



FIT5147 Assignment #2
Data Visualization Project
Air Crash Investigation

Submitted by:
Rohan Singh

Student ID: 30042496

Email: rsin0021@student.monash.edu

Table of Contents

1. Introduction	1
2. Design.....	1
2.1 Sheet1: Ideas Sheet.....	1
2.2 Sheet2.....	3
2.3 Sheet3.....	4
2.4 Sheet4.....	5
2.5 Sheet5 : Realization	6
3. Implementation.....	7
4. User Guide	9
5. Conclusions.....	14
6. References and Bibliography	15
7. Appendix	16

1. Introduction

Air industry has evolved over the years to become one of the safest modes of travel today. The chance of one dying while traveling by airways is mere 1 in 9821. This evolution has come through many failures and repeatedly learning from them. Therefore, this interactive data visualization application is intended to enlighten the general public about the journey of these failures (air crashes) over the years.

The aim of the application is to present the data to the user in the form of visualizations which are easily interpretable by the common public and also involve high user interaction at the same time.

The application enables the user to interact with the crash data by allowing the user to :

- Choose the time period of analysis for crash trend.
- Visualize the top 10 worst flight operators with most number of crashes.
- Visualizing the worldwide percentage distribution of crash fatalities on an interactive choropleth map.
- Explore the causes of crashes in the world and the top 5 countries with the most crashes by varying the controls of the interactive world cloud.

The data used for analysis is the collection of worldwide crashes over a span of 100 years, i.e from 1908 to 2009, which has been taken from the Kaggle repository <https://www.kaggle.com/ruslankl/airplane-crashes-data-visualization/data>.

2. Design

The design process for the application involves the use of five design sheet methodology. Here's the summary of ideas of each design sheet:

2.1 Sheet 1: Ideas Sheet

The flow of ideas was structured in five stages, namely

1. Ideate: Brainstorming Ideas

- The first thing to consider is the platform on which the application would be built. I considered two alternatives:
 - ✚ R Shiny: To create an application Dashboard using R.
 - ✚ D3 : To create a web application using java script and produce enhanced visualizations.

- The next thing to consider was the kind of plots to be included for the data visualization. Here are the alternatives that were considered;

Line charts, Choropleth maps, Bar Chart, Histograms, Scatter Plot, Word Cloud, Box plots, heat map, chord map, collapsible tree.

2. Filter:

- I chose to create my application using R shiny dashboard, since I have a better familiarity of the language R as compared to D3. Also the features of D3 library can be accessed in R using r2D3 package.
- Next, I filtered the plots into two categories:
 - ✚ For Continuous Data: Choropleth map, Histogram, Word Cloud
 - ✚ For Discrete Data: Line chart, scatter plot, bar graph, box plot.

3. Categorize:

Now, I categorized the data for which I wanted to create the interactive plots. Here are the alternatives that were categorized for the plots:

- ✚ Crash Trends : Yearly and Monthly
- ✚ Plot for percentage distribution of crash fatalities.
- ✚ Visualizing the crash Causes
- ✚ Plot for top 10 Worst flight operators with their crash count
- ✚ Plot for top 10 crash ridden countries

4. Combine

The above categorizations of data were combined with the plots as follows:

- ✚ Line graph with time control for showing crash trends.
- ✚ A Choropleth map displaying the worldwide percentage distribution of crash fatalities.
- ✚ Word Cloud for displaying crash Causes of different countries and across the world (involves lot of data wrangling).
- ✚ Bar plots for displaying the top 10 worst operators and top 10 crash ridden countries.

5. Questions (3 ideas to carry forward):

Here are the 3 ideas that would be carried forward in sheet 2,3 and 4:

- ✚ How does the crash trend varies from 1908 to 2009?

- ✚ What is percentage distribution of crash fatalities across the world?
- ✚ What are most common causes of air crashes and which are the worst flight operators?

2.2 Sheet 2:

- **Information:** This sheet discusses the representation of crash trends and the number of crashes by the top ten worst operators.
- **Layout:** Creating a tab named 'Crash trends' which shows two plots:
 - a. Line plot showing Yearly Crash Trend with slider control to vary the time period of analysis on x axis featuring labels on mouse hover, displaying the year of crash and the number of crashes in that year
 - b. A collapsible tree having each leaf corresponding to each flight operator and the size of the leaf being proportional to the crashes by each operator.
- **Operations:** The list of operation involves:
 - ✚ Generating a gg plot for line chart with years on the x axis and number of crashes on the y axis.
 - ✚ Making the line chart interactive by implementing a reactive slider that can vary the plot output with respect to time.
 - ✚ Each point on the plot displays a label on mouse hover, showing the year of crash and the number of crashes in that year.
 - ✚ Creating a collapsible tree for displaying the number of crashes by each flight operator.
 - ✚ Each click on the node opens up another node (user engagement) while the size of each node is proportional to the magnitude of crashes by each flight operator.
- **Focus/Parti:**
 - ✚ We can easily have a localized view of each data point and the time period using the slider and the labels options.
 - ✚ The leaf size of each operator easily distinguishes the operators with more number of crashes as compared to the other. Also the display of data on clicking draws user attention.

- **Discussion:**

- **Pros of using these plots:**

- ✚ The line chart is a good representation of trends and changes.
 - ✚ Line charts are very simple for the user to read
 - ✚ Collapsible trees promote user interaction.
 - ✚ Collapsible trees are good to show branched data.

- **Cons of using these plots:**

- ✚ For very sparse data points, line chart may not represent the actual values between two distant data points very well.
 - ✚ A collapsible tree may become very cumbersome for the user if the data represented is too much, since user has to click each node to explore the next node.

2.3 Sheet 3:

- **Information:** This sheet discusses the representation of percentage distribution of crash Fatalities across the world.
- **Layout:** Creating a tab named 'Fatalities' which shows a choropleth map of the world representing the percentage distribution of crash fatalities. The darkness of the colour palette is proportional to the percentage of the crash fatalities.
- **Operations:** The list of operation involves:
 - ✚ Generating a choropleth map for the world.
 - ✚ Adding boundaries for each country using the shape file data.
 - ✚ Highlighting each country on mouse hover, and add a label displaying the percentage of crash fatalities in that country, Year of maximum crashes for that country and the number of fatalities occurred in that year.
 - ✚ Data wrangling for extraction of the data shown on the labels.

- **Focus/Parti:**

- ✚ The highlight feature emboldens the country border and increases its opacity. User can easily see the crash fatalities' data for any country on a single map just by mouse hovering.

- **Discussion:**

- **Pros of using this plot:**

- ✚ The choropleth map is a very good representation of continuous data like percentages and ratios
 - ✚ User can access the world data in a very limited space and distinguish between the crash percentages with respect to color contrast.
 - ✚ The highlight and labels feature provides the user great data accessibility.
 - **Cons of using this plot:**
 - ✚ It can be difficult to distinguish between lighter shades of colour.
 - ✚ It assumes that whole region has a uniform value, but there could be some variations.

2.4 Sheet 4:

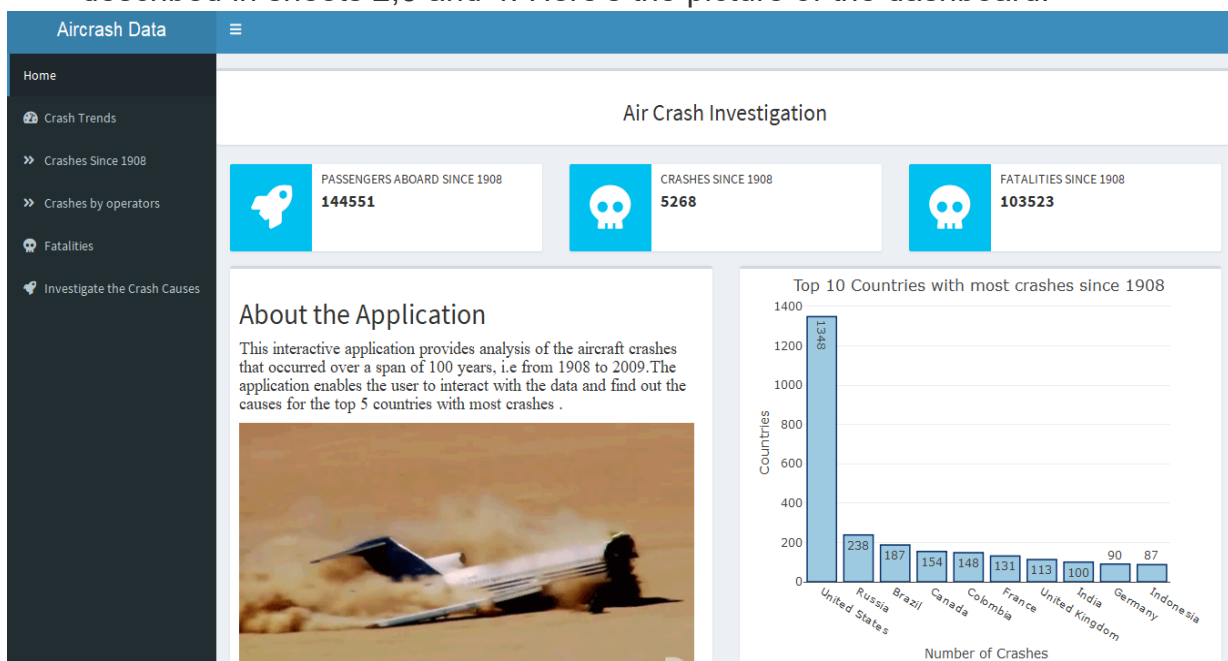
- **Information:** This sheet discusses the representation of the main causes of crashes across the world and in top 5 crash ridden countries.
- **Layout:** Creating a tab named 'Investigate the crash causes' which shows a Word cloud with a drop down menu to select the country to investigate and two slider controls, one for varying the words in the word cloud and the other for fixating the minimum frequency of the words.
- **Operations:** The list of operation involves:
 - ✚ Wrangling Data to extract the causes of the crash for each country and for the entire world.
 - ✚ Creating a word cloud by text preprocessing like tokenization, removal of stop words and including meaningful bigrams.
 - ✚ Adding a dropdown menu for selecting different countries to investigate.

- ✚ Making the word cloud dynamic by adding the slider control for varying the minimum frequency and the number of words in the word cloud.

- **Focus/Parti:** The interactive word cloud lets user explore the causes related to crashes by varying the frequency and the number of words in the word cloud. Furthermore, user can do this analysis for different countries by selecting the country of his choice from the drop down menu.
- **Discussion:**
 - **Pros of using this plot:**
 - ✚ The interactive word cloud provides a great deal of user involvement by the providing user an option to vary the frequency and number of words in the word cloud.
 - ✚ Word cloud is a very good tool to summarize vast amount of text data and hence it is perfect to represent the causes of the crashes.
 - ✚ It is very simple to understand.
 - **Cons of using this plot:**
 - ✚ It is almost impossible to remove all the non-context words.
 - ✚ It is not necessary that the word cloud data is able to show the entire picture correctly due to some inaccuracies.

2.5 Sheet 5: Realization

- **Information:** This sheet combines all the designs discussed in Sheets 2,3 and 4 on a single shiny dashboard.
- **Layout:** We create a dashboard which has tabs to access each of the designs described in sheets 2,3 and 4. Here's the picture of the dashboard:



- **Description:**

- ✚ The dashboard contains three tabs and 2 sub tabs (Crashes since 1908 and Crashes by operators under Crash Trends), each one corresponding to each of the design sheets 2,3 and 4.
- ✚ The front page is the home page for the app which displays some overall statistics of the data which are related to the plots linked to the side bar tabs.
- ✚ There is some text on the dashboard that gives an overall view about the application.

- **Software requirements:**

- ✚ The application was created by using R shiny dashboard
- ✚ Libraries Used: **shiny, shinydashboard, plotly, ggplot2, gridExtra, data.table, maps, mapproj, rgdal, leaflet, gpclibPermit, maptools, RColorBrewer, ggmap, ColorPalette, htmltools, tm, wordcloud, memoise, dplyr, collapsibleTree**

- **Estimates of Cost and Time:**

- ✚ R is an open source free language, so there was no capital involved in making of this application.
- ✚ The designing and implementation of this project would approximately take a week.

3. Implementation

Here is the sequence of steps followed in order to design this interactive application:

I. Implementing the dashboard (Home Page)

First a dashboard is designed by using the '**shinydashboard**' library in order to integrate all the visualizations on one platform. This dashboard consists of several tabs, each of which leads to different visualizations.

The first page of the dash board is the home page which consists of overall crash statistics, a barchart representing the top 10 countries with most number of

crashes and a brief description of the application. The home page is built to give the user an overview of the data and introduce the purpose of the application

II. Implementation of the “Crash trends” tab

The Crash trends tab contains two sub-menu items:

- i. **Crashes Since 1908:** This tab displays a line graph showing the crash trend from 1908 to 2009. The line graph is implemented with a slider control that enables the user to analyse any localized time period between 1908 to 2009.
- ii. **Crashes by operators:** This tab is designed to see the top 10 flight operators that were responsible for the most number of crashes. It represents this via a collapsible tree using the **collapsible tree** library. Each node of the tree represents the flight operator and the radius of each node is proportional to the magnitude of the air crashes committed by each flight operator. User has to click on each node to explore the associated node data.

III. Implementation of the “Fatalities” tab

After visualizing the data for the crash trends, we must know how much loss of lives was caused by these crashes world-wide. So, in order to visualize the percentage distribution of the crash fatalities worldwide, we decide to implement an interactive a choropleth map. This map highlights the country region over which the user places the cursor and shows the label that displays the percentage of crash fatalities in that country, Year of maximum crashes for that country and the number of fatalities occurred in that year.

The following libraries were used for the implementation of this choropleth map: **maps, mapproj, rgdal, leaflet, gpclibPermit, maptools** and **RColorBrewer**.

IV. Implementation of the “Investigate the Crash Causes” tab

In this section, we let the user to explore the causes of the air crashes by implementing a dynamic word cloud, whose attributes can be varied by the user according to his/her needs of exploration. The word cloud displays the air crash causes for the entire world and for the top 5 most crash ridden countries. The minimum frequency of the words and the number of words in the word cloud can be varied by the slider controls given for each of them. This provides great deal of user engagement and makes it intriguing.

The following libraries were used for the implementation of interactive word cloud: **tm**, **wordcloud**, **memoise**.

V. List of all the Libraries used

The following libraries were used for the implementation of this app:

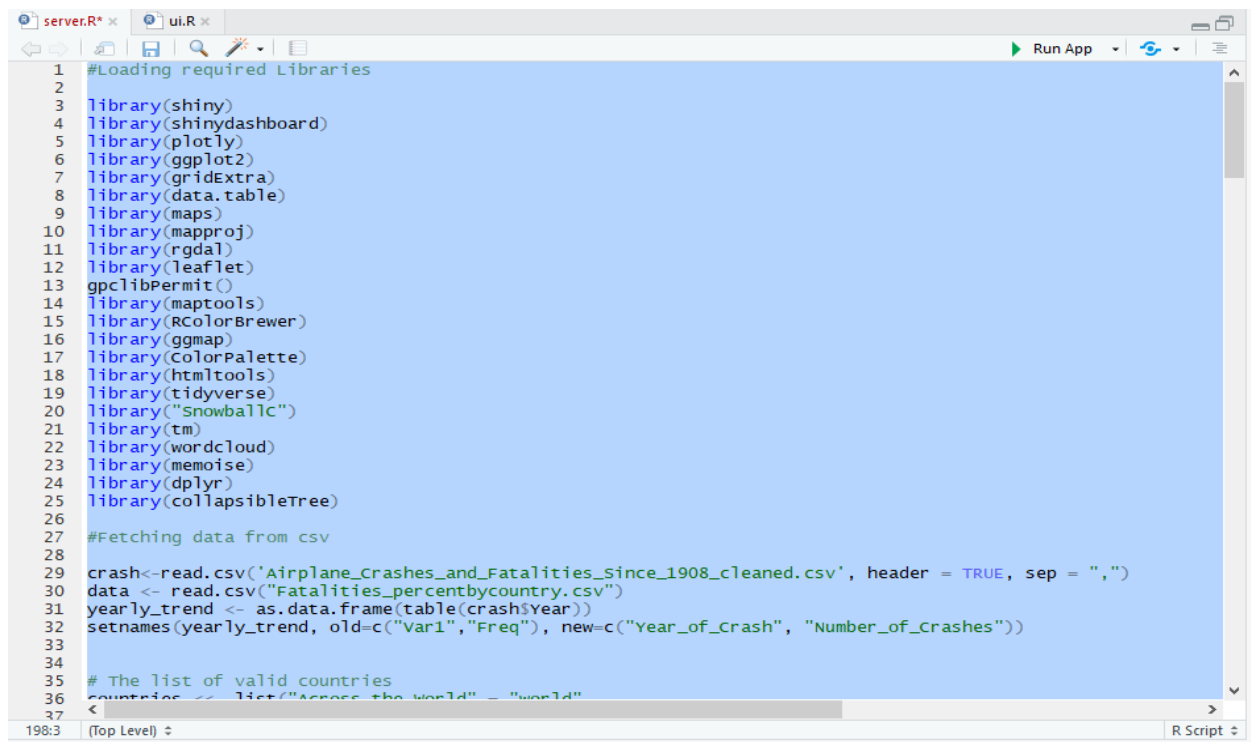
shiny, shinydashboard, plotly, ggplot2, gridExtra, data.table, maps, mapproj, rgdal, leaflet, gpclibPermit, maptools, RColorBrewer, ggmap, ColorPalette, htmltools, tm, wordcloud, memoise, dplyr, collapsibleTree

4. User Guide

***Note: The application is designed with respect to the Aspect ratio of Monash Systems. So Please run it on a Monash System Screen to get perfect view of tabs.**

Here are the instructions for viewing and exploring this narrative visualization application:

- i. Install all the required libraries mentioned in the list above.
- ii. Open server.ui and select the entire code and press **Ctrl+Enter**

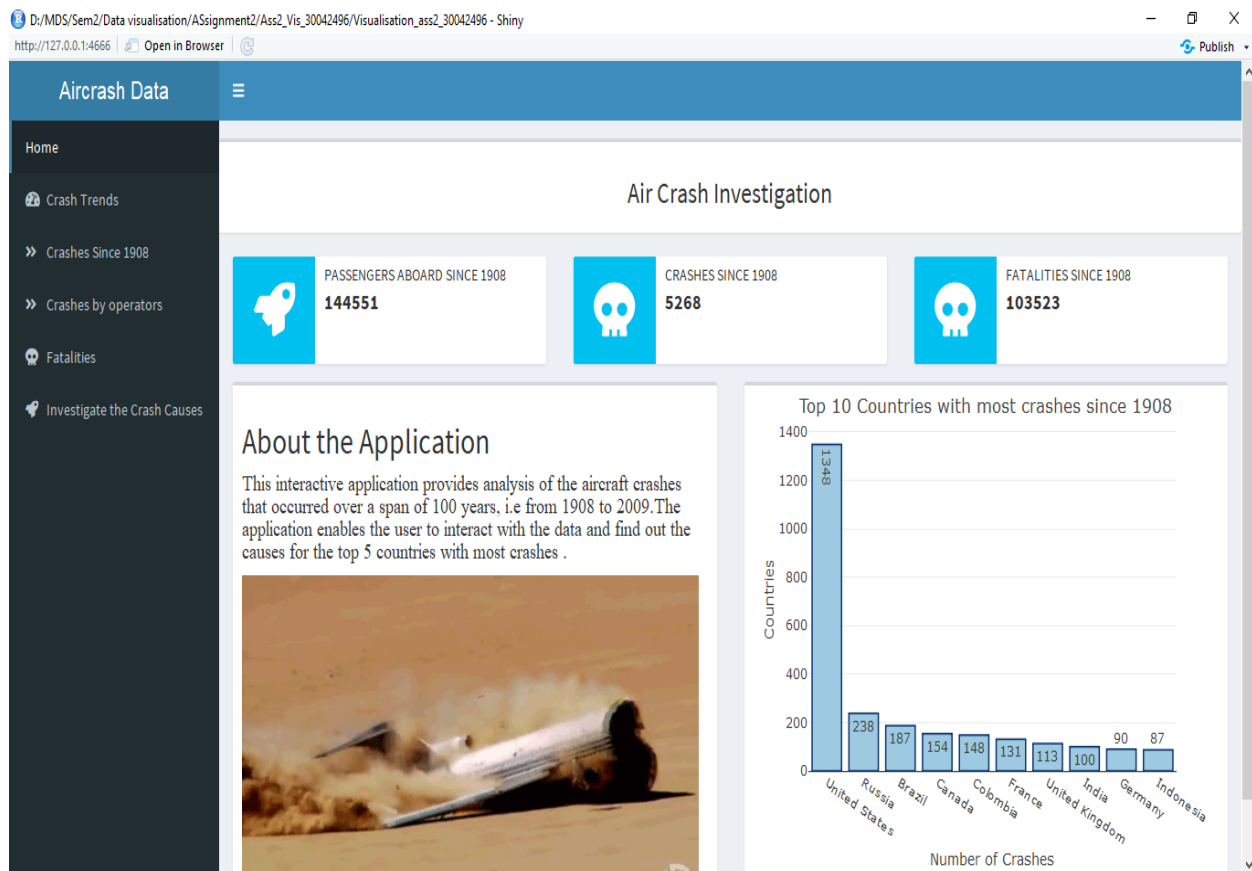


```

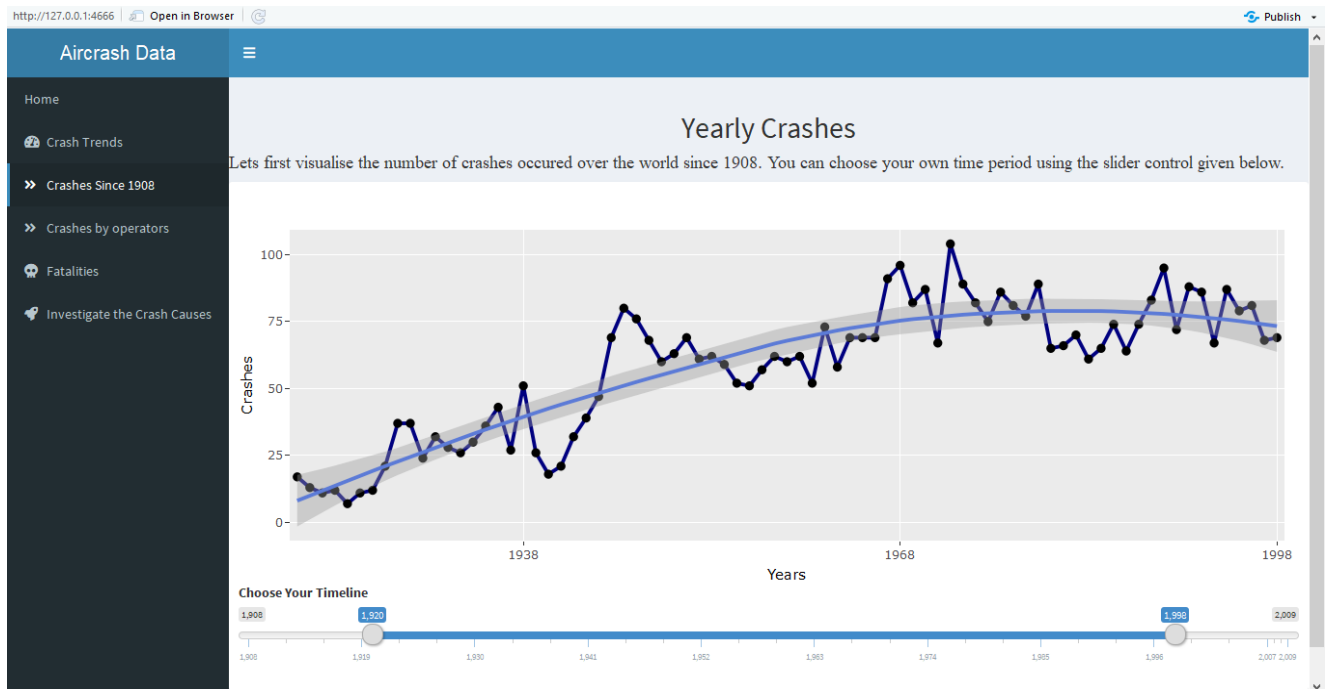
1 #Loading required Libraries
2
3 library(shiny)
4 library(shinydashboard)
5 library(plotly)
6 library(ggplot2)
7 library(gridExtra)
8 library(data.table)
9 library(maps)
10 library(mapproj)
11 library(rgdal)
12 library(leaflet)
13 gpclibPermit()
14 library(maptools)
15 library(RColorBrewer)
16 library(ggmap)
17 library(ColorPalette)
18 library(htmltools)
19 library(tidyverse)
20 library("SnowballC")
21 library(tm)
22 library(wordcloud)
23 library(memoise)
24 library(dplyr)
25 library(collapsibleTree)
26
27 #Fetching data from csv
28
29 crash<-read.csv('Airplane_Crashes_and_Fatalities_Since_1908_cleaned.csv', header = TRUE, sep = ",")
30 data <- read.csv("Fatalities_percentbycountry.csv")
31 yearly_trend <- as.data.frame(table(crash$Year))
32 setnames(yearly_trend, old=c("Var1","Freq"), new=c("Year_of_Crash", "Number_of_Crashes"))
33
34
35 # The list of valid countries
36 countries <- list("Across the World" = "world")
37

```

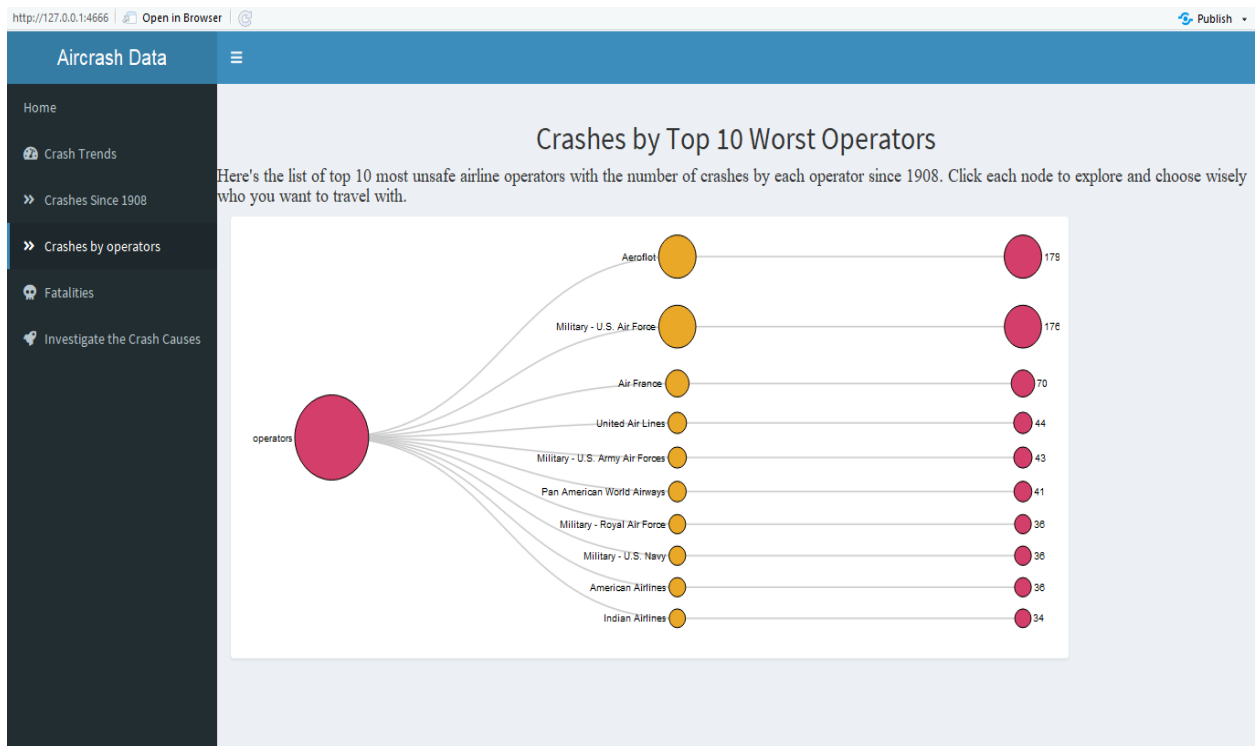
- iii. Now the application is ready to run. So, click on the Run app button
- iv. You will see the home page which shows the overall statistics of the aircraft crash data, an interactive bar plot showing the number of crashes in top 10 crash ridden countries from 1908 to 2009 and a brief description about the application.



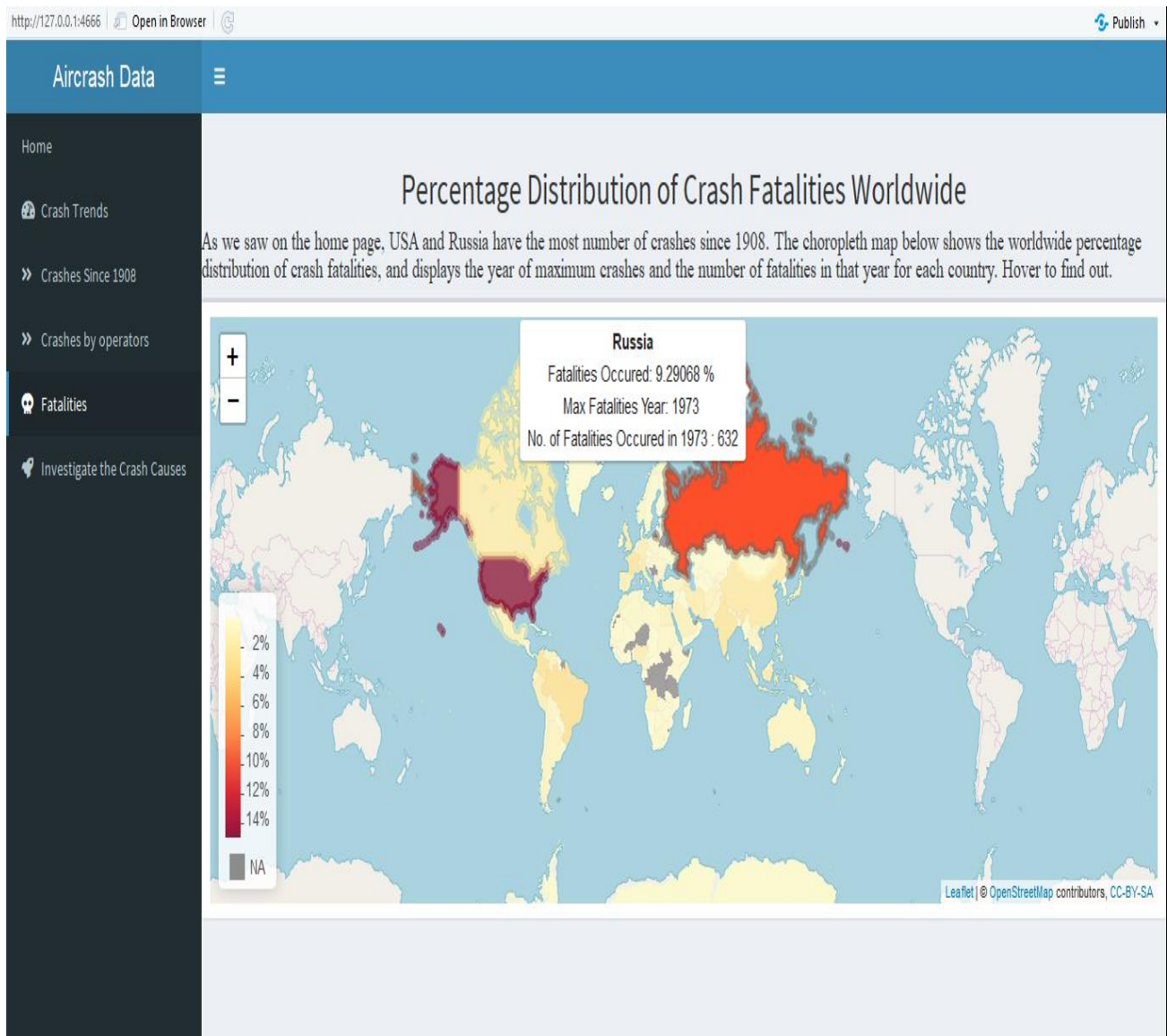
- v. Click on the tab named “Crashes since 1908”. It shows the yearly crash trend from 1908 to 2009. You can vary the timeline using given slider control.



- vi. Now click on the next tab named “Crashes by operators”. This shows a collapsible tree where the user has to click on each node to find out the number of crashes caused by each flight operator. The radius of each node is proportional to the number of crashes by each operator.



- vii. The next tab is the “Fatalities” tab which shows the percentage distribution of crash fatalities world wide. Hover on the different countries to find out the details of fatalities for each country.



5. Conclusion

- ✚ The intended audience for this interactive application was the general public. Hence, the challenge was to implement the application in such a way that it presents the visualizations that are easily understandable and engaging at the same time.
- ✚ The application is able to achieve both the goals. We produced the plots that were easy to interpret by the user and they are engaging user to play with the data at the same time. For eg: the word cloud implementation, where the user can explore the crash causes for several countries and vary the number and frequency of words.
- ✚ The application also succeeds to provide a deep insight for a vast amount of data in a very compact and concise visualization. For eg: A single choropleth map conveys the data about the crash Fatalities over entire world by just hovering the cursor on any country.

- **Reflection**

- ✚ The project introduced the new application building platform known as r shiny. It made me learn about building reactive plots using R and also strengthened the concepts of Data wrangling using Python.
- ✚ I got to learn and explore about several advanced data visualization plots like the tree maps, choropleth maps, chord diagrams, Radar charts, etc .
- ✚ The plots generated in R shiny could have been made more interactive by using the D3 platform. Hence the application could have built using D3 for achieving better results.
- ✚ Another good Idea would have been make a comparative analysis of each flight operator based on the crash trends for each of them.

6. **References and bibliography:**

The following learning resources were used in the making of this project :

- <https://shiny.rstudio.com/gallery/>
- Lecture and Lab Materials from Week1 to Week 12
- <https://www.kaggle.com/ruslankl/airplane-crashes-data-visualization/data>
- https://rpubs.com/apubh/airplane_crash_analysis

7. Appendix

Sheet-1 IDEAS SHEET

* IDEATE :

(i) Choose platform → R shiny → More Like an Application
 → D3 → More interactive
 → Web Page

(ii) Plots: Line charts, heat map, choropleth map, Bar chart, scatter plot, word cloud, Box plots, Chord map,

* FILTER :

(i) Choose Platform → R shiny Dashboard
 ↳ Use of R2D3 Package

(ii) Plots :

- ↳ Plots For Discrete Data: Scatter plot, Box plot, Line charts, bar graph.
- ↳ Plots For Continuous Data: Choropleth, word cloud, Histogram

* CATEGORIZE :

↳ CRASH TRENDS → Yearly
 → Monthly

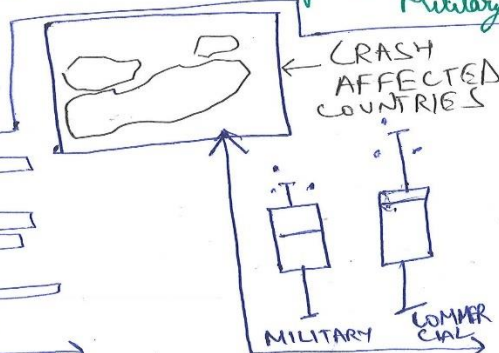
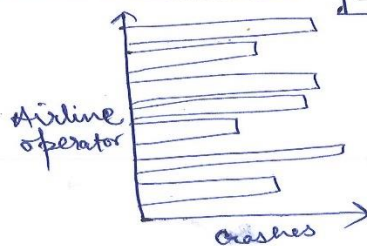
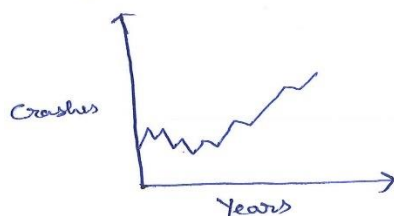
↳ Fatalities % age

↳ Top 10 Worst operators

↳ Top 10 Crash affected Countries

↳ Causes of Crashes
 ↳ Military Vs Non Military

* COMBINE :



Questions :

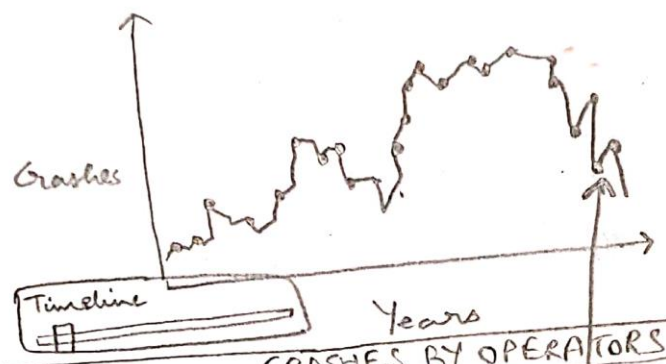
- ↳ How crash trend varies from 1908 to 2008?
- ↳ Which countries have worst rate of Fatalities?
- ↳ Which flights operators are worst to travel?
- ↳ What are the most common causes of Aircrashes?

Ideas to carry forward → ①, ③, ④

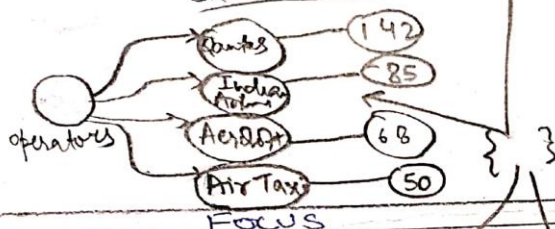
SHEET 2

LAYOUT

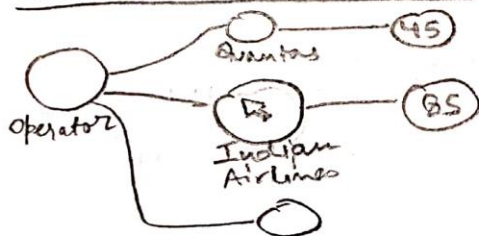
TRENDS OF AIRCRASH



CRASHES BY OPERATORS



Click Each Node to Explore



Mouse-over



No. of Crash in the : 52
year
% age of fatalities : 20%

INFORMATION

TITLE	TREND OF AIRCRASH
AUTHOR	Rohan Singh (30042496)
DATE	28/5/2019
SHEET	2
TASK:	To visualise the crash trend on user specified time period

OPERATIONS

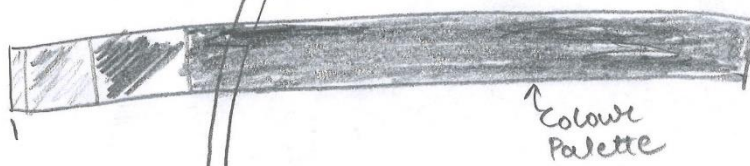
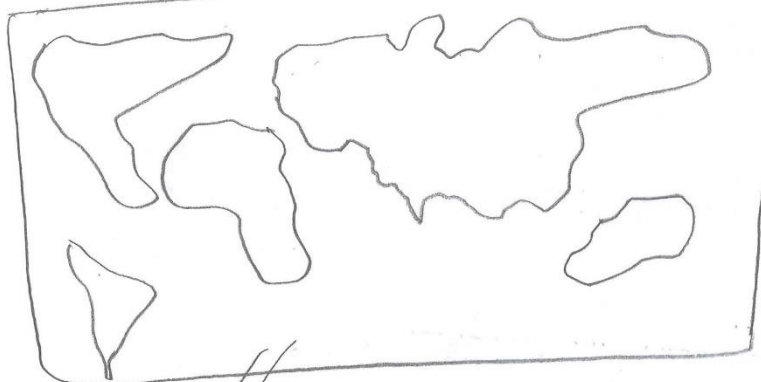
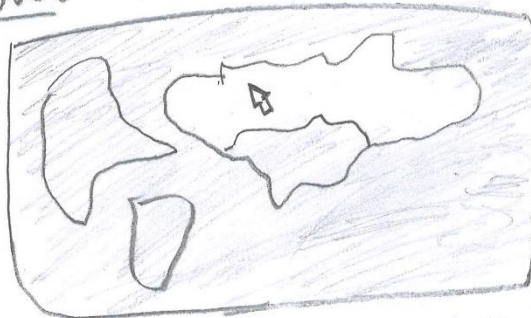
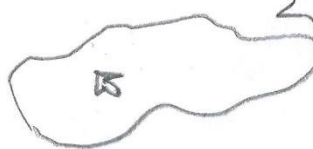
- ↳ Draw a line chart using ggplot to display yearly trends of crashes
- ↳ A slider shows trends for a specific time period.
- ↳ Implement Mouseover on each point on line.

DISCUSSION

- Using slider for Years axis provides user interaction
- The Mouse over feature gives more information about each data point
- Nodes have to be clicked to access Data.

SHEET 3LAYOUT

Percentage of fatalities all over the world

FOCUS• On Mouseover• On Mouseclick → (click brings up details)

%age of fatalities: 15%
 Max Fatalities Year: 2005
 Min Fatalities Year: 2000

INFORMATION

TITLE	Percentage of Fatalities all over the world.
AUTHOR	Rohan Singh (30042496)
DATE	28/05/2019
SHEET	3
TASK	Visualise the percentage of fatalities all over world

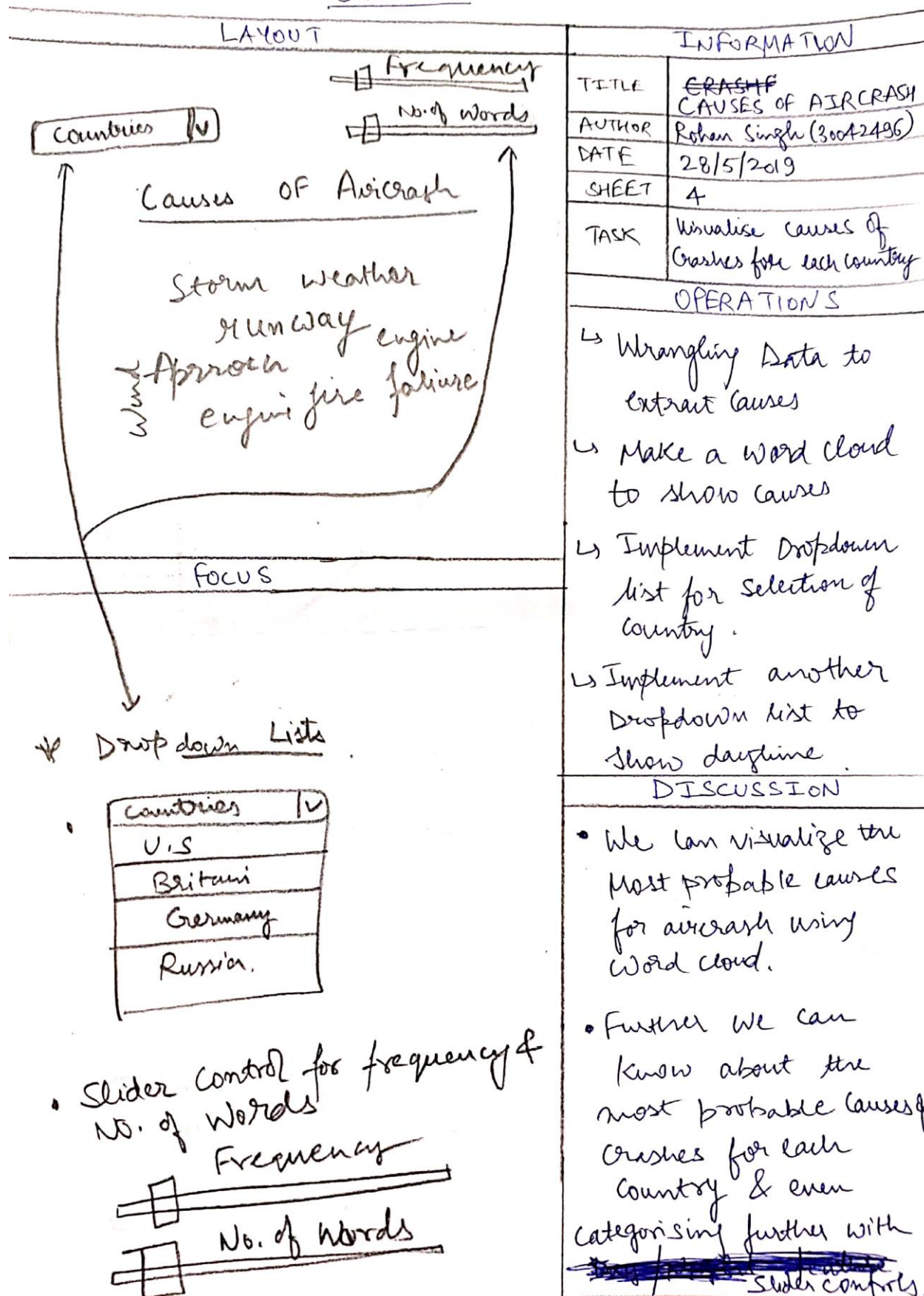
OPERATIONS

- ↳ Build a choropleth Map to show Fatalities percentage
- ↳ Mouse over function
- ↳ Mouse click function
- ↳ Vary the Opacity of regions.

DISCUSSION

- A choropleth Map effectively shows continuous data
- On mouse click over any country we can get details of %age of fatalities,
 - Maximum Fatalities year for that country
 - Minimum Fatalities Year for that country.

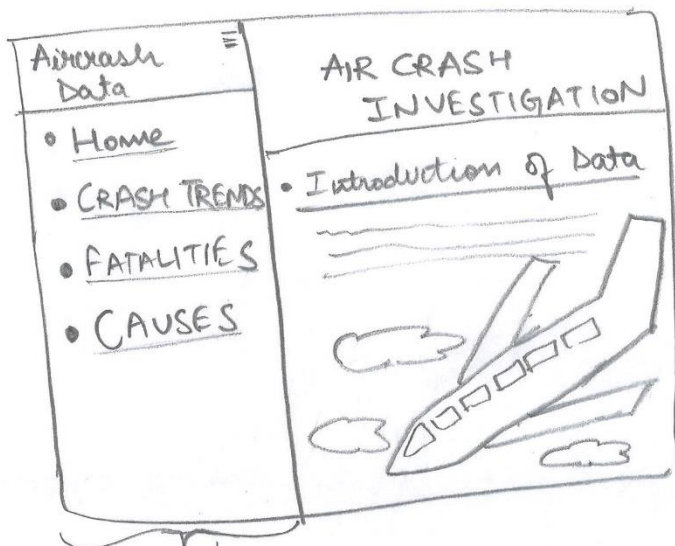
SHEET 4



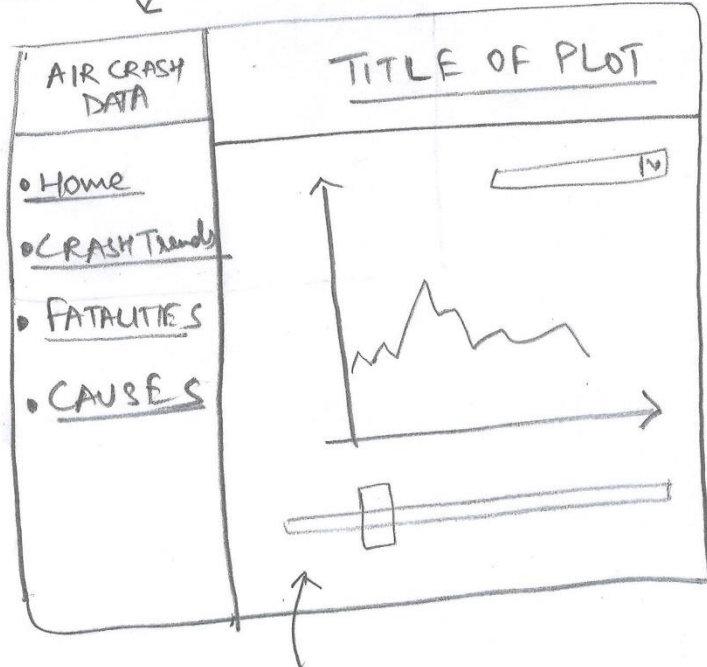
SHEET 5: REALISATION

FINAL LAYOUT

• DASHBOARD



• LINKS



Plots From
Sheets 2, 3 & 4

Information

TITLE	AIR CRASH INVESTIGATION
AUTHOR	Rohan Singh (30042496)
DATE	28/5/2019
SHEET	5 → Realisation
Task	Combine / Integrate the plots into one single application

Description

- ↳ Use Shiny Dashboard for integration
- ↳ Display details about data (Introduction)
- ↳ Provide Links to plots on the Left.

Software requirements

- ↳ Using R
- ↳ Libraries
 - Shiny Dashboard
 - ggplot
 - Leaflet
 - word cloud

Estimates of Cost & Time

- ↳ Integration of the plots into one single dashboard
- ↳ Efficient way to show plots