COMP7507 Visualization and visual analytics

Visualization of big plane accidents

Project Report

Group 15

WANG, XICHEN 3035561974

FUNG, KAYUE 3035074634

HE, ZHAOXUAN 3035561699

Live demo: https://sprayandpray.github.io/Visualization_Plane/

Contents

1. Foreword

The goal of this project is to identify trends/patterns of big plane crashes, by visualizing data of plane crashes over time and locations. We are looking forward to finding some inspirations by analyzing the number of crashes distributed in different time, location and so on. The death rate of the crash is our point as well. We managed to derive the relationship between the rate and specific factor.

In this report, we would talk about our data firstly, focusing on the preprocessing procedure. Next, we are going to compare the visualization tool we use and the popular ones, and discuss why we choose it. The major of the report is to show our work. We would list the products and dive deeper in the data of each chart. After summarizing the work, we try to show the insights found in the charts and explore the tips in terms of flight crashes. Finally, the contribution of members and the future effort will be given, and the limitation of the project would be analyzed.

2. Data

2.1 Source

Data of plane crashes is collected from the database of “Aviation Safety Network” at https://aviation-safety.net/database/. This database is updated daily to store the most recent plane accidents records since 1901. Necessary information of plane crashes such as time, location, number of fatalities, can be accessed from the database.

One of the major weaknesses of this database is that data is not aggregated into a single file for downloads. As a result, a Google Web Crawler Extension named “Web Scraper” is used to scrap all plane accidents in the website. As it was later found that the scrapping process was inefficient (See discussions in Session 3.1), only the plane accidents in recent 20 years was scrapped. This data will be used throughout this project as the main source of visualization.

Apart from scraping data of plane accidents from the database, data of the causes of plane accidents were also scraped from the same database for visualization. However, the cause data is aggregated from all plane accidents since 1910, instead of the recent 20 years. As a result, the 2 scrapped datasets are not combined, and will be used separately for different visualization topics.

2.2 Features and data processing

Tabular data of the crashes and causes will be transferred and interpreted to explore the patterns of plane crashes. The actual attributes scraped can be retrieved in the following website:
https://aviation-safety.net/database/record.php?id=20190112-0
Pre-processing is conducted to remove irrelevant attributes, refine format of attributes and identify necessary information from strings. The final attributes used can be found in Table 1:

Table 1: Dimension of plane accident data:

Number	Attribute	Description
1	Accident location country	Country which the accidents take place
2	Date of accident	Date when the accident occurs
3	Airline of plane	Airline of plane
4	Crew death	Total number of fatalities of crew members
5	Crew total	Total number of crews onboard
6	Passenger death	Total number of fatalities of passengers
7	Passenger total	Total number of passengers onboard
8	Nature	The purpose of the flight
9	Total airframe hours	Total flight hours of the plane

Regarding the cause data, the aggregated frequency of cause was directly scrapped from the Aviation database. The data only contains the cause and its frequencies.

3.Tools we use

3.1 Web Scraper

WebScrapper.io, a Google Web Scrapper Extension, was used to scrape data from the Aviation Safety Network Database. The extension provides an easy interface for users to select the necessary information by clicking and highlighting the web data, reducing the needs to write long programs for scraping data.

The major weaknesses of this scrapper is the speed for scraping data. The time for scrapping one data record is at least 2 seconds. During the scrapping, any error will terminate the scrapping and all previously scrapped result can be lost. However, if all plane accident data is scrapped, it will take more than 6 hours to scrap all data. Therefore, only accidents in recent 20 years was scrapped.

3.2 Tableau public server

During the initial stage of the project, both D3 and Tableau were explored to identify the best tools for visualization. Eventually, Tableau was chosen as the main tool due to the following reasons:

Advantages of Tableau over D3:

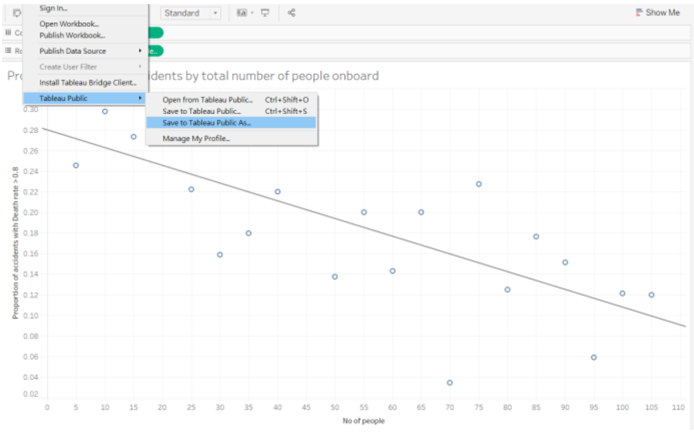
1) Much flatter learning curve. Once Tableau is installed, graphs can be created by simple drag-and-drop operations.
2) Users can define new columns in the interface without repeatedly processing the data.
3) Contains most of the necessary build-in templates which can be directly used without the need to study how to configure the parameters.
4) Graphs can be loaded faster as they are run in the server.
5) Support a wide variety of data input format, such as csv, SQL database.

Disadvantages of Tableau over D3:

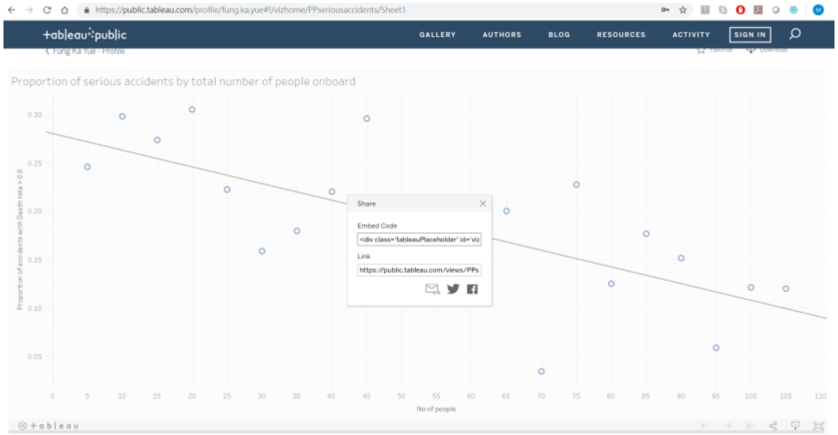
1) Highly interactive graphs such as Sunburst are not available.
2) Limited customizability. Some operations may not be realizable, e.g. Stacking multiple graphs of different types into a single worksheet.
3) Graphs must be pushed again into Tableau Public Server after each changes.

Once a graph has been created in Tableau, it can be pushed into the Tableau Public Server for generating an embedded JavaScript code:

Saving the graph to Tableau Public:



Generating JavaScript code in Tableau Public:


https://public.tableau.com/profile/fung.ka.yue#!/vizhome/PPseriousaccidents/Sheet1

Share

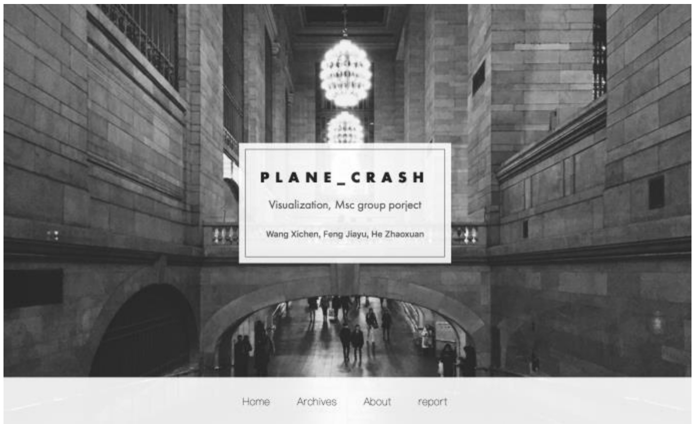
Embed Code: <div class='tableauPlaceholder' id='viz

Link: https://public.tableau.com/views/PPs

The graph will be automatically loaded from the Tableau Server once this code is embedded into the source codes.

3.3 Website developer

Node.js brings incredible generating speed. Hundreds of files take only seconds to build. Hexo is a fast, simple & powerful blog framework, which helps us input command to deploy website to GitHub Pages. We also use Hexo-hiker as our template framework. D3.js is open-source and has a large user community. Visualizations are fully customizable. Possible to create highly interactive displays of data.



PLANE_CRASH

Visualization, Msc group porject

Wang Xichen, Feng Jiayu, He Zhaoxuan

4. Design Features and Rationale of our Design

At the very beginning of our project, we found a website that supplied the data of flight crash. We were pretty interested in this topic and came up with a lot of questions to the data and we would like to explore more information about it, so we chose it as the topic of our project. With the advice of TA, we raised some questions:

(1) How did the accident number change in these years?
(2) Where did the planes crash? Could we find some feature of location?
(3) Are there some factors directly influencing the accident amount? What are they?
(4) Why did the aircrafts crash? Could we find some pattern of the reason?
(5) In what extent, the passenger could survive from the crashed flights?
(6) If there is difference between the survive rate of crashing in ocean and land?
(7) Could we find some dangerous flight route or flight company?

According to the questions, we divided our visualization work into four part: Glimpse of Distribution of Plane Crash, Further Analysis of Accidents, The Correlation of Factors with Death Rate, Accident Cause Analysis.

The first part focuses on the temporal and spatial feature of the crashes. We set three graphs here. The first graph shows the crash amount in 20 years in a world map, which could show the location distribution in a intuitive way. But in the other hand, the drawback of the spatial map is that it is hard to distinguish the non-outstanding ones. Therefore, we add a treemap here to illustrate the data in detail. In addition, we create another spatial map with an interactive time axis to show the data of each year. We handle the question (1) and (2) here.

The second part introduces the factors directly connected to the accident amount and the number of deaths. We would discuss on the relationship between crash amount and the death number here. We handle the question (3) and (7) here.

The third part concentrates on the death rate. We had assumed many reasonable factor and chose three of them based on the test results. We set scatter diagrams and heat diagram here because they are the most proper ones to fit a regression and find the co-relationship. We handle the question (5) here.

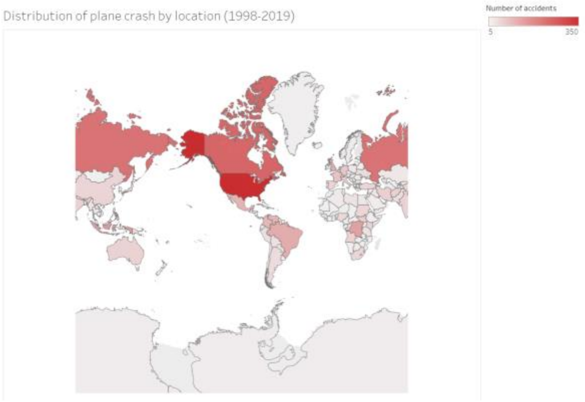
The final part tries to figure out what caused the crashes and in what extent we could avoid it? We classified the causes into three layers and managed to use a hierarchical sunburst diagram to show the causes more systematically. We handle the question (4) here.

5. Visualization

5.1 Glimpse of Distribution of Plane Crash

I. Crash distribution by Location (map)

Graph 1: Crash distribution by Location

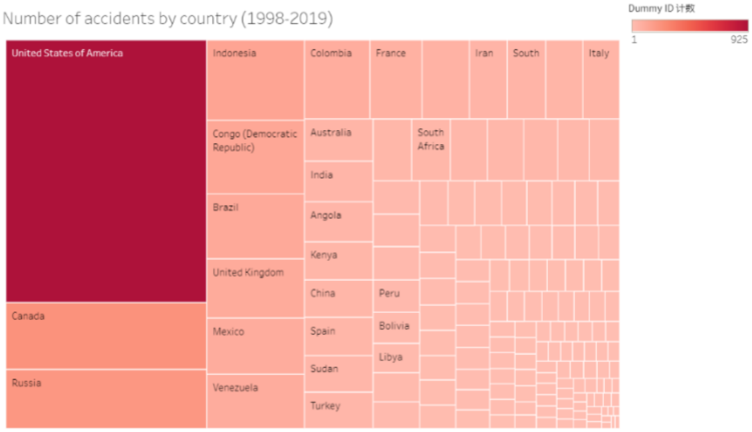


The graph 1 illustrates the distribution of plane accidents across locations over the recent 20 years. Higher color intensity in the graph represents higher frequency of accidents. The most eye-catching part of the map is the United States in deep red, considering that the number of crashes located in the United States were much higher than that of other countries. One of the possible reasons is that the United States had the largest throughput within the past 20 years, and the big base number increased the crash amount.

Following the United States (925 crashes), Canada and Russia revealed 234 and 210 crashes respectively. Apart from these, four countries, Congo, The United Kingdom, Brazil, Indonesia reported over 100 crashes throughout the 20 years. It should also be noted that China is not in the top list in the number of plane crashes, even though it is the second largest economy in the world.

II. Crash distribution by Location (tree map)

Graph 2: Crash distribution of Location

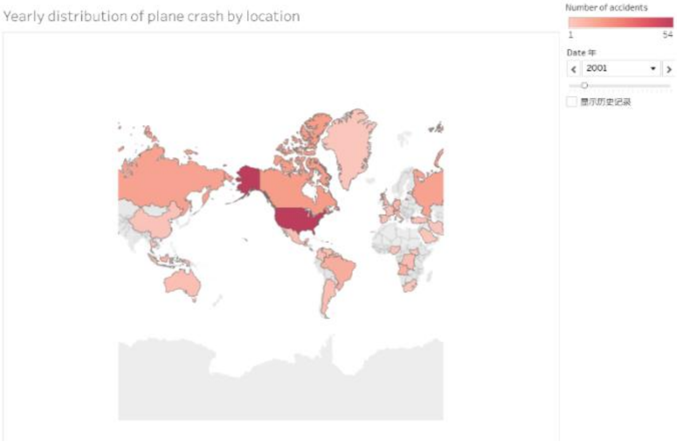


This is the Tree map version of the previous graph. We could observe the gap in number of crashes between United States of America and other countries more clearly. The number of accidents in the United States alone is about that of the next five countries combined.

Some of the country names are not displayed in the tree map due to limited spaces inside the squares. Nonetheless, users can still identify the countries by moving the cursor to the squares. By removing the names, we can also avoid stacking the names in small spaces for clearer visualization.

III. Crash distribution of Location in each year

Graph 3: Crash distribution of Location in each year



Graph 3 is the crash distribution in each year with an interactive time axis. Interested users can find our Project webpage and choose the year to check the data more precisely. After looking at the crash data in every year, we could find that the United States is always the country with the most number of crashes while the records of other countries fluctuated in a reasonable range. (~1-10) It can also be seen that China still has a small number of plane accidents each year.

IV. Methodology tried and questions

In graph 1, we set the threshold 5-300 to limit the color performance. Because the number of America is over 900, nearly the sum of other countries, it is hard to distinguish them. And all we could see is the red America and a grey map without a max threshold. In the other hand, it would be a light pink map without the min threshold.

We had tried to use two color in the amount axis, which is the max number of color used in map in Tableau, and it worked poorly because the middle part between two colors is white. Most countries would be printed in white if we used two-color attribute.

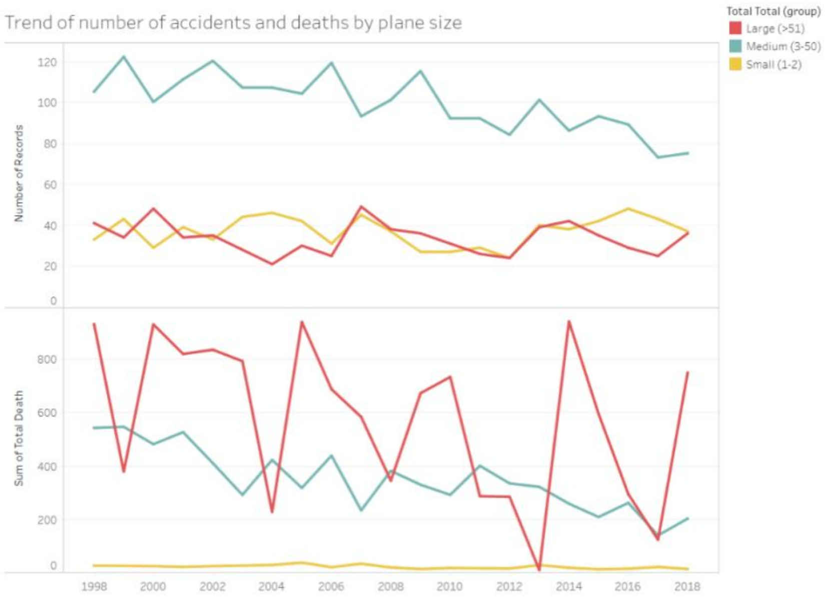
V. Insights

We could derive the objective data from the distribution graph and learn about the outline of plane crash. We could not only find the increase of accidents but imply that the number of flight route and aircraft is increasing as well. We could also get some information about the economic development situation of some countries. The data we gain here also help with the following work.

5.2 Further Analysis of Accidents

I. Relationship of number of accidents and deaths by plane sizes

Graph 4: Trend of number of plane accidents and deaths by plane size



Graph 4 is a combination of two line graphs which displays the number of accidents and deaths of each plane sizes each year. These graphs used to study: 1) Whether larger planes are more susceptible to accidents, 2) Whether accidents of larger planes are more catastrophic (more deaths) compared to smaller planes.

Plane size is defined by the number of people onboard (crew + passengers). A total 3 plane sizes (small, medium, large) is created and represented by each color in the graph. The definitions and colors of each plane sizes are as follows:

1) Larger planes (Red) -- Number of people onboard > 50
2) Medium planes (Blue) -- 50 >= Number of people onboard > 2
3) Small planes (Yellow) -- Number of people onboard = 1 or 2

The first graph shows that medium plane has the largest number of accidents among all plane sizes. (~100) The number of accidents of large and small planes are similar and fluctuates each year. (~30) Moreover, the graph also shows that the number of accidents of medium planes are steadily decreasing, while that of the large and small planes stay the same. This suggests that the total number of accidents is slowly decreasing every year.

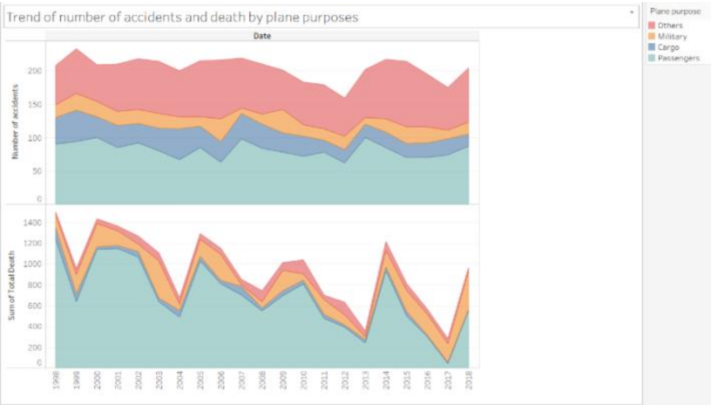
When the total number of deaths is concerned, it is found that most fatalities come from accidents of large or medium planes. As shown in the second graph, the number of deaths from accidents of small plane accounts for only a small proportion of total deaths each year. The trend is also stable for accidents of small planes.

Although the number of deaths from accidents of large planes can be significantly higher than that of medium and small planes, the variance of the number of deaths is much higher than that of medium and small planes. For instances, in 2013-2015, the number of fatalities from large planes is even smaller than that of small planes in 2013. However, the number surges in 2014 again to >800 fatalities. Given that the number of accidents of large planes are relatively stable, the fluctuation of total number of deaths might suggest that accidents of large planes have high fluctuations on death rate.

II. Relationship between number of accidents and deaths by flight nature

Graph 5: Trend of number of accidents and deaths by purpose of flight

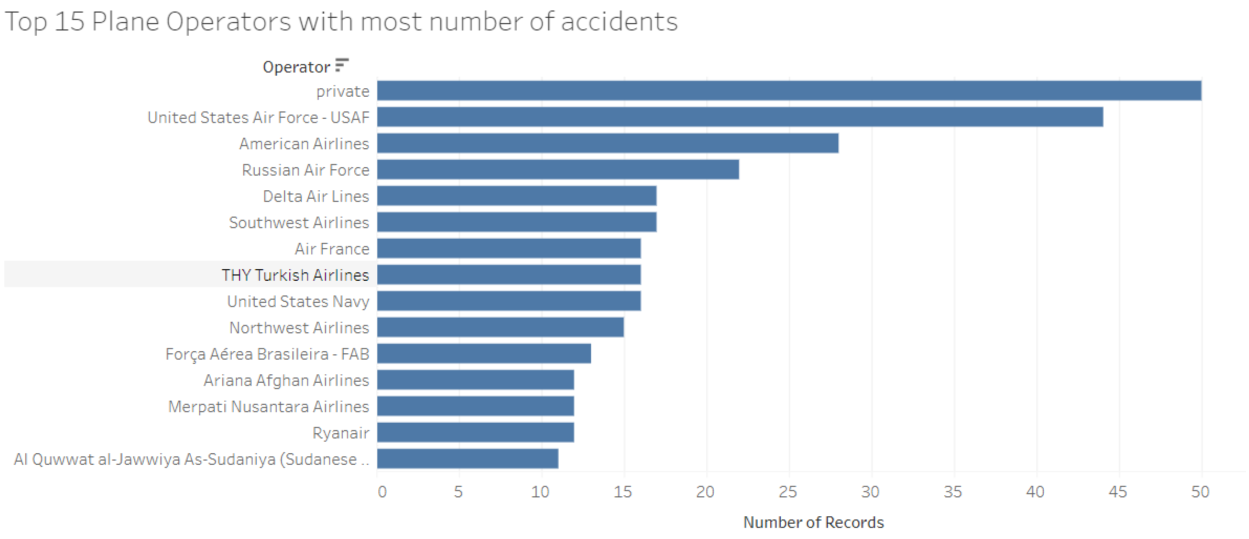
Similar to the previous graph, this graph illustrates the relationship between the number of accidents and deaths over different flight natures each year. 4 common purposes (Passengers, Military, Cargo, Others) are selected for visualization.



The first graph shows that the number of accidents by different flight purposes. It is found that passenger planes account for ~50% of all flights. The trend of the proportion remains stable throughout the years.

While “Others” and “Passengers” have similar proportions of accidents each year, the trend of total number of deaths differ significantly from that of the number of accidents. A large proportion of deaths comes from passenger planes. While “Military” contains the least number of accidents over the years, it has the second highest proportion of number of deaths. These observations may suggest that passenger and military planes are more susceptible to deaths than any other plane natures.

III. Plane operators with the greatest number of accidents



Graph 6: Trend of number of accidents and deaths by purpose of flight

This graph demonstrates the top 15 plane operators that have the highest number of accidents in recent 20 years. Among the 15 plane operators, “private” has the highest number of accidents. This is intuitive because all private planes are grouped into the airline “private. It is also found that 6 operators (~40%) comes from the United States. This aligns with our earlier observations that the United States have the highest number of accidents every year.

IV. Methods we tried and questions

Graph 4: Trend of number of plane accidents and deaths by plane size:

Graph 4 was originally replaced by “number of passengers” instead of plane sizes. The graph is later found to be inappropriate for visualization, because the number of accidents and deaths starts to fluctuate between 0 as number of passengers increases. This is due to the large range of the number of passengers in large planes. The number of passengers can range from 50 – a few hundred in ordinary planes. There can be no records for some numbers of passengers, which leads to 0 accidents and deaths. However, this visualization is not ideal, because the fluctuation from 0 to a number provides no insight on the relationship between number of accidents and number of passengers.

By grouping passenger numbers into small, medium and large plane sizes, the number of accidents for each group increases. The graph provides more useful information for the relationship between passenger capacity and plane size.

Graph 6: Trend of number of accidents and deaths by purpose of flight

The variable “Operators” is sparse and contains more than 1000 distinct values in the dataset. However, one may be more interested in the airlines which has more accidents. Only the top 15 operators are selected to remove operators which has only a few accidents.

V. Insights

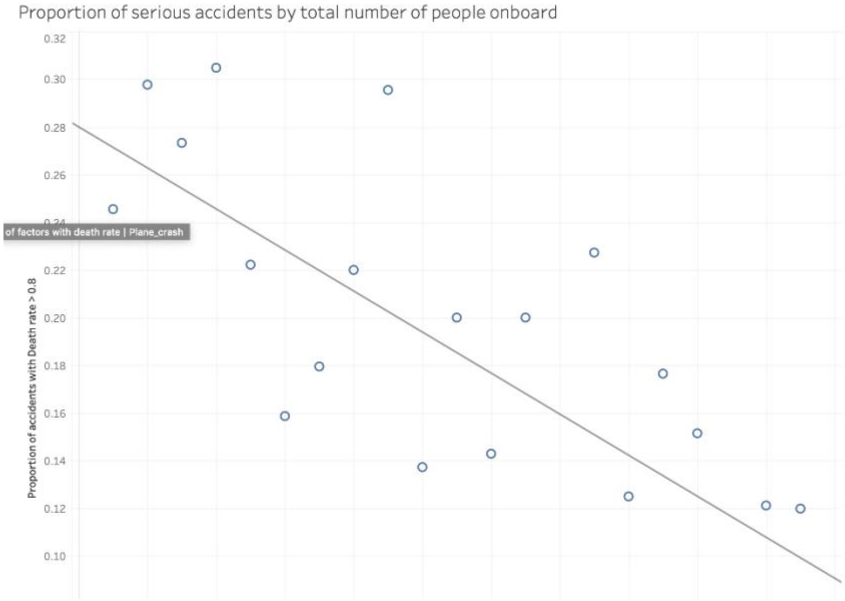
From Graph 4, it is found that most deaths come from middle and large planes. It is also found from Graph 6 that most deaths come from Passengers and Military flights. To reduce the number of deaths in plane accidents every year, one should focus on reducing the number of deaths in large and middle-sized passengers’ planes.

The significant proportion of deaths over all deaths in large planes support the intuition that accidents in larger planes tend to be more catastrophic. While the number of middle-sized planes is the highest, the total number of deaths is comparable to that of large planes. The high variance of total deaths in large planes further suggest that deaths in large plane accidents can be affected by a few catastrophic accidents. Therefore, it is essential to prevent accidents in large planes to reduce number of deaths.

5.3 The Correlation of Factors with Death Rate

I. Proportion of serious accidents by total number of people onboard

Graph 7: Proportion of serious accidents by total people onboard

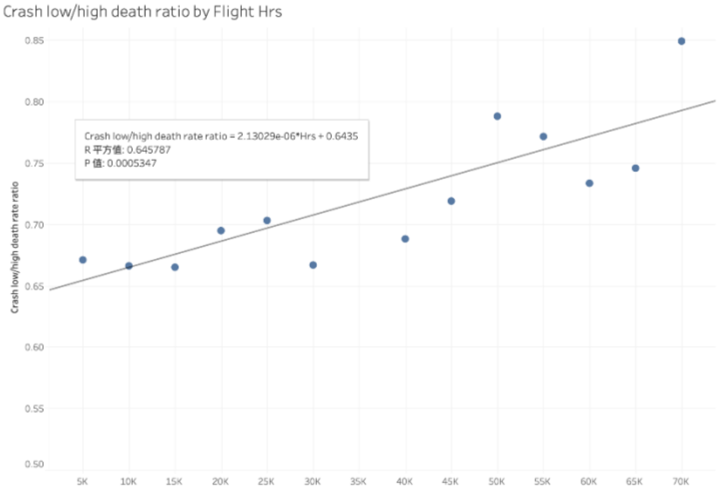


The graph below illustrates the relationship between the severity of plane accidents and number of people onboard. X-axis and Y-axis contain the number of passengers and severity respectively. Severity of each accidents is defined by its mortality rate. A mortality rate of >80% is considered as catastrophic.

It is suggested that proportion of catastrophic accidents decreases as the number of people onboard increase. Large planes tend to have fewer accidents of large mortality rate. One of the possible reasons for this observation is that catastrophic accident is defined by proportion of mortality rate. As small planes have fewer passengers, accidents in small planes are more likely to be classified as catastrophic.

II. Ratio of low death accidents / high death accidents by flight hours

Graph 8: Crash low/high death ratio by Flight Hours

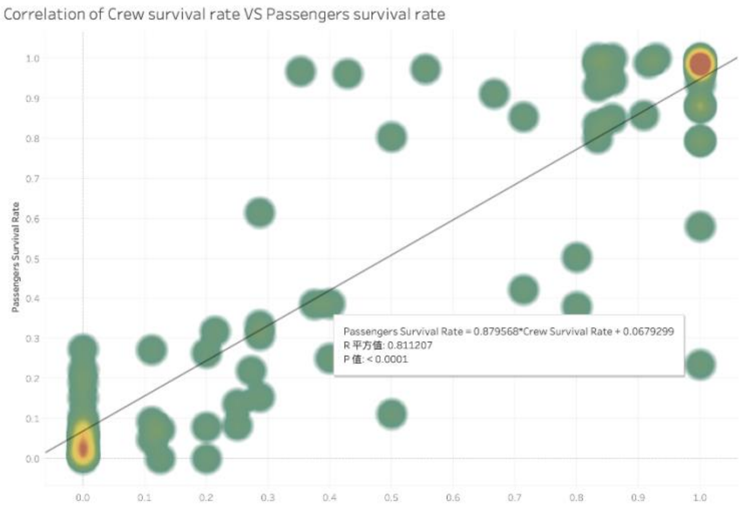


The graph shows the relationship of flying hours with the occurrence of catastrophic accidents. Y-axis represents the ratio of number of non-catastrophic and catastrophic accidents. (Ratio of number of accidents with small mortality rate and high mortality rate)

Plane with more flying hours tend to have fewer catastrophic accidents than new planes. It can be suggested from the finding that planes with more flying hours tend to be more stable. New planes can be more susceptible to errors because errors may not be found within a short flight time.

III. Correlation of Crew survival rate VS Passengers survival rate

Graph 9: correlation analysis



Correlation of crew and passenger survival rate is analyzed with the following graph. Survival rate, which is defined by number of fatalities / total number of people onboard, is calculated for each accident. Crew and passenger survival rate are then plotted to analyze their correlations.

There is a positive correlation between crew and passenger survival rate. This suggests that passengers are less likely to survive in an accident if most crews cannot survive. The key message is that survivability of crew is important to that of the passengers. To reduce the number of fatalities of passengers, it is essential to increase the survivability of crews.

IV. Methods tried and questions

Scatter plots was originally used instead of density plots. However, density plot is eventually used because it is found that most accidents cluster around (0,0) and (1,1). It is difficult to visualize individual points if most points lie on the same spots in the graph. By replacing individual points by density, intensity of colors can better reflect the above observations.

Moreover, a filter constraining on the minimum number of crews and passengers is set to improve the calculation of crew and passenger survival rate. For instance, if there is only one passenger in the plane, the survival rate will be either 0 or 1, increasing the variance of the points. The filter can better generalize the survival rate and provide a better fit.

V. Insights

Contrary to the belief that planes which have higher number of flight hours are more susceptible to accidents (due to old engines and gears), it is found that the opposite is true. Planes with more flight hours tend to be more stable. One of the possible reasons maybe that planes with lower flight hours are more susceptible to potential errors of new architectures.

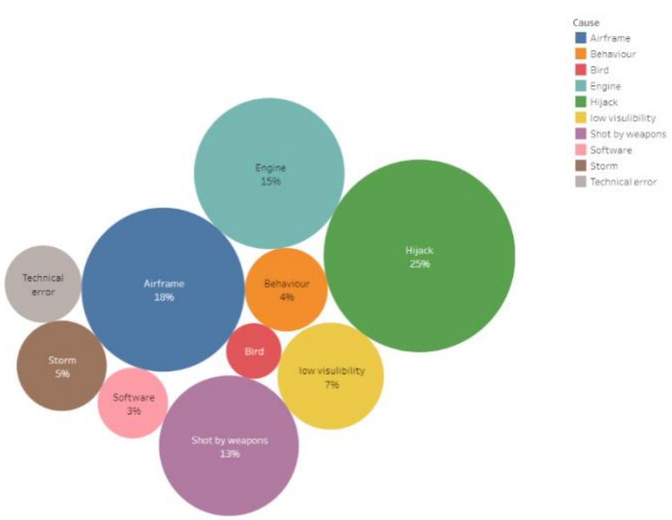
Moreover, while it is found in Session 5.2 that larger planes tend to have more deaths in an accident, Graph 7 suggests that larger planes tend to have fewer catastrophic events. This suggests that catastrophic events in large planes can be rarer, but accidents can be much more serious if happened.

To understand which directions, we should take in reducing deaths, the next session analyzes the frequency of causes of accidents.

5.4 Accident Cause Analysis

I. Causes of crashes graph

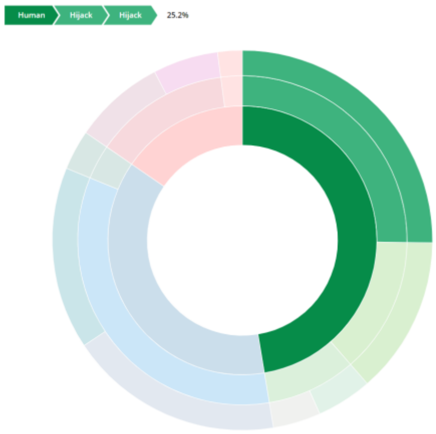
Graph 10: Causes of crashes graph



The diagram mainly shows the popular causes of the crashes. It could be seen from the diagram that Hijack is the most frequent reason which accounts for 25% of the total. Airframe gear broken, Engine and Shot by weapons contribute 18%, 15% and 13% respectively.

II. Causes of crashes sunburst diagram

Graph 11: Causes of crashes sunburst diagram



This is the Sunburst Graph of the cause of the accident. It is hierarchical. We summarize the three main subjects, then divide it and sort the collected causes into secondary and tertiary tags. The size of the roulette represents the proportion of the same level. Move the pointer to see the data and scale represented by each tile clearly.

III. Method we had tried and questions

The data we crawled from the website was in mass, so we had to preprocess it by Python and format it to suit the requirement of D3 and classify the genre by manual work. We tried to create a sunburst diagram by Tableau at first. After 4-hour trying, it failed because of the wrong parameter setting, it could only generate a ellipse rather than a circle. Therefore, we managed it by D3.

IV. Insights

As Hijacking is the most likely causes for plane accidents, it is suggested that precautions should be done to reduce the number of accidents. Given that the large passengers plane has the greatest number of deaths in all accidents (as suggested in Session 5.2), to reduce the number of deaths, more safety measures should be conducted on large passengers' plane to reduce number of deaths.

Contrary to the belief that weather can be a serious threat to planes, it is found that weather only constitutes a small portion of accidents among all accidents. Most of the accidents occur due to human errors or malfunctions of airplane gears. As suggested by the graphs, 15% of accidents are related to malfunctions of engines. Therefore, one observation is that more reliable engines in large airplanes are very detrimental in reducing the number of deaths in plane accidents.

6.Limitation and Future Work

This visualization is probably a baby step - more of a proof of concept than a fully-fledged product. We believe most users who want to know more about airplane crashes would like benefit from having visual aids when scanning the data, creating demand for visualization tools like ours. This is a huge motivation for us to keep improving.

6.1 Knowledge Limitations

Historically, many visualization frameworks became fashionable only to go out of favour as time passed. In the same manner, no single visualization would appeal to all of the diverse user preferences. For instance, technical analysis is highly recommended by some researchers, while at The same time criticized by others who would favour data analysis.

In our opinion, some airplane company have making a lot of research to avoid accidents due to respective reasons. Our tool has demonstrated that certain analysis ideas can be visualized in intuitive ways; there is a long road ahead to mine raw static data into useful visual tools.

6.2 Data Limitations

Accurate description and fundamental data are both hard to obtain. Due to the nature of the aircraft accident itself, people can hardly get the full information about the accident, some inferred reasons may deviate from reality. For example, some accidents are inexplicably lost and cannot be traced back to the cause. This situation often occurs before 1980s, and that’s why we choose the data from 1989-2019.

What’s more, we couldn’t collect the data about the total amount of flights in each year. So we could only find some trend about the accident amount, not the accident rate.

6.3 Technical Limitations

Due to tight time constraints and poor experience in front-end, we only managed the data to use a part function of tableau. Plans for adding more user interactions were also scraped. We should add more animations to guide users to more quickly browse the expressions they want. Priorities would be on developing a fix for these issues, in Addition to fine-tuning the visual appearances of the chart.

6.4 Future Work

During the initial planning stage, we spent a significant amount of time discussing an important interaction element, of which we did not have sufficient time to implement - Animation shows changes in accidents for different reasons each year, such that the user can select multiple stocks and industries and compare their total returns. This interesting feature is certainly high up in the list of future work.

7.Contribution

	Wang, Xichen	Fung, Ka Yue	He, Zhaoxuan
Proposal	✓	✓	✓
Data collection		✓	
Data processing		✓	✓
Website	✓		
Table adaptation	✓		✓
Visualization	✓	✓	✓
Table analysis	✓	✓	
Explanation		✓	✓
Report	✓	✓	✓
Presentation	✓	✓	✓

8.Reference

1.使用密度标记进行构建（热图）https://onlinehelp.tableau.com/current/pro/desktop/zh-cn/buildexamples_density.htm

2.A fast, simple & powerful blog framework

https://hexo.io

3.Accessing the Web Server in Tableau Server

https://community.tableau.com/docs/DOC-6141

4. Apache

https://wiki.apache.org/httpd/FAQ

5.hexo + github pages 搭建博客样式

https://blog.csdn.net/sinat_37781304/article/details/82729029

6.Refused to execute script from 'URL' because its MIME type.

https://blog.csdn.net/csdn_tingou/article/details/82852217

7.如何用 github 来展示你的前端页面

https://www.jianshu.com/p/85a479eab55c

8.jQuery 教程 www.runoob.com/jquery/jquery-tutorial.html

9.设定站点建立时间 EditNew Page https://github.com/iissnan/hexo-theme-next/wiki/设定站点建立时间