INM433_PRD1_A_2016-17 Coursework

Karunakar Kotagaram, City University London, UK

Project Title: Airplane Crashes explanation geographically across the Globe

Figure (1): Common factors causes to aircraft accidents

Abstract — The aviation safety aircraft accident database contains each aircraft investigation details that has occurred from 1900 to till present. This project work is to visualise the data with analytical techniques to search aircraft accident trends, correlation, clustering, and real factor relationships in the accidental data. This analysis is chronological approach, doing analysis based upon historical crash data dividing into different geographical epochs to predict which aircraft, operator, region, country is in most safe or least safe. The goal of analysis is mainly to visualise the factors that causing aircraft crashes.

1. MOTIVATION, DATA, RESEARCH QUESTIONS

This analysis was motivated by the premises that how the research to be effectual based on crash frequency from past decades make to follow the better safety standards. Recently, the shrewdness of airplane crashes is dreadful in the aviation history. When we plan to travel by airlines, first we think about which airline is very safe and will also analyse the statistics about that airline type and operator. In the recent air crash accidents like Malaysian plane MH17 accident occurred in Ukraine country border, and missing MH370 is still mystery, we can't believe about airplane journey is very safe. Predicting the factors for airline accidents is very complicated so if we have basic information about airline that helps us to find which aircraft is very safe to travel. I work on this project with intention of finding the factors causing the aircraft incidents over the period and safety information for passenger for each region of the world by doing the explanatory analysis by using visualisation techniques.

Catastrophic accident was happened in November 2016 for Colombia plane carrying Chapecoense football team. Overall plane crash rates in the year 2016 far better than to previous years. We need to look at too many factors including aircraft condition, amount of fuel, pilot's behaviour, the interaction with air-traffic controllers and the predominant weather etc. Based on root cause analysis of previous accidents and to take precaution of aviation safety, there must be some real factors should be learned that will be helpful to make aircraft flying safer in the future.

The source of original dataset was collected from aviation safety network databases. The dataset which I have used is collected from Aviation Safety Network website. ASN is established in year 1996, the information was retrieved will be in the form of CSV which contains 4216 observation of 22 variables.

Currently available vast amount of data in the website is about airline accidents in the tabular format but they are not clearly visualising the accidents data with different key factors. The goal of this aircraft data visualisation will help us to answer the following research questions.

- What are the factors best describing geographic differences for the airplane accidents?
- Do these factors describes geographic differences in accidents across the globe?
- Which region, year, location, country, operator, and type of airlines have had a higher chance of accidents/fatalities all over the world?
- Identify which airline operator, route and type is safety to travel?

2. TASKS AND APPROACH

2.1 Identify the probability of plane crashes over the period by using line graph, pie charts and bar charts.

In the database, there are more than 4126 observations with 22 attributes, most of the observations have had multiple relationships to each attribute, using single select SQL query or other tools is not straight forward and it can be difficult to handle it. In order to overcome this situation, we need to use the computational methods to visualize the data with mathematical calculations using line graph (Figure 2), bar chart (Figure 5 & 6), pie chart (Figure 3 & 4). The data visualization charts were created by using Python, R, and Tableau that how the data correlated to airline accidents.

2.2. Predicting accidents by using linear model

Designing and to explore the aircraft dataset using visualization tools at once is not an easy task. This project explains how to pull the related attribute details and investigate for probability of trends and correlation between the event of accidents. Visualization approach is best for investigation and to understand what factors influence that increase the probability of accidents. Linear regression model is the best fit (Figure 8 & 9) to predicting the incidents based on the year.

2.3. Visually characterise aircraft accident based on location.

We do geospatial analysis to estimate and show the routes that planes usually met with accidents based on geo locations (longitude and latitude values). Accident locations have been divided in to different regions on the map, each region has its own colors and the size of the place represents the amount of crashes have been happened within that region. This analysis will help to have a quick view of plane information from multiple regions (Figure 10, 12 & 13).

2.4 Identify the most common factors causes of aircraft accidents using statistical methods.

As per historical information plane crashes be likely to break down into five main categories such as pilot error, weather, mechanical failure, sabotage, other ways of human error. It is always challenging to decide that the accident is related to a single factor but it also will have multiple chain of events, to analyse these chain of events is puzzling task (Figure 14, 15 ,16 &17).

2.5. Visualising the factors that influences accidents using clustering techniques.

Airline travelling is the safest comparing to other modes of travel but sometime it has risk as accident can occur any time due to combination of factors like bad weather, human intervention, fault of the pilot, and malfunction equipment. Using clustering techniques data can be grouped in similar categories to identify the factors influencing the fatal crashes and fatalities (Figure 18,19 & 20)

3 ANALYTICAL STEPS

3.1 Aviation accidents historical trends.

3.1.1 Accidents trend by year, month, day

First, we will need to break the aircraft historical data down by year. The below graph shows the number of fatal plane accidents per year, month, and day. From the below visualisation, we can observe that plane crashes trend decreased since 1970. The crash data by month trend shows December month has the highest plane crashes occurred, and April month has the lowest accidents. According to the trend observed by days, 31st is the lowest and 8th is the highest crashes rate.

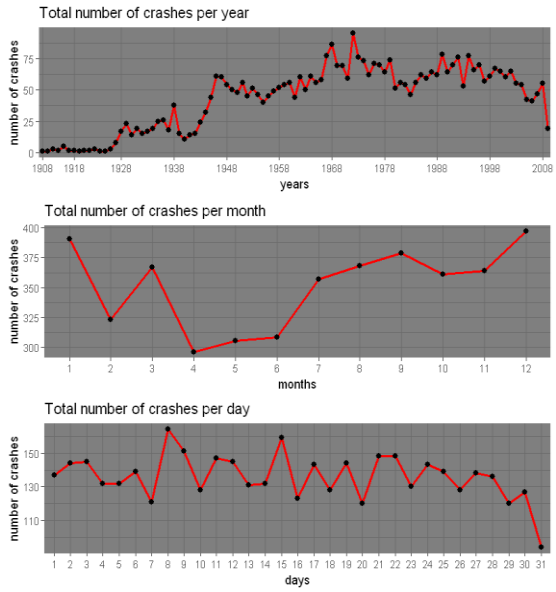


Figure (2)

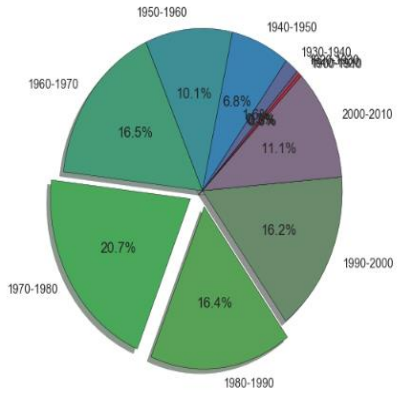


Figure (3): Accident rates by year (1920-2010).

Figure (3) shows the number of accident rates from the period of 1920-2010, the highest accident rate is 20.7% during 1970-1980. However, in the last 40 years plane crash rate brought down due to the development of new technology in the field of aviation safety.

3.1.2 Accidents by region, country, operator, and type

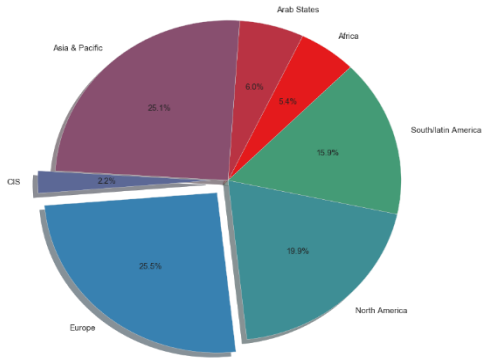


Figure (4)

Figure (4) shows the number of accidents with probability chance rates by the region. After normalising the data, it concludes that highest accidents rates occurred in USA (North & South American region).

Figure (5) shows the top 10 countries involved in the highest rate of accidents. US has the highest number of accidents 1321 and Russia is the second highest of 170, and Brazil is in third highest of accidents in total of 158. However, the lowest number of accident and fatalities are Australia and Guinea.

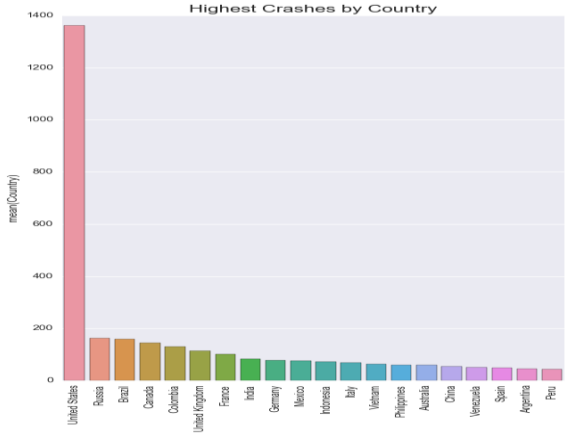


Figure (5)

Figure (6) shows the top 10 operators involved in aircraft fatalities and accidents. Each wedge represents the number of incident for each type of aircraft, bigger slice represents the higher volume of accidents. The total number of aircraft accidents was normalized by operator by diving with number of fleets. US Air Force has highest accident rates, this operator probably due to the age of the aircraft or war or lack of maintenance.

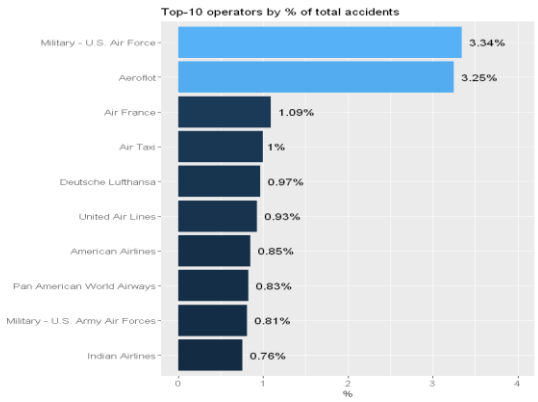


Figure (6)

Table 1 Shows top 5 Aircraft accidents by operator

Aircraft operator	***Number of accidents***
Military-US Air force	*141*
Aeroflot	*137*
Air France	*48*
Air Taxi	*44*
Deutsche Lufthansa	*43*

Table (1)

Figure (7) shows the airline type Douglas DC-3 has highest probability of plane crashes, DHC-8 is the second

highest one, whereas Douglas DC-8 is the lowest accident rate.

Type	*Accidents*	*Probability*
Douglas DC-3	*275*	*6.54%*
DHC-6 Twin Otter 300	*66*	*1.57%*
Douglas C-47A	*58*	*1.48%*
Douglas C-47	*46*	*1.09%*
Douglas DC-4	*34*	*0.81%*

Table (2)

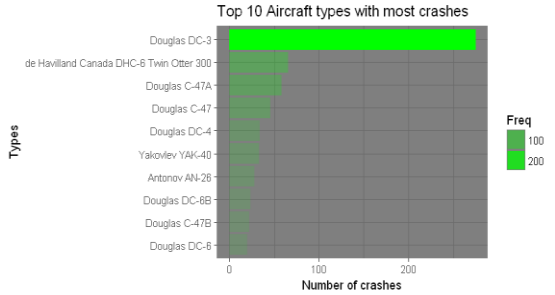


Figure (7): Top 10 aircraft types by percentage of total accidents.

3.2 Predicting Accidents by using linear model

Linear regression is most important statistical method for estimating the accidents by year. Here accidents are dependent variable and year is independent variable. Ordinary least square method was used to estimate the aircraft accidents. The formula Y = M * X + B. Regression model can be determined based on R square value, here R square values is approximately 0.93% so we can conclude that our model is best fit.

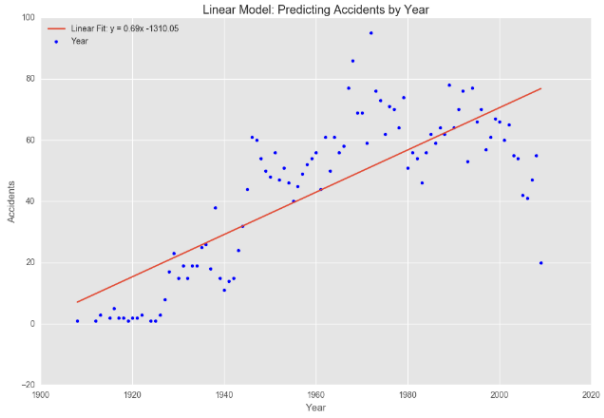


Figure (8) linear regression model predicting accident by year

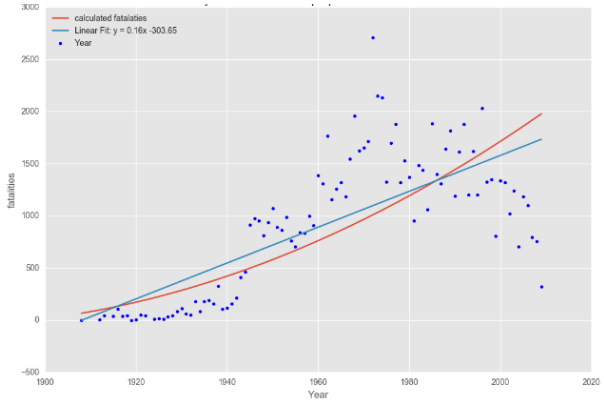


Figure (9) Calculate fatalities based on year and accidents using linear model

3.3 Visual analysis on plane crashes based on locations

3.3.1 Accidents based on locations

To find the location of aircraft crashes we should use the longitude and latitude for getting the flight from aviation database. Figure (10) shows the approximate locations of the plane crashes across the world.

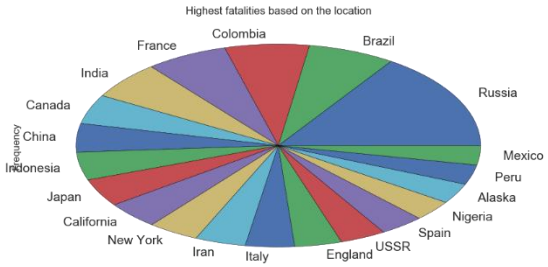


Figure (10)

Top 5 dangerous locations

Location	*No of accidents*
Sao Paulo, Brazil	*15*
Moscow, Russia	*14*
Manila, Philippines	*13*
Anchorage, Alaska	*13*
Bogota, Colombia	*12*

Table (3)

Figure (11) shows the location of crashes across the world that are highlighted in the green color dotted spots.

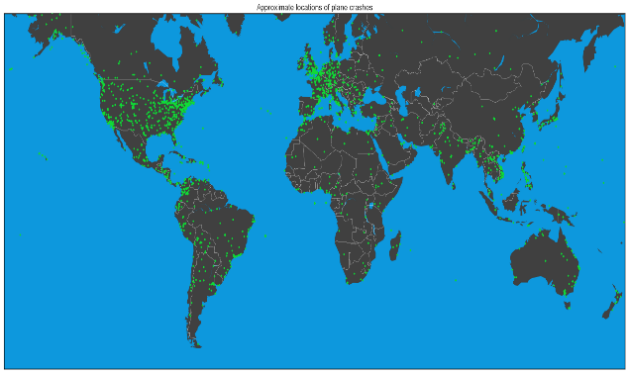


Figure (11)

Top 10 countries based on accidents

USA	***1321***
Russia	*170*
Brazil	*158*
Canada	*143*
Colombia	*129*
UK	*109*
France	*102*
Mexico	*100*
India	*83*
Germany	*76*

Table (4)

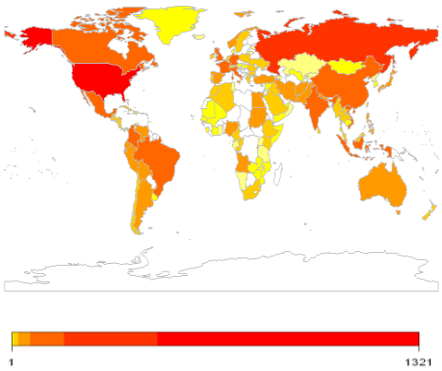


Figure (12)

We are considering 25 geographical regions from the dataset since 1908. Figure (12) shows that countries shaded by the frequency based on number of incidents. Here, USA has the high volume of accidents as 1321 so it has shaded with red color. Nigeria has the smallest number of accidents as 39 so this region has shaded orange color on the world map. In second place, Russia has highest number of accidents is 170, and third place is Brazil, it has 158 accidents. The geographic figure shows aircraft accidents from past 100 years, this chart can quickly get the plane crash information for specific years as well as operator and types of the airlines.

3.3.2 Most dangerous departure and destination cities.

Analyse the number of crashes in a city as either arrival or departure flights, perhaps this help to find the most dangerous route and cities of the location. To find the dangerous route first we should normalise to the number of flights arriving and to departures between cities. After observing the below table statistics based on ranking we can conclude that that New York city route has more risk factor.

Top 5 risk factor routes are in the below table.

Count	***Departure City***	***Destination City***
106	*New York*	*Paris*
102	*Paris*	*Prague*
96	*Lakehurst, NJ*	*s.t louis, mo.*
83	*lympne, England*	*Rotterdam, the Netherlands*
80	*Toulon*	*Algiers*

Table (5)

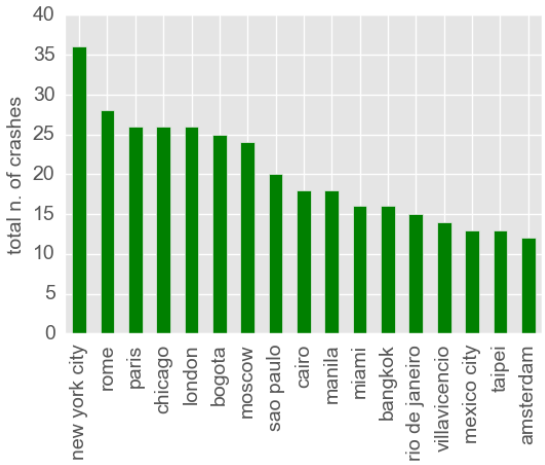


Figure (13)

3.4 Causes of aircraft accidents.

Figure (15) shows the percentage of each factor causes for aircraft accidents. However, accidents can occur a result of a mixture of several causes so the cumulative count is not equal to number of accidents. In the pie chart *figure (15)* show 57% of accidents occurred (in total of 1558) due to pilot error. In other causes, the second highest percentage is engine failure caused 369(13%), the poor weather conditions in total of 358(13%).Therefore, pilot error, poor weather, and engine failure accounts for more than 75% of all the aircraft accidents.

Factors cause for accidents	**Number of occasions**
pilot error	1558
engine failure	369
poor weather	368
stall	248
turbulence	106
on fire	105

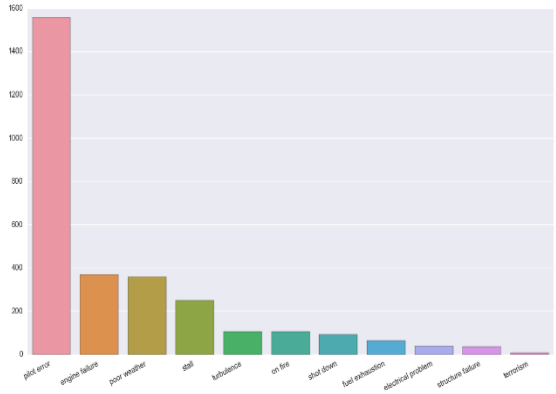


Figure (14) top 10 factors causing plane crashes

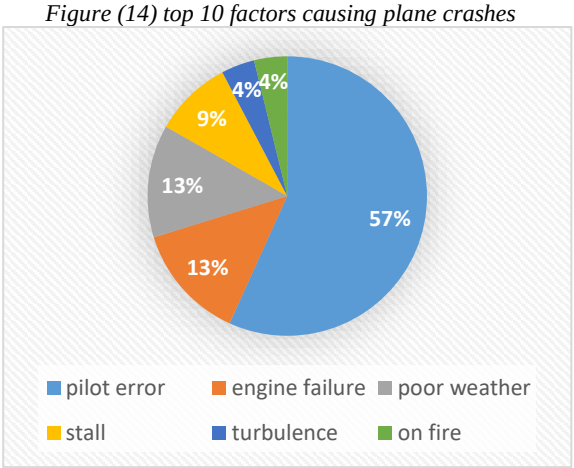


Figure (15)

During the take-off and landing, pilot error can be classified into four factors are operational error, judgement error, procedure error, misguidance error. Operational errors normally include lack of information in declaration, whereas judgement error late decision making to abort take-off and landing. Procedural error is some mistakes to follow the procedure, misguidance can occur while doing the practice in training.

The breakdown the factors that cause of pilot error as below figures shows, 328 procedural errors (26%), 356 operational errors (23%), 624 judgment errors (28%), and 250 misguidances (23%).

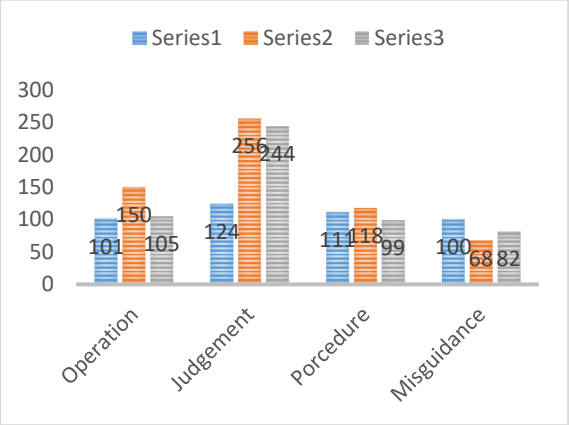


Figure (16)

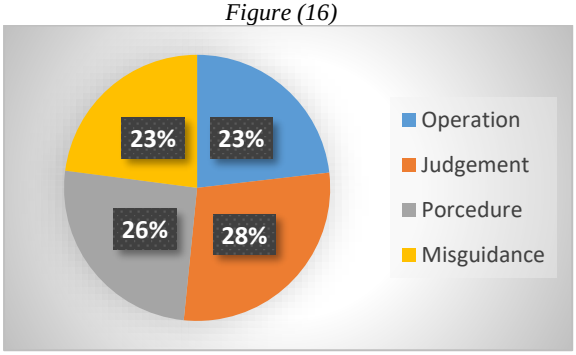


Figure (17)

3.5 Evaluate factors in generalisability of explanatory models using K-means clustering techniques.

We can generate the groups for each similar factor by applying K-means clustering techniques. In R used Factoextra package to print the list of clusters based on each factor. Figure (18) shows how far each cluster comparting to others by using Euclidian distance matrix. With this initial part of analysis, we can identify some groups such as pilot approach of the runway, engine, weather conditions, altitude are most like to be causes of aircraft accidents.

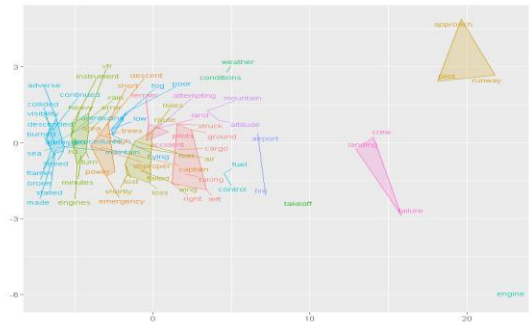


Figure (18)

Based on the factors on the figure 18, pilot error is the highest correlation factor so human error caused by pilot factor is constant with the fact of 57% percent of aircraft crashes.

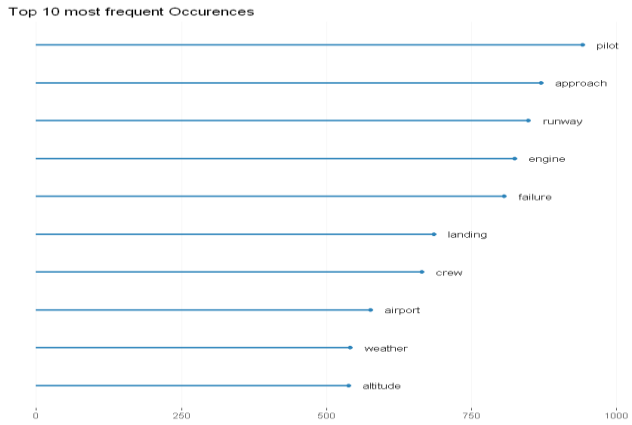


Figure (19)

We are dividing the data into 7 (C1 is for Cluster1 etc.…) different groups by using K-means clustering model. Table (10) shows the number of aircraft crashes and number of fatalities for each cluster

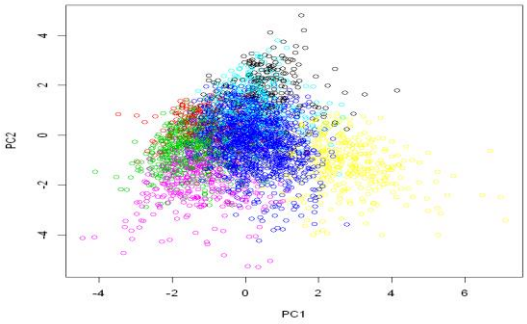


Figure (20)

Figure (20) shows the clustering the data by using PCA (principal component analysis). Here we have specified number of clusters to be K = 7, found the different values for each cluster starting from C1, C2, C3 etc.…. To visualise and evaluate the data point can be used PCA method, the figure (20) shows the result on PC's as seven different clusters such as blue, yellow, red, orange etc. For each cluster, we can summarise data based on different factors and can be identified the causes of the crashes.

	C1	C2	C3	C4	C5	C6	C7
Crashes	252	355	541	2026	369	329	358
Fatalities	4516	7609	10442	44167	7648	8199	5185

Table (6)

Example: Cluser5(369): Plane crashed by engine failure

4. FINDINGS

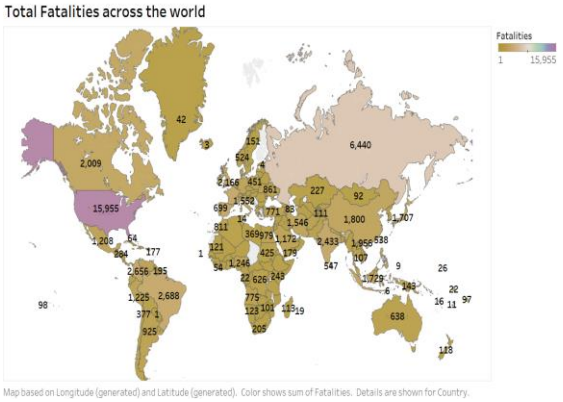
We have analysed the past trend of accidents occurring across the world. Also, we have used the mathematical methods. This study concerned on factors that relates to the aircraft accident over the period, country, route, type, operator and with geographical location. A total of 4126 aircraft accidents were used. Finding proves that US region has highest accident rates. Pilot error is highest number contributing factor for aircraft accidents. This study also discovered correlation between number of

plane crashes and operator, type of airline and region on world map. Remain a question what is best way of maintaining the aircraft safety, it is crucial factor while travelling for the people to trust the safety of airline regardless of locations, routes, regions of the world.

We examined the flight data visualization in Figure 15, there is 57% chances causing the fatal accidents due to pilot error. The second highest chance of fatal incident (13%) is due to bad weather condition and engine failure. Also, we determined most dangerous routes that shows in figure (13) is from New York city to Paris on the historical aviation dataset. Analysed the most dangerous airline type and operator based on country, region, and year. Figure (6) shows the Military-US Air force is the danger operator and Figure (7) shows the Douglas DC-3 is the type of airline that has highest probability of accidents (*6.54%).*

5. CRITICAL REFLECTION

Our goal is to find statistics of air crash accidents. Visualising the data in better way that helps to find out what factors are influencing to increase the likelihood of fatal aircraft incidents. There are many reasons involved with airline incidents so visualising the information in different perspective with different patterns is most important. As per our analysis most of the aircraft accidents were triggered by pilot error but other factors like weather, engine failure, air traffic controller error, lack of maintenance factors also contributing significantly. In past decade history says advanced technology or tools which will helps to find the root cause of the fatalities/accidents to decrease the airline incidents. However, airline accidents continuing in some regions, for example MH370 missing flight still mystery so in this area further investigation and other new technology or tools required to prevent this to be happened again in forthcoming days. In these cases, visualisation tools and techniques can be used broadly to look the data in diverse perception. Apart from the above, weather conditions will also be involved in major influence like Air Asia.



Catastrophes accident was brought down due to bad weather so to handle this kind of situation air craft designers can be used this visualisation information to build the new aircraft and to give better training to the pilot to handling these situations if it occurs in the future. It is also very important to determine the past incidents to improve the aviation safety. Pilot error mostly occurring when approaching and landing so pilot's workload should not be increased as it has highest chances of accidents can occur. During the approaching and landing must approach properly and give them best training that will help pilots to communicate effectively with air traffic controller to deal with critical weather conditions.

This project work has identified some factors that caused of aircraft crashing but it is not a common occurrence and estimated of crashes. There are number of factors that will applicable for other domains like air temperature but there might be other reasons some variables are not clear to apply in the other domains. This analysis is chronological approach as we are doing analysis based upon geolocations and faceting scatter plot by the regions. In some other domains, this data may be more important or may not useful for other analysis. The data was collected is very limited and it is aviation historical data. By using the linear regression model, it is bit complex to make conclusion of predicting air flight accidents as it can't give accurate results. For analysing historical aviation data, Data Mining techniques (like neural networks) will be the suitable for the aircraft investigation.

REFERENCES:

1) Airplane Crashes and Fatalities Since 1908 | Socrata. Available at: https://opendata.socrata.com/w/q2te-8cvq/y34g-bnf3?cur=lF2pfOEgPdy&from=root
2) Aviation Safety Network > Available at: https://aviation-safety.net/
3) Chapecoense plane crash: The victims, the survivors and those left behind - BBC News. Available at: http://www.bbc.co.uk/news/world-latin-america-38155840
4) Database index. Available at: http://www.planecrashinfo.com/database.htm
5) How odd is a cluster of plane crashes? - BBC News. Available at: http://www.bbc.co.uk/news/magazine-28481060
6) 2014. Pilots reveal death-defying ordeal as engines failed on approach to Chek Lap Kok. *South China Morning Post*. Available at: http://www.scmp.com/magazines/post-magazine/article/1491534/pilots-reveal-death-defying-ordeal-engines-failed-approach
7) R: Visualizing Plane Crash Data using Some Interesting Graphs/charts Including googleVis and Great Circle ones | Anandh Shanmugaraj | LinkedIn. Available at: https://www.linkedin.com/pulse/r-visualizing-

plane-crash-data-using-some-interesting-shanmugaraj
8) RPubs - Plane Crash Analysis. Available at: https://rpubs.com/pgirish/plane_crash_analysis
9) Team 7: Visualizing Aircraft Accidents | Visualization Design. Available at: https://visualizationdesign.wordpress.com/2015/06/02/team-7-visualizing-aircraft-accidents/
10) The story behind a century of aircraft crashes. Available at: https://deltanomics.wordpress.com/2016/11/23/the-story-behind-a-century-of-aircraft-crashes/
11) What Really Causes Plane Crashes? (It's Not What You Think). Available at: https://www.yahoo.com/style/what-really-causes-plane-crashes-its-not-what-125605316542.html
12) Why do planes crash? Expert explains five most common reasons for airliner disasters | Daily Mail Online. Available at: http://www.dailymail.co.uk/sciencetech/article-3600784/Why-planes-crash-Expert-explains-five-common-reasons-airliner-disasters-one-ten-caused-terrorism.html
13) Aopa, T. et al., 2016. 2014-2015 GA ACCIDENT SCORECARD.
14) Broach, D., Joseph, K.M. & Schroeder, D.J., 2003. Pilot Age and Accident Rates Report 4 : An Analysis of Professional ATP and Commercial Pilot Accident Rates by Age.
15) Fox, T. et al., Visualizing the FAA Aviation Accident Database.
16) Hansen, M., Mcandrews, C. & Berkeley, E., 2008. HISTORY OF AVIATION SAFETY OVERSIGHT IN THE UNITED STATES.
17) Hansman, R.J., 2014. Analysis of Impact of Aircraft Age on Safety for Air Transport Jet Airplanes. , (October), pp.1–19.
18) Herrera, J. & Vasigh, B., 2011. A Basic Analysis Of Aging Aircraft, Region Of The World, And Accidents. *Journal of Business & Economics …*, 7(5), pp.121–132. Available at: http://journals.cluteonline.com/index.php/JBER/article/viewArticle/2299.
19) Jain, S., 2015. Recovery of Crash Site Web Application. , (August).
20) Kenny, D.J., 2016. 25th Joseph T. Nall Report: General Aviation Accidents in 2013.
21) Tonyleather, *The Deadliest Airplane Accidents in History*, Available at: http://www.environmentalgraffiti.com/mass-murder/news-most-fatal-plane-crashes-past-100-years?image=
22) Agrawal, S. & Siddiqui, S., 2014. TaleSpin.