APAN 5205: Applied Analytics Frameworks and Methods - Project Proposal

Cluster Analysis on Survival Rate on Air Disaster

Group Name: APAN5205 Geniuses

(Group: Keqi Yu, Shuchen Yao, Chang Yuan, Qi Shi)

27/02/2022

I. Overview

Over 100 years, a large number of airplanes crashed due to different causes. As time and technology arise, people put more effort into more complex means of transportation. All sources of transportation include ground, water, and air for the most common uses for people to travel from place to place. One concerning question that appears to be most eye-catching is how safe is the transportation we choose? Undouptfully, transportation disasters are happening every second around the world, and the most fatal with the lowerest survival rate post-disaster is traveling by aircraft. The safety of mankind while in the air requires high attention and precautions; especially for those large airplane companies and airports that have the responsibility of ensuring people's safety while in the air. Therefore, comparing the data from the twenty century to the twenty-first century, is there a rising number of airplane crashes due to more people traveling or decreasing number of airplane crashes due to technological improvement with more advanced aircraft manufacturers? These data and analysis on a variety of facts and factors that cause airplane crashes can be an asset to airports, aerospace companies, governments seeking to improve the passenger's air-safeness, own reputations, and decrease the number of deaths caused by air crushes to wider people.

II. Research Question

What are the similarities and differences between airplane crashes with high survival rates and crashes with high fatalities? The segments will then be explored to discover the possible differences based on the properties of aircraft (Aircraft model, Capacity) or causes of crashes(Damage Type, incident category).

III. Data Cleaning Summary

A. Data Source:

We will use Aircraft Accidents, Failures & Hijacks Dataset from Kaggle that shows worldwide accidents and hijackings involving passenger airliners, corporate jets and military transport aircraft since 1919. Airliners are considered here aircraft that are capable of carrying at least 12 passengers. The original dataset is from the ASN Safety Database, updated daily, contains descriptions of over airliner, military transport category aircraft and corporate jet aircraft safety occurrences since 1919.

In the raw dataset, there are 23519 incidents recorded in the dataset and 23 variables including incident information, descriptions of aircraft, and information about onboard passengers. The detail of the variables is shown in Appendix.

B. Data Preparations:

In order to make the data more suitable for analysis, the following data cleaning procedures were made.

1. Missing Values:

By glancing the dataset, there are many missing values displaying as '?', '-' and ''. After filtering these out, we replace these values with NA. According to Table 2: Missing Variables in reference, a few variables have more than 50% missing values: "Time", "Aircraft_Engines", "Ground_Casualties", "Collision_Casualties". I only removed Aircraft_Engines and Time, because we thought the variables "Ground_Casualties" and "Collision_Casualties" record casualties causing by hitting the ground and collision. Although there are only 1% incidents causing casualties on ground and by collision, we cannot ignore these fatalities.

2. Keeping useful variables:

We deleted eleven variables that are not relevant to this research question. The rest of the variables (Aircraft_Damage_Type, Aircraft_Model, Aircraft_Operator, Incident_Category,), help to analyze plan crashes, and compare segments of different crashes with different survival rates in order to generate any similar patterns.

3. Changing Data Types:

For variables Collision_Casualties, Ground_Casualties, Onboard_Total, these character data are actually describing number, so we extract the number out and change them to numeric data. For NA, we changed to 0. For incident_date, we changed character to date format.

4. Adding more variables:

Since our research question is to investigate the survival rate, we create survival rate by calculating the ratio of total survivors(total occupants - total onboard fatalities) and total occupants. In addition, we calculate the total fatalities which were resulted by hitting the ground or collision.

C. Variables:

After data cleaning process, the cleaned dataset consists of 23519 observations, and 17 variables, as shown in the table below.

Table 1: Variables

Aircaft_Damage_Type	Aircaft_Model	Aircaft_Operator	Aircraft_Phase
Collision_Casualties	Fatalities	Fatalities_Onboard	Ground_Casualties
Incident_Category	Incident_Date	Incident_Location	Incident_Month
Incident_Year	survival_rate	Survivors_Onborad	Total_casualities
Total_Occupants	Aircaft_Damage_Type	Aircaft_Model	Aircaft_Operator

D. Data Visualization:

The overall aircraft crash data plots are provided below.

Figure 1: Number of Incidents/Year shows the number of incidents occurred per year from 1919 to 2022. As we can see from the table, despite that there is a peak around Year 1943 to Year 1947, the number of incidents happened each year are similar. However, with increasing transportation needs, the number of flights are increasing dramatically from 1919 to 2022, which indicates that ratio of incidents may drop dramatically over time.

Figure 1: Number of Incidents/Year
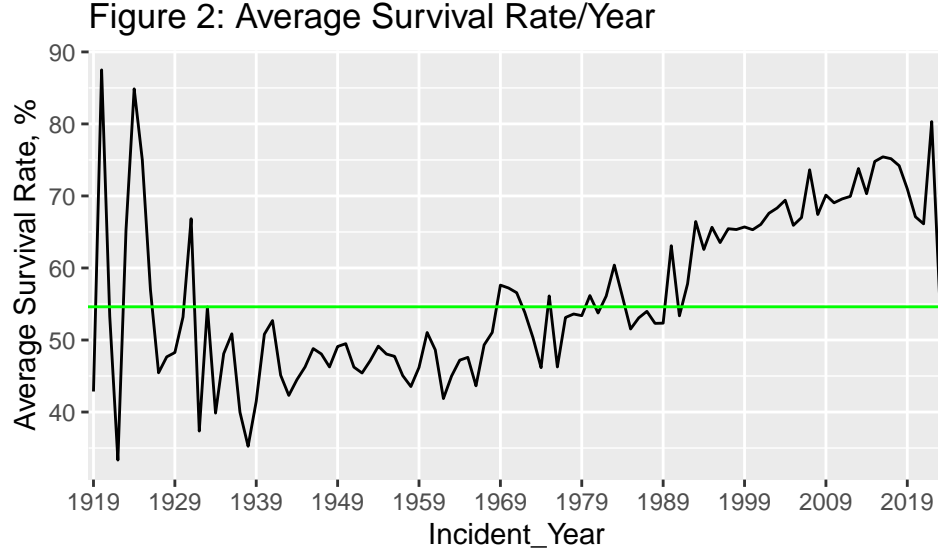
Figure 2: Average Survival Rate/Year
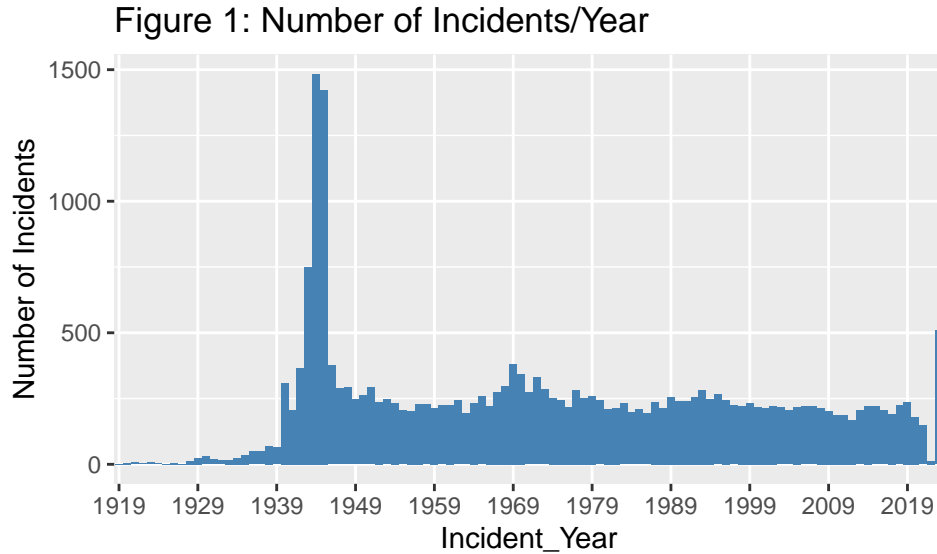
Figure 2: Average Survival Rate/Year shows the average survival rate for each year. The green line represents the mean of survival rate from 1919 to 2022, which is 54.61%. As we can observe from Figure 2, from 1919 to 1929, the trend appears to fluctuate widely, while for almost 40 years from 1929 to 1969, the average survival rates become flatter but all fall below the mean. Then, after 1969, the survival rate has continued to increase. However, there is a sudden drop in recent years, which worth more investigation.

Figure 3: Travelers Aboard, Survivors and Fatalities Pe
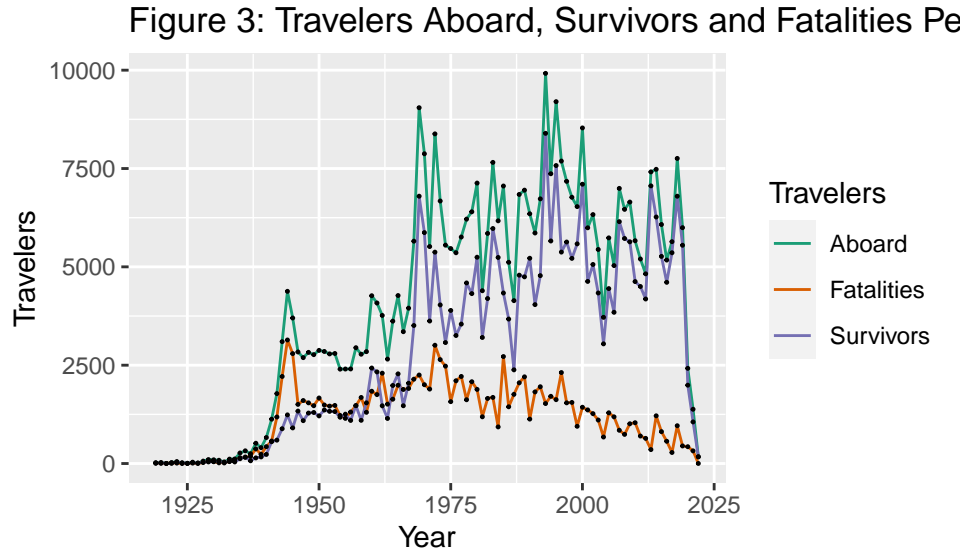
As we can see from Figure 3, the total number of travelers aboard airplane crashes declined in the past two decades. In the last 50 years, fatalities also declined as would be expected with the decline of travelers involved in crashes. The gap in the difference between the number of fatalities and survivors has increased over the years. This indicates that the survival rate over time has became more and more higher.

IV. Analysis

In order to solve our research question, first, we decided to cluster the data to segment crashes into low and high survival rates in order to discover the possible differences between survivable crashes and fatal crashes. The method we used is K-Means Clustering, which attempts to find groups that are most compact, in terms of the mean sum-of-squares deviation of each observation from the multivariate center (centroid) of its assigned group. This disorganized approach to clustering produces similar quality of clusters to hierarchical clustering

but much faster.

Based on the rules of k-means, it doesn't allow the missing value, so we omit all the NAs from the dataset.

The k means clustering analysis technique applied to find similarities between airplane crashes based on: the number of people aboard, fatalities, survivors and survival rates. Twenty year intervals were applied to the clustering analysis model in the event technical advances that took place over the years may have contributed to increasing or decreasing crash events.

We tried four different K means cluster sizes of 3, 4, 5, and 6 were made along with a visualization comparison.

Figure 4: K means cluster sizes of 3, 4, 5, and 6
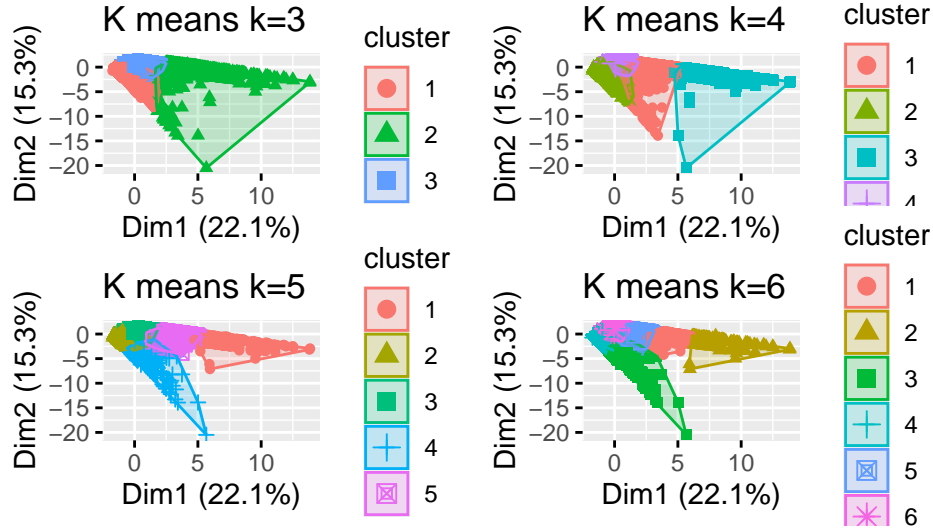
As shown in Figure 4, all four clusters provided distinctive groups with little overlap when depicted in a 2 dimensional space. To aid in determining the optimal number of clusters to us, the elbow method using the total within sum of squares plotted against the number of values of k was applied. We compare the total within sum of squares for a set of 10 cluster solutions. Ideal number of clusters is inferred from a sudden change in the line graph. As we can see from the Figure 5: Optimal number of cluster, three clusters will be used to segment the airplane crashes.

Figure 5: Optimal number of cluster
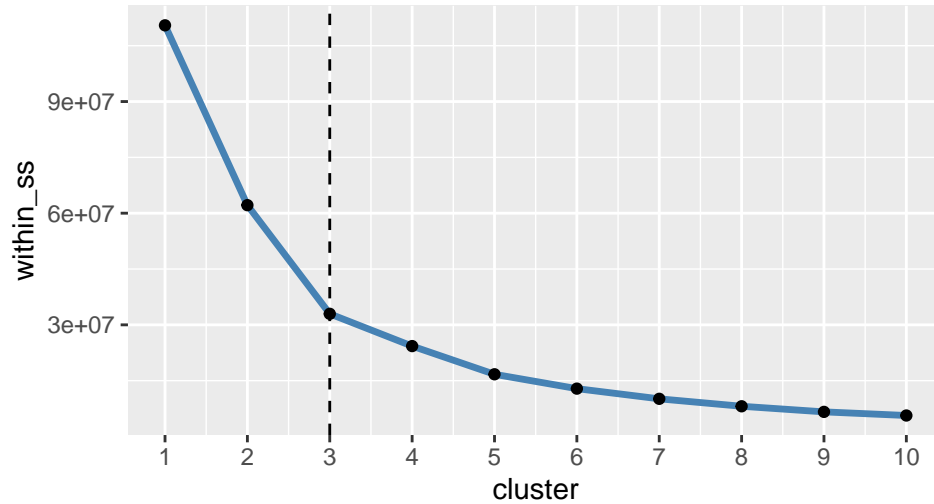
In order to examine the cluster result of airplane crash, we generated a list of mean values for travelers aboard, fatalities, survivors and survival rates. Moreover, to identify the capacity of aircraft, the minimum

and maximum values for travelers aboard would also be included.

Table 2: Aboard, Fatalities, Survivors and Survival Rates

Cluster	Plane_Crashes	Max_Aboard	Min_Aboard	Mean_Aboard	Mean_Fatalities	Mean_Survivor	Mean_SurvivalRate
1	6928	200	1	16.38842	15.5399827	0.8484411	4.308984
2	1177	524	86	161.39762	9.9167375	151.4808836	95.632957
3	8537	98	0	14.15462	0.6962633	13.4583577	89.780642

Inspecting the mean values by Aboard, Fatalities, Survivors and Survival Rates, the clusters can be defined by the number of travelers aboard and their survival rates.

- Cluster 1 - travelers ranged from 1 to 200 with an avg. of 16 aboard and had a low mean survival rate of 4%
- Cluster 2 - travelers ranged from 87 to 524 with an avg. of 162 aboard and had a high mean survival rate of 96%
- Cluster 3 - travelers ranged from 0 to 98 with an avg. of 15 aboard and had a high mean survival rate of 93%

Note: The average survival rate for all crashes in the data set was 54.61%.

These clusters can be referred to as follows:

Small-Midsize Passenger High Fatality = Cluster 1 Large Passenger High Survival = Cluster 2 Small Passenger High Survival = Cluster 3

The next step is to evaluate survival rate for each cluster as per their cluster profile. Eventually, We would like to know if the survival rate is influenced by any significant factors.

Figure 6: Plane Crashes Per Year by Cluster
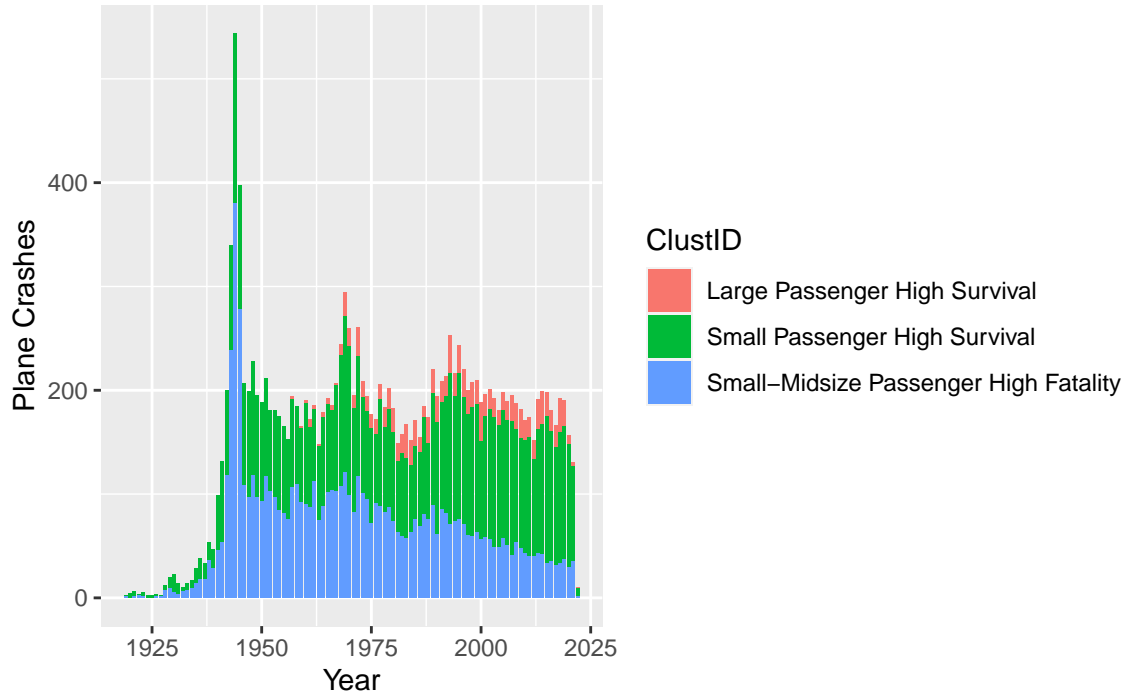
The chart clearly shows that Small Passenger Crashes involving planes with less than 98 occupants historically make up the bulk of plane crashes. There is a peak during Year 1940 to Year 1945. Moreover, Small-Midsize

Passenger planes with less than 200 occupants are half less than the Small Passenger Plane. Large passenger plane crashes averaging more than 162 occupants begin to appear in the 1960's.

Travelers Aboard Airplane Crashes Per Year by Cluster
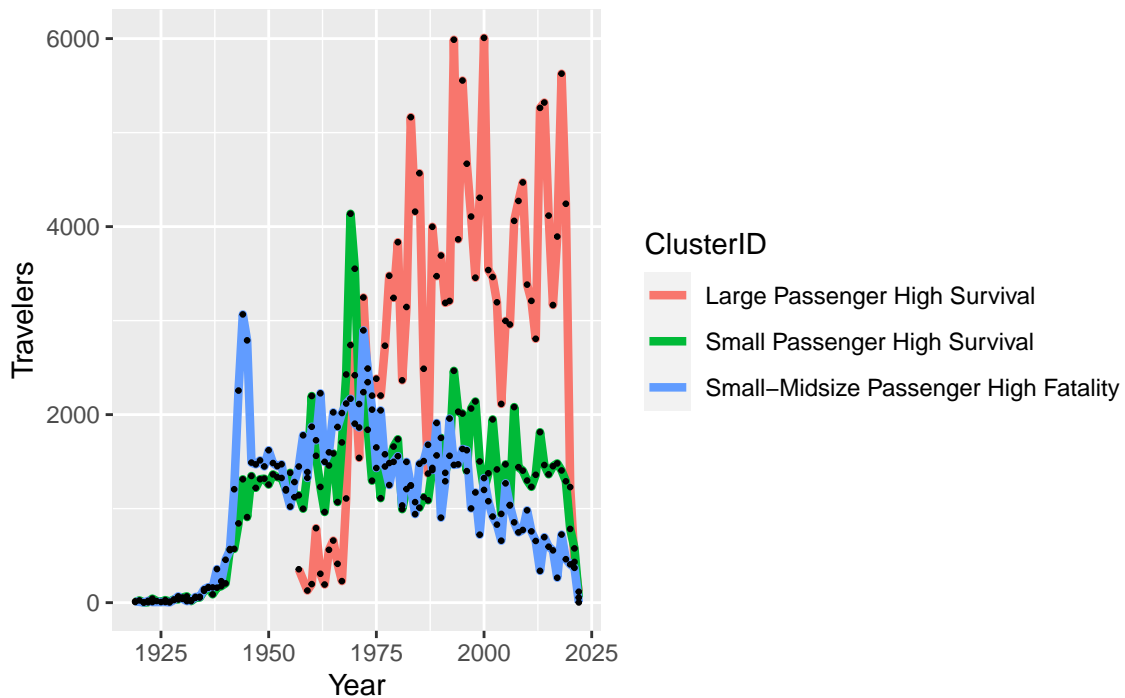
As the years progressed and large capacity planes increased, the annual number of travelers aboard large airplanes that crashed began to exceed the annual number of travelers aboard small to midsize and small planes that crashed. This is especially evident from the 1970's onward.

## Airplane Crash Fatalities Per Year by Cluster
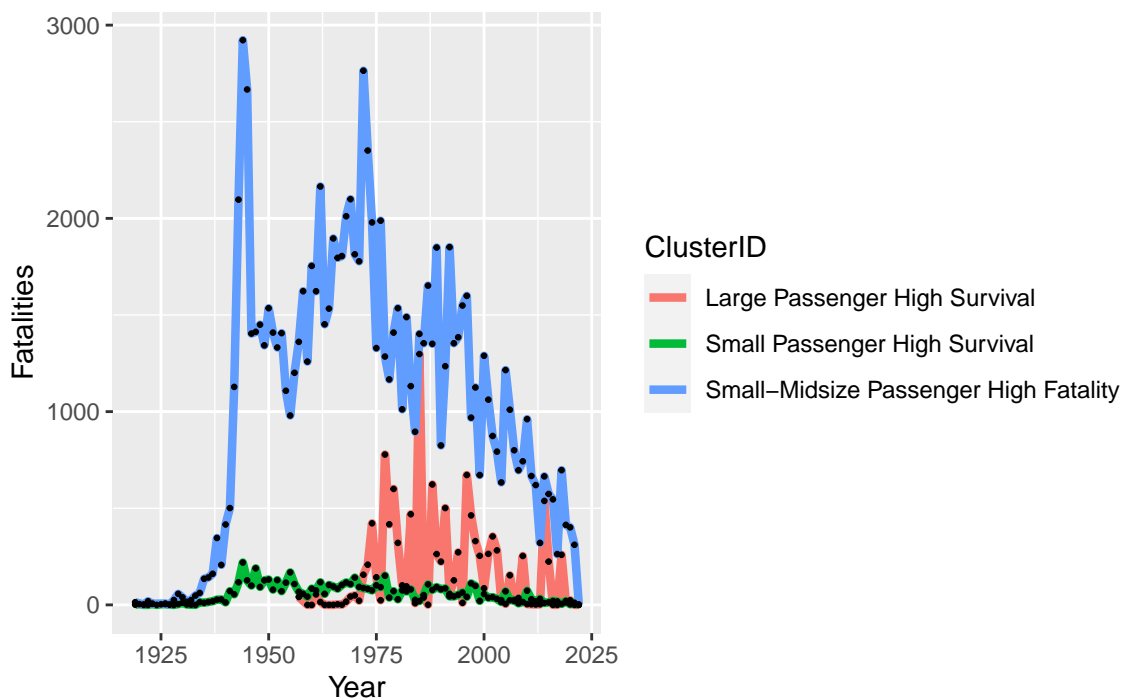
As expected small plane crashes with high survival rates ("Small Passenger High Survival") had the lowest annual fatalities.
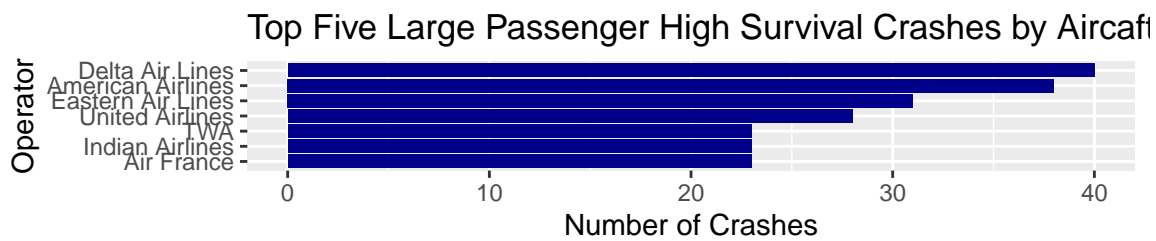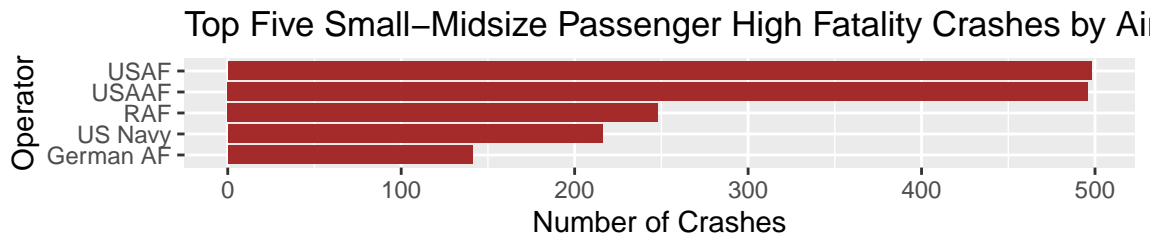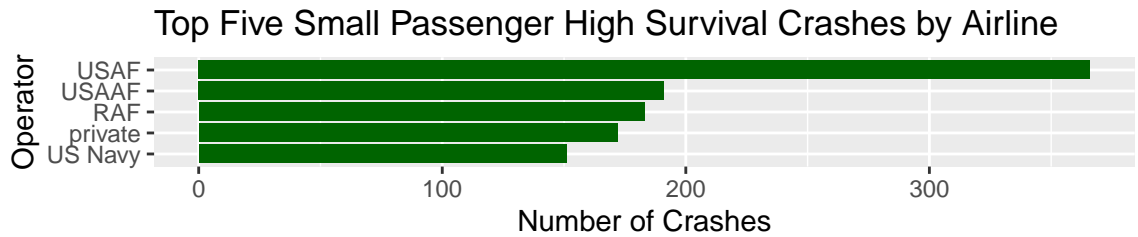
Also, as expected small-midsize plane crashes with high fatality rates ("Small-Midsize Passenger High Fatality") had the highest annual fatalities over years.

Starting in the 1960's we also see large plane crashes represented by the "Large Passenger High Survival" cluster (red line) begins to raise but still recognize as low fatality rate.

Comparing Survival Rates on Three Size Passenger Planes
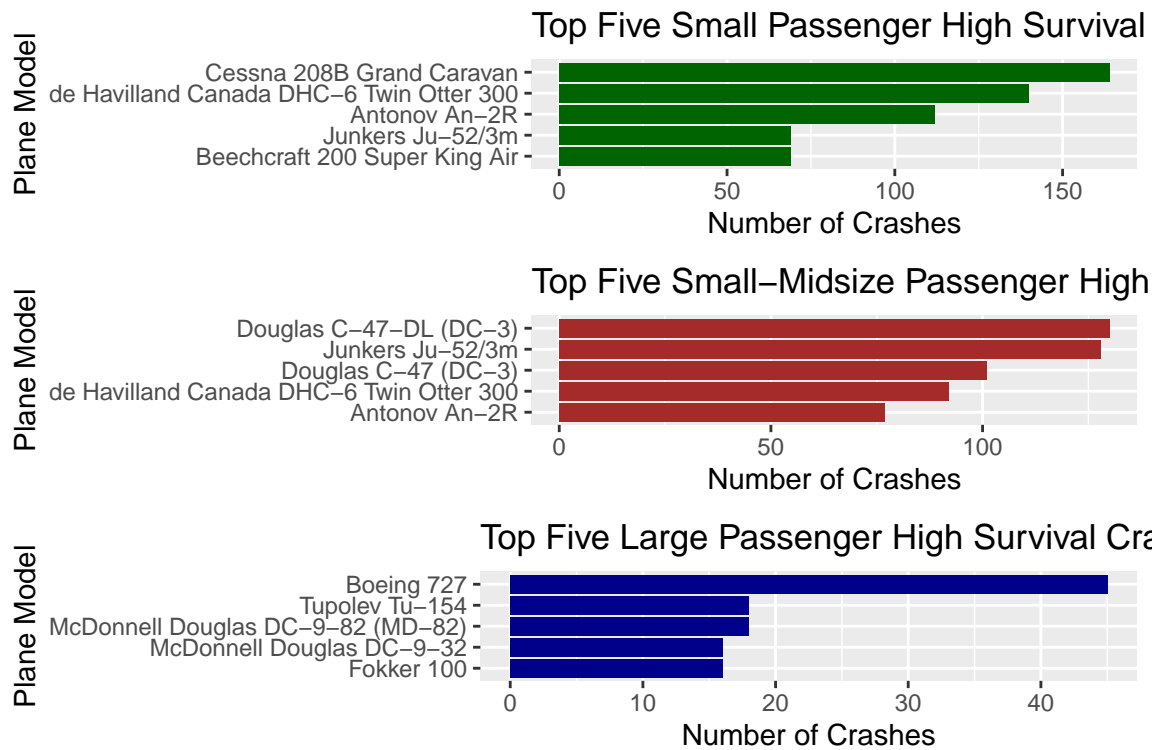
- American made Cessna 208B Grand Caravan and Canadian made de Havilland Canada DHC-6 Twin Otter 200 models appeared most frequently among the top 5 small passenger plane crashes with high survival rates.

- American made Douglas models and the German Junkers Ju-52/3m models made up the top 5 small-midsize passenger plane crashes with high fatalities.

- For large passenger airplane, the model most frequently involving in the air crashes, are Boeing 727.

Top Five Small Passenger High Survival Crashes by Airline

Top Five Small–Midsize Passenger High Fatality Crashes by Ai

Top Five Large Passenger High Survival Crashes by Aircaft

As we can see from the plots, both USAF and USAAF Airline appeared most frequently among both the top 5 small passenger plane crashes with high survival rates and high fatality rates.

Based on this result, we cannot determine whether airline operator will affect the survival rate or not.

Large passenger plane are mostly operated by Delta Airlines and American Airlines.

Top Five Small Passenger High S

Top Five Small–Midsize P

Top Five Large Passenge
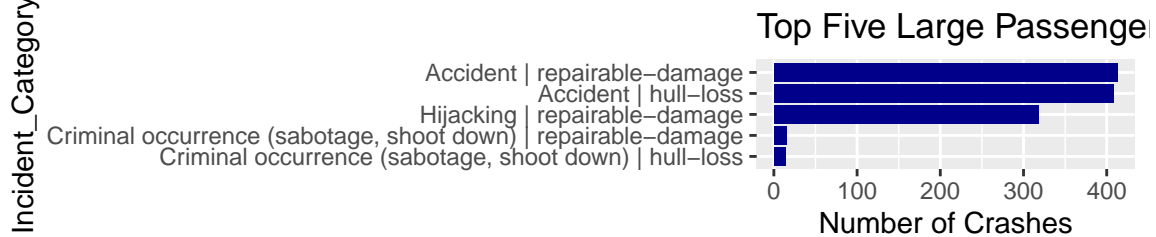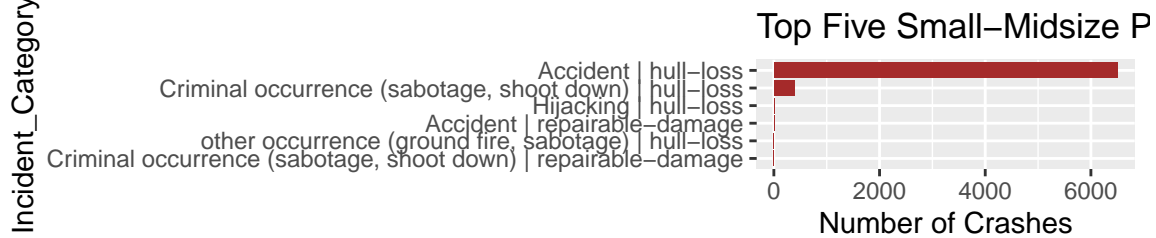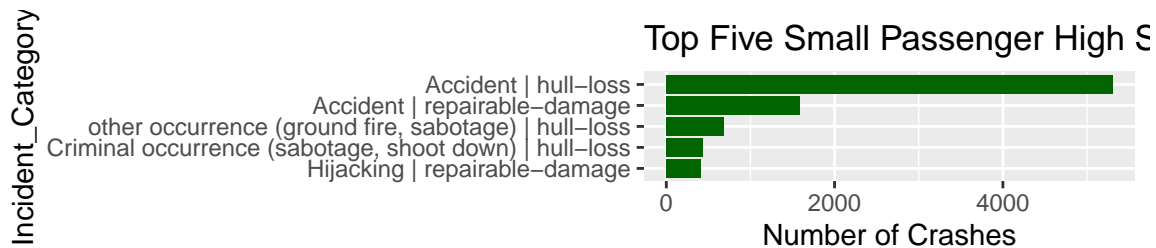
Small passenger clusters and small-midsize passenger both have the highest accidents of a hull loss.

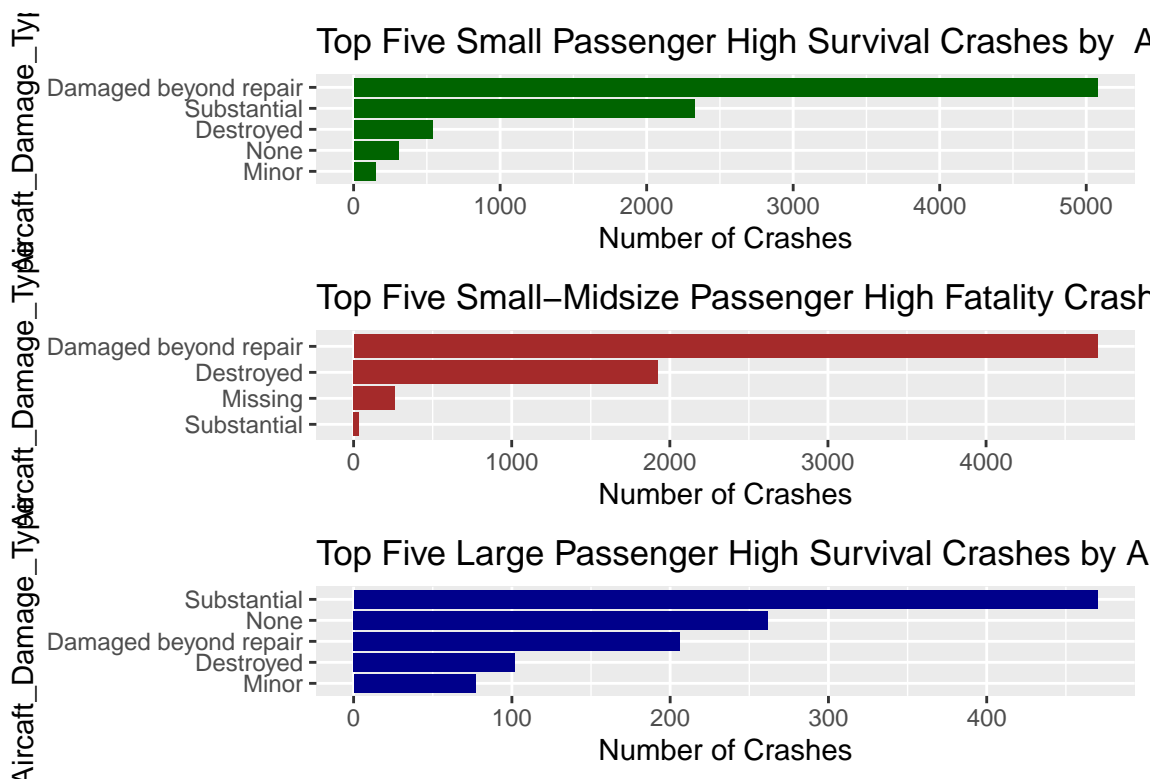Compared to large passenger clusters, the proportion of hijacking is significantly higher than in small and small-midsize aircraft. In addition, large passenger also has the highest proportion of repairable damage.

Top Five Small Passenger High Survival Crashes by A

Aircraft_Damage_Type	
Damaged beyond repair	
Substantial	
Destroyed	
None	
Minor	

Number of Crashes: 0, 1000, 2000, 3000, 4000, 5000

Top Five Small–Midsize Passenger High Fatality Crash

Aircraft_Damage_Type	
Damaged beyond repair	
Destroyed	
Missing	
Substantial	

Number of Crashes: 0, 1000, 2000, 3000, 4000

Top Five Large Passenger High Survival Crashes by A

Aircraft_Damage_Type	
Substantial	
None	
Damaged beyond repair	
Destroyed	
Minor	

Number of Crashes: 0, 100, 200, 300, 400

By comparing the 3 clusters, all the cluster aircraft_damage_type "Damaged beyond repair" ranked the highest, except for the large passenger cluster.

Small-midsize passenger has the highest destroyed damage type compared to the other two clusters.

The large passenger cluster contains the highest substantial damage type.

V. Conclusion

Although the occurrence of air crashes contains many unknown factors, some of which are factors that cannot be predicted by human beings, we can still rely on scientific and technological means to help human beings reduce the occurrence of air disasters. The findings provided results to answer the question and discover reasonable patterns. In this instance, cluster analysis was used as a tool to further explore different segments under different factors that successfully identified patterns in three groups of survival and fatality rate and contributed insights to causes to similar yet different survival and fatality rate.

By identifying different clusters, we analyzed whether there are any significant factors that influence survival rate. We separated all plane crashes to 3 groups: Group 1 which travelers ranged from 1 to 200 has low mean survival rate of 4%, Group 2 ranged from 87 to 524 travelers with an average of 162 aboard had a low mean survival rate of 96%, Group 3 with travelers ranged form 0-98 with an average of 15 aboard had a mean survival rate of 93%. In addition, clearly small passenger crashes with 98 occupants are at most among all crashes, and large passenger plane crashes with more than 162 occupants appear after 1960.

Furthermore, for factor aircraft model, American made Douglas models appears most frequently in the top 5 small-midsize passenger plane crashes with high fatalities. For factor aircraft operator, USAF and USAAF are most frequently involved in the crashes for both high survival rate and high fatality rate, so it is hard

to see the differences. For incident category, there is also not obvious difference between two groups. For aircraft damage type, damaged beyond repair makes up the most crashes, but there is not obvious pattern for differences.

Despite the clear pattern of the findings, there are some limitations that can be spotted while analyzing the result. For instance, before 1960 there were nearly no large passenger plane crashes, this could be due to the reason that before 1960, the aero technology was not mature enough to carry that large amount of passengers, therefore, most crashes would focus on small-sized passengers. Not only that, there were also military aircraft models included in the data, since military aircraft can only be occupied by one to two people, therefore, that could also be part of the reason why small-mid sized crashes accounted for more before 1960.

There are other considerations to keep in mind that when interpreting the result of cluster analysis, it does not help to predict or model based on current data; it helps to find a pattern from current data through visualization. The use of the findings could help to have a clearer view on how survival rates change and spread under different factors in order to help readers understand what can be reduced and what can be improved to eventually enhance the safety of passengers while in the area

VI. Recommendation

Based on the result we analyzed, we can conclude that an airplane with a large capacity has a higher survival rate. On the other hand, an airplane with a lower capacity has a lower survival rate. Therefore, we recommend that passengers try to choose the largest aircraft possible when traveling. In spite of the result-based recommendation, logical reasoning should support our recommendation. Indeed, the row data not only reports commercial flights but also reports military, corporate jet, and training aircraft. Therefore, the recommendation is not necessarily true that larger aircraft are safer.

With the development of science and technology, the models of aircraft are also constantly updated. The latest aircraft have a faster flight speed and can reach longer flight distances than before, also accompanied by higher safety performance. As clarified by Julie O'Donnell, a spokesman for Boeing, where the 50s and 60s fatalities occurred roughly every 200,000 flights, today "fetal accidents [occur] less than once in every two million flights". National Transportation Safety Board also proves Julie O'Donnell's explanation, in fact, passengers on flights now have a 95% chance of surviving an accident now.

From our observation of the data and the above, we can see that the cause of the aircraft impact may not only be a problem with the aircraft itself but may be accompanied by many human factors. There were 130 hijackings between 1868 and 1972. "Some situation goes really bad, One of the most infamous pre-9/11 hijackings was TWA Flight 847, in which Hezbollah terrorists hijacked a plane bound from Athens to Rome, sought out passengers with "Jewish-sounding" names, and forced a two-week-long hostage situation at an airport in Beirut." But since 911, there have been only 50 hijackings in the last 20 more years, and none of them happened in the United States. That is because the Airline Company changed the security and airplane construction. It is another way to minimize the chance of air flight accidents by human factors.

VII. References

Table 3: Missing Variables

	x
Incident_Date	0.0
Aircaft_Model	0.0
Aircaft_Registration	3.5
Aircaft_Operator	0.0
Aircaft_Nature	2.0
Incident_Category	0.0
Incident_Cause.es.	0.0
Incident_Location	2.0
Aircaft_Damage_Type	0.0
Date	0.0
Time	30.5
Arit	0.0
Aircaft_Engines	27.5
Onboard_Crew	0.0
Onboard_Passengers	0.0
Onboard_Total	8.5
Fatalities	0.0
Aircaft_First_Flight	11.5
Aircraft_Phase	0.0
Departure_Airport	19.5
Destination_Airport	19.5
Ground_Casualties	49.5
Collision_Casualties	49.5
Total_Occupants	14.5

```
## 'data.frame':    23519 obs. of  23 variables:
##  $ Incident_Date      : chr  "03-JAN-2022" "04-JAN-2022" "05-JAN-2022" "08-JAN-2022" ...
##  $ Aircaft_Model      : chr  "British Aerospace 4121 Jetstream 41" "British Aerospace 3101 Jetstream
##  $ Aircaft_Registration: chr  "ZS-NRJ" "HR-AYY" "EP-CAP" "RA-64032" ...
##  $ Aircaft_Operator   : chr  "SA Airlink" "LANHSA" "Caspian Airlines" "Cainiao, opb Aviastar-TU" ..
##  $ Aircaft_Nature     : chr  "Domestic Non Scheduled Passenger" "Domestic Scheduled Passenger" "Dome
##  $ Incident_Category  : chr  "Accident | repairable-damage" "Accident | repairable-damage" "Accident
##  $ Incident_Cause.es. : chr  "Airplane - Engines, Airplane - Engines - Prop/turbine blade separation
##  $ Incident_Location  : chr  "near Venetia Mine..." "Roatán-Juan ..." "Isfahan-Shah..." "Hangzhou-X
##  $ Aircaft_Damage_Type: chr  "Substantial" "Substantial" "Substantial" "Destroyed" ...
##  $ Date               : chr  "Monday 3 January 2022" "Tuesday 4 January 2022" "Wednesday 5 January
##  $ Time               : chr  "08:10" "ca 12:00" "17:07" "04:40" ...
##  $ Arit               : chr  "03-JAN-2022" "04-JAN-2022" "05-JAN-2022" "08-JAN-2022" ...
##  $ Aircaft_Engines    : chr  "2 Garrett TPE331-14GR-805H" NA "2 CFMI CFM56-3C1" "2 Soloviev PS-90A"
##  $ Onboard_Crew       : chr  "Fatalities: 0 / Occupants: 3" "Fatalities: 0 / Occupants:" "Fatalities
##  $ Onboard_Passengers : chr  "Fatalities: 0 / Occupants: 4" "Fatalities: 0 / Occupants:" "Fatalities
##  $ Fatalities         : int  0 0 0 0 0 0 0 0 0 2 ...
##  $ Aircaft_First_Flight: chr  "1995-05-19  (26 years 8 months)" "1985" "1992-09-18  (29 years 4 month
##  $ Aircraft_Phase     : chr  "Landing (LDG)" "Landing (LDG)" "Landing (LDG)" "Standing (STD)" ...
##  $ Departure_Airport  : chr  "Johannesburg-O.R. Tambo International Airport (JNB/FAOR) , South Afric
##  $ Destination_Airport : chr  "Venetia Mine Airport (FAVM) , South Africa" "Roatán-Juan Manuel Gálvez
##  $ Ground_Casualties  : chr  NA NA NA NA ...
##  $ Collision_Casualties: chr  NA NA NA NA ...
```

```
##  $ Total_Occupants    : num   7 19 116 8 0 NA 3 NA 12 2 ...
```
|| || || ||