Mortality and survivor rates on plane crash keywords based off k-means clustering algorithm and text analysis

Nicholas Paisley

DSSA 5302-491

9 August 2021

Abstract:

On May 21, 1958, a bill was introduced to create an independent Federal Aviation Agency to provide the safe and efficient use of national airspace. On August 23, 1958, a new independent Federal Aviation Agency responsible for civil aviation safety was enacted.

Ever since the enactment of the FAA, there have been many advances in aviation safety, however crashes still do occur. After exploring the data, 10 unique terms involving the reasons for crashes were selected. With these unique terms, the question was asked, was the mortality and survivor rate of passengers in these crashes lower or higher since the enactment of the FAA?

Ten clusters were identified by using k-means clustering technique on a matrix obtained by a text corpus created using text analysis. These clusters each had multiple words present within their frequent terms of what caused the crash, however, one unique term from each cluster was then picked and analyzed to see if the mortality and survivor rate of passengers aboard these types of crashes was higher or lower before or after FAA enactment. These unique terms were filtered by date, fatalities and survivors and then normalized by number of passengers aboard to create a percentage that showed the mortality and survivor rate of a passenger that was involved in said crash.

The results showed that out of the 10 unique terms selected from each cluster, for the reason of the crash, 6 of the 10 showed a lower mortality and higher survivor rate, while 4 of the 10 showed a higher mortality and lower survivor rate since the enactment of the FAA.

Introduction:

Aviation leaders believed the airplane could not reach its full commercial potential without federal action to improve and maintain safety standards. At their urging, the Air Commerce Act was passed in 1926. This landmark legislation charged the Secretary of Commerce with fostering air commerce, issuing and enforcing air traffic rules, licensing pilots, certifying aircraft, establishing airways, and operating and maintaining aids to air navigation.

In 1934, the Department of Commerce renamed the Aeronautics Branch the Bureau of Air Commerce to reflect the growing importance of aviation to the nation. In one of its first acts, the Bureau encouraged a group of airlines to establish the first air traffic control centers to provide en route air traffic control. The Bureau had no direct radio link with aircraft, but used telephones to stay in touch with airline dispatchers, airway radio operators, and airport traffic controllers. Although en route Air Traffic Control became a federal responsibility, local government authorities continued to operate airport towers. A 1931 crash that killed all on board, including popular University of Notre Dame football coach Knute Rockne, elicited public calls for greater federal oversight of aviation safety. Four years later, a plane crash killed U.S. Senator Bronson Cutting of New Mexico.

To ensure a federal focus on aviation safety, the Civil Aeronautics Act in 1938 was established. The legislation enacted the independent Civil Aeronautics Authority (CAA), with a three-member Air Safety Board that would conduct accident investigations and recommend ways of preventing accidents.

On June 30, 1956, a Trans World Airlines Super Constellation and a United Air Lines DC-7 collided over the Grand Canyon, Arizona, killing all 128 occupants of the two airplanes. The collision occurred while the aircrafts were flying under visual flight rules in uncongested airspace. The accident dramatized the fact that, even though U.S. air traffic had more than doubled since the end of World War II, little had been done to mitigate the risk of midair collisions.

On May 21, 1958, Senator A. S. "Mike" Monroney introduced a bill to create an independent Federal Aviation Agency to provide for the safe and efficient use of national airspace. Two months later, on August 23, 1958, the President signed the Federal Aviation Act, which transferred the Civil Aeronautics Authority's functions to a new independent Federal Aviation Agency responsible for civil aviation safety.[1]

Ever since the enactment of the FAA, there have been many advances in aviation safety, however crashes still do occur. With all of these advancements throughout the years, like the National Airspace System (NAS) and the Aviation Safety Research Act of 1988, did the FAA help with the mortality rate of passengers aboard planes that were in a crash scenario? After exploring the data, 10 unique keywords involving crashes were selected. With these unique terms, was the mortality and survivor rate of passengers in these types of crashes lower or higher since the enactment of the FAA?

[1] https://www.faa.gov/about/history/brief_history/

Methodology

This data was taken from Kaggle.com. The user that posted this data had the following to say, "At the time this Dataset was created in Kaggle (2016-09-09), the original version was hosted by Open Data by Socrata at the at: https://opendata.socrata.com/Government/Airplane-Crashes-and-Fatalities-Since-1908/q2te-8cvq, but unfortunately, that is not available anymore. The dataset contains data of airplane accidents involving civil, commercial and military transport worldwide from 1908-09-17 to 2009-06-08"[2].

The data set started out with 13 variables and had 5268 records within it.

Those variables are the following:

- Date: Date of accident
- Time: Local time, in 24 hr. in the format hh:mm
- Location: Location of the accident
- Operator: Airline or operator of the aircraft
- Flight: Flight number assigned by the aircraft operator
- Route: Complete or partial route flown prior to the accident
- Type: Aircraft type
- Registration: ICAO registration of the aircraft
- cn/In: Construction or serial number / Line or fuselage number
- Aboard: Total people aboard

[2] https://www.kaggle.com/saurograndi/airplane-crashes-since-1908

- Fatalities: Total fatalities aboard
- Ground: Total killed on the ground
- Summary: Brief description of the accident and cause if known

To check the data veracity and the data integrity, random crash data was chosen and then independently researched to see if the entries were accurate. After checking the data, it was confirmed to be in good condition and be accurate on the crashes.

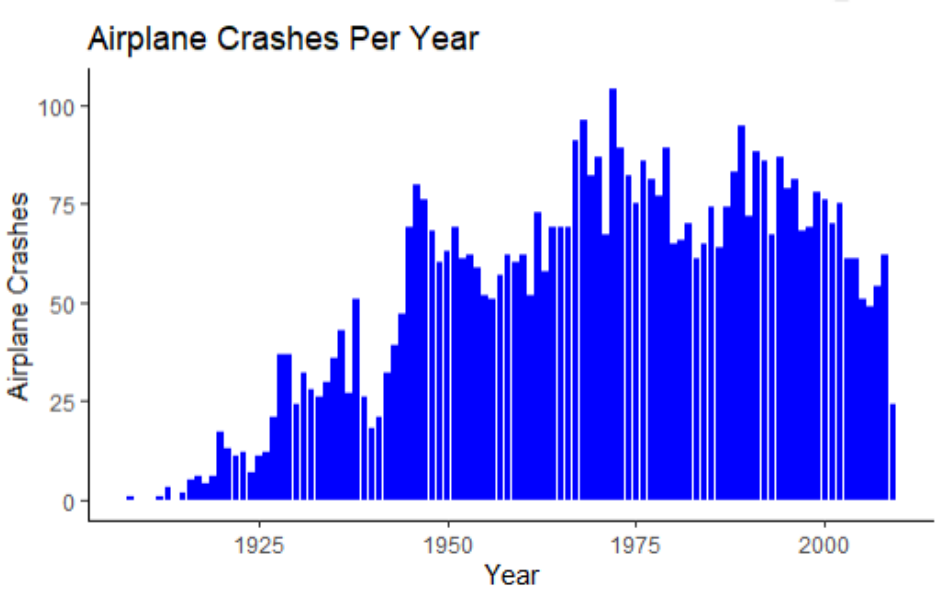
There was no initial question when data exploration began. Data exploration was completed to attempt to find some sort of pattern and/or a curiosity about the data to propose a question.

The data exploration began in R studio using the following libraries: ggplot2, tidyverse, lubricate, tm caret, crayon and reshape2. There were some columns that were missing a plethora of data and that were not going to be utilized within the exploration and analysis. Therefore, the following variables were removed; Flight, Registration and cn/ln. There were also issues with the "Time" variable where there were some of the times that were either, missing and or improperly inputted. Those times were cleaned and converted back to the standard of the hh:mm format. However, even though the "Time" variable was cleaned, this data was not utilized, but kept in case of further study.

With the removal of some variables, some new variables were created to help visualize the data a little more clearly. Because the dataset had both an "Aboard" and "Fatalities" variable, a "Survivors" variable was created to visualize how many people survived the plane crashes (Abroad – Fatalities). Also, a "Year" variable was created to help show crashes per year and not

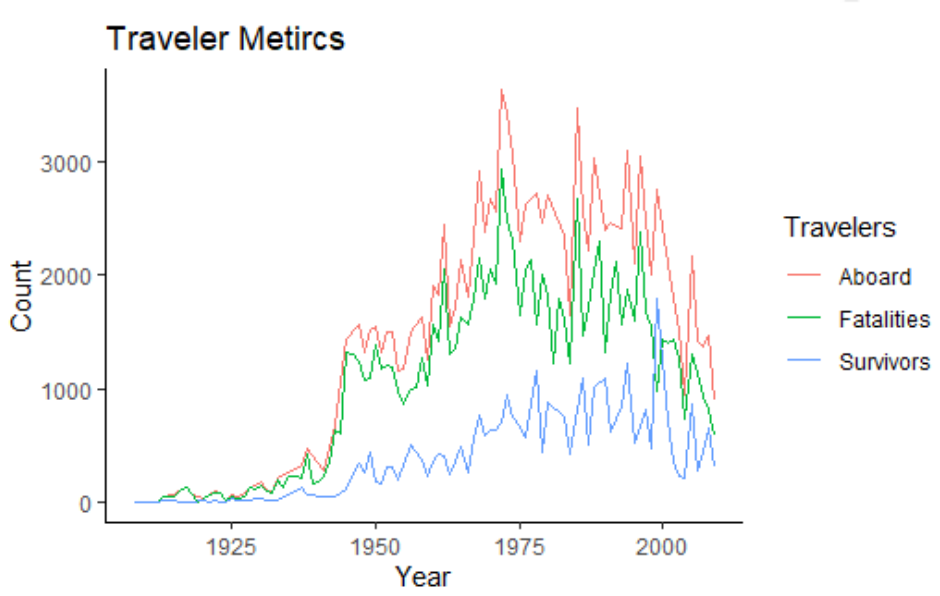
on a date-by-date basis. This was done by changing the date format to “"%m/%d/%Y” and then saving off the year into a separate data frame.

The first question that was explored was if there was any noticeable data variation in the crashes per year.



Looking at the graph, unfortunately, that was not the case. There was really no data variation that was able to be seen that could not be explained logically. So, there was no question yet to be pursued at this moment.

The next question pursued was if there was any correlation between the travelers “Aboard” “Fatalities,” and “Survivors.”



Again, there was no big correlation between the travelers so, again, no questions were solidified.

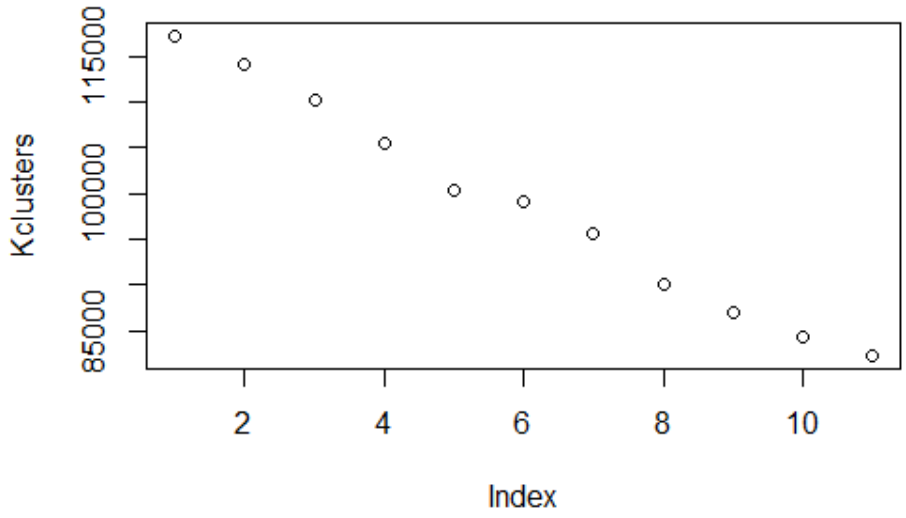
The next exploration was in the "Summary" of the crashes. Were there any pinpointable words that caused the planes to crash? This question was explored by creating 10 clusters that were identified by using k-means clustering technique on a matrix obtained by a text corpus created by use of text analysis.

Text analysis is the process of transforming unstructured text into a structured format to identify meaningful patterns and new insights[3]. The text analysis code was taken from Kaggle, however, was edited and revised for my exploration of data[4]. Text analysis was used to remove punctuation, to change all words in the "Summary" variable to lower case, to change the "Summary" varaiable to a plain text document (removing all bold, underlines and italicized

[3] https://www.ibm.com/cloud/learn/text-mining
[4] https://www.kaggle.com/saurograndi/text-analysis-cluster-analysis

words and converting the to normal text), and removing all terms whose sparsity was greater than 92.5%.

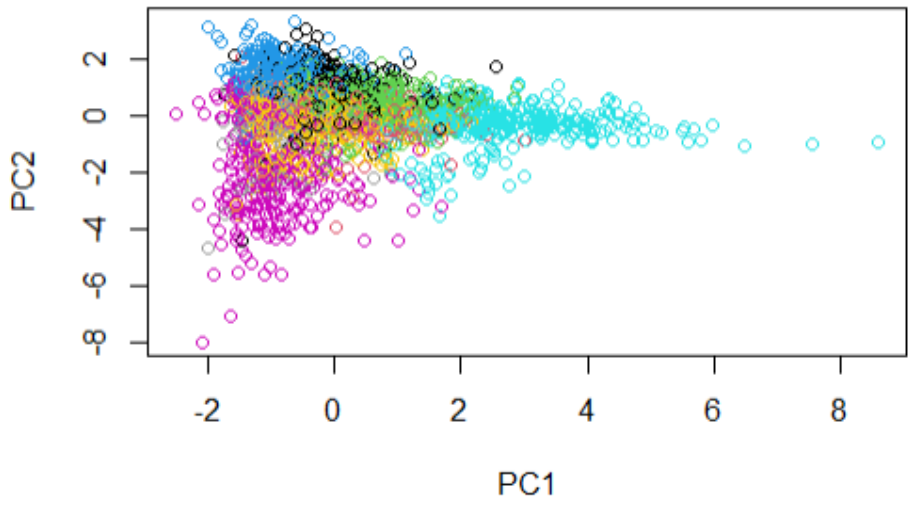
Frequent terms that appeared at least 100 times were then found. Rows where there were frequent terms present, got summed up into 1 number. This showed that the terms were present within then columns. All rows that had a sum 0 frequent terms in a column were then removed. The data was then normalized and put into a matrix. The matrix was then prepared to produce a k-means.



*****NOTE: THE GRAPH CENTERS START AT 2. THEREFORE INDEX 1 IS CENTER = 2*****

The kmeans-cluster graph, if read does say to pick the center of 6, however, because of the data margin with the frequency of words, the ending clusters would have a lot more terms in each. Therefore, 10 centers were chosen to allow more data variety within the frequency terms.

The k-means cluster was then plotted on a principal component graph to show the cluster based on their similarities (in this case the frequencies of words).



The most frequent words were then printed out from each cluster.

```
50 most frequnt terms in cluster 1:
aircraft  control  crashed  failure  flight  pilot  plane
50 most frequnt terms in cluster 2:
aircraft  crashed  plane  taking
50 most frequnt terms in cluster 3:
aircraft  altitude  approach  conditions  crashed  crew  engine  failure  flight
ground  landing  miles  mountain  pilot  plane  route  runway  struck  takeoff  taking
weather
50 most frequnt terms in cluster 4:
aircraft  crashed  crew  engine  failure  flight  landing  pilot  plane  takeoff
taking
50 most frequnt terms in cluster 5:
aircraft  approach  conditions  crashed  flight  mountain  pilot  plane  route
weather
50 most frequnt terms in cluster 6:
aircraft  approach  attempting  cargo  crashed  land  landing  pilot  plane  runway
struck
50 most frequnt terms in cluster 7:
aircraft  airport  altitude  approach  attempting  conditions  crashed  crew  engine
failure  flight  ground  land  landing  miles  pilot  plane  runway  struck  taking
weather
50 most frequnt terms in cluster 8:
aircraft  approach  crashed  crew  failure  landing  pilot  plane  runway  struck
takeoff
50 most frequnt terms in cluster 9:
aircraft  crashed  miles  route  takeoff
50 most frequnt terms in cluster 10:
crashed  mountain  struck
```

With this clustering of frequent terms, the question then arose, from the most unique term from each cluster, did the enactment of the FAA show decreases in fatalities and increases in survivors rates from the people who were aboard during the crash?

Certain keywords were deemed too similar to use due to the frequency that they appeared in each cluster (i.e. aircraft, crashed and plane). From here the most unique keyword was chosen to be analyzed from each cluster. The keywords that were chosen from each cluster were the following:

Cluster 1: Control

Cluster 2: Taking

Cluster 3: Altitude

Cluster 4: Engine

Cluster 5: Conditions

Cluster 6: Cargo

Cluster 7: Weather

Cluster 8: Crew

Cluster 9: Route

Cluster 10: Mountain

To answer this question the following was completed, for each of the clusters. The word from the cluster was grepped via a string detection command. A rbind was done on the “Survivors”, “Aboard” and “Fatalities” to create a new data frame and then creating a new row for all 3 of the listed variables named “Travelers”. Any empty space within this data frame was removed and then graphed with a red and green background rectangle showing before and after FAA

enactment. This was completed to see if there were any patterns appearing with the travelers and the mortality rates between them.

Another graph was created by going through the data frame and grepping the selected word from the cluster. The data frame with the selected word was then split into 2 separate data frames. The first data frame included everything before the date the FAA was enacted (8-23-1958), and the second data frame includes all the data from the enaction of the FAA to the present. The 2 data frames were then purged of all the NAs and empty rows. The data frames were then split once more into before and after FAA enacted and then normalized for both “Fatalities” and “Survivors” by using the “Aboard” variable.

These graphs officially show if the FAA being enacted had any effect on passenger mortality rates for the keyword that was chosen from each cluster for the reason for the crash.

This procedure was then completed 9 more times for each of the keywords from each cluster.

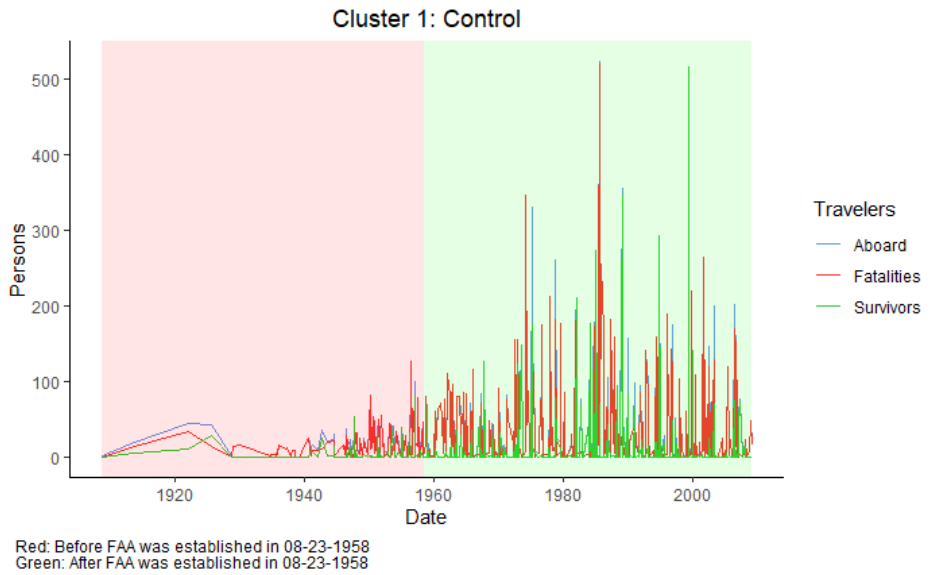
Data:

Each cluster has one unique term that was used in an analysis to see if the morality and survivor rate of the passengers aboard was higher or lower before the FAA was enacted.

The first graph that is going to be presented is a visual representation of the “Survivors”, “Fatalities” and “Abroad.” This is to show if there are any outlier events that may contribute to the mortality rate of the passengers. It will also show the increase in passengers per flight and show an accurate representation of “Survivors”, “Fatalities” and “Abroad”.

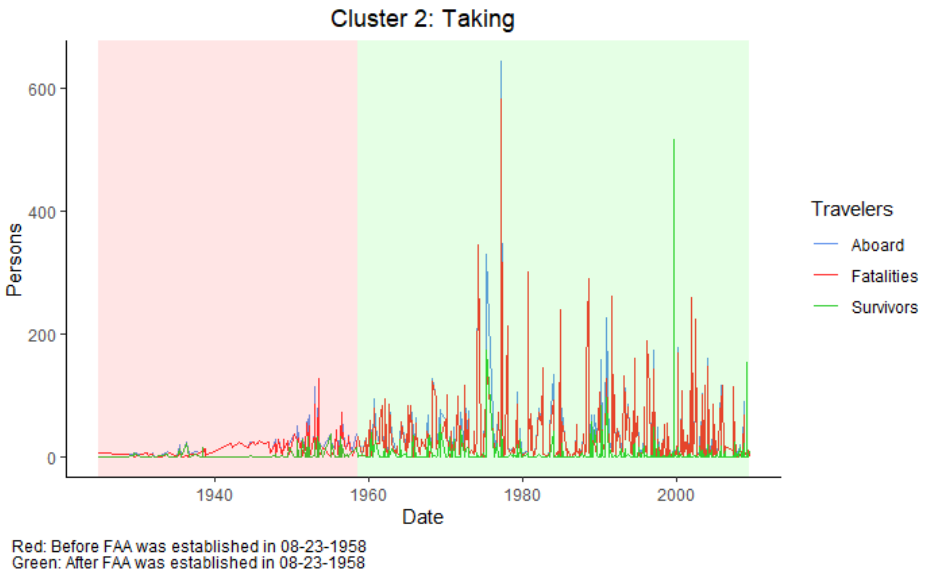
The second graph that is going to be presented is going to show the mortality and survivor rate for passengers before and after the FAA was enacted.

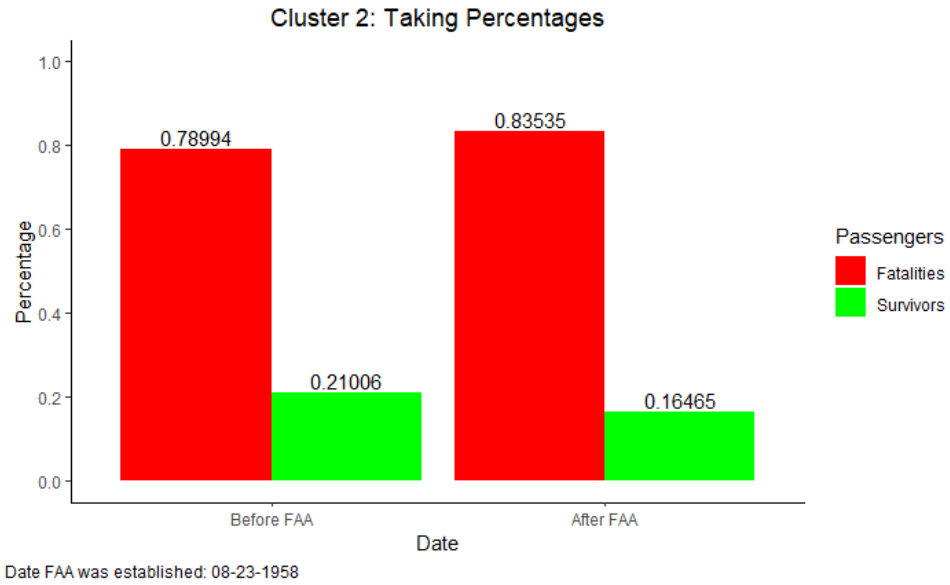
Cluster 1: Control



It was determined that the mortality rate of passengers decreased, and survivor rate increased when the FAA was enacted when there was a “control” issue involving a plane crash.

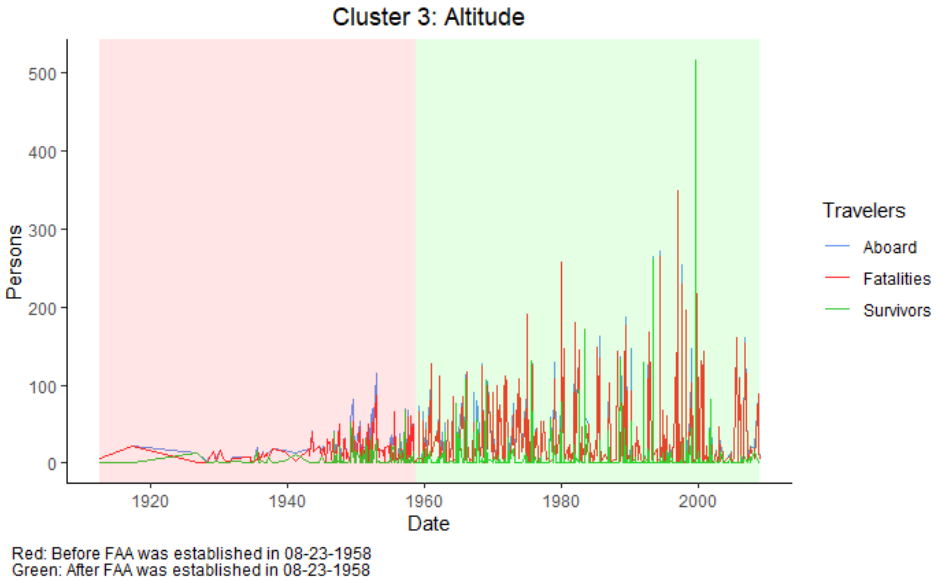
Cluster 2: Taking

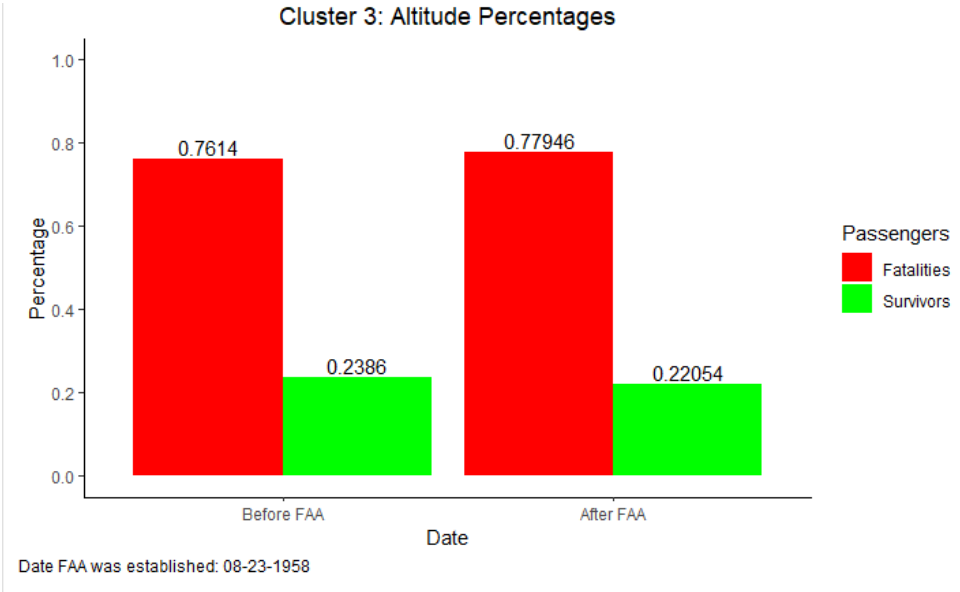




It was determined that the mortality rate of passengers increased, and survivor rate decreased when the FAA was enacted when there was a “taking” issue involving a plane crash.

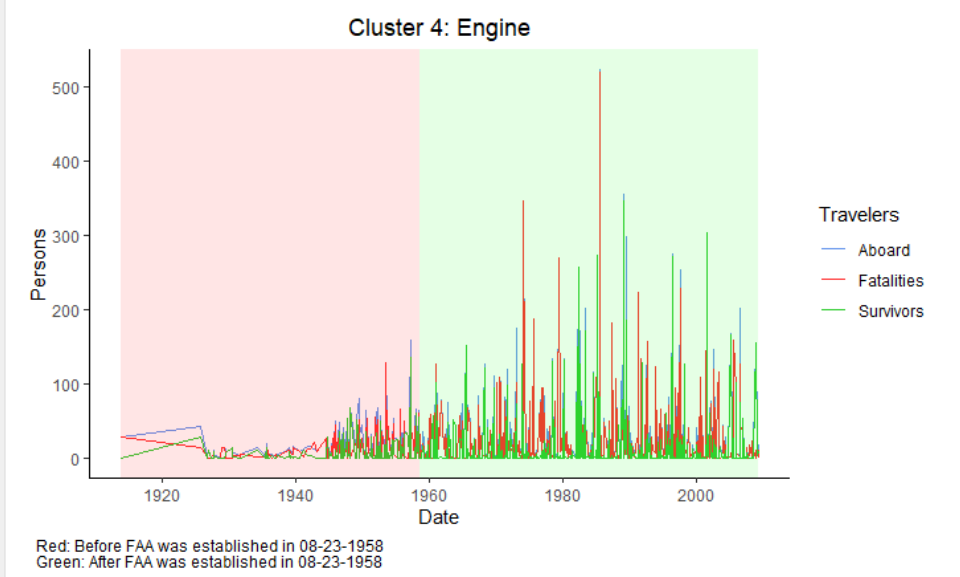
Cluster 3: Altitude

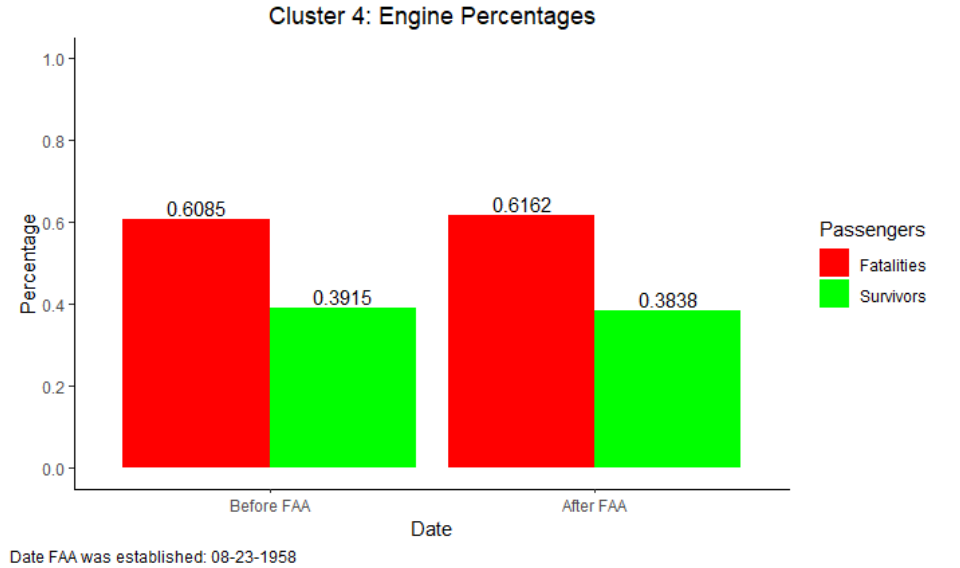




It was determined that the mortality rate of passengers increased, and survivor rate decreased when the FAA was enacted when there was a “altitude” issue involving a plane crash.

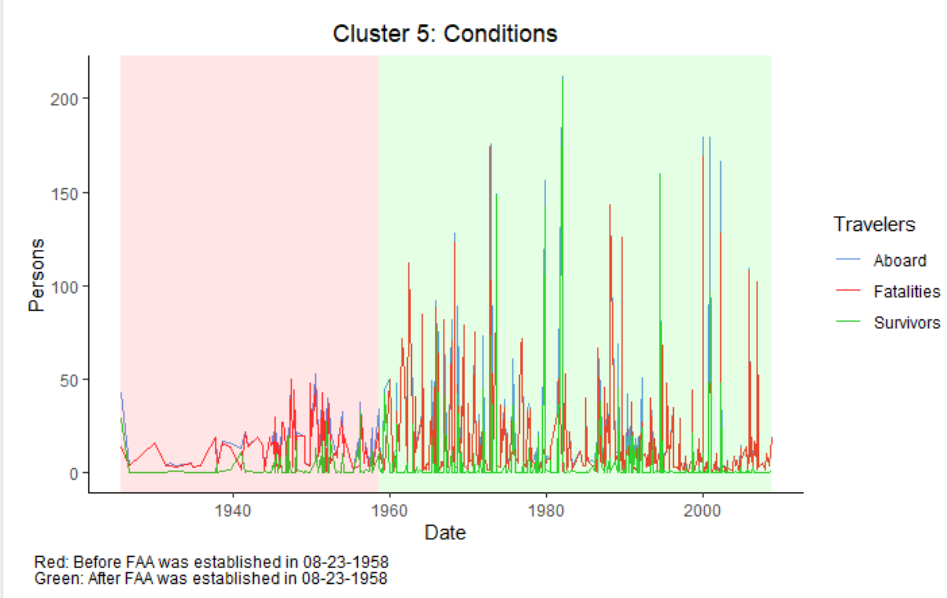
Cluster 4: Engine

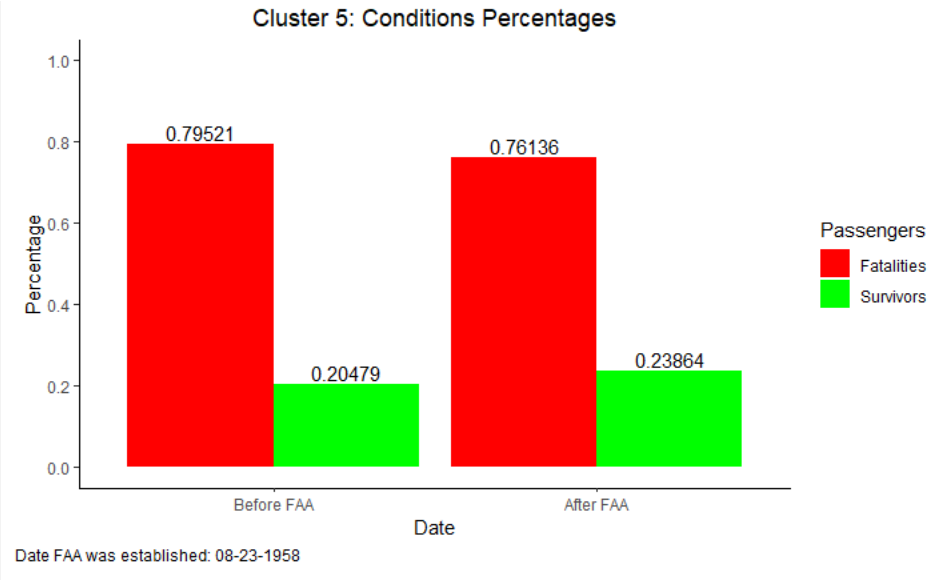




It was determined that the mortality rate of passengers increased, and survivor rate decreased when the FAA was enacted when there was a “engine” issue involving a plane crash.

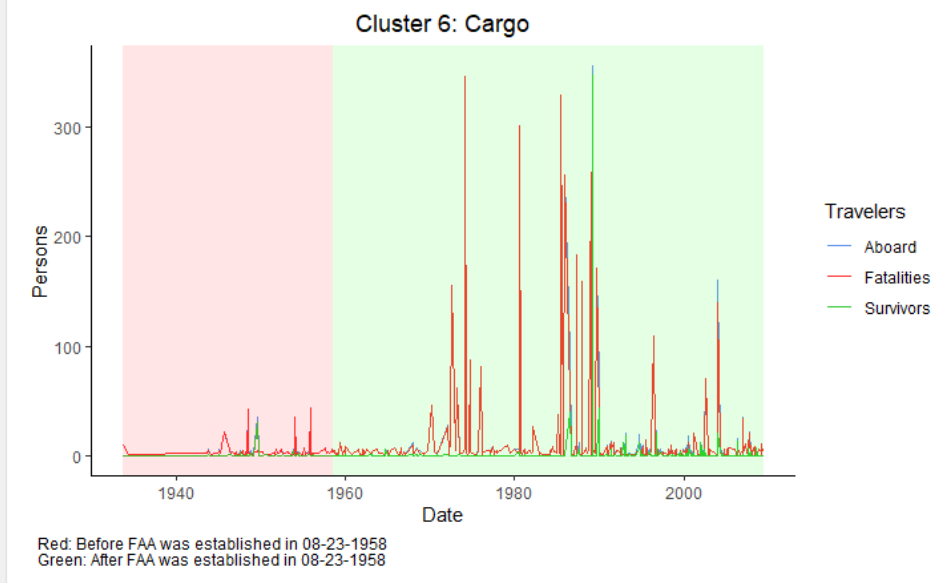
Cluster 5: Conditions

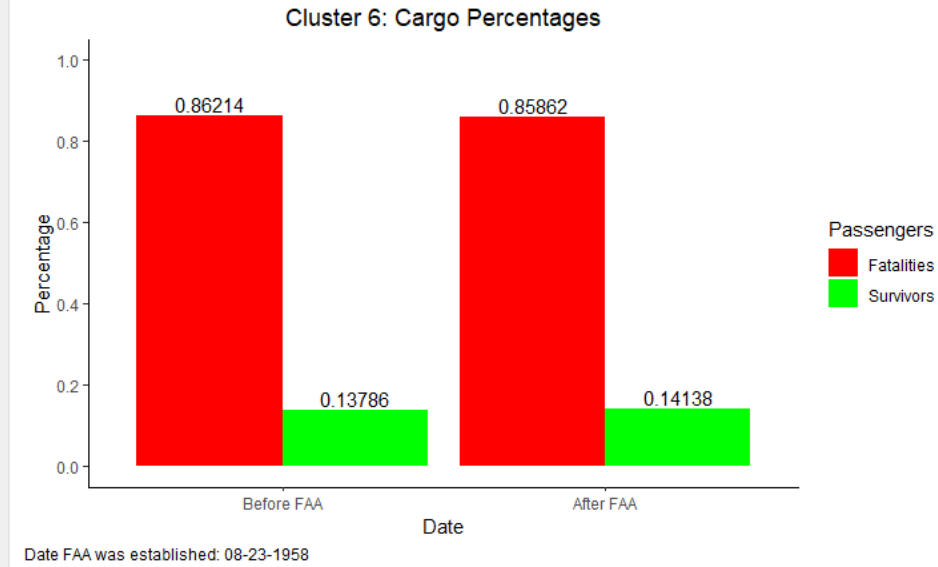




It was determined that the mortality rate of passengers decreased, and survivor rate increased when the FAA was enacted when there was a “control” issue involving a plane crash.

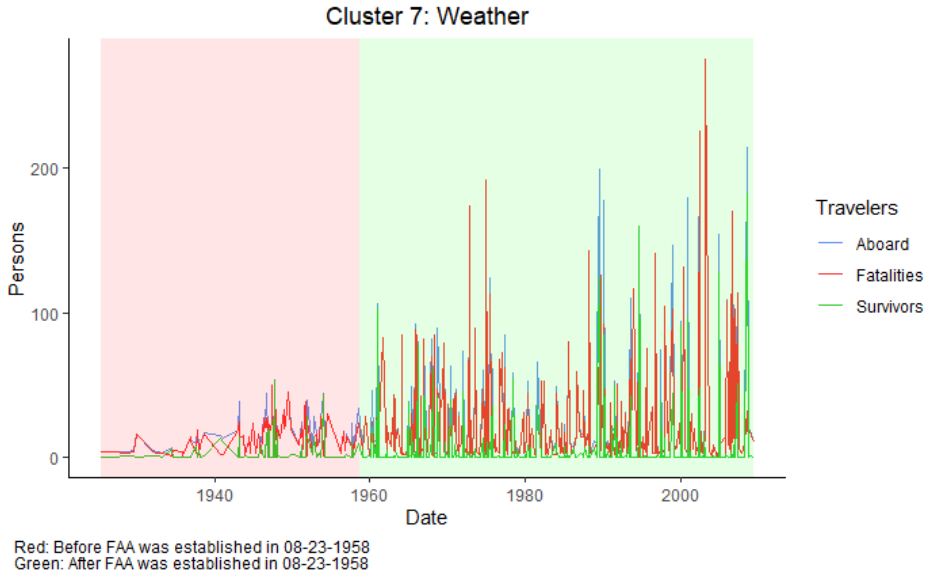
Cluster 6: Cargo

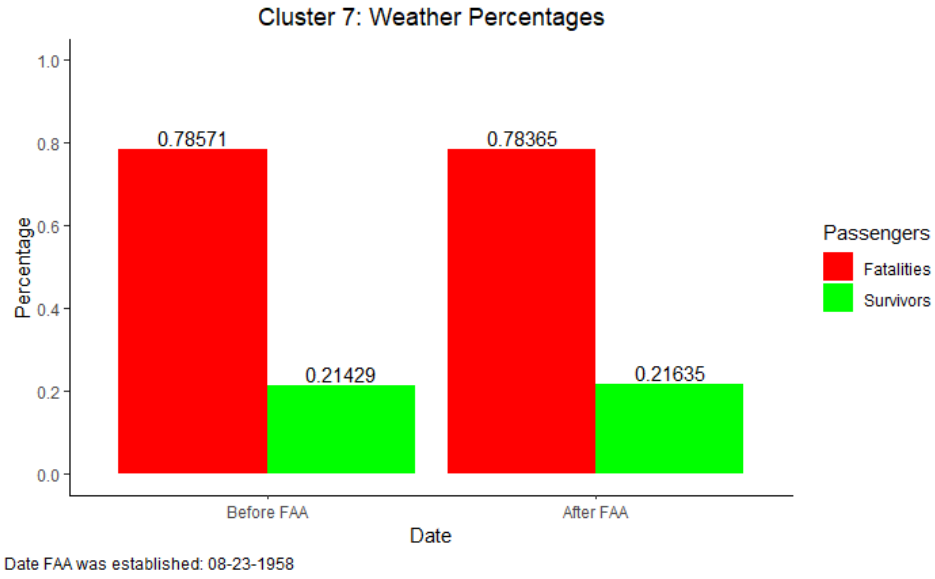




It was determined that the mortality rate of passengers decreased, and survivor rate increased when the FAA was enacted when there was a “cargo” issue involving a plane crash.

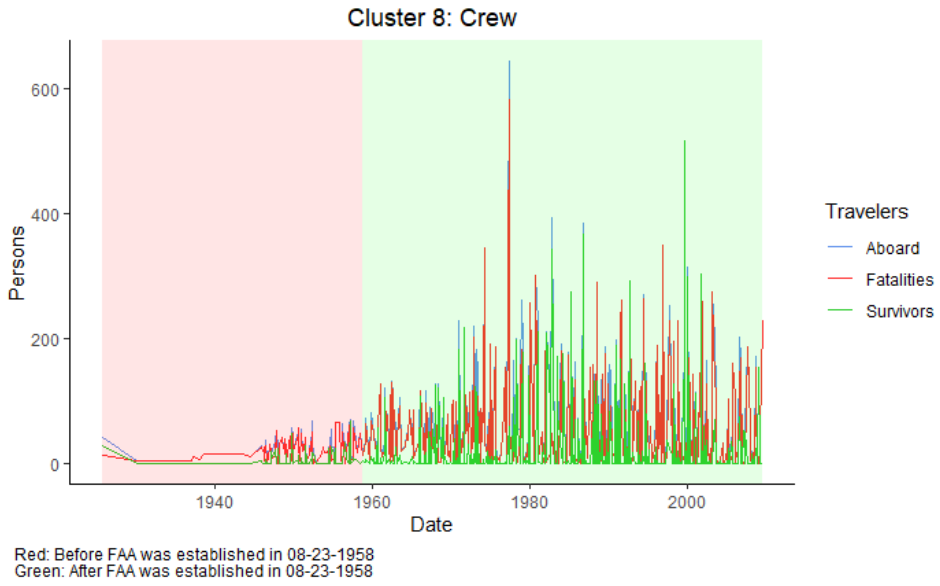
Cluster 7: Weather

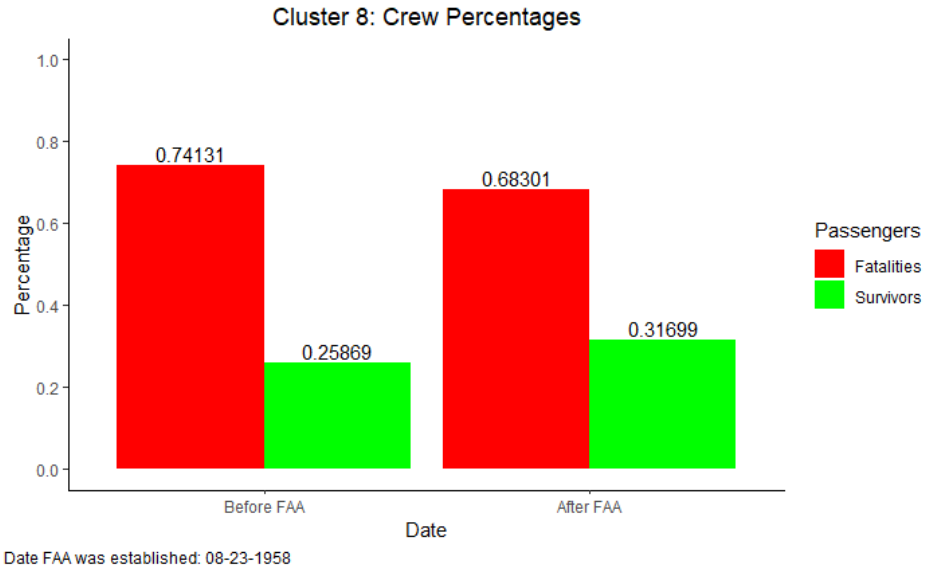




It was determined that the mortality rate of passengers decreased, and survivor rate increased when the FAA was enacted when there was a “weather” issue involving a plane crash.

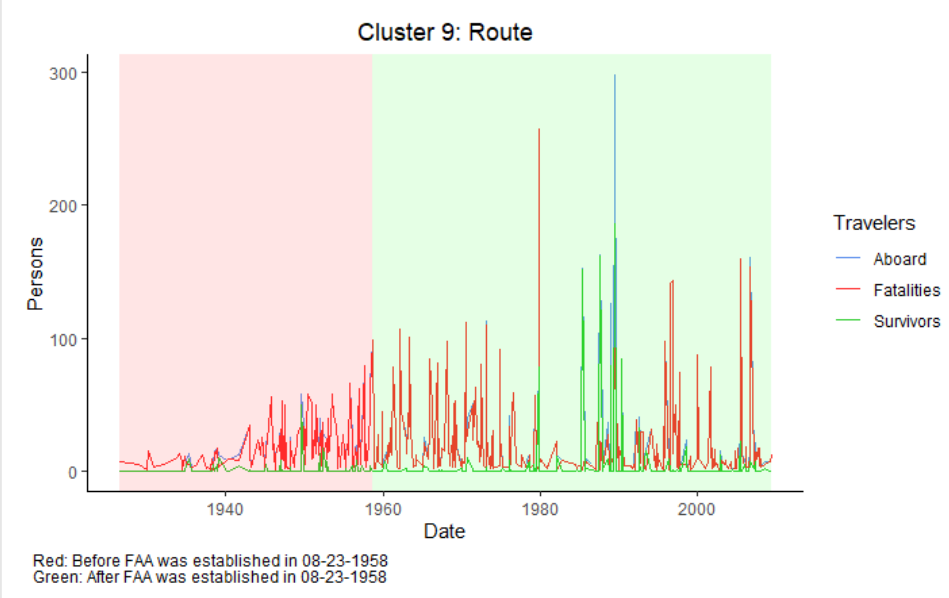
Cluster 8: Crew

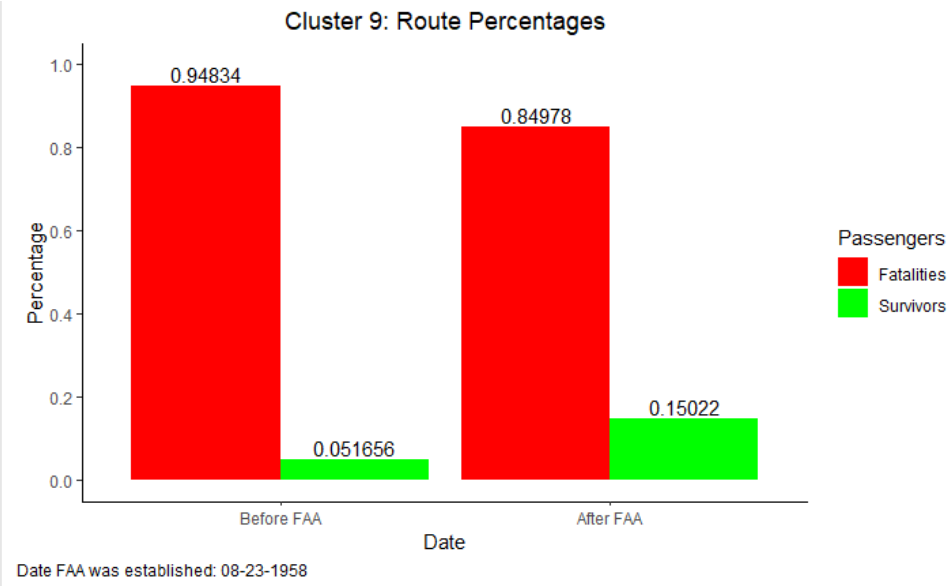




It was determined that the mortality rate of passengers decreased, and survivor rate increased when the FAA was enacted when there was a “crew” issue involving a plane crash.

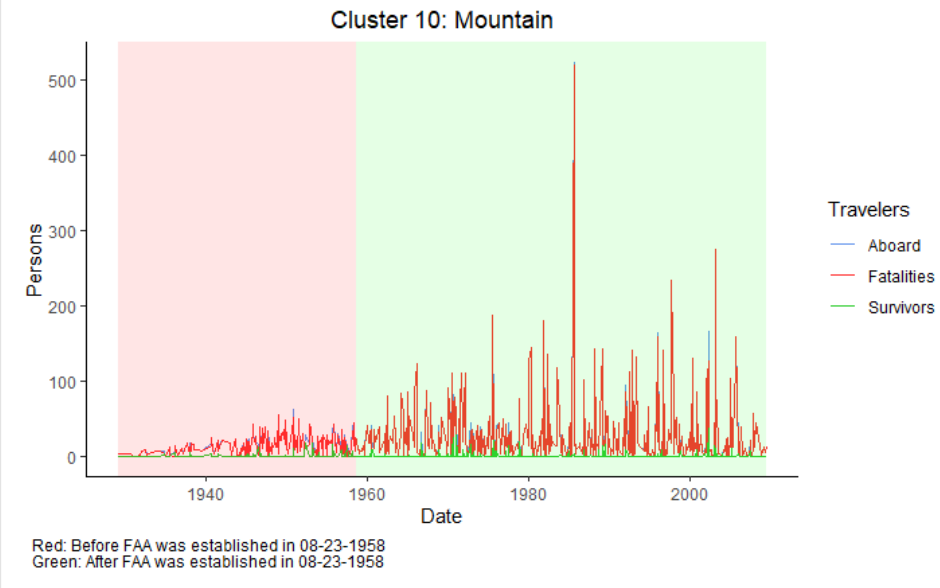
Cluster 9: Route

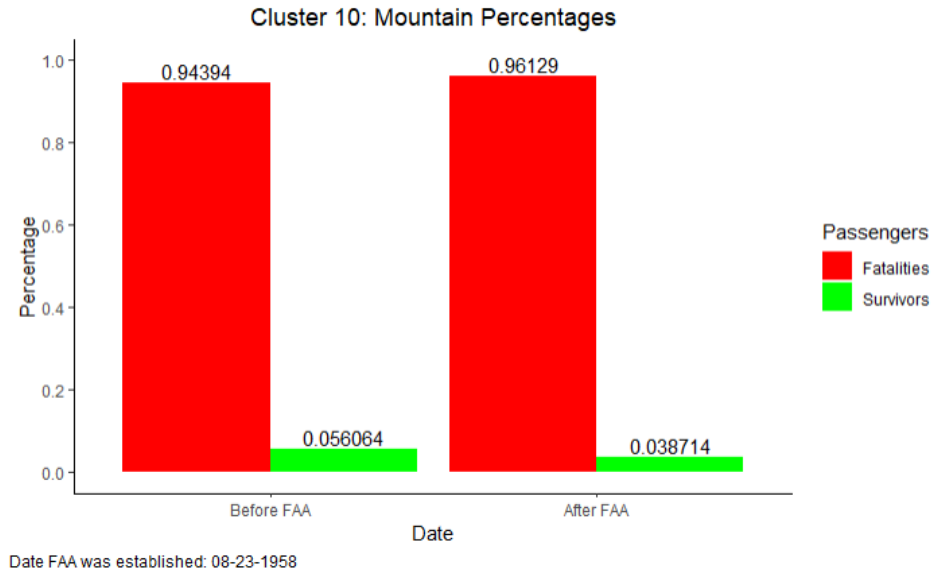




It was determined that the mortality rate of passengers decreased, and survivor rate increased when the FAA was enacted when there was a “route” issue involving a plane crash.

Cluster 10: Mountain





It was determined that the mortality rate of passengers increased, and survivor rate decreased when the FAA was enacted when there was a “mountain” issue involving a plane crash.

Conclusion:

The question that was asked was, after the data exploration, 10 unique keywords involving crashes were selected. With these keywords, was the mortality and survivor rate of passengers in these crashes lower or higher since the enactment of the FAA?

It is clear, that even after the FAA was enacted, "Mountain" showed the highest mortality rate (≈95%)**.** Unfortunately, when it comes to a mountain, there is really not much that could be done to avoid a tragic scenario. It is also shown that the highest survival rating is with engine (≈39%). However, it is shown that after FAA enactment, the mortality rate increased slightly. Because of this increase, there is a foreseeable problem that can be addressed, which is better maintenance and more checks and balances.

"Weather", "Cargo", "Altitude", "Control", and "Engine did show less than a 2% difference in mortality and survivor rate percentages. However, the following graphs, "Crew" and "Route", did show distinct changes in the mortality and survival rates after the FAA was enacted. When it comes to "Crew", which had a positive increase in the survival rate of passengers by ≈5.8%, it is most likely due to the fact of better training of airline staff. When the staff is knowledge and they know what to do in emergency situations, more lives are going to be saved. When it comes to "Routes", which had a positive increase in the survival rate of passengers by ≈9.9%, it is most likely due to better technologies to assist pilots. These statistics show the positive part of the FAA guidelines, regulations and technological advancements.

"Taking" is a result that is hard to interpret. Unfortunately, because of the clustering, "Taking" was one of the terms that was chosen. "Taking" did show an increase in mortality rate since the

FAA has been enacted by ≈4.5%. However, since "Taking" is such a broad term, it is hard to pinpoint exactly what the main issue that caused this increase was. Additional research and studies would need to be conducted to reach a viable conclusion.

"Cargo" was another intriguing result. As conveyed via the graphs, a mortality rate of ≈85% of the passengers has occurred. This could be attributed to the fact that cargo planes have different rules and regulations than passenger planes. Pilots that fly cargo planes also usually fly at non-peak hours (usually during the night) and into smaller airports with less than average grade instrumentation. This is another issue that need to be addressed.

Overall, the results concluded that from the 10 clustered choice keywords, since the FAA was enacted, 6 showed the results of the mortality rate decreasing and the survivor rate increasing and the other 4 showed the mortality rate increasing and the survivor rate decreasing.

There were some flaws within the study. One of the flaws was how the clusters were produced. According to the k-cluster graph, the appropriate number of centers should have been 6. There was a discretionary decision to make the clusters higher to help show more variability in frequent terms from the k-means clustering technique on a matrix obtained by the text corpus. There were also cluster 2 that had less desirable frequent terms within it. Also, the non-removal of outliers from the visualization before and after FAA enactment. By removing the outliers, it would create a more distributed graph and may allow for better data analysis.

The future work that could be completed on this question is creating less or more clusters to produce more specific terms. The sparsity of the words being taken out being greater or less than the 92.5% is also a variable of change that can be done. Also, filtering out time of day and

type of aircraft with the unique cluster word to see if more crashes involving the unique term happen during a certain time of day and if a particular aircraft has more issue than others. These issues can also be investigated to see if there was more of a hardware, software, pilot, tower and/or other issue.

Appendix A:

Code is found at the following repository.

https://github.com/npaisley93/school

Appendix B:

Data is found at the current locations:

https://github.com/npaisley93/school

https://www.kaggle.com/saurograndi/airplane-crashes-since-1908