Trend of Flight Safety and Detection of Causes of Fatal Air Accidents Using Text Mining Techniques

Kaiwei Qian

2019/5/4

1 Introduction

There are 6 sections in this report. Data and the ways how they are collected from Internet are described in Section 2. Exploratory analysis is performed in Section 3. In Section 4, the probability of being involved in an air accident is studied for the past 10 years (2009-2018), followed by Section 5, where text mining techniques are tried to extract the keywords from the cause of air accidents. The last section gives the conclusion of the project.

2 Data

2.1 Aircraft accidents dataset

The data used in this study is scraped from the **Bureau of Aircraft Accidents Archives** (http://www.baaa-acro.com/crash-archives) for the past 10 years (2009-2018). The Bureau of Aircraft Accidents Archives (BAAA) was established in Geneva in 1990 for the purpose to deal with all information related to aviation accidentology.

Scraping involves two steps.

First, After **read_html()** reads the string-type HTML from the response, **html_node(xpath = '/html/body/div[1]/div[4]/div/section/div[2]/div/div/div[3]/div/table')** locates the table that contains the information of the air accidents, such as Date, A/C Type, Location, Fatalities and Registration. Detailed explanation of these variables are shown in Table 1.

If you visit their website, you'll find that there is a plus sign at the end of each row of the table. Clicking on it leads you to the page that briefly summarizes the air crash. It provides more detailed information, such as Flight Phase and Circumstances. So, I choose to keep the hyperlinks to this page. **html_nodes(xpath = '/html/body/div[1]/div[4]/div/section/div[2]/div/div/div[3]/div/table/tbody/tr/td[8]/a/@href')** locates the column of the hyperlinks belonging to and **html_text()** reads the data from that column.

In the second step, the spider visits each page that contains a brief summary of air crashes. I choose to keep the records of Flight Phase, Flight Type, Site, Circumstances and Probable Causes (if exists) for further study. Circumstances and Probable Causes are mainly used to support the text mining part of this project. XPath is used to read the data from the HTML.

Note that instead of air crashes, this dataset include all the records of air accidents, which means the plane may not be necessarily hit land or water and be damaged. For example, some mechanical problems during the flight are also deemed as air accidents.

Table 1: Description of the variables in BAAA Air Accidents dataset.

Variable Name	Data Type	Description
Date	Date	The date when the accident occurred

Variable Name	Data Type	Description
A/C Type	character	The type of the aircraft
Location	character	The location where the accident occurred
Fatalities	numeric	Number of deaths caused by the accident
Registration	character	A unique ID for the airplane
Flight Phase	character	Flight phase during which the accident occurred
Circumstances	character	A breif description of the circumstances where the accident occurred
Flight Type	character	The type of this flight; can be private, commercial and so on
Site	character	Description of the location where the air crash occurred
Probable cause	character	Possible causes given by the investigation by NTSB

2.2 Flight statistics

The flight statistics dataset is scraped from **Statista**. The number of flights performed by the global airline industry from 2009 to 2018 and the number of scheduled passengers boarded by the global airline inudstry from 2009 to 2018 are collected from https://www.statista.com/statistics/564769/airline-industry-number-of-flights and https://www.statista.com/statistics/564717/airline-industry-passenger-traffic-globally, respectively.

In each of websites mentioned above, there are two tables, which contain the number of scheduled commercial flights and the number of passengers and the number of scheduled passengers boarded from 2004 to 2019, respectively. For each of the websites, **html_node(xpath = '//table')** locates the table, and **html_table()** reads the table from its HTML structure.

Only the records from 2009 to 2018 are used in this project in order to keep them consistent with the air accidents dataset in terms of time range. Moreover, the numbers from **Statista** are only for the commercial flights while the air accidents dataset also contain the records for private, cargo and other flights. So, it's essential to keep the difference of these two data sources in mind and avoid causing confusion.

3 Exploratory Analysis

3.1 Aircraft Type

As shown in Table 1, the air accidents dataset contains 1343 observations and 10 variables. The 10 variables are Date, A/C Type, Location, Fatalities, Registration, Flight Phase, Circumstances, Flight Type, Site, Probable cause.

Barplots are employed to see what types of aircraft is involved in most air accidents and what types of aircraft cause the most fatalities. After grouping the data by **A/C Type**, I sum the **Fatalities** and count the number of air accidents for each type of aircraft. **gather()** and **facet_wrap()** are used to generate these two plots.

Top 10 aircraft types are shown in Figure 1. **Boeing 777-200** has caused the most deaths and **PZL-Mielec AN-2** has been involved in the most accidents in the past 10 years. Interestingly, the A/C Type which causes hundreds of deaths are not among the types which are most frequently involved in accidents, except for **Lockheed C-130 Hercules**. One possible reason for this phenomenon is that **Boeing** and **Airbus** produce commercial airplanes that can contain hundreds of people and a handful of fatal air crashes can lead to hundreds of fatalities.

In conclusion, the capacity of different aircrafts must be taken in to consideration when we evaluate the flight safety.

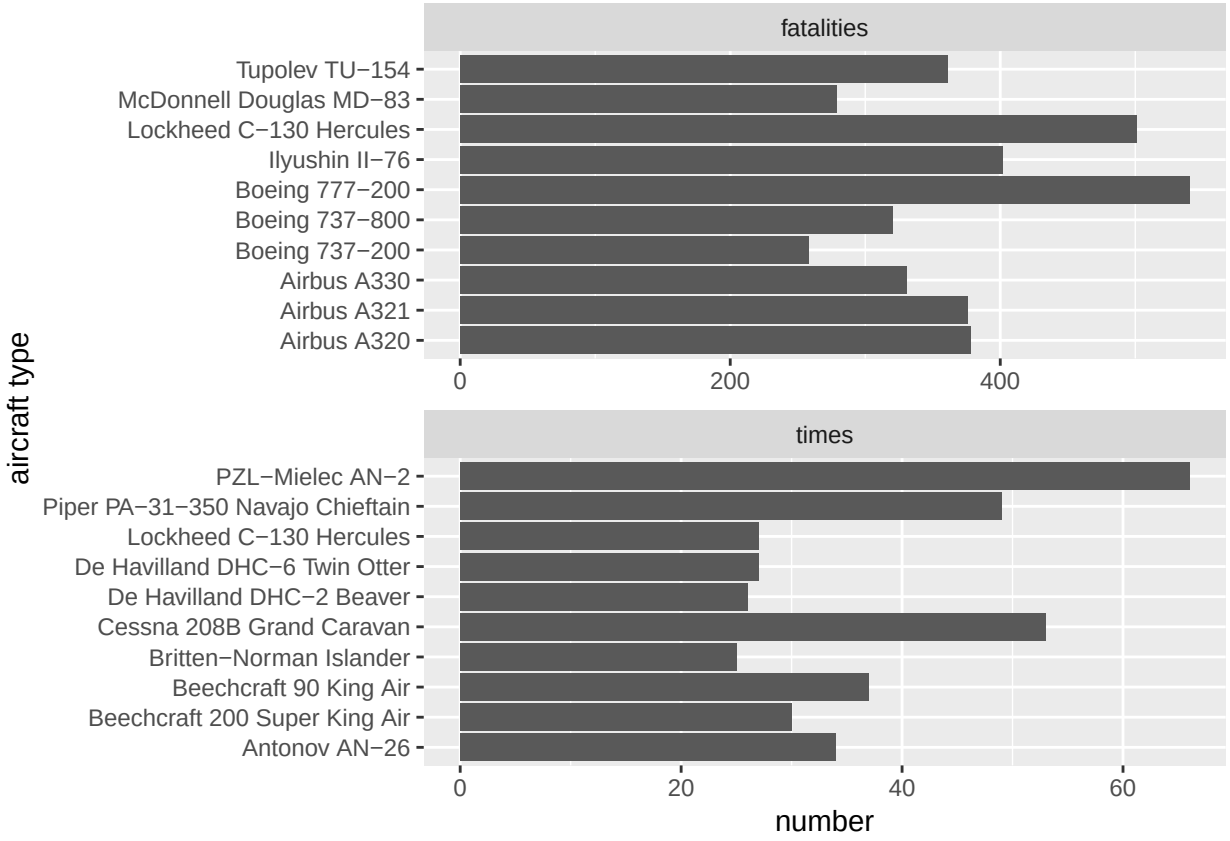

Figure 1: Frequency and Fatalities of Air Accidents by Aircraft Type.

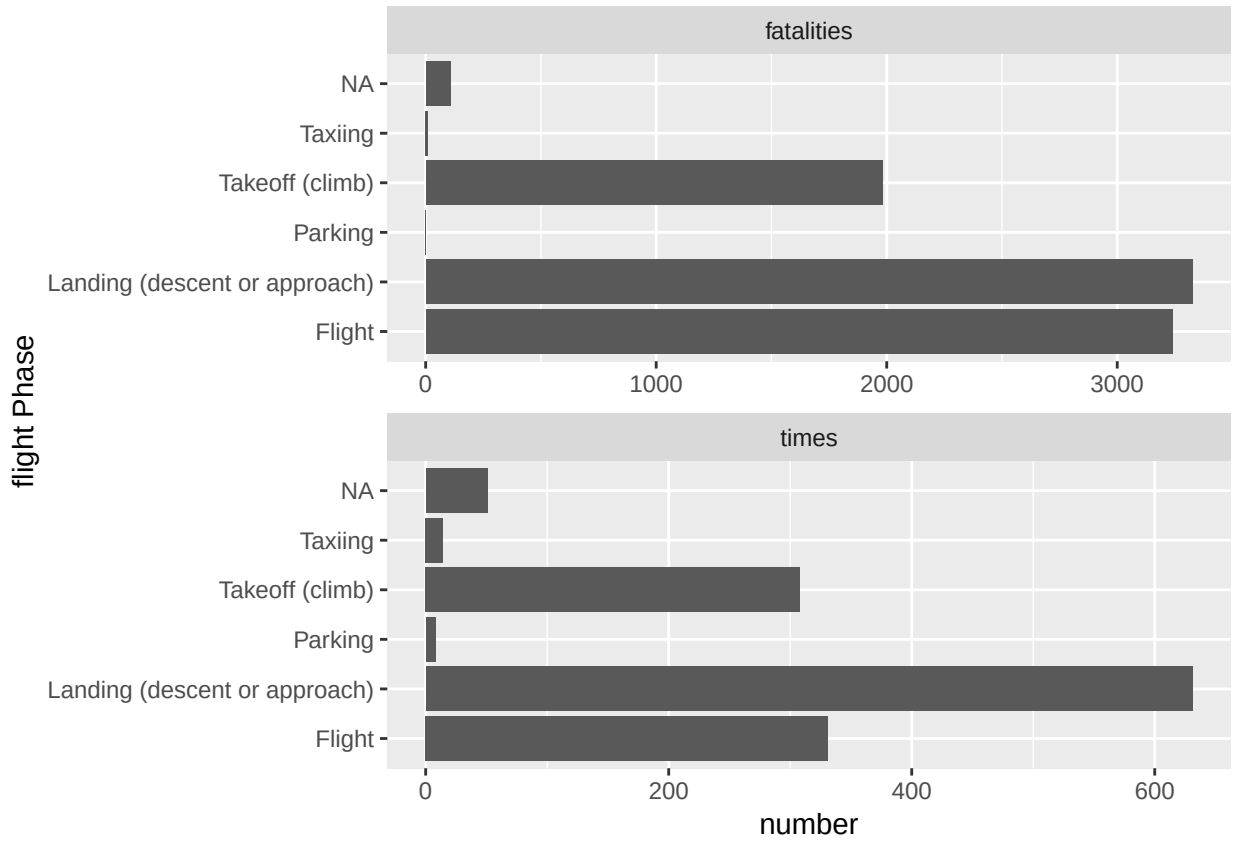

Figure 2: Frequency and Fatalities of Air Accidents by Flight Phase.

3.2 Flight Phase

Barplots in Figure 2 shows that nearly half of the accidents that occur during the landing phase. Though only 1/4 of the accidents happen during the flight, the fatalities caused in this phase are as many as that those caused during the landing phase.

When the plane is landing, it has lower speed and lower altitude. So, even the airplane accidents during the landing or approaching phase occur more often than during flight, the chance of surviving must be higher than the accident in the sky. What's more, landing or approaching means that the plane is near an airport, where the people are ready for any emergency situation.

4 Time Series for Commercial Flights

In this section, records are grouped by year, and I'm interested in the trend of the fatalities and number of air accidents from 2009 to 2018. In addition, the number of scheduled flights and the number of passengers boarded are used to normalize the air accidents dataset in order to achieve fair evalution of flight safety.

Figure 3 indicates that the fatalities related to air accidents fluctuate around 800 per year, and the number of air accidents has deacreased from 160 to 110 from 2009 to 2018. The problem is whether fluctuating fatalities caused by air crashes each year indicates no improvement in flight safety.

Figure 4 shows that the number of commercial flights scheduled and number of passengers each year has been increasing in the past ten years. Therefore, it's more appropriate to evaluate the flight safety by the ratio of fatalities to the number of passengers aboard and **fatal accident rate**, which is the ratio of number

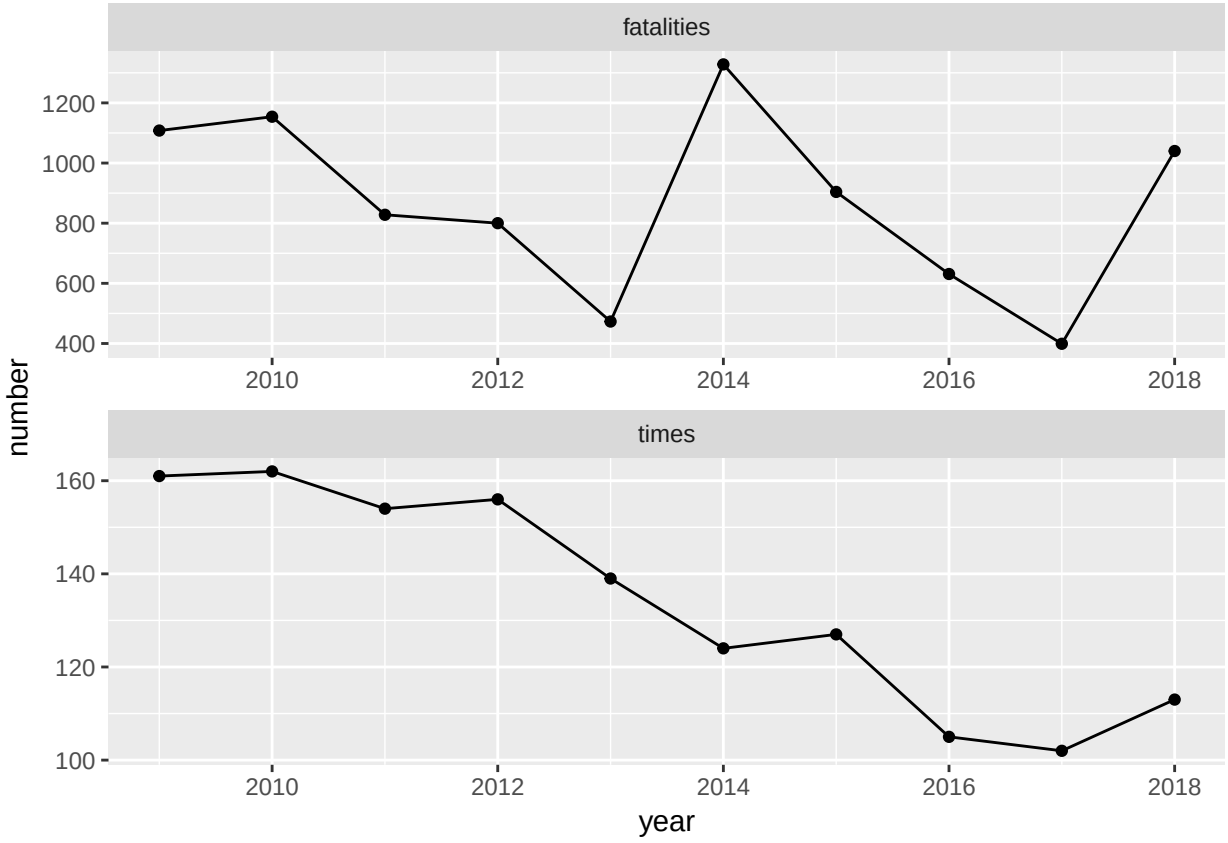

Figure 3: Number of Air Accidents and Fatalities per Year

Figure 4: Number of Commercial Flights and Passengers per Year.

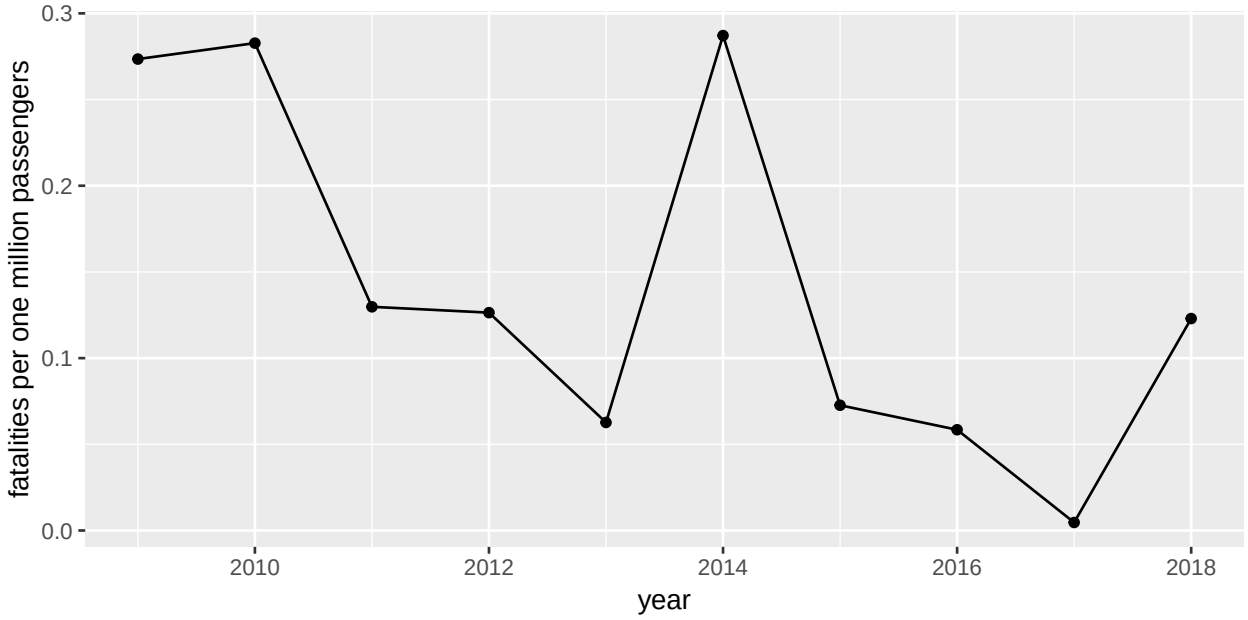

Figure 5: Fatalities per Passengers Boarded per Year.

of fatal air accidents to the total number of flights, because they're normalized and can fairly reflect the true trend of flight safety.

Note that the data of the number of flights and passengers are only for commercial flights. Therefore, only the air accidents with the flight type of "Scheduled Revenue Flight" can be used to calculate those ratios for commercial flights.

4.1 Fatalities per Passengers Boarded

Figure 5 shows a downward trend for the number of victims per one million passengers for commercial flights. 2014 and 2018 are two abnormal years because the number of fatalities do not follow the overall trend for these two years. In 2018, there were two major air crashed that caused more than 300 deaths. One of them is the Boeing 737 MAX 8 that crashed in Jakarta and killed 189 people. In 2014, there were three air crashes that involved Boeing and Airbus. These three major accidents killed 162, 298 and 239, respectively.

4.2 Fatal Accident Rate

Figure 6 indicates that the number of air accidents that caused deaths per one million flights has been decreasing in the past ten years. However, the decreasing fatal accidents rate does not necessarily mean that the number of victims related to air crash also goes down. It's clear that Figure 6 fails to capture the extremely high fatalities in 2014 and 2018. Though the number of crashes was normal in 2014 and 2018, it is because those accidents involved high-capacity commercial airplanes.

Furthermore, Figure 7 indicates the strong correlation between the death rate per one million passengers and the fatal accidents involving Boeing and Airbus, and its correaltion is 0.7206942.

str_match() is used to select the fatal accidents that involved Airbus and Boeing along with **filter()**. Regular expressions used here are **\bBoeing\b** and **\bAirbus\b**. Two variables included in Figure 7 are scaled with respect to their mean and standard deviation, respectively. Since there was no fatal accident related to these two giant manufacturers in 2014, the number of accidents for that year should be set equal to 0 manually, instead of leaving it N/A.

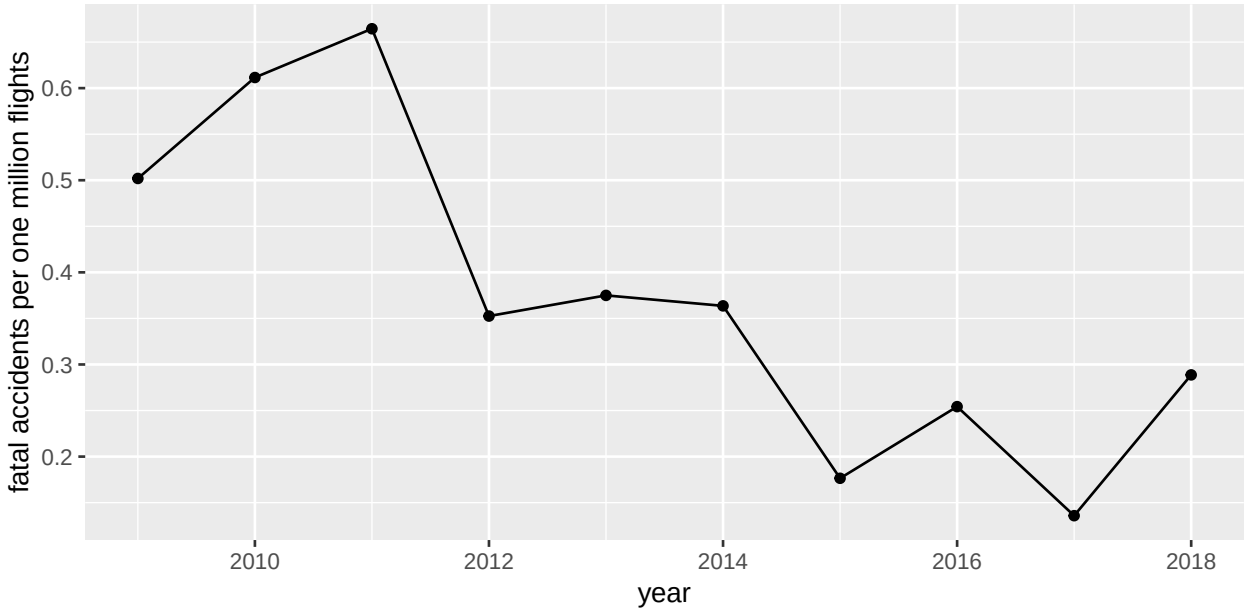

Figure 6: Fatal Accident Rate per Year.

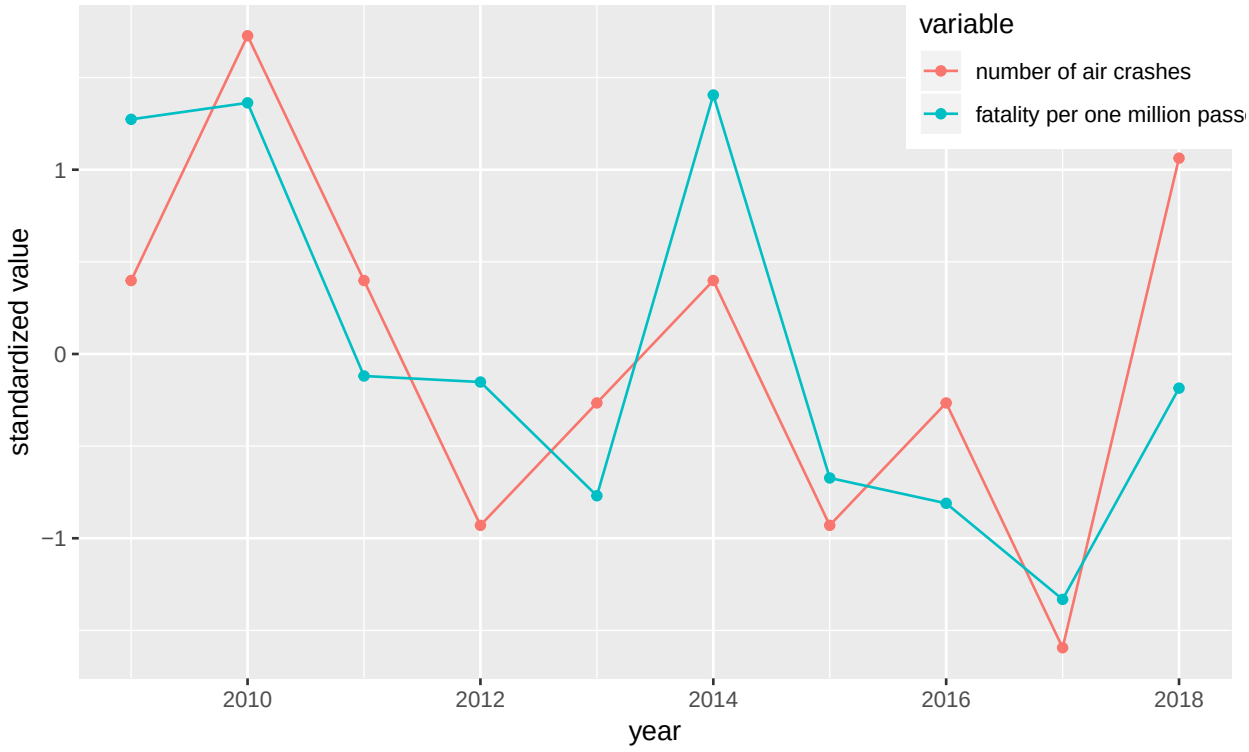

Figure 7: Relation between the Number of Fatal Air Crashes Involving Boeing and Airbus and Fatality per One Million Passengers per Year

Figure 8: Word Cloud for Top 30 Frequent Words in "description"

In conclusion, except for some extreme years, the commerical flights have become safer over the past ten years in terms of fatality rate per passenger. In addition, the rate of fatal accidents keeps going down. However, air crashes that involve high-capacity planes, such as Boeing and Airbus, lead to significant increase in the fatalities.

5 Text Mining

BAAA provides information of the circumstances when the air accident happened and there would be one section called "Probable cause" if the investigation had been done. I'm interested in the key factors that lead to fatal air accidents and try to use text mining techniques to detect them.

Since the investigation of some fatal accidents has not been completed, there is no **Probable cause**, i.e. its value is N/A. Instead of leaving a lot of blanks, **ifelse()** is used to substitute the **Circumstances** for **Probable causes** for those unclosed investigations.

Since commercial flights are more related to our daily life, I decide to investigate the causes of air crashes with **Flight Type** of "Scheduled Revenue Flight". **filter()** helps select those air accidents related to commercial flights.

N-gram model, including n = 1, is the major tool I use in this section.

5.1 Single Word

str_detect() helps select the air accidents related to Boeing and Airbus. The accidents with and without fatalities are included. After using **unnest_tokens()** to tokenize the **desciption** and removing the stop words, I group the records by word and count their frequency. Top 30 words are selected to present in a word cloud (Figure 8).

The bigger the word is, the more often it appears in the **description**. The top 3 largest words, "aircraft", "flight" and "crew", almost contain no information of the accidents because they're commonly used in articles related to plane. The fourth most frequent word, "approach", may indicate that the air accidents occur during the approach, the flight phase shown to have the most air accidents (Figure 1).

Figure 9: Word Cloud for Top 20 Frequent Bigrams in "description"

On the other hand, the words that appear less frequently can be the main reasons for air accidents. "weather" may indicate bad weather, "control" may be part of "loss of control", and "fire" may imply that the there was fire on the plane. These words make more sense than the words mentioned in the last paragraph. However, they can be from different paraphrases and cannot indicate the true causes. So, the complete paraphrases seem more favorable to me. Therefore, I'd like to try bigram and trigram in the following parts.

5.2 Bigram

For bigram and trigram model, I don't want to filter out any phrase that contains some stop words. Though they mean nothing when they're single words, they can make up phrases that provide information. For example, "loss of control" consists "of" and it's a stop word. If I remove all the phrases that contain stop words, I must lose some valuable information. Therefore, I decide remove all the stop words from **description** before converting the **descripiton** to bigram or trigram using **unnest_tokens()**.

First, **unnest_tokens()** is used to split the **description** into single words. Then, **anti_join()** is utilized to remove the stop words. After that, **spread()** along with **unite()** helps recover the description but with no stop word.

Without stop words, we lose the structure of phrases. However, sometime we can recover the original meaning using our knowledge. Though the biggest word, "flight crew", in Figure 9 is uninformative, "weather conditions" and "poor weather" are two phrases with high frequency in the **description** of fatal air accidents, which implies that bad weather may be highly related to air crash. It may not directly cause an air crash, but it can cause problems for the pilots, such as low visibility, and increase the chance that human make mistakes.

5.3 Trigram

Similarly, we can achieve the frequency of trigrams. It is hard to plot a word cloud for trigrams. So, I decide to list some of the important phrases here.

Among the top 20, the phrases that make sense include "poor weather conditions" (1), "controlled flight terrain" (3), "instrument meteorological conditions" (4), "aircraft hit ground" (8), "minimum descent altitude"

(8), “post impact fire” (8) and “auto feather unit” (17). The number in the parentheses is the rank for each trigram.

“poor weather conditions” ranks the first. It is consistent with our results in **Bigram**. “controlled flight terrain” means controlled flight into terrain, or **CFIT**, which occurs “when an airworthy aircraft under the complete control of the pilot is inadvertently flown into terrain, water, or an obstacle”. In addition, “instrument meteorological conditions” indicates low visibility. “post impact fire” means the fire after air crashes. Moreover, “auto feather unit” implies low engine power. These are all factors that can induce fatal air accidents.

For the details, please see: https://www.skybrary.aero/index.php/Main_Page#operational-issues.

5.4 TF-IDF

Table 2: Top 10 Phrases Related to Fatal Air Crashes Given by TF-IDF

phrase	tf_idf	rank
civil aviation authority	0.0004461	1
controlled flight terrain	0.0003824	2
destroyed impact forces	0.0003824	2
impact forces post	0.0003824	2
instrument meteorological conditions	0.0003824	2
findings contributing factors	0.0003187	6
flight control inputs	0.0003187	6
flight idle gate	0.0003187	6
minimum descent altitude	0.0003187	6
post impact fire	0.0003187	6
standard operating procedures	0.0003187	6

TF stands for Term Frequency, and IDF is short for Inverse Document Frequency. Generally speaking, TF increases the weight of the words with more appearance. However, IDF reduce the importance of words that appear in too many documents.

Since I’m interested in the determinant factors for fatal air accidents, I split the records into two document, fatal and non-fatal. Also, I choose to use three-word phrases to calculate TF-IDF. However, TF-IDF gives similar results to **Trigram** (Table 2).

6 Conclusion