Data Visualization
Project - Air Crashes since 1908
Process Book

Bhargav Sai Gajula	**bgajula@clemson.edu**	**C26845700**
Shreyash Dhumale	**ssdhuma@clemson.edu**	**C47391387**

Project Link - https://github.com/bhargavsai2/Air_Crashes_Visualization

Website Link - https://bhargavsai2.github.io/Air_Crashes_Visualization/

CONTENTS

Overview

The project's goals are twofold: first, to increase public awareness about the safety of air travel and encourage informed debates on this important subject; and second, to give aviation experts, regulators, and researchers actionable insights, ultimately resulting in safer skies for everyone.

Motivation

The project's motivation stems from a strong conviction in the ability of data to effect good change. Our world has changed as a result of air travel, which has brought people closer together and created previously unheard-of opportunities for trade, travel, and cultural interaction. However, this incredible development was not without risk, and air travel safety is still a major worry.

We are driven by a common vision of safer skies and a society where flying is not only practical but also incredibly safe. We are motivated to take on this project by the innumerable lives that are touched by plane catastrophes, the tenacity of survivors, and the commitment of aviation experts.

We are motivated by the notion that, via data narrative and visualization, we can contribute to a safer, more knowledgeable aviation industry, where each flight serves as a monument to the ability of data to save lives and pave the way for a better future for all.

Related Work

We have taken references from a few projects, one of them being this image. The dashboard here, though complex, gave us an idea as to what components are important to be conveyed in a data visualization project. Another couple of blogs had a few ideas of the data they used to get to their respective visualizations like page 1 and page 2.

Questions

1) **How does the trend of air crashes look over the period of the dataset?**
   To observe the trend of air crashes throughout the period of the dataset i.e from 1908 to 2023, initially we thought of using a time series line graph to show the total crashes occurring using a 5 year interval in the x-axis. Eventually it occurred to us that not the crashes but the deaths/survivals of individuals would give a way better insight about the crashes and would in fact be the best way to go ahead with.

2) **How do various countries compare in terms of air safety?**
   We had envisioned observing the crashes happening around the world categorized by countries in a world map. It uses a world map as the base, with countries marked in different shades of color. The intensity of the color represents the frequency of crashes in each country. This design visually communicates which countries have had more air crashes over the years, allowing viewers to quickly grasp geographical patterns in air crash occurrences.

3) **Which aircrafts are the most dangerous?**
   According to the crashes of an aircraft model, we wanted to visualize a treemap of the aircraft model which seemed to have been in the most crashes. The visualization gives a clear, comparative picture of the most crash-prone aircraft models, thereby making it simple to determine which models have been involved in the most occurrences.

Dataset

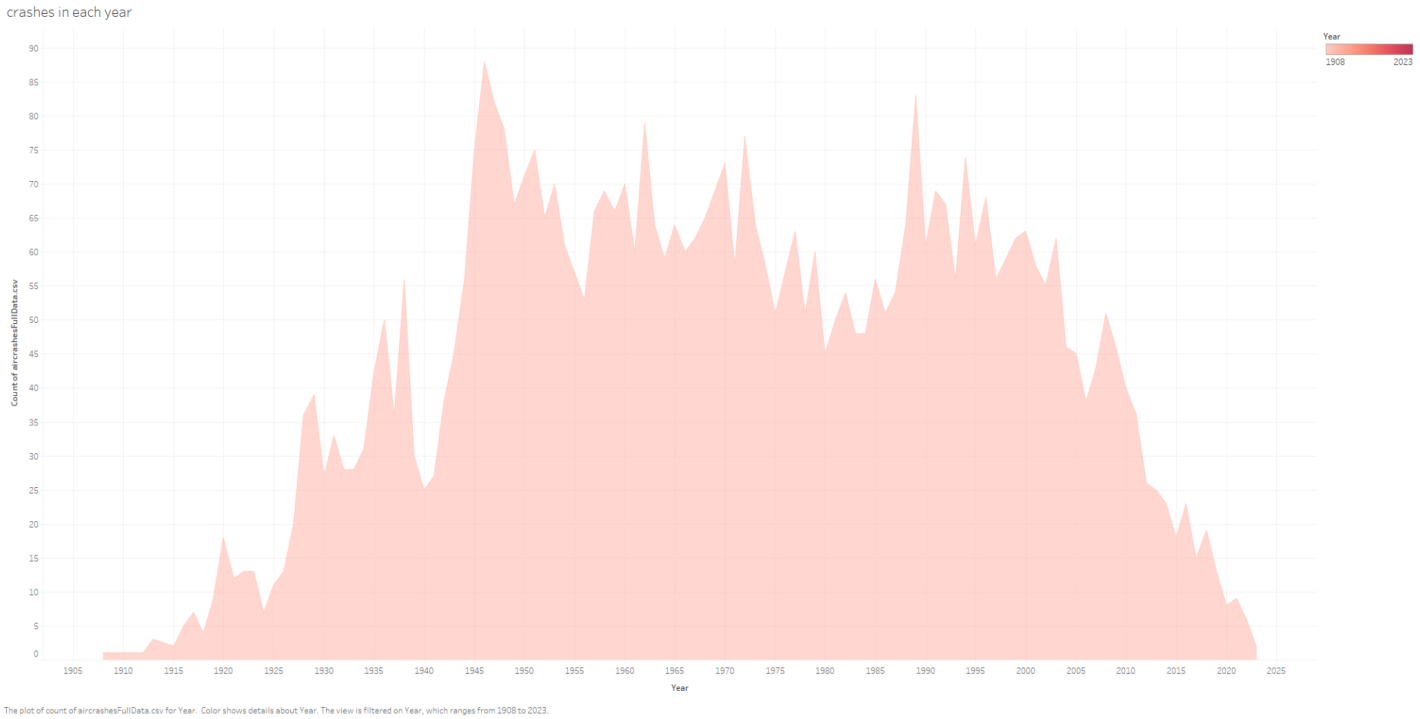
While searching for data for the aircraft crashes, we landed on this dataset on kaggle having records of air crashes from 1908 to 2023. This huge dataset gave us enough data to play around and decide on which visualizations would answer our questions giving desired results.

With a dataset of this magnitude, comes null-values, negative values, irregular values which contribute nothing to the visualization's importance. This is where we scraped off negative and null values so that they don't affect the visualizations. We removed several records having regions rather than countries as their geographical values, which wouldn't help plot the crashes in the World map.

Exploratory Data Analysis

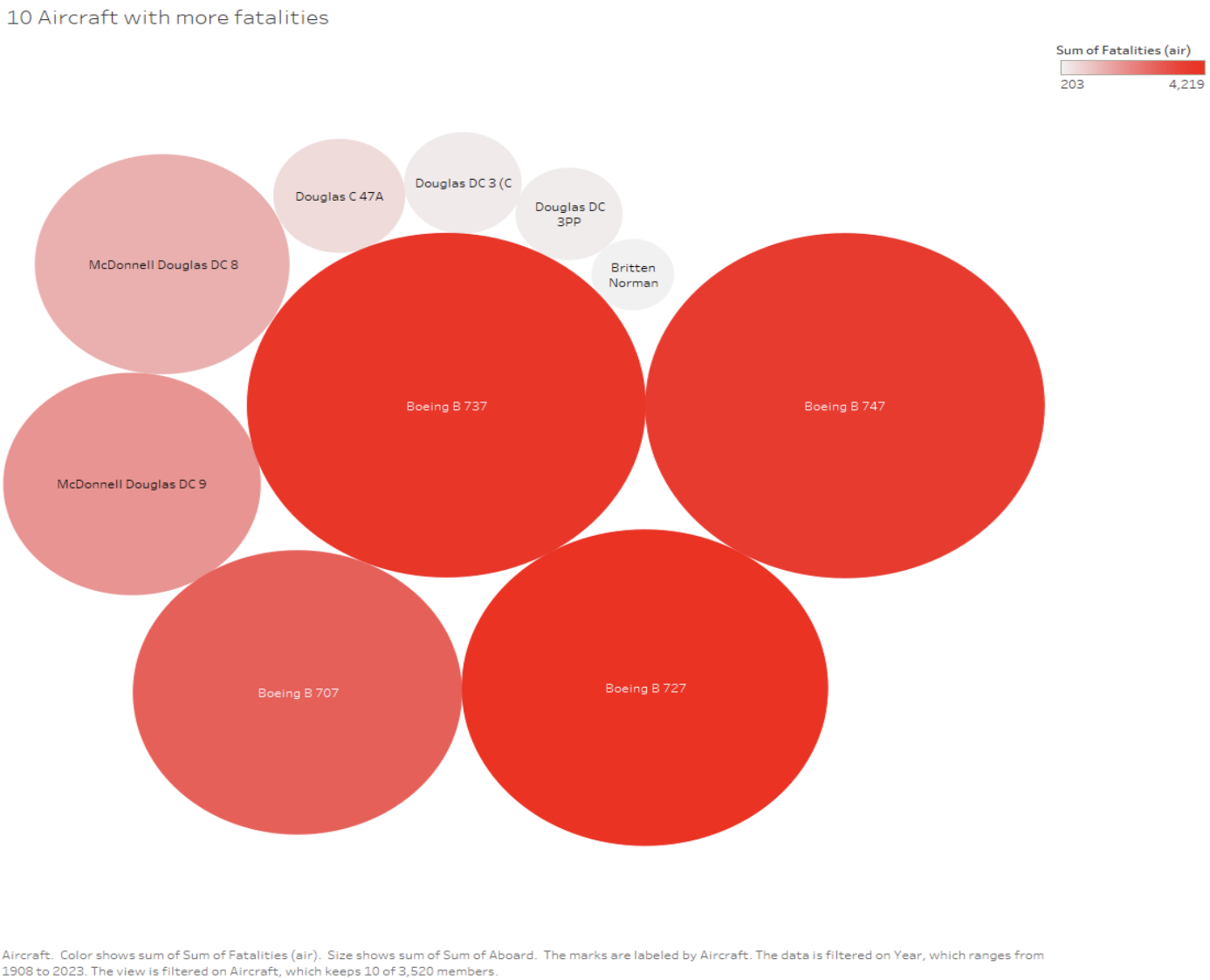
As mentioned above we first thought of using a time series line graph to observe the trend of crashes throughout the period of this dataset. While working with the line graph and taking insights from the professor, we concluded that just showing crashes with respect to time does not give strong insights about the data. Rather, if we show the people who survived and the people who died as a comparative bar graph, it will be more insightful for the viewers.

Below are the initial visualizations we used -



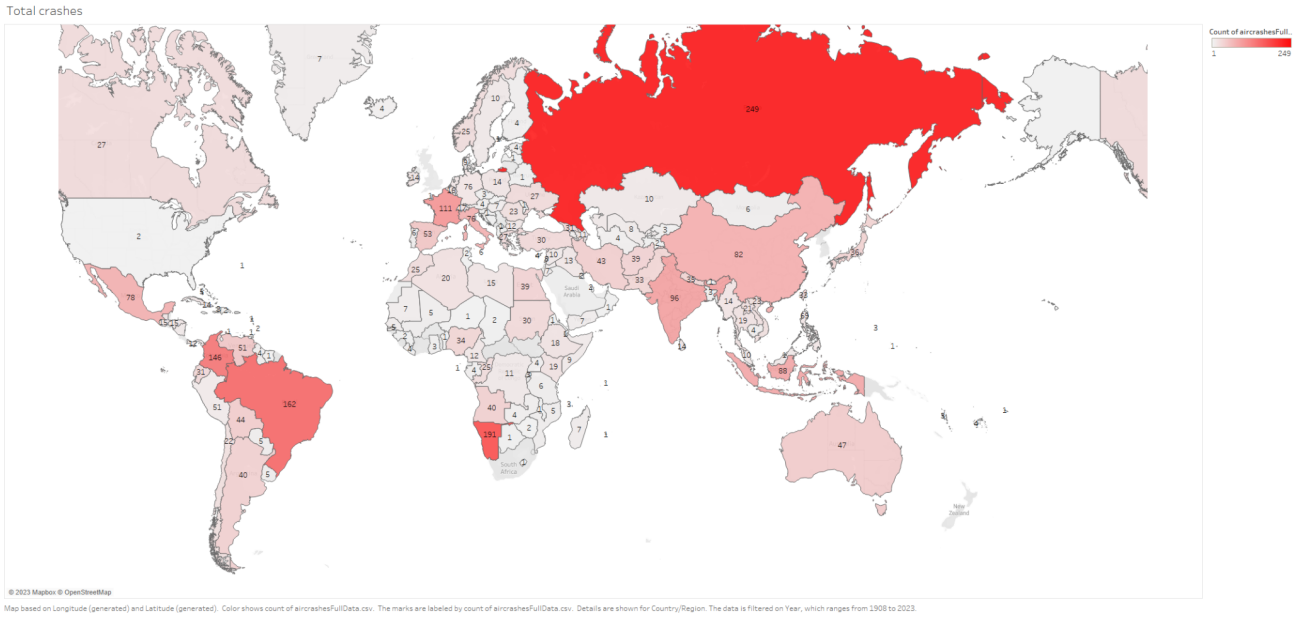
Visualization 1

The time series line graph will display the crashes per year and help us infer the trend of the crashes throughout the years.



Visualization 2

A Bubble graph displaying the 10 aircrafts with the most casualties in the dataset.



Visualization 3

A Geospatial map showing us the geographical distribution of crashes, having also incorporated heat maps to highlight the area with the higher density of crashes.

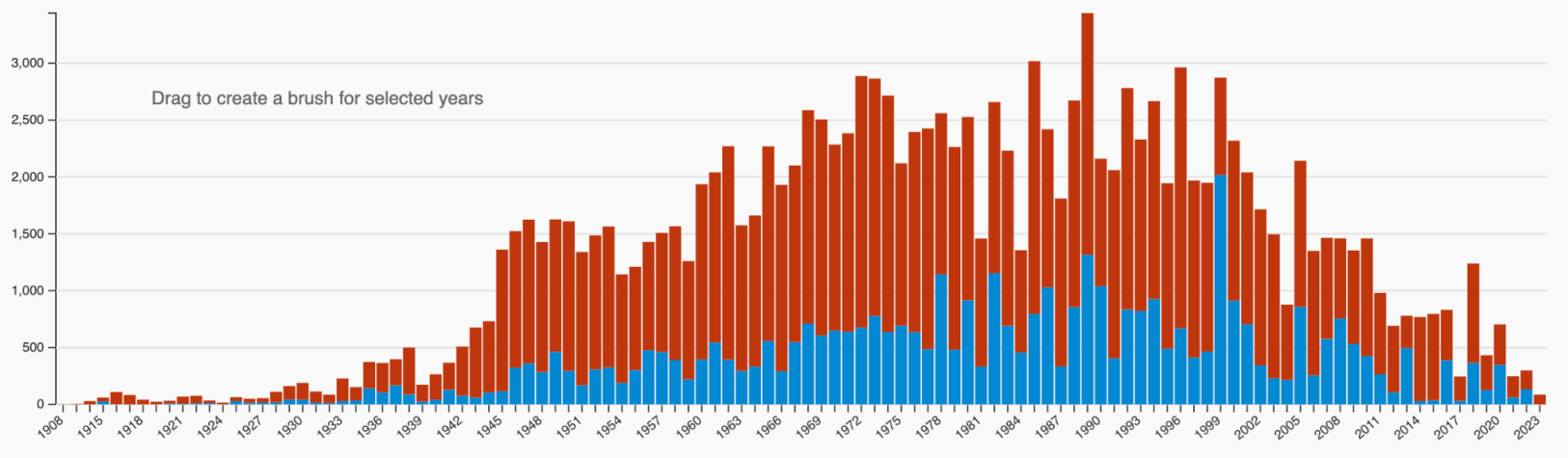
Design Evolution

As we constantly kept taking feedback and inputs from the professor and working on making the final website a fully informational and conventional dashboard. A lot of initial ideas were changed into something more insightful. The first initial line graph was rather vague considering what we finalized now, having a double stacked bar chart showing deaths/survivals which gives more context about the crashes. Secondly, a donut chart was added showing the total deaths vs survival comparison of the whole data. Later it was made interactive, by only showing the period of years selected by the brush that the viewer would make to view the selected period of years. It was also decided to make the donut chart **“on-demand”** so that the written information only appears when the mouse is hovered over the graph. All the four visualizations were made interactive to the brush that the viewer would make to view their desired period of time.

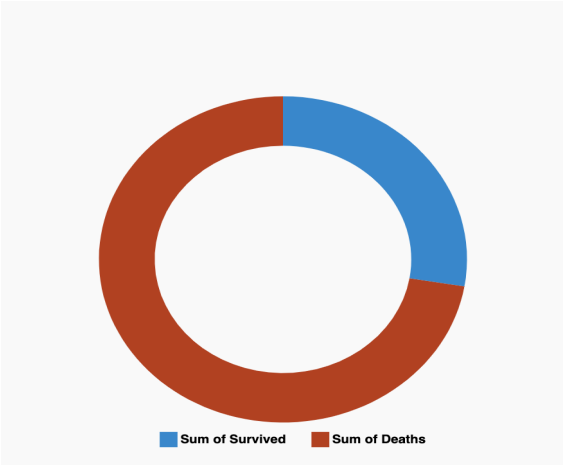
Implementation

Here is the final implementation of our project explained with images -

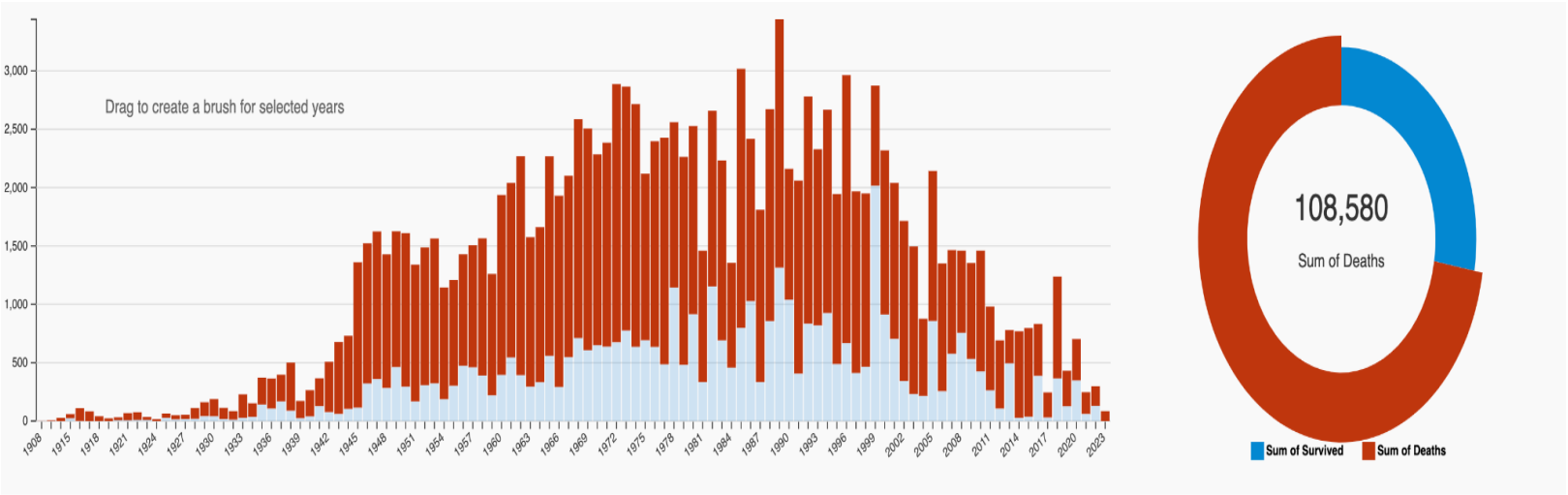
The first and probably the most important part of the dashboard is the stacked bar graph. Showing the number of deaths as red and the number of survivors as blue spread over all the years in the dataset. Mouse-hover here as everywhere shows the value of the pointed value and year of the graph.



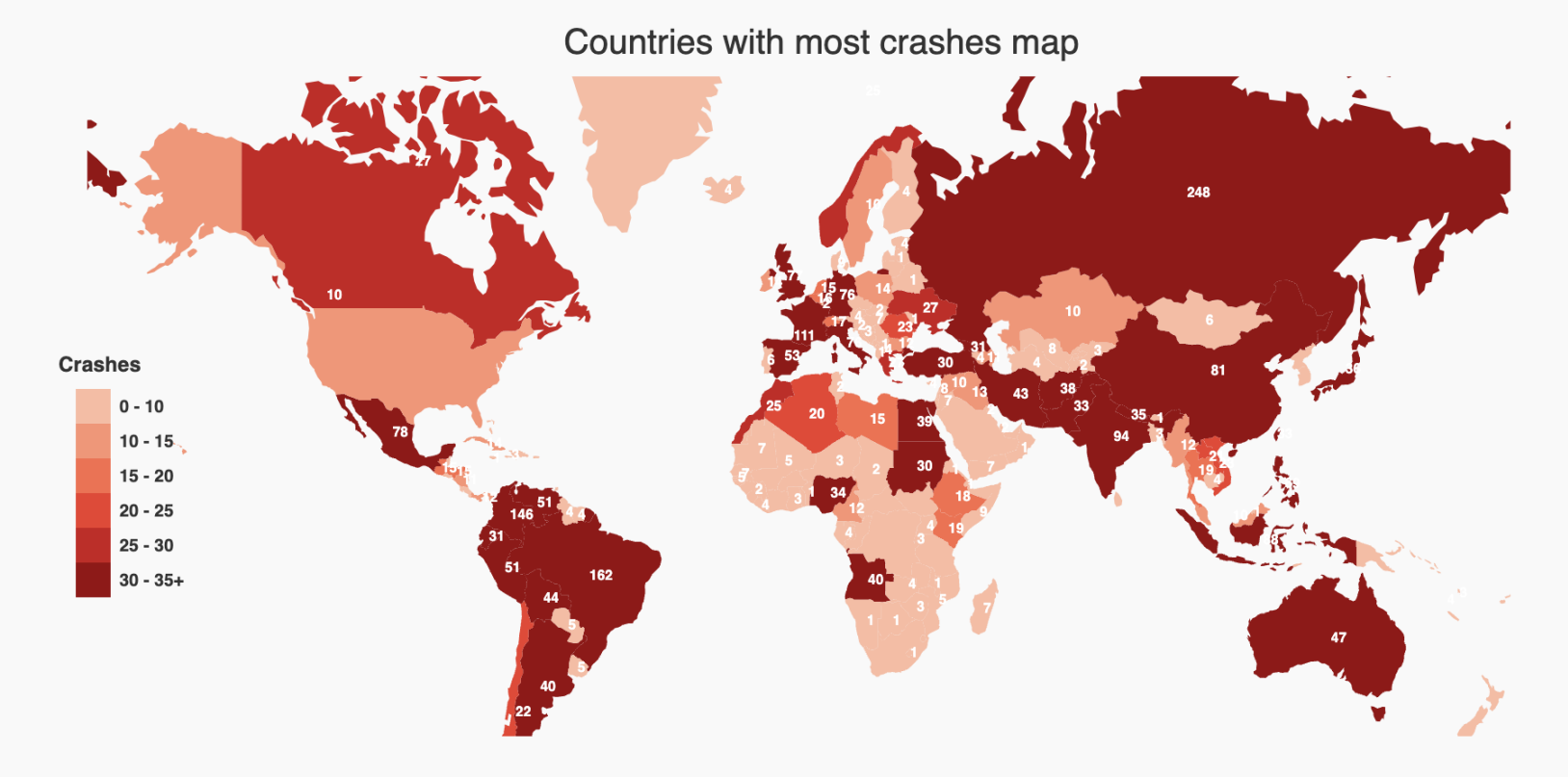
Next is the Donut chart showing us the total number of deaths and the sum of survived using respective red and blue colors. The first image shows only the donut with no labels because of the on-demand feature, the second image shows the values when you hover the mouse over it.



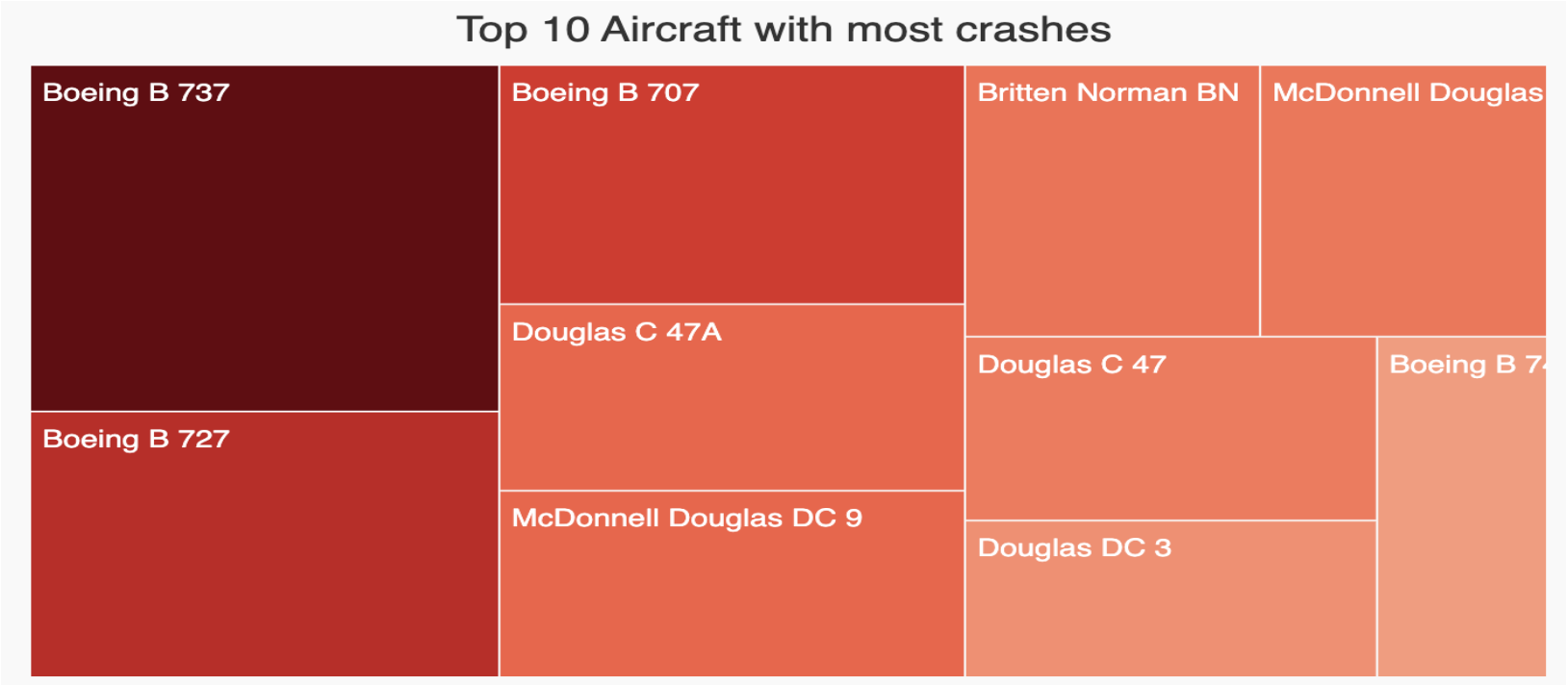
Now this mouse hover also changes the stacked bar graph by only showing the component you have the mouse over with. Having the mouse pointed on the red part of the donut chart makes the blue bars discolored in the bar chart and only shows the related red bars.



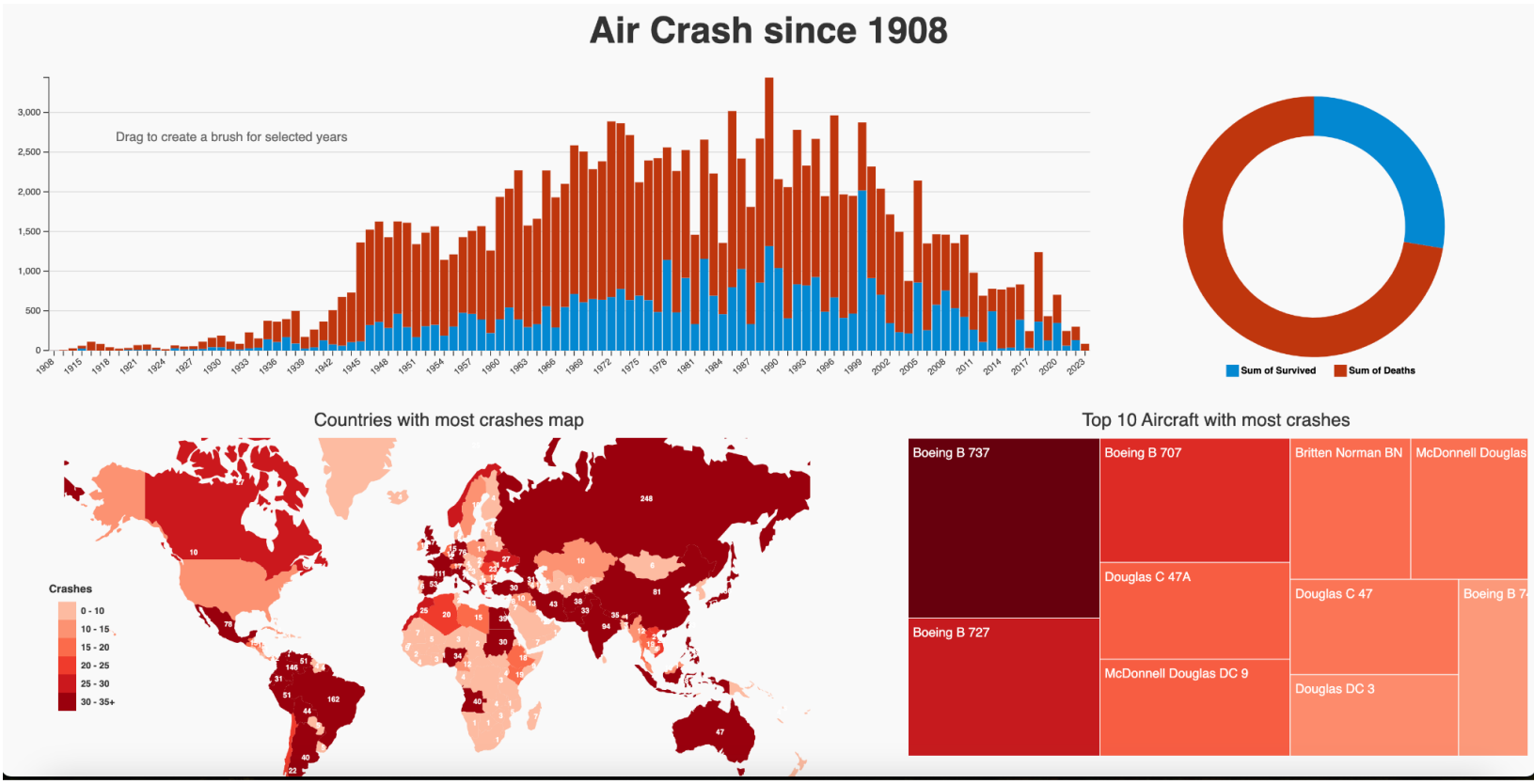
Next is the world map, showing the crashes geographically according to the countries they occurred in. Mouse over works on this visualization as well, showing the countries name and the total crashes. Now the heat map has been given certain limits of number of crashes and has been color coded too. So as we can see the tint of red increases as the number of crashes increase. This map is also affected when we select a selected number of years from the bar chart, resulting in showing crashes during only this period of time.



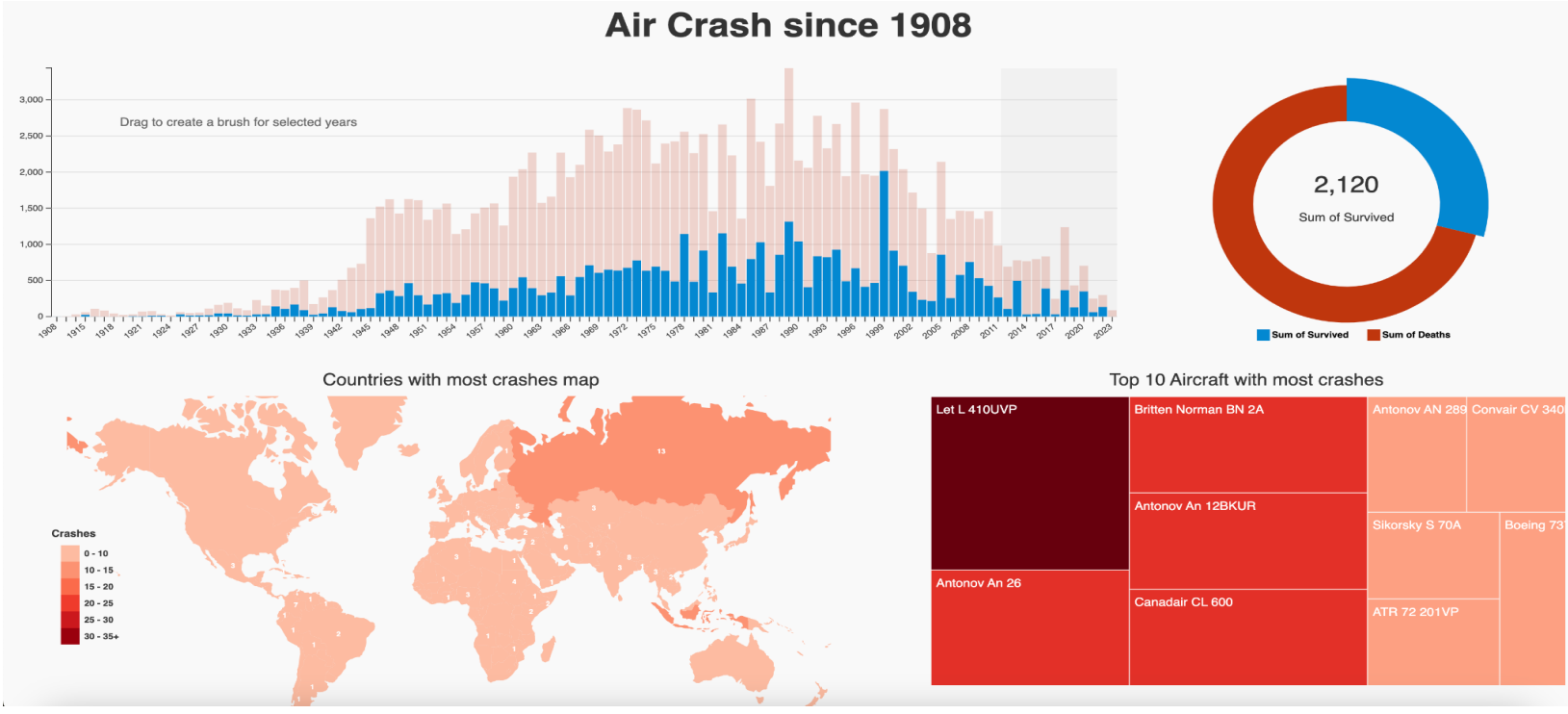
And finally the tree node shows the top 10 aircraft models having the most crashes in the dataset in hierarchical order, the aircraft with the most number of crashes has the biggest and the darkest rectangle and the one with the least of out of the 10 is the lightest red and smallest rectangle as well.



The next two images are of the final dashboards, one being the dashboard as a whole and one showing all the visualizations connected while interacting with the area selected by the brush in the stacked bar graph.



Final Dashboard (Above) vs Final Dashboard with Interaction (Below)



Evaluation

By performing all these visualizations, from the proposal to changing visualizations and reaching this final dashboard with very important takeaways from the data has been a journey through immense knowledge.

Insights gained:

1) Trend Analysis: Visualization 1, the enhanced bar chart, revealed significant historical trends in air crashes, including identifying specific years with high incident rates. This granularity allowed us to observe not just annual but also seasonal and daily patterns.
2) Survival vs. Fatality Rates: Visualization 2, the donut chart, clarified the proportion of fatalities versus survivors in air crashes. This perspective was vital for understanding the overall impact of these incidents on human life.
3) Geographical Impact: Visualization 3, the geographical map, provided clear insights into geographical patterns in air crashes. It highlighted regions with higher incidences, prompting questions about regional safety protocols and geographical risk factors.
4) Aircraft Model Analysis: Visualization 4's treemap offered a direct comparison of aircraft models involved in crashes, indicating which models might require more rigorous safety checks.

Overall, the final visualizations answered all our questions mentioned below.

- How does the trend of air crashes look over the period of the dataset?
- How do various countries compare in terms of air safety?
- Which aircrafts are the most dangerous?

Effectiveness:

1) Interactivity: The ability to interact with the data, especially the brushing feature in the bar chart and clickable elements in the map and treemap, enhanced user engagement and understanding.
2) Clarity: Each visualization offered clear, distinct insights, with color coding and design elements effectively highlighting key data points.
3) Comprehensiveness: Together, the visualizations provided a holistic view of the air crash data, covering historical trends, geographical patterns, and specific aircraft model insights.

Future Scope:

1) Real-Time Data Integration: Incorporating real-time data feeds could make the visualizations more dynamic and current.
2) Additional Data Variables: Including more variables, such as weather conditions or mechanical details, could uncover new correlations and insights.
3) User Customization: Allowing users more customization options in terms of data filters and display preferences could enhance the usability and relevance of the visualizations.

Conclusion

The visualizations proved effective in dissecting complex air crash data into comprehensible, actionable insights. They collectively facilitated a deeper understanding of trends, regional risks, and specific aircraft safety, proving valuable for safety experts and policymakers. Future enhancements focusing on real-time data, additional variables, and user customization will further improve their utility and impact.

References

[1]https://nycdatascience.com/blog/student-works/exploring-aviation-accidents-from-1908-through-the-present/
[2]https://medium.com/@Fehintiti/airplane-crashes-and-fatality-since-1908-7c4c2ec8a124
[3]https://datavizblog.files.wordpress.com/2016/04/fatal-airliner-crashes.jpg
[4]https://en.wikipedia.org/wiki/Asiana_Airlines_Flight_214