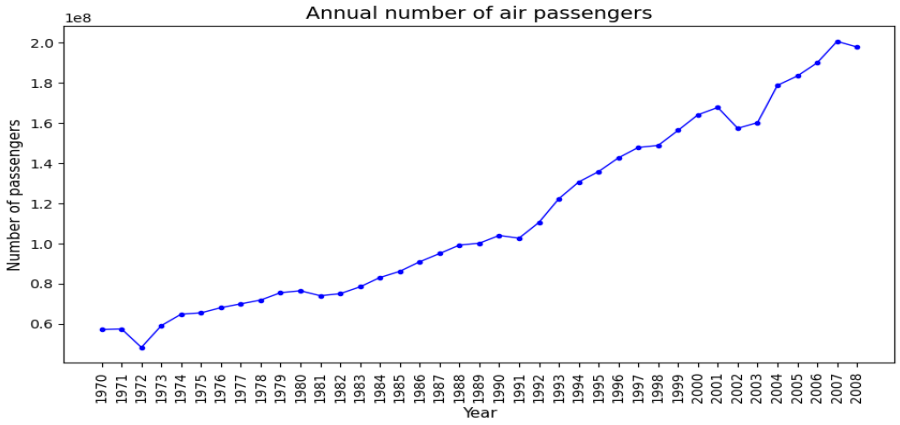


Figure 7. Total number of air passengers per year

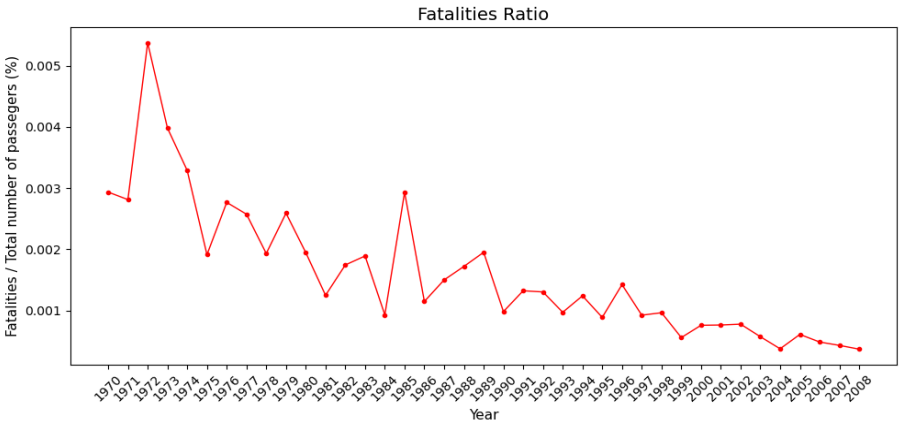


Figure 8. Fatalities/total number of air passengers ratio per year

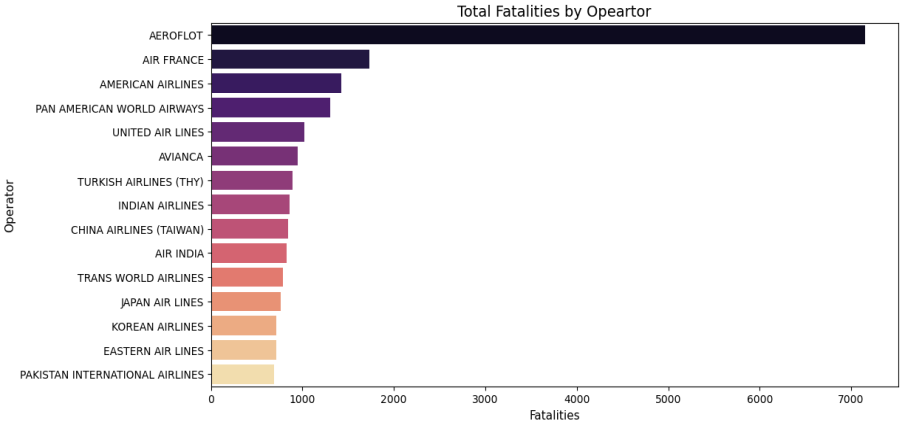


Figure 9. Crash breakdown by the operator

2.3. Clustering

Out of 4469 crashes, 4131 have a description in the Summary column. K-means clustering was used to analyze the Summary of flight descriptions and to identify whether the crashes could be categorized into major clusters. PCA was used for crash data in order to reduce the dimensionality of the dataset, while still retaining the most

important information. In other words, PCA was used to identify the key features or variables that best describe the differences between different crashes. In the context of analyzing airplane crashes, the descriptions of each crash can contain many variables such as the type of aircraft, location of the crash, weather conditions, number of fatalities, etc. Analyzing these variables individually can be difficult, especially when there are many variables involved.

To elaborate further, K-means clustering was used to group together similar crash descriptions based on the frequency of specific words or phrases. An elbow diagram was plotted to determine the optimal number of clusters, and it was found that the data could be best represented by three clusters. while PCA was implemented to identify the most important words that differentiate the different groups. The first and second principal components for the three clusters are plotted in Figure 10. As it can be seen, the clusters are separated from each other with various colors and the combination of K-Means and PCA was found to be effective to sperate the crash data into three clusters.

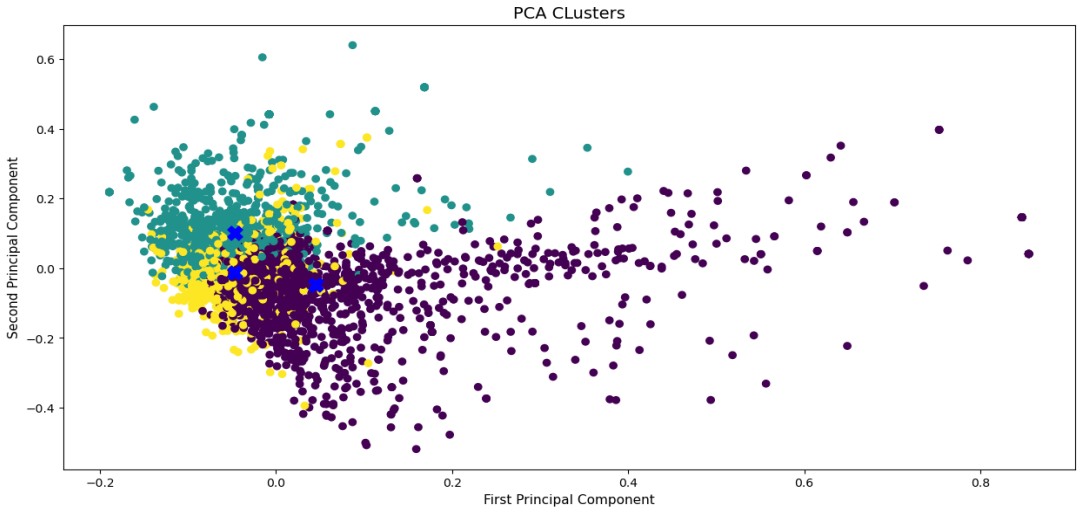


Figure 10. First and second principal components for three K-means clusters

In order to better understand each cluster and the difference between them, Table 2 is presented to show the population of the cluster, aboard/fatalities ratio, top 10 repeated words in each cluster and a word cloud for better visualization. The findings are fascinating and are described below.

Cluster 1: This cluster contains the largest number of crashes, with 2157 incidents. The aboard/fatalities ratio is 0.79, indicating that a relatively high number of people were killed in these crashes. The top ten words associated with this cluster include "crashed," "mountain," "route," "aircraft," "flight," "weather," "pilot," "conditions," "plane," which suggests that many of the crashes in this cluster were caused by environmental factors such as poor weather conditions, difficult terrain, or pilot error.

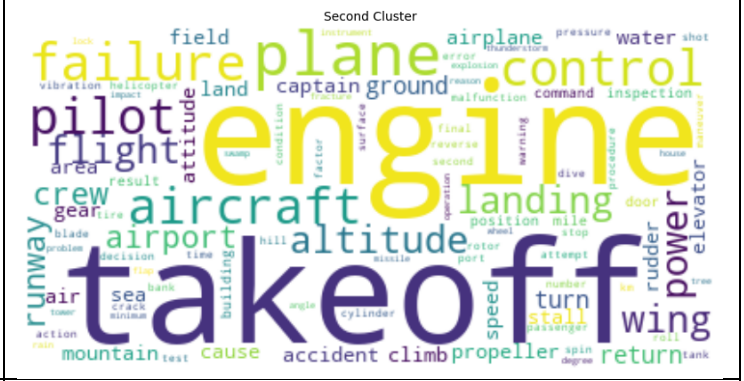
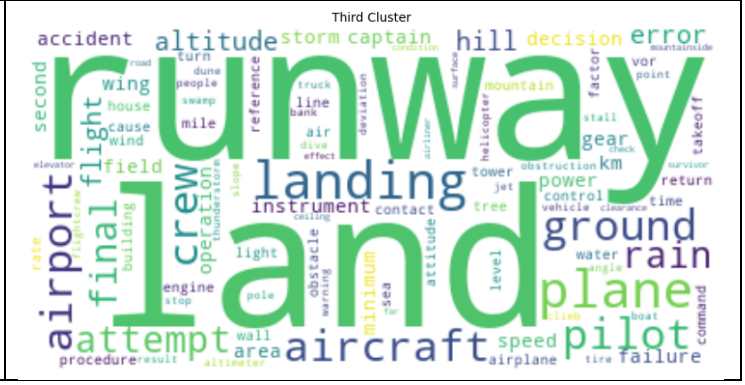
Cluster 2: This cluster contains 1040 crashes, with an aboard/fatalities ratio of 0.61. The top ten words associated with this cluster include "engine," "takeoff," "plane," "crashed," "cargo," "failure," "aircraft," "taking," "shortly," and "control." This cluster may be characterized by crashes that occurred shortly after or during takeoff , perhaps due to mechanical failure or problems with cargo or control systems.

Cluster 3: This cluster contains 934 crashes, with an aboard/fatalities ratio of 0.59. The top ten words associated with this cluster include "runway," "approach," "crashed," "land," "attempting," "landing," "short," "plane,"

"aircraft," and "fog." This cluster may be characterized by crashes that occurred during the approach or landing phase of a flight, possibly due to poor visibility or problems with the aircraft's landing gear.

In summary, by examining the aboard/fatalities ratio and the top ten words associated with each cluster, we can infer that Cluster 1 is likely characterized by crashes caused by environmental factors or pilot error, Cluster 2 by mechanical failure or cargo/control issues, and Cluster 3 by problems during the landing phase of flight or issues related to visibility. The three cluster were able to categorize the crashes into the main three stages of an airplane flight including, takeoff, fly, and landing. This is fascinating how the data was able to suggest the three main categorizes and differentiate the crashes based on the flight descriptions.

Table 2. Crash clusters' description and word cloud

Cluster No. & Info	Ten Most Repeated Words	Word Cloud
1 2157 Crashes Aboard/Fatalities Ratio = 0.79	crashed mountain route en aircraft flight weather pilot conditions plane	
2 1040 Crashes Aboard/Fatalities Ratio = 0.61	engine takeoff plane crashed cargo failure aircraft taking shortly control	
3 934 Crashes Aboard/Fatalities Ratio = 0.59	runway approach crashed land attempting landing short plane aircraft fog	

3. Conclusion

The analysis conducted on the aviation crash dataset found that there were 5240 crashes recorded, out of which 781 were military crashes while 4469 were non-military crashes.

The report also highlights several findings based on the data analysis, such as the months of December and August having a higher frequency of airplane accidents compared to other months, with one possible explanation being the peak travel seasons in many parts of the world during these months. Thursday and Friday were found to have a higher frequency of airplane accidents compared to other days of the week, with one possible explanation being the popularity of these days for business travel. The hours of the day with the highest historical airplane accidents were found to be between 7 AM and 6 PM, which covers the peak hours for air travel.

The report also suggests that the ratio of fatalities per total number of air passengers can be a useful metric for assessing the safety of air travel, which has decreased significantly over the years due to advancements in technology and safety measures in the aviation industry. The breakdown of airplane crashes by operator found that Aeroflot was responsible for the highest number of crashes among the operators, which may be due to factors such as the Soviet-era prioritization of production quotas over safety standards, inadequate pilot training, and insufficient maintenance of aircraft.

K-means clustering and PCA were used to analyze the flight descriptions and to identify whether the crashes could be categorized into major clusters. It was found that the data could be best represented by three clusters, and Table 2 was presented to show the population of the cluster, aboard/fatalities ratio, top 10 repeated words in each cluster and a word cloud for better visualization. It was remarkable how the data was able to classify airplane crashes into the three main stages of a flight: takeoff, flight, and landing, using the three clusters. The ability to differentiate the crashes based on their flight descriptions was fascinating and found to be the novelty of my analysis.