**Airplane Safety Project**

Cody Alliston, Collin Fernandez, Kareena Kothari, Zack Nakasone

DATA 424 – Spring 2022, Rhonda Crate

April 24, 2022

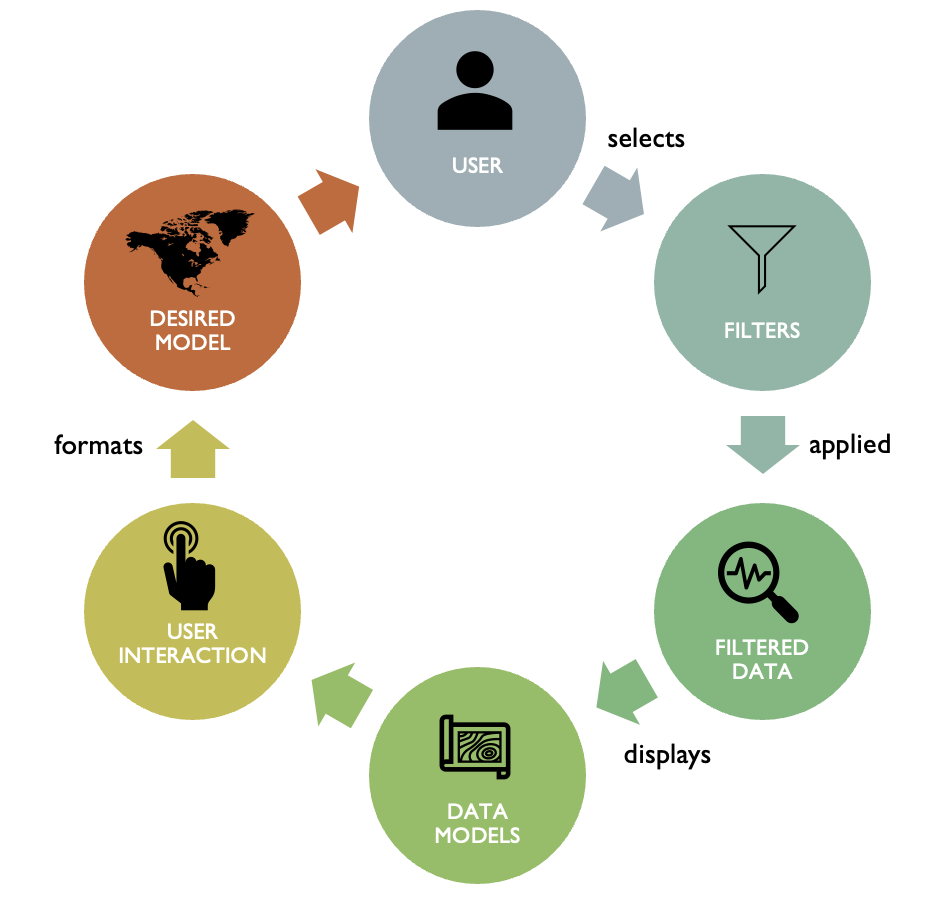
**Abstract**

The initial goal of this project was to produce a shiny web application to determine hotspots for accidents using data from the National Traffic Safety Board. The scope evolved into creating visuals with interactive filters to see where there are factors that may be related to each other and contribute to accidents or accident severity. The application allows for filtering of data to easily show factors contributing to accidents in certain areas. The application focuses on smaller aircraft, with airline accidents being filtered out due to their frequency being lower and fatalities being higher. A rating called a Severity Index is assigned to each accident. Techniques such as clustering and other factor analysis were applied to the data to determine if there were common themes relating to accidents and higher severity index ratings. Factors analyzed included broad phase of flight (takeoff, landing, taxi, cruise, etc.), month, weather conditions, whether the aircraft was personally or professionally manufactured, and more. These factors can be grouped in different ways to see where they are related.

***Introduction***

This project was drawn to understand accident hotspots in the United State and analyze the internal and external causes of accidents involving smaller aircraft. The objective was to visualize these hotspots and attribute relationships in a shiny web application, which would also allow user input to further filter the data used in the visualizations. The team focused on identifying which attributes contributed to severe or fatal accidents. Key components of this analysis were conducted via factor analysis, heat map visualization, and the calculation of a severity index to focus on the more severe accidents.

***Data Flow Diagram.*** The planning phase of this project took several weeks. After the initial mentor meeting, the scope of the final application was to create an interface with filters and visuals. A user would have the ability to adjust filters to create visuals with their desired output. These visuals can be used to identify trends in the data, whether they be related to location or other factors.



The final application still follows this general flow. Users can select their own filters to apply to the default data model and create visuals to compare factors that they are interested in. The dynamic nature of the application allows for efficient display of attribute relationships and effective visualizations, enabling a user to form conclusions of their own.

**Literature Review**

The project is based on and involved similar analysis to a research article which analyzed hotspots of aviation accidents within Florida (Li, 2018). While there was no other specific literature to be reviewed for this project, it should be noted that the dataset from the National Traffic Safety Board (NTSB) was not perfect. While robust, it is also incomplete. In recent decades there has been some standardization to the kinds of data that are collected for accidents, but older data does not include information that would be relevant for this kind of analysis or visualization. Latitude and longitude data is only available for more recent incidents. For the purposes of this project, this worked out as the focus was on more recent incidents.

There were also several inconsistencies regarding input format in the dataset. Many columns contained different input formats for the same kinds of data. These columns had to be cleaned before they were usable for analysis or visualization. Recent standardization has helped to ensure completeness of accident collection data, however the format for inputting the data has not yet been standardized.

**Methods**

***EDA and Data Processing***

***Initial Data Cleaning.*** Accident data from the NTSB was freely available for use but needed cleaning and filtering before it was usable because it was inconsistent and incomplete. It had attributes that were deemed irrelevant to our analysis, which were then removed. After, the data was filtered to only include the analysis of airplane accidents in the US. Since the project was also focused on smaller aircraft, larger aircraft were first filtered out by make. After filtering out large aircraft easily identified by some manufacturers, it was found that there were still several incidents with large numbers of fatalities or injuries. This seemed odd, as there would be a smaller number of people affected if the aircraft involved was smaller. The data was then sorted based on the number of accidents and looked at for further cleaning on a case-by-case basis. Several of these occurrences were extremely high for the smaller aircraft we wanted. For example, one accident had over 200 people involved because a commercial aircraft was not filtered out. Not only was this the wrong kind of aircraft, but the sheer size of the incident, although minor regarding fatalities, became an outlier because of the number of uninjured passengers. This resulted in also removing specific aircraft models, even if they had a manufacturer that mainly focused on smaller aircraft.

***Data Cleaning: Airport Attributes.*** Some analysis and visualizations factor in the airport code and other airport attributes that may not already be in the dataset. Here, another dataset, the Global Airport Database (GADB) (Partow, 2022), was merged with the current one. This was to make each airport code in the dataset only represent an airport’s International Air Transport Association (IATA) code. The airport’s IATA code was associated with its name, latitude, and longitude which were added from the GADB in conjunction with the fixing of the airport codes. Map visualizations can use either all the airports in the GADB within the US or the airport longitude or latitude values within the new dataset to only include those airports that have recorded accidents.

***Data Cleaning: Latitude and Longitude.*** Accurate latitude and longitude values are essential for accurate map visualizations. There were several steps involved in cleaning the latitude and longitude data, beginning with removing any NA values. This largely removed any accidents before the year 2000, since most aircraft now record their latitude and longitude. The data then had to be converted to numeric type as some values were formatted as alphanumeric. Next, values had to be verified. Latitude values must be between -90 and 90, but for the purposes of this project, they must also be between 0 and 90 to represent the area of only the US. Longitude values must be between -180 and 180, but a range of –180 to 0 was implemented. Some of these incorrect values were due to human error and were adjusted accordingly, while any other accidents were just removed entirely with any other missing values were filtered out at this point. This filtering helped completely visualize the data for a map, because only about 2000 of our accidents were plotted before, but now all 18000 are accurately represented.

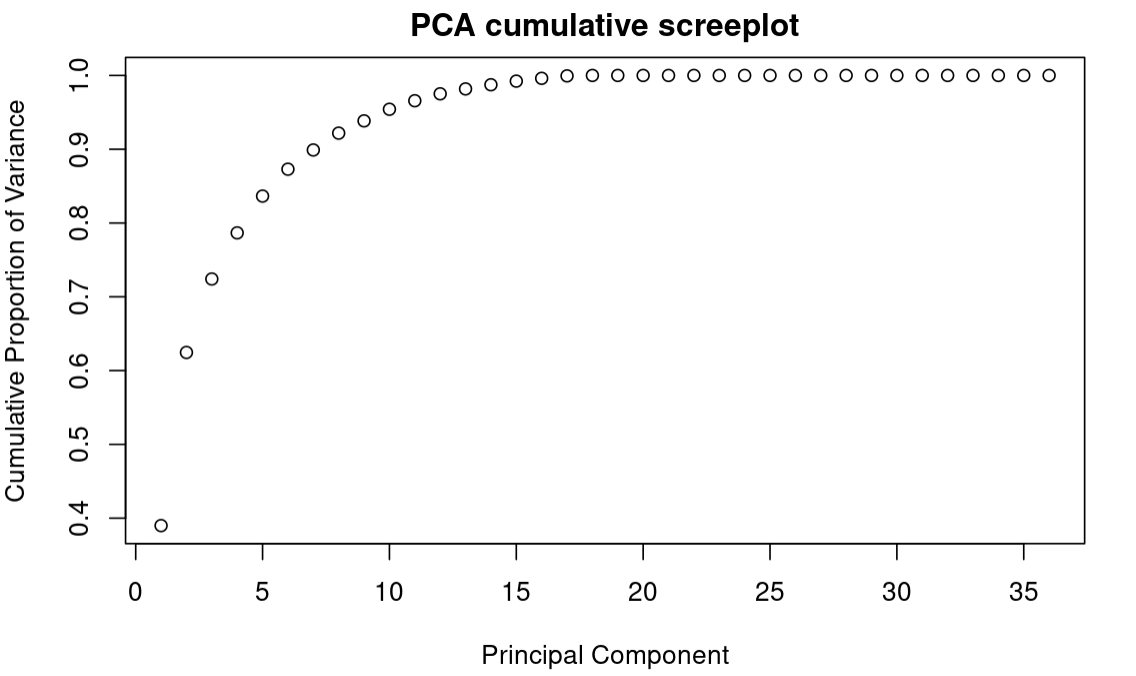
***Other Data Cleaning.*** There were several other issues with the dataset. For example, many of the values were not entered consistently. In some entries, fatalities were entered as an integer and some entries were alphanumeric. This was the case for all injury severity columns. Non-numeric characters had to be removed and in some cases NA or NULL values had to be changed to zero. Date values were also split up to analyze incidents based on month. It was found that there were several incidents that shared an event id, but after further investigation it was determined that the duplicate event id was because there were multiple aircraft involved in those accidents, meaning that accident number is the unique identifier of the dataset.

***Severity Index.*** Without assigning a measure to the degree of damage caused by an accident it would be difficult to effectively show differences regarding the severity of accidents. The only method for comparing accidents would be simply counting the numbers of different kinds of accidents. By assigning a value to each accident, accidents can be compared to one another on an individual or group basis. Much of the analysis was focused on the Severity Index which assigns a numeric value to each incident based on factors like injuries and death. In this way more emphasis can be placed on accidents that had more of an impact on life without discounting those accidents that may only have caused minor injuries or even no injuries at all. The severity index is calculated as follows (Li, 2018):

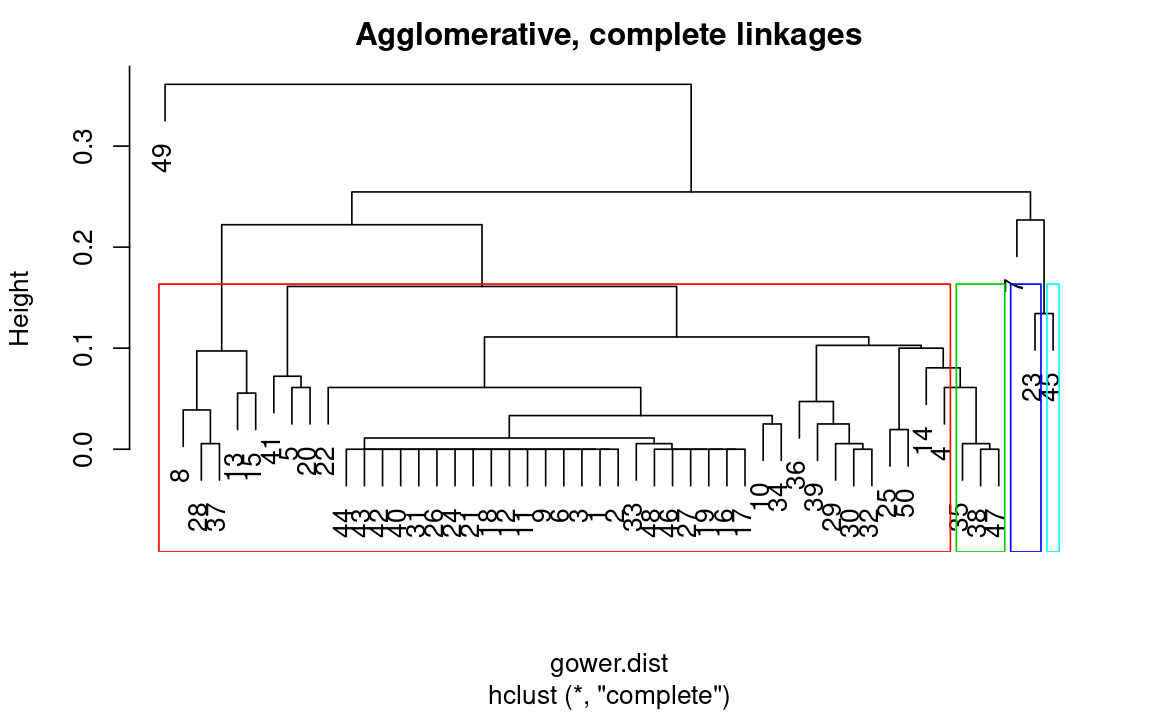
X1 represents the total number of fatalities, X2 represents the number of serious injuries, X3 represents the number of minor injuries, and X4 represents the number of uninjured involved in the accident. Each variable is multiplied by a constant corresponding with the severity variable: 3.0 for fatalities, 1.8 for serious injuries, 1.3 for minor injuries, and 1 for uninjured. In this way minor accidents with lower numbers of fatalities are scored lower than major accidents with higher fatalities.

***Results***

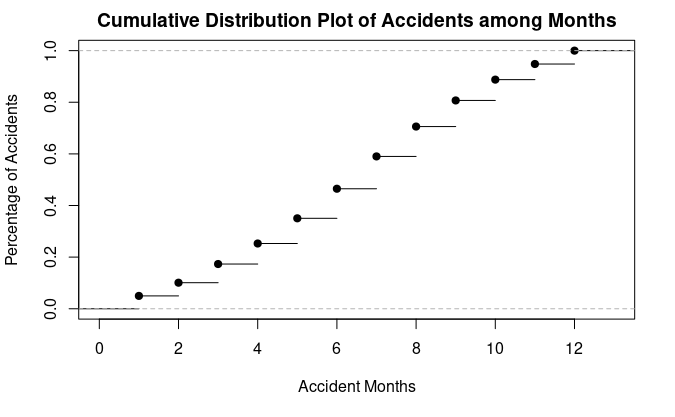
***Principal Component Analysis.*** We wanted to perform Principal Component Analysis on the categorical data, so we one hot encoded the data and created separate columns for each nonnumeric column. In the screen plot we see that the first 10 components count for more than 90% of the variance seen in this data. In the Biplot we can see that Injuries results in the largest distance.



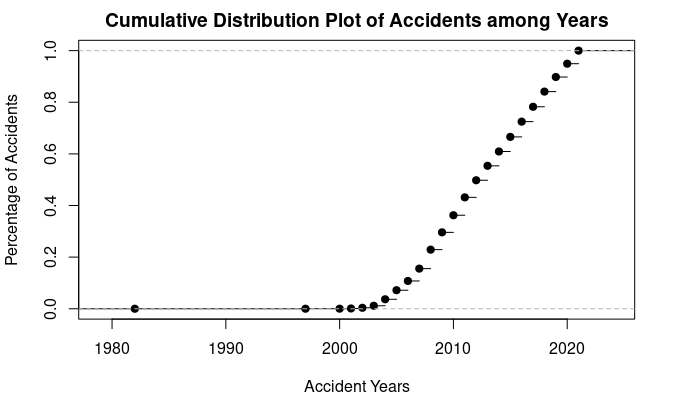
***Hierarchical Clustering.*** As another exploratory analysis method, we used hierarchical clustering with the Gower method. Looking at the clusters we see a large group for the first cluster, and very small alternative clusters. Analysis of one hot encoded data is extremely difficult, while checking clusters 3, and 4 I found that fatal injuries occurred in those clusters, while cluster 1 generally did not have any fatalities. In the future we may have to reformat the data back so that it can be more easily understood.



***Cumulative Distribution of Time (Months and Years):*** When determining which factors contribute substantially to accidents. Cumulative distribution describes the probability distribution of different types of variables. Those variables include continuous, discrete, and random. Further analysis of the continuous variable, time was necessary in anticipating the correlated relationships of the variables.

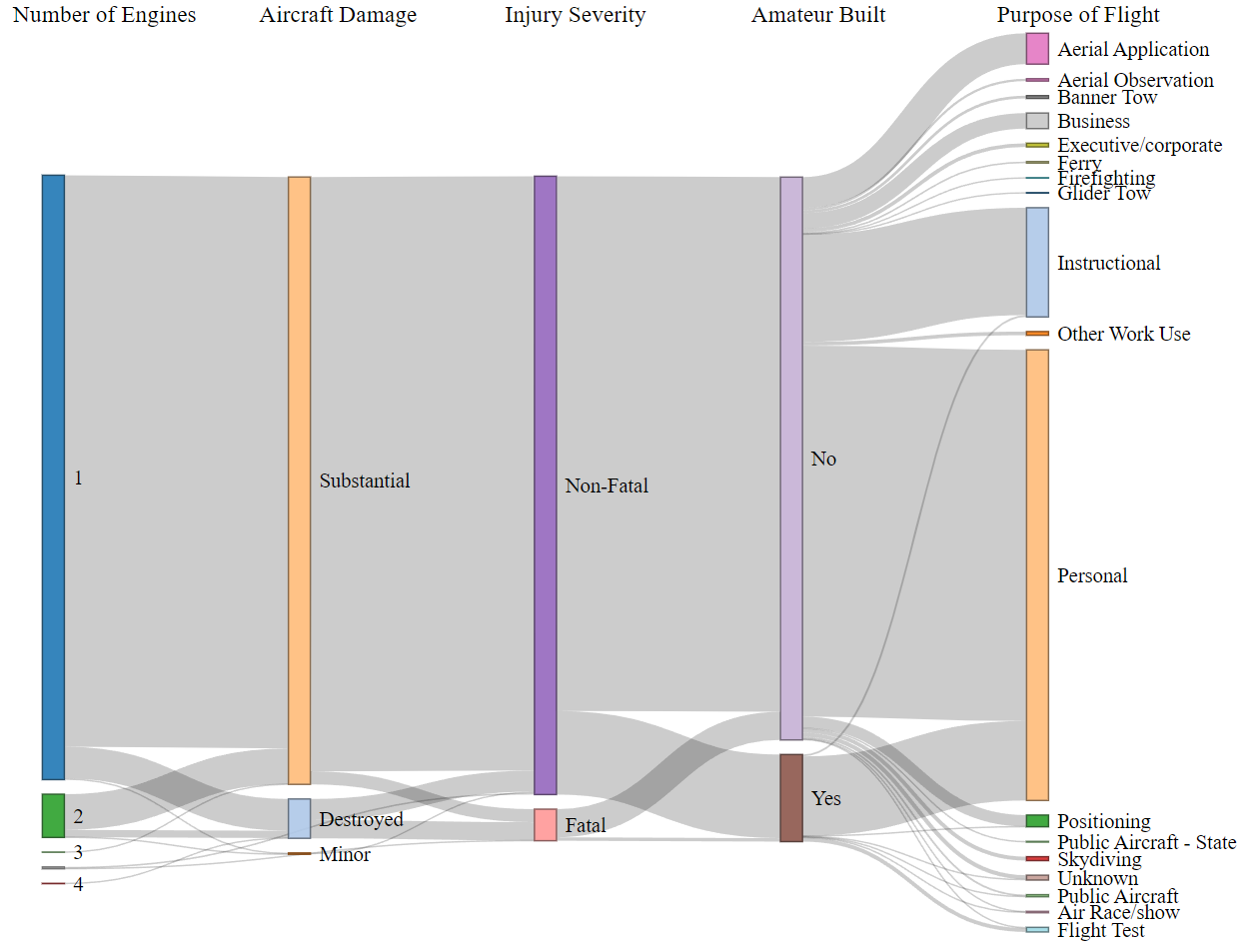


As seen, depicted in the above figure: Cumulative distribution plot of accidents among months. We were able to understand that there was a higher probability of accidents of smaller air carriers occurring during the months during colder seasons, than warmer ones.



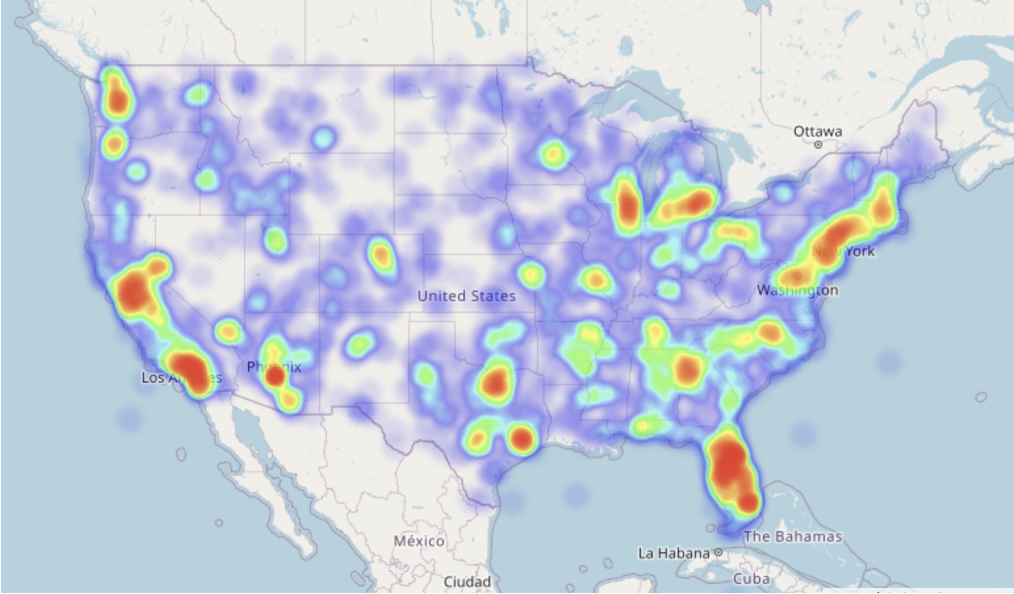
In comparison, the second plot depicted, regarding cumulative distribution of accidents among years. This plot illustrates those accidents in relation to smaller air carriers, occurred in the late 2000’s, compared to earlier.

***Sankey Plot.***  An interactive Sankey plot is available in the interface. It allows users to select column factors to see how they relate to one another. The plot below compares the following user-selected columns:

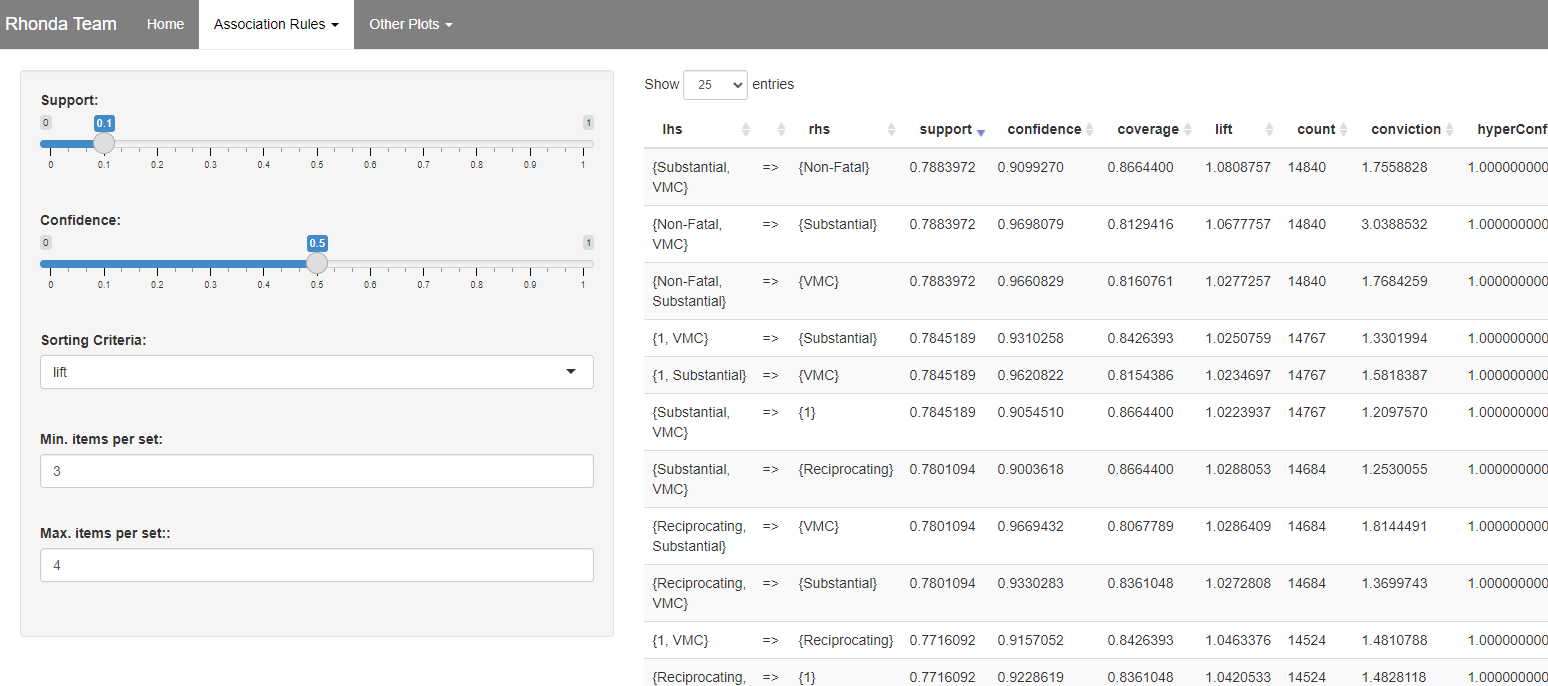


The plot allows the user to change the columns to see where there are strong correlations. For example, we can see that most amateur-built aircraft (the brown section in the fourth column) were involved in accidents when the purpose of flight was personal use. In the application, users can move the factors in the columns by dragging them into different positions, which aids in analysis of specific factors and creating a more useful and direct visual.

***Heat Map.*** A heat map using all the data is helpful in visualizing areas where crashes are more common. There are too many points to use markers when working with the entire dataset. When the set is broken into smaller subsets, markers become more practical. The interactive map allows users to select accidents by phase of flight. The heat map then displays the hotspots based on those phases. Markers show airport locations in the user interface.



***Association Rules Mining.*** A Data table will show the rules generated from running market basket analysis on the categorical variables for mining. Users will be able to adjust the arguments from the UI and sort the data for further exploration. When we sort the list by Support, we can see that there is a relation with VMC and substantial aircraft damage. Which is seen in more than 14,000 incidents. We can sort with the highest support and confidence to find valuable correlations between variables and their groups. However, there were no association rules found which led to a fatal result. Maybe sorting out fatal crashes and then performing association rules mining might be more effective for studying fatal crashes.



***Looking Forward***

The methods and visuals used in this project can be reused and repurposed in the future for different datasets where similar information is collected. The same techniques could be applied to subsets of this dataset, perhaps with a focus on commercial airline accidents or privately built aircraft. It could also be applied to datasets involving vehicle or boat accidents if there was enough information available. Filtering could be useful to examine subsets like commercial trucking accidents and their severity and, if available, applying latitude and longitude values to a heat map to look for hot spots. If latitude and longitude were available for boating accidents, then heatmaps could again help identify hotspots where other landmarks like cities or highways may not be as helpful.

***Conclusion***

This study quantified the contributing factors, these include, broad phase of flight, number of engines, damage, injury severity, amateur built, purpose of flight, and location. Results illustrate that a substantial amount of smaller air carriers' accidents occur during the landing phase of flight, in coastal states (Washington, California, Florida, etc). Our analysis aligns with the prior studies. The calculation of severity index was a crucial element in creating effective visualizations. As illustrated in the Sankey plot, and heat map. Alternatively, the heat map, the severity index is depicted in the highlighted (red areas) across the U.S. In comparison, the Sankey plot depicts various types of contributing factors of accidents. These factors include number of engines, aircraft damage, injury severity, amateur built, and purpose of flight. Separately the base plot illustrates the influence of number of accidents on the factor, broad phase of flight. Our analysis revealed there were several other relationships of various attributes contributing to accidents.

***Executive Summary***

The primary goal of this project was to create an app that could allow for analysis of airplane accidents in the United States. This analysis could be used with different air carriers to analyze the severe hotspots for their aircraft and determine what key attributes contributed to those accidents. The main components of the data that were analyzed and explored were injury severity, aircraft damage, phase of flight, number of engines, engine type, weather condition, whether the aircraft was amateurly built, and the calculated severity index. The team thought that the main areas of concern for accidents would be the coastal states and within the summer months since there would generally be more flights. The results demonstrated that while the coastal states did account for the greater number of accidents, they did not produce the most severe accidents. Weather condition was also thought to be a main contributor to an accident. Only one state, Alaska, seemed to be egregious in the number of accidents they produce and the severity of them with them being in the top ten states for both. The attributes that were determined the most contributing to an accident were engine type and the number of engines. Overall, the app this project produced and the analysis performed should help a user identify problem areas and the attributes that contribute them.

**Bibliography**

Li, Yafei and Chen Liang. (2018). The Analysis of Spatial Pattern and Hotspots of Aviation Accident and Ranking the Potential Risk Airports Based on GIS Platform (hindawi.com)- Journal of Advanced Transportation. Accessed April, 2022.

url: https://www.hindawi.com/journals/jat/2018/4027498/

NTSB Aviation Accident Database & Synopses.

url: <https://www.ntsb.gov/Pages/AviationQuery.aspx>

Partow, Arash. (2022). The Global Airport Database. Retrieved from

https://www.partow.net/miscellaneous/airportdatabase/index.html

Schweinberger, Martin. 2022. The Language Technology and Data Analysis Laboratory (LADAL). Brisbane: The University of Queensland, School of Languages and Cultures. url: https://slcladal.github.io/index.html (Version 2022.04.03).