A Look on Airplane Crashes: Visualization

Beatriz Sofia Mesquita Gonçalves (115367)
MSc in Data Science
Information Visualization
Department of Electronics, Telecommunications and Informatics
University of Aveiro
beatrizs.mesquita@ua.pt

***Abstract*—This work allows the visualization of data from a dataset on aviation accidents from 1908 to 2009, available on the *Kaggle* website [1]. In this way, the use of this application allows a statistical visualization of how aviation safety has evolved over the years, which areas had more accidents in this period of years, as well as distinguishing the percentage of accidents in different types of flight operators.**

***Index Terms*—aviation, statistical visualization, aviation safety**

I. Motivation and Objectives

Fear of flying is still very common, despite constant developments in aviation safety. Between 33% and 40% of all people experience some form of anxiety when it comes to flying. 60% of sufferers experience generalized anxiety during the flight. Between 2.5% and 5% of the population have crippling anxiety, a genuine fear of flying that is classified as a clinical phobia. [6] [5] Therefore, the purpose of this application is to answer simple questions about aviation safety, in order to reassure the user about this subject with statistical facts. These factors are also coupled with my interest in front-end web development, and this paper serves as an introduction to it.

II. Users and Questions

A. Characterization of the users and their context

In the low-fidelity prototype, we defined a persona. It was defined as being a woman with age between 25-40 years old, who is a writer. Her motivation is to travel so that she can get inspiration for her stories. The problem is that she has never flown before and she has questions about the safety of this kind of transportation. So, basically, what she needs is a statistical point of view of air disasters, which corresponds to the goals of this app. However, although the app is geared towards this persona, anyone can use it. The ages of users who responded to the usability test were between 22 and 27 years old, and their level of computer skills ranged from low to high.

B. Characterization of the users and their context

The main questions the application should answer are, for example:

- How has flight safety evolved?
- Where do most of the accidents happen?
- In what type of flight do most accidents happen?

III. Dataset

The dataset used is called "Airplane Crashes Since 1908" and is available from *Kaggle*. It provides a history of airplane crashes around the world from 1908 to 2009. Below is a list of the variables available in the dataset:

- **Date**: date of the accident;
- **Time**: local time of the accident;
- **Location**: accident location;
- **Operator**: airline or aircraft operator;
- **Flight**: flight number assigned by the operator;
- **Route**: partial or total route taken before the accident;
- **Type**: aircraft type;
- **Registration**: aircraft ICAO registration;
- **cn/In**: serial number/construction number;
- **Aboard**: total on board (passengers + crew);
- **Fatalities**: total fatalities on board (passengers + crew);
- **Ground**: total of people killed on the ground;
- **Summary**: a brief description of the accident and causes (if known).

[1] However, knowing the type of persona to whom we are presenting this project and, in order to satisfy the questions that may arise, we can discard some of these variables from our exploration. In addition, many of them had most of entries with null values. Therefore, the variables that ended up being explored from this dataset were:

- **Date**

- **Location**
- **Operator**
- **Fatalities**

The data was extracted into a `.csv` file, available in the folder `datasets/datacleaning.ipynb`.

IV. Visualization Solution

In order to implement our proposed solution we started by defining a low-fidelity prototype (on paper). This prototype was used for testing and user feedback in Information Visualization classes and with colleagues. Next, the previous feedback was implemented in order to develop the high-fidelity prototype. The System Usability Scale (SUS) test was conducted remotely with five colleagues, using google forms, available in the main folder with the name `SUS_VIapp.pdf`.

A. Low fidelity prototype and user feedback

As mentioned earlier, a fairly simple low-fidelity paper prototype was defined, which eventually underwent some changes.

We start by prototyping the home page, which is divided into four sections: Timeline, Survival Probability, Accidents by Country, and finally Flight By Type. This can be seen in Fig.1.

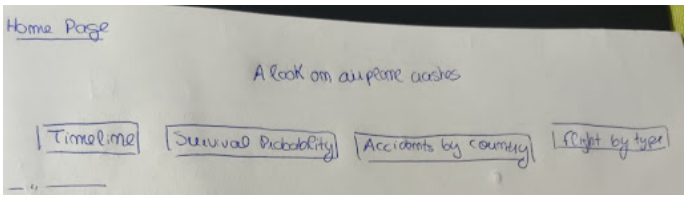

Fig. 1. Low-fidelity prototype, homepage.

Clicking on the first section, ”Timeline”, gives us access to the first view. Here we first have a drop-down where the user can choose whether they want to see information about crashes or fatalities. Depending on what is chosen, a bar chart is displayed where we see the temporal evolution of the same, i.e., the temporal evolution of crashes and fatalities from 1908 to 2009, as we can see in Fig.2

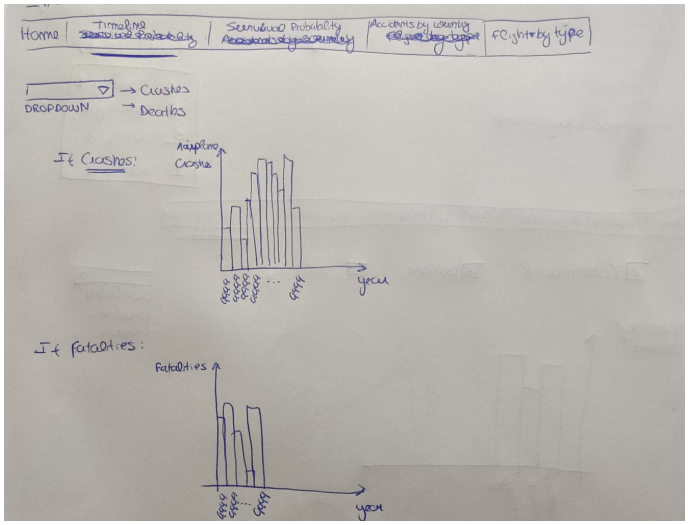

Fig. 2. Low-fidelity prototype, timeline.

In the second section, we have ”Survival Probability”, where the idea of making a simple pie chart was presented. We can see the comparison of the average survival probability with the average fatality probability, represented in Fig.3.

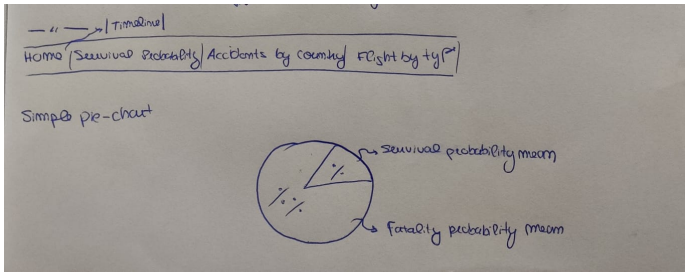

Fig. 3. Low-fidelity prototype, survival probability.

Then, opening the ”Accident by country” section, the idea presented was to have a representation of the world map, which would be fixed and show spheres of different sizes depending on the density of crashes in a certain country. Besides the map, a drop-down would appear next to it where we could choose to see the countries with more crashes, with less crashes, as well as the countries with more fatalities and less fatalities. Depending on the choice in the drop-down, the output would be a list with the top five countries that meet the condition. All this can be seen in Fig.4.

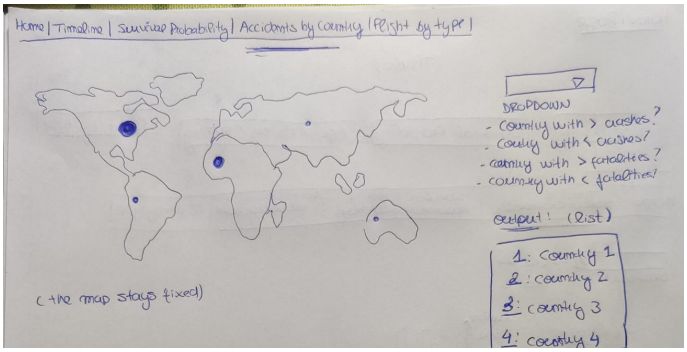

Fig. 4. Low-fidelity prototype, survival probability.

Finally, by choosing the ”Flight by type” section, the user is presented with three options to choose from: a view on military, commercial, and private flights. The user can choose one option, two, or all at the same time. In addition, we also have a drop-down where we can choose to have a statistical analysis of the percentage of crashes, i.e. the number of crashes over the total number of flights presented in the dataset, or the percentage of fatalities, i.e. the number of fatalities over the number of people aboard a given flight. Depending on the option chosen, we can see a multi-value bar graph, where the percentage of crashes or fatalities, depending on the drop-down, are analyzed for each type of flight chosen. All this is represented in Fig.5.

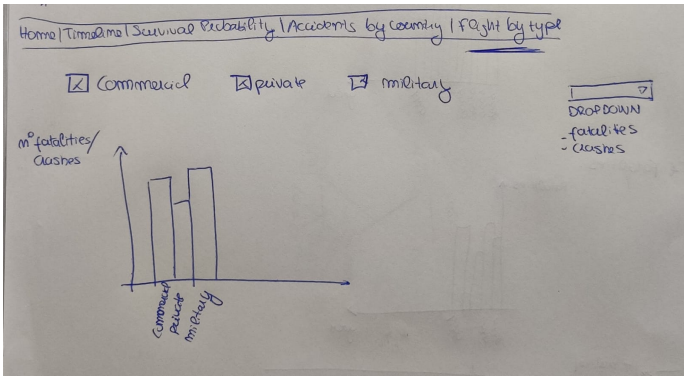

Fig. 5. Low-fidelity prototype, flight by type.

After presenting this low-fidelity prototype in class, some suggestions were made by Professor Beatriz Sousa Santos and Professor Paulo Dias to be implemented in the functional prototype:

- In the first section, ”Timeline”, since we are talking about a temporal evolution, it makes more sense to make a line chart instead of a bar chart.
- In the ”Survival Probability” section we have to take into account that using a pie chart can be ”dangerous” in visualization. This visualization would not be the best for this purpose. Of course, in this case, the average survival probability, compared to the fatality probability, is very small. This would end up having the opposite effect on the user of the main purpose of this application. We would end up discouraging the user to fly when we want the complete opposite.
- In the ”Accidents by country” section, where the world map is presented, instead of having a fixed map, it would make more sense to let the user move the mouse over the map and have information about accidents or fatalities about the country they want. Also, perhaps a color scheme is more appropriate for distinguishing accident or fatality densities rather than spheres of different sizes.
- Lastly, still in the ”Accidents by country” section, since the dataset gives us information about all US states, it would be interesting to do a more detailed visualization only on USA states, for example.

After the presentation, we started the tests with our colleagues, having done a total of three tests. As suggested by the Professors, it was suggested that the pie chart on the comparison of the mean probability of survival and fatality be removed, this being the most criticized aspect.

B. Functional prototype/ Final Version

After these suggestions made during the usability tests of the low-fidelity prototype, we then move on to the implementation of our functional prototype.

We start, then, with the home page, where four sections are presented, which differ from the sections that were shown in the low-fidelity prototype. As suggested, the ”Survival Probability” section has been removed, for the reasons we’ve already seen. In addition, a visualization of the number of accidents in the USA was also added, in addition to just the world map, as we can see in Fig.6.

The user can then interact with the home page, choosing the section he wants to explore. Starting with the ”Timeline” section, clicking there, the user is directed to the page represented in Fig.7. D3 was used to fetch the data from the `.csv` files, as well as to draw the graph. A drop-down is then displayed, where the user can choose whether he wants to see information about the number of crashes or fatalities over the years. Depending on the chosen option, a line graph is then represented, where the y-axis represents the number of crashes or fatalities (depending on the option) and the x-axis represents the years, from 1908 to 2009. Of course, if the user changes the option in the drop-down, the graph axes adapt to the chosen data. Furthermore, we have a time slider on the x-axis, where we can choose the range of years we want to see, in case the user is interested in a certain period in detail.

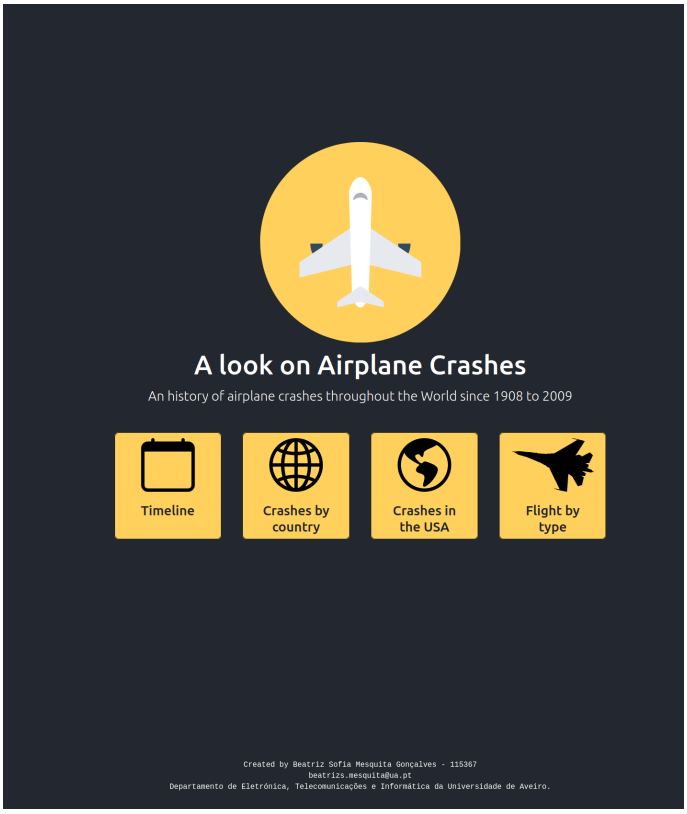

Fig. 6. Functional prototype, homepage.

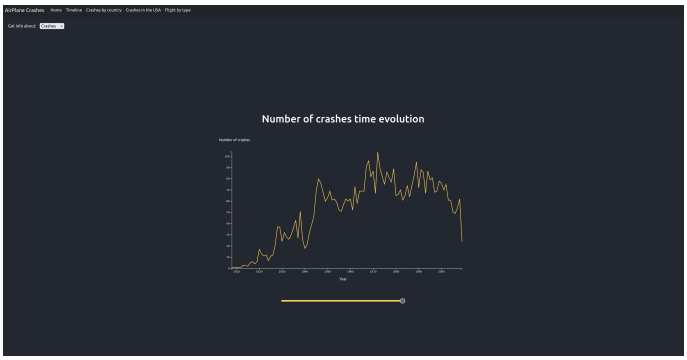

Fig. 7. Functional prototype, timeline for crashes.

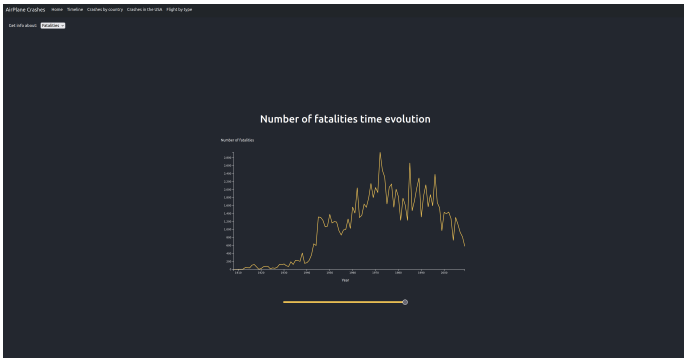

Fig. 8. Functional prototype, timeline for fatalities.

We can draw some conclusions from these line graphs. We don't have the total number of flights made by each airline to date so we have to be careful not to jump to conclusions. Taking this into account, we can see that the number of crashes in the first years is very low, precisely because, at that time, perhaps there was not so much registration of the flights made. The first commercial plane appeared in 1914. [3] The second known passenger aircraft carried 8 passengers and was produced between 1925 and 1933. Well, it is normal that the first concern when these planes began to appear was not safety and the immediate evolution of the same. When cars appeared, for example, the first concern was not their safety, but their performance. We can see a large increase in crashes and, consequently, fatalities, between 1940 and 1950. Now, DC-3 planes were produced in large numbers during World War II and sold as war surplus after 1945. At that time, commercial planes were almost always repurposed bombers for civilian use and perhaps this is a good explanation for this increase. [2] However, we see that more recently, from 1995 onwards, the number of crashes and fatalities has been decreasing. Now, we know that the number of flights has been increasing over these years and never the opposite, due to the greater demand and need for this means of transport. Therefore, we know that the number of flights increases and the number of crashes and fatalities decreases. Well, this is a direct indication that security has been improving over the years, even though we don't know exactly how many flights there were. This may be due to advances in technology and, of course, learning from past tragedies.

Moving on to the next section, "Crashes by Country", we can see a drop-down where the user can choose between seeing statistics about crashes or fatalities. D3 was again used to grab the data from the `.csv` files, both fatalities, and crashes, as well as the name of the locations. The only difference from the others is that, besides the data files, we have to get a `json` that has the identification of the zones, either states or countries, and the respective polygons or shapes that allow them to be drawn with `SVG`. The user can hover the mouse and see the name of the country and how many fatalities and crashes we have in each zone. As we move the mouse over the zones, these are highlighted from the rest. The map is represented with a color scheme, where we have its legend on the left edge of the map, changing depending on the option that the user chooses. Here, contrary to what was predicted in the low-fidelity prototype, we didn't do a drop-down with the top five countries with the most/least crashes/fatalities. Instead, a box was made on top of the map where we can extract information about the country and the number of crashes/fatalities in it, depending on the chosen option, and the color scheme allows us to know their density. Therefore, analyzing the map, we can conclude that the country with the most crashes was the USA, with 1436 crashes, and the country with the fewest crashes was Antarctica, Botswana, and Namibia, with one crash. Out of curiosity, we can see that Portugal recorded 19 crashes in this period. As for fatalities, we can see that the country with

the most fatalities was, again, the USA, with 16553 fatalities, and the country with the least fatalities was Belize, with 8 fatalities, for example. Portugal recorded a number of fatalities equal to 747. We can see this in Fig.9 and Fig.10.

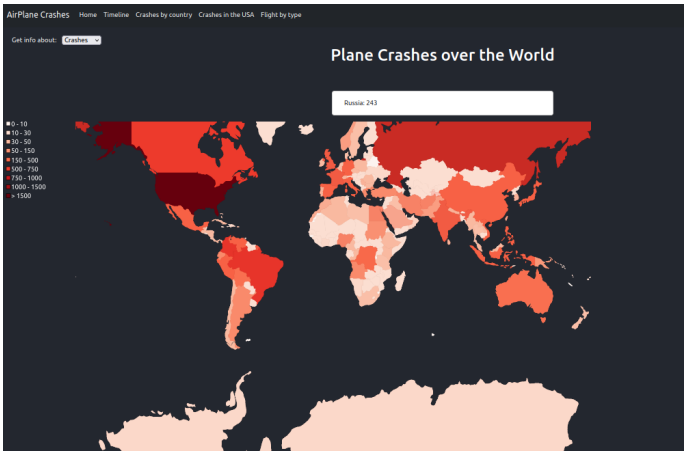

Fig. 9. Functional prototype, crashes by country.

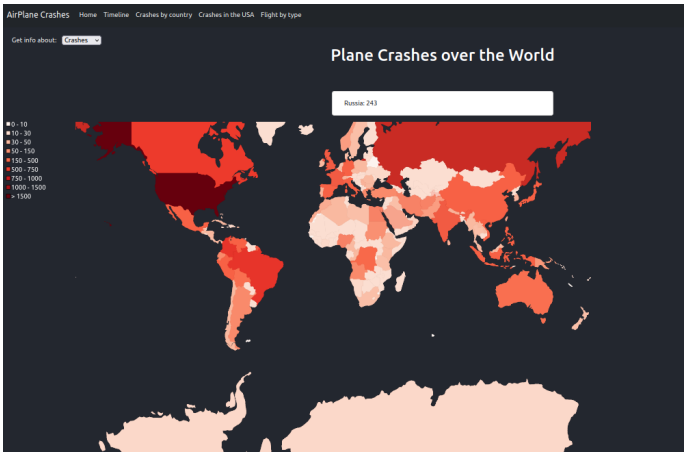

Fig. 10. Functional prototype, fatalities by country.

In the section that follows, ”Crashes in the USA” we have a visualization of the same genre as the previous section. However, instead of a world map, we have a map of the USA and therefore a more detailed view of this country. This is because we had data on all states in this country, allowing us to reach this detail. Again, we have a drop-down where the user can choose between getting information about crashes or fatalities. The user can also interact with the map in the same way as in the previous section. We can take from here that the state with the most crashes was Alaska with 177 crashes and the country with the least was North and South Dakota, with 2 crashes. Regarding fatalities, the state with the most fatalities was California with 1754 fatalities and the states with the least were South Dakota and Delaware with 8 fatalities. This section is represented in Fig.11 e Fig.12.

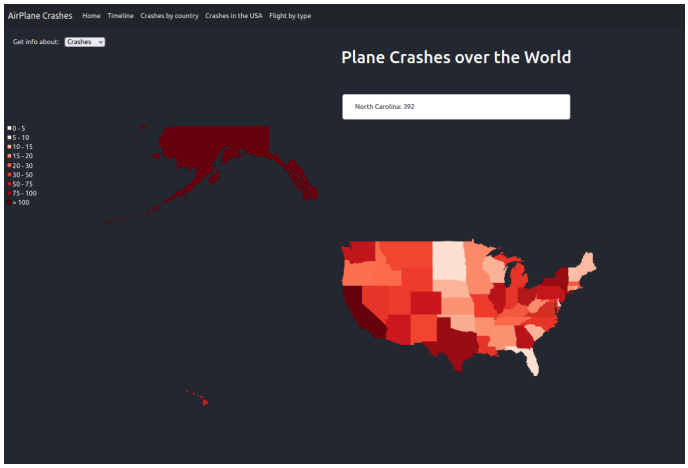

Fig. 11. Functional prototype, crashes in the USA.

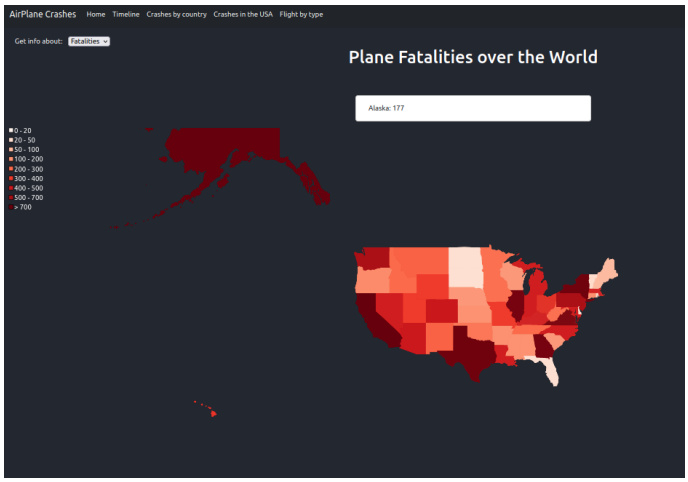

Fig. 12. Functional prototype, fatalities in the USA.

Finally, moving to the last section, ”Flight by type”, we can see a bar graph that allows us to compare the percentage of crashes by the operator. By percentage of the crashes, we mean the quotient of the number of crashes by the total number of crashes recorded in the dataset. This visualization allows the user to choose the type of operators that he wants to compare or if he wants to see only one, for example. The types of flights that the user can choose to obtain information or compare are: military, private, or others. In others, commercial flights, carriers, or air mail are inserted, because these classes had different formats in the dataset that do not allow us to distinguish them through an algorithm. The color scheme also adapts to the number of options chosen by the user. Again, D3 is used to take information from the `.csv` files and to draw the graph. Therefore, in this way, on the x-axis of the bar chart, we have the type of flight, accompanied by a legend of the color scheme used in the chart, on its right margin. On the y-axis, we have the percentage of crashes. Contrary to what was proposed in the low-fidelity prototype, a drop-down was not made between the percentage of crashes and

the percentage of fatalities. This is for the reason that we have already mentioned, the percentage of fatalities in an accident, compared to the percentage of survival, will be very high, as is natural. This would end up having a negative effect on the user, which is not intended. We can see that the type of flight with the highest percentage of crashes is the ”others”, which is normal, since this includes more than one type of flight, including commercial ones that have a large number of flights, as well as carriers, with over 80% crashes. On the other hand, the type of flight with the lowest percentage of crashes is the private one, which was also expected due to the reduced number of flights that these must make, compared to the rest, with less than 1.6%. As for military flights, we have a percentage of crashes approximately equal to 15%. This can be seen in Fig.13.

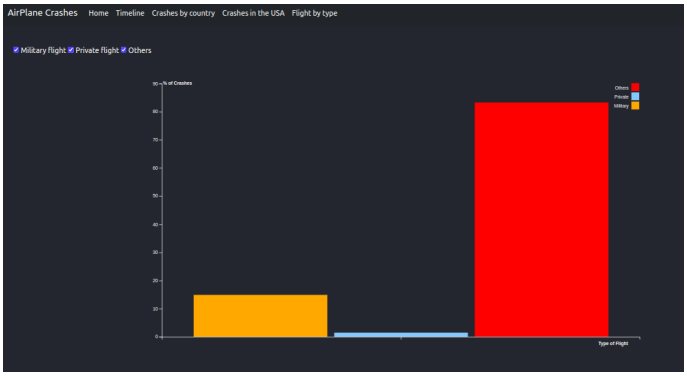

Fig. 13. Functional prototype, flight by type.

V. Implementation Challenges

There were some implementation issues that managed to be overcome with some research. Since this was my first time using web language and D3, there was a certain period of adaptation, which I quickly overcame, offering me a good and practical introduction to these. There was some difficulty in inserting legends in the graphics, as well as implementing mouse-over functions in the maps. However, as stated, these were overcome through some research and viewing of examples. This research was facilitated by the fact that the JavaScript language and the D3 library are widely used, with the material available on the internet being very vast, as well as vast forums with users' doubts. Sources other than those given in classes were explored and therefore, some methods used in these sources ended up being replicated. However, it is important to point out that, at no time, the content of these sources was plagiarized. Code excerpts were taken, later adapted to our objective and our data, explicitly referenced in the document `references.txt`, present in the main folder. Finally, another difficulty felt was obtaining the countries from our dataset. Many of these countries were badly written and with varied formats. It was then necessary to analyze them one by one and correct the names, as well as standardize the format.

A. Evaluation of the functional prototype

To evaluate the functional prototype, a SUS questionnaire (System Usability Scale), represented in Fig.14, was applied to five different people. As seen previously, the answers to our questions, as well as the score, are represented in the PDF file `SUS_VIapp.pdf` that can be found in the main folder. The score obtained was 95%. [4]

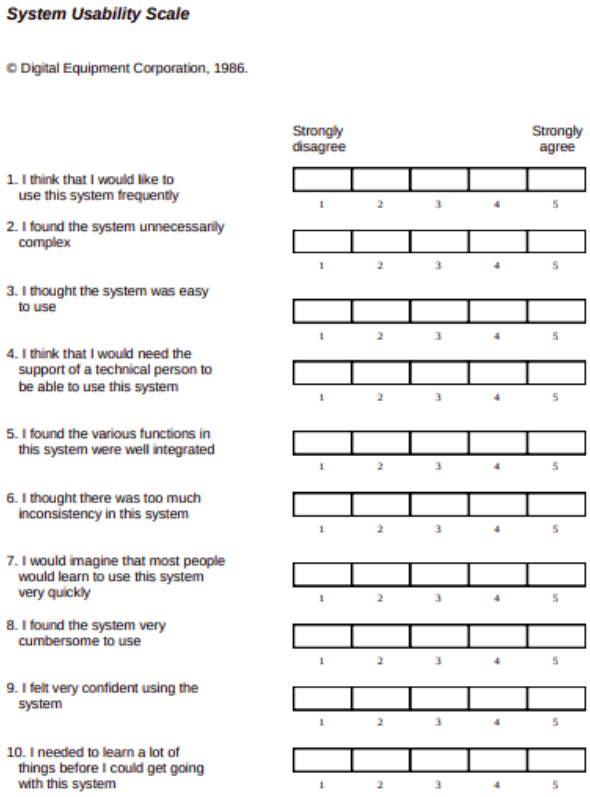

Fig. 14. Functional prototype, SUS.

VI. Conclusion and Future Work

This project allowed me to acquire new notions of the D3 library, as well as JavaScript and HTML. It also allowed us to have a better idea of how much the choice of visualization can affect the user's perception of the theme, for example, as we saw with the case of the pie chart with the probability of survival that was presented in the low-fidelity prototype, but which ended up for not being applied due to this very reason.

Taking into account future work, some aspects that I would like to implement are:

- Since we have information about several cities, an interesting idea would be to allow the user to interact with the map, zooming in on it, in order to have more detailed information about them;

- To avoid discouraging the user with the probability of surviving a fatality, a good alternative would be to try to look for more data, in this case, on road accidents,

for example, and compare with the number of aviation accidents, normalizing the data, and showing that in fact, it's safer to fly;

- Another aspect to improve would be, for example, when we make the timeline, we present normalized data, allowing the user to draw more direct conclusions from it;

- In the case of the maps, instead of two sections it could have been made just one, since the visualization technique is the same, and create a button that would then show the USA map instead of the world map, for example

- Finally, a funny conclusion would be to use the data on deaths on the ground, which we ended up removing at the beginning, and from these show the areas where it is more likely to be hit by a plane.

REFERENCES

[1] Airplane crashes since 1908. https://www.kaggle.com/datasets/saurograndi/airplane-crashes-since-1908, note = Accessed: 2022-11-20.

[2] Lancaster test bed images. https://www.lancaster-archive.com/lanc_photos_testbed.htm, note = Accessed: 2022-12-01.

[3] Sikorsky s-22 ilya muromets 1913. http://www.aviastar.org/air/russia/ilja_muromets.php, note = Accessed: 2022-12-01.

[4] John Brooke. SUS: A quick and dirty usability scale. 1995.

[5] Bert Busscher, Philip Spinhoven, and Eco J. C. de Geus. Psychological Distress and Physiological Reactivity During In Vivo Exposure in People With Aviophobia. *Psychosom Med. 77(7):762-74*, 2015.

[6] Robert D. Dean and Kerry M. Whitaker. Fear of Flying: Impact on the U.S. Air Travel Industry. *Journal of Travel Research, 21(1), 7–17*, 1982.