Visualization of Business Intelligence Insights into Aviation Accidents

Aviation Accidents Discovery: Actionable Insights and Data Visualization

Loe Piin Piin[1], Sarasvathi Nagalingham[2]
Faculty of Data Science and Information Technology, INTI International University
Nilai, Negeri Sembilan, Malaysia[1, 2]

***Abstract*—Despite the recent tragic loss activity, flying is often said to be the safest form of transport, and this is at least true in terms of fatalities per distance travelled. The Civil Aviation Authority reports that the death rate per billion kilometers travelled by aircraft is 0.003, which is much lower than the rates of 0.27 for train travel and 2.57 for vehicle travel. Despite the fact that safety has been the aviation industry's top focus for the last century and a half, accidents involving aircraft continue to be a source of horror even in the present day. Hence, the aim of this project is to identify the major causes and reasons that led to accidents in the aviation industry and to carry out research, finding, design, build and suggest a Business Intelligence (BI) solution to the problem. Throughout the project, it will discover problems, both elementary and critical which needs to be corrected or changed in order to prevent major negative happenings and improve the current situation in a positive way. Tableau will be the primary BI tool used in this process. Data visualization is the graphic depiction of information and data. Data visualization tools offer an easy approach to data analysis that observe and find patterns, outliers, and patterns in data by employing visual elements like charts, graphs, and maps. The project will also cover the initial to building and deployment stage of the BI solution to improve and prevent further accidents.**

***Keywords*—Aviation; accidents; business intelligence; prediction; dashboard visualization; data analysis**

I. Introduction

A. Business Intelligence Methodology

Businesses can improve their decision-making processes with the help of Business Intelligence (BI), which showcases current and historical data within the context of the company's operations. BI may be used by analysts to establish the business's performance and competitive benchmarks, which will help businesses operate more seamlessly and productively and also uncover market trends that can be improved upon [14]. Every BI project has to specify a systematic methodology and strategy in order to provide decision-makers with the ability to adopt, implement, and integrate sustainable management practices across the company. In addition, it may improve the likelihood of accomplishment, reduce the amount of time and effort required, eliminate redundant procedures, and guarantee accurate reporting and analysis. As a result, the Polar's BI methodology will be used, which consist of five major steps. It helps organizations by delivering meaningful insights and develops new solutions to meet the problems based on recommended steps. Fig. 1 depicts the stages that are included in the phase of constructing a Polar's BI methodology:

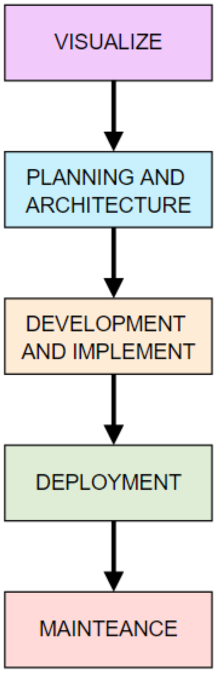


Fig. 1. Diagram for polar BI methodology

- Visualize: This step will involve defining the business requirement, which in our case, will be to identify trends and causes that leads to aviation accidents. It establishes a roadmap which will define the information clearly.
- Planning and Architecture: This step involves detailed business, data, and requirements by the system to produce the implementation plan accordingly including detailing the architecture of the system along with design specifications to provide solution. It is vital to define each requirement obtained from visualization and define a proper scope. It includes clearly defining the basic software and hardware requirements.
- Development and Implement: This phase is where the requirements are transformed into reality with a tested software production. Data will be populated into the

defined data structure through ETL process. After which database must be developed and software testing has to be conducted to ensure smooth functioning of the system and customer satisfaction. Finally, extensive documentation has to be prepared before deployment.

- Deployment: This is the implementation phase where the system will be finally implemented into the production environment to facilitate information exchange and migration of database structures and components. It gathers the readiness of the users and infrastructure, and the deployment plan will be shared with the initial set of users. The product would be deployed, and the system will be checked for correctness and accuracy.
- Maintenance: After the deployment, the system will need proper maintenance to ensure long term functionality and flow. This will include administration control, regular monitoring and performance updates and tuning.

B. Overview of Boeing Company

Boeing Co., Major U.S. firm that is the world's biggest aviation organization and principal creator of business fly vehicles. It was established by William E. Boeing (1881-1956) in 1916 (as Aero Products Company). In the last part of the 1920s it turned out to be essential for United Aircraft and Transport Corp., however it reappeared as a free element in 1934 when that organization was separated to conform to antitrust regulation [1]. Boeing spearheaded the advancement of single-wing planes during the 1930s; its B-17 Flying Fortress (first flown 1935) and B-29 Superfortress (1942) assumed noticeable parts in World War II.

It is additionally a main maker of military airplane, helicopters, space vehicles, and rockets, a standing essentially improved with the organization's procurement of the aviation and safeguard units of Rockwell International Corporation in 1996 and its consolidation with McDonnell Douglas Corporation in 1997. Previously Boeing Airplane Company, the firm accepted that its present name in 1961 to mirror its venture into fields past airplane make. Central commands were in Seattle until 2001, when Boeing migrated to Chicago.

Boeing Company's constituent specialty units are coordinated around three primary gatherings of items and administrations business planes, military airplane and rockets, and space and correspondences. Boeing fabricates seven particular groups of business airplane, which are gathered in two offices Renton, and Everett-in Washington State and one office in California. The Renton plant fabricates the restricted body Boeing 737 and previously constructed the 757 airplanes (ended in 2004), while the wide-body Boeing 767 and 777 airplanes and a predetermined number of the plane, to a great extent suspended 747s is gathered at the Everett plant. The 787 airplanes are collected at the Everett plant and at an office in North Charleston, South Carolina. Boeing Business Jets, a joint endeavor of Boeing and General Electric Co., makes and markets business jets in view of the 737-700 aircraft as well as VIP forms of the 747, 777, and 787 carriers.

C. Challenges Faced by Boeing Company

It is critical for businesses to be able to identify and forecast existing and future challenges that they may encounter. Identifying difficulties early on may assist the firm in developing a better solution sooner rather than later, which can help the company lower its losses. Aside from that, businesses may transform issues into chances for growth, allowing them to become more profitable in the long run. Being the largest producer of commercial and military aircraft and having the highest revenue compared to it' competitors, the company has faced and faces several challenges.

First, Boeing being the leading company in the aviation sector has many responsibilities and must abide government-imposed norms without fail as well as manage their business operations without defect. Certain time sudden laws and restrictions can cause the company to change its business layout in order to adjust. For example, in 1977, Boeing faced travel cost competition when the US administration brought down air travel restrictions which promoted more newer airlines to offer the same services at very competitive prices by the year 1999. Another scenario of rapid change was during the Covid pandemic which affected many businesses but demanded a complete shutdown of aviation travel.

As the making of an airplane is directly related to safety, it has to undergo various tests in order to prove its feasibility and safety mechanisms. These include various additional costs and takes a long time. When Boeing developed the '787 Dreamliner' to save fuel and minimize stopovers, the airplane was made to be 50% carbon fiber-reinforced plastics, 12% titanium and almost 80% reduction of the use of aluminum which was primarily used in the old plane models. When it decided to use Lithium-ion batteries there was a concern that it would overheat to a level where it would stimulate fire. In order to solve this issue, it had to undergo various tests to determine its feasibility.

Besides, there is a high risk of the parts not being suitable and face trouble when interlocking since the company has partnerships and contracts with many other production companies for different parts of the plane. Therefore, they have to be returned back in order to be reproduced which increases the initial cost and also is very time consuming. Also, the aviation industry is very much dependable on service and the satisfaction of passengers. Therefore, the airline production companies have to meet the demands of the purchasing airlines in order to produce the model that will meet current consumer expectations as well as balance business costs.

Technological advancement is one of the major challenges of the industry. In order to implement newer software, sometimes with additional measures, the construction plan of the flight model has to be altered in order to compile it with the newer changes. These may take longer times as it must be planned and the whole company has to be suited with the model in order to train the employees such as engineers, software developers and other employees. Changes like this cannot be executed frequently as it is a very complex process, and it will also require additional training for existing pilots all around the world. Suppose the technical advancement is not considered to be great, airline will switch companies to buy airplanes in order to save their pilots from additional training which takes up time as well as money.

D. Opportunities of Boeing

Boeing has always been at the top among aerospace-based organisations. However, Boeing still faces some disparity with other modern gaming companies. Based on the studies found on the current environment of the company and gaming industry, there are some key business opportunities that can be implemented into the standard operation procedure of the company.

The first opportunity that presents itself to the organisation is in the field of technology application. For example, the sharing of digital information among air traffic control, the flight deck, and the operations centre of an airline to enhance the effectiveness of flight routing and ensure passenger safety. An electronic flight bag software that utilises next-generation communications may provide information on alternative flight paths to pilots in the event that the conditions of the weather so need. In addition, connected cabin technologies that enhance common areas, such as galleys and restrooms, as well as monitor environmental factors, like as temperature and humidity allow automatic adjustments. Lastly, the installation of cameras will provide a greater number of passengers a view of the world beyond the window of the aeroplane.

Over the last several decades, safer and more dependable designs have been responsible for most of the progress gained in lowering accident rates and boosting efficiency. This breakthrough has been made possible by advancements in engines, systems, and structures. Furthermore, design has long been acknowledged as a component in minimizing and managing human mistake. When Boeing embarks on a new design activity, previous operating experience, operational objectives, and scientific understanding establish the human factors design criteria. To examine how well alternative design solutions fit these objectives, analytical approaches such as mockup or simulator assessments are performed. A human-centered design philosophy underpins this work, which has been confirmed by millions of flights and decades of experience.

When there is a big change in the design of a plane it has to undergo various tests which consume time and large sum of money. Since bringing lengthy changes to a plane is necessarily not a desired output in the short term, the company can adjust only the required mechanisms and not change the whole flight design which further causes workers to consume more time and gain used to the working process. This can be achieved by analyzing the areas which can be improved only to the level where it doesn't require additional guarantee tests and slowly bring changes over time to maintain balance in the business operations and revenue as well as providing a safe machine to its consumers.

Finally, the 787 can travel farther distances than its predecessors because to its improved fuel efficiency. As a result, it has enabled the creation of more than 50 new nonstop routes throughout the globe. The design and construction of the Dreamliner are both ground-breaking. New composite materials are used in the construction of fifty percent of the principal structure of the Boeing 787, which includes the fuselage and the wings. Both the speed of the aircraft and its fuel efficiency may be boosted thanks to the design and construction of the wings.

II. Proposing BI Solutions

Business intelligence is a term that encapsulates the processes and methods used to collect, store, and analyze data from business operations or activities in order to improve performance [10]. All of these factors combine to provide a full image of a company, allowing customers to make better, more decisive decisions. In this section, the proposed business intelligence solutions assisting the business opportunities and solving the problems would be discussed.

Accident or crash rate is a calculation between the number of aircraft crash that unfold in a given period (crash frequency) and number of flights conducted within the same period. BI tools enable the accident rate to be calculated faster. In addition, the metrics for calculating the crash rate can be analyzed individually too to provide more insights, such as the crash frequency. Thus, the company can predict the probability of crash happening in the following years and put up a countermeasure to reduce it.

Through BI tools, the company will be able to discover factors that contribute to airplanes crashes. It allows the company to correlate the cause of the accidents, either its human errors or machine malfunctions, and any other piece of information together. Based on the piece of information founded, the factors that lead to aircraft crash can be grouped based on certain criteria to provide better insight which can also be used for comparison purposes. Therefore, a more specific study of aircraft crash on each criterion can be performed and a plan to address each category problem can be addressed.

Predictive maintenance is a proactive approach to machine maintenance, which is made possible through utilizing BI tools. It provides the company to organize schedules based on continuous condition monitoring. To minimize additional costly failures, broken parts are fixed or replaced if unfavorable patterns are discovered. Lower maintenance costs, longer equipment life, less downtime, increased production capacity, and improved safety are just a few of the advantages the company can expect.

III. Dataset Analysis

To describe how BI can improve Boeing performance, the NTSB aviation accident dataset was used as our data source. NTSB is an independent U.S. government investigation agency that identify and report on aviation accident. The NTSB aviation accident database contains information on civil aviation accidents and selected incidents that occurred in the United States, its territories and possessions, and international waterways. For our research, the dataset that we use contains information on aviation accidents from 1962 to present [5].

The reason the records are up to date is made possible by the continuous update of the accidents into the database. First, a preliminary report will be available online within a few days of an accident. When available, information is added, and when the investigation is complete, the preliminary report is replaced with a final description of the accident and its probable cause. Full narrative descriptions may not be available for cases prior to 1993, those under revision, or those in which the NTSB did not have primary investigative responsibility. It's worth noting that the information isn't limited to just commercial jets. On September 18, 2002, data from 1962 to 1982 were added to the aircraft accident database. The structure and type of data given in previous briefs may change from that contained in subsequent reports.

An aviation accident is described as an event occurring due to aircraft operation happening between the time people boards the aircraft with the aim of flight and the time people depart, in which people are mortally or severely injured, the aircraft sustains severe damage or structural failure, or the aircraft goes missing or becomes totally inaccessible, according to Annex 13 of the Convention on International Civil Aviation. An aviation incident is defined in Annex 13 as any occurrence, other than an accident, related with the operation of an aircraft that affects or has the potential to influence the safety of operation. Government agencies such as the FAA and the NTSB examine accidents and occurrences.

A. *Description of Dataset*

Our dataset is obtained NTSB from aviation accident database containing information about civil aviation accidents within the United States and its territories including international water from the year 1962. The dataset comes in a single data sheet and consists of 31 columns describing details about the recorded aviation accident [5]. The table below shows the detailed description of each attribute in the dataset and what it represents (Table I).

TABLE I. DESCRIPTION OF DATASET

No	Attribute Name	Data Type	Description
1	Event ID	ID	Unique alphanumerical for each of the row about the accident
2	Investigation Type	ID	Conveys if the scenario was an accident or incident
3	Accident Number	ID	Specific number given to the accident/incident
4	Event Date	Date	Date at which the accident occurred, displayed in the format "YYYY-MM-DD"
5	Location	String	Location at which the accident took place
6	Country	String	Country the accident took place
7	Latitude	Numeric (Continuous)	Angle of latitude which ranges from 0 degree at the Equator to 90 degrees (North or South) at the poles
8	Longitude	Numeric (Continuous)	Longitude which is the measurement east or west of the prime meridian
9	Airport Code	String	The airport code, which is unique for each and every airport in the world and is usually represented by three letters or four letters
10	Airport Name	String	The name of the Airport
11	Injury Severity	Nominal	Severity of the injury
12	Aircraft Damage	Nominal	The damage level to the aircraft
13	Aircraft Category	Nominal	Type of aircraft, ranging from Airplane, helicopter, glider, etc.
14	Registration Number	ID	Registration number of the aircraft
15	Make	Nominal	Brand type of the aircraft
16	Model	Nominal	Model number of the aircraft
17	Amateur Built	Boolean	True or false statement confirming if the aircraft is amateur built, meaning if they were built by people for their own education and recreation and not officially by an enterprise
18	Number of Engines	Numeric (Discrete)	Number of engines the aircraft had
19	Engine Type	Nominal	Type of engine the aircraft had
20	FAR Description	String	Description by Federal Aviation Regulations
21	Schedule	String	Represents the aircraft's schedule
22	Purpose of Flight	Nominal	Purpose of the flight, whether personal, instructional or others
23	Air Carrier	String	Airline company of the aircraft
24	Total Fatal Injuries	Numeric (Discrete)	Total number of fatal injuries due to the accident
25	Total Serious Injuries	Numeric (Discrete)	Total serious injuries due to the accident
26	Total Minor Injuries	Numeric (Discrete)	Total minor injuries due to the accident
27	Total Uninjured	Numeric (Discrete)	Total number of people uninjured after the accident
28	Weather Condition	Nominal	Condition of the weather at the time of accident
29	Broad Phase of Flight	Nominal	Phase of flight when the accident occurred
30	Report Status	Nominal	Status of the accident when reported
31	Publication Date	Date	Publication date of the accident

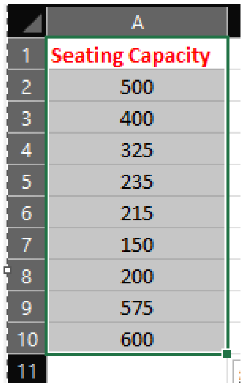
B. *Data Preprocessing*

With the dataset obtained from NTSB, we went through several rounds of data cleaning and preparation to make sure the data is in a clean format and is ready to be used for our BI dashboards. To perform data cleaning and preparation, we have used two programs, namely Microsoft Excel and R Studio. The programs were used to look at the state of the dataset (i.e. number of missing values for every column). We will go further and explain the techniques that we used in these two programs to aid our data cleaning and preparation process [6].

Microsoft Excel is an excellent software owned by Microsoft incorporation. It serves the purpose to organize, clean, arrange and simplify the data. Among the hundreds of features it has, some are being mentioned below.

1) *Conditions based formatting:* If you want to format the data based on conditions, then MS Excel is the best tool. For instance, you want to see only the Boeing Planes with a

seating capacity of less than 300 people. Then you can apply these conditions and see them. To use this feature, you need to select the desired column. In this case, the Fig. 2 demonstrates Column "Seating Capacity" is selected:



	A
1	Seating Capacity
2	500
3	400
4	325
5	235
6	215
7	150
8	200
9	575
10	600
11	

Fig. 2. Select "Seating Capacity" column

Once you have selected the column go to "Home" option in toolbar and click on "Conditional Formatting" as shown in Fig. 3.

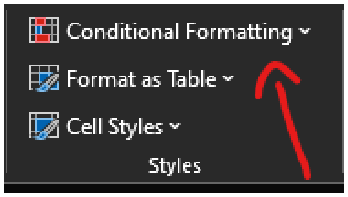


Fig. 3. Selection of "Conditional Formatting"

Once you hover the mouse over it, you can see the following options as shown in Fig. 4.

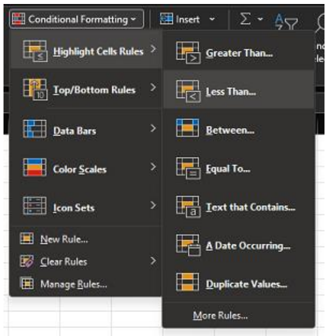
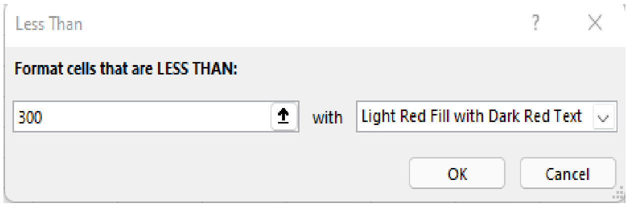


Fig. 4. Options in "Conditional Formatting"

In this case I have selected "Less Than..." as a condition. Once we click on "less than..." as shown in Fig. 5.



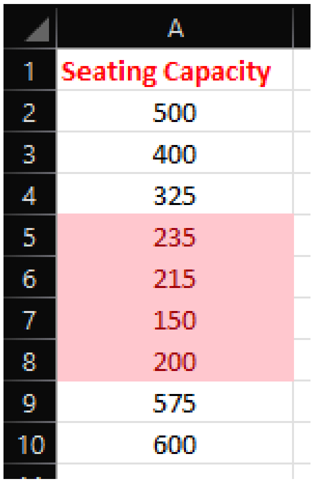
Less Than

Format cells that are LESS THAN:

300 with Light Red Fill with Dark Red Text

Fig. 5. Options in "Less Than..."

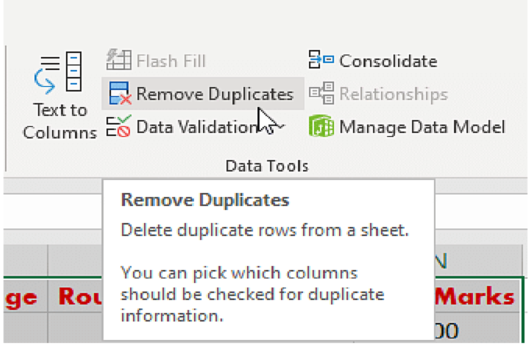
Here we can put the desired value and then the results will highlight in the column as depicted in Fig. 6.



	A
1	Seating Capacity
2	500
3	400
4	325
5	235
6	215
7	150
8	200
9	575
10	600

Fig. 6. Results of "Conditional Formatting"

2) Removing duplicate columns: There is a high probability that a data might have duplicate columns. MS excel provides a fantastic feature to remove them. You don't have to select each column in Dataset to remove them. Instead, there is an amazing feature in Excel which lets you remove duplicate columns automatically as shown in Fig. 7.



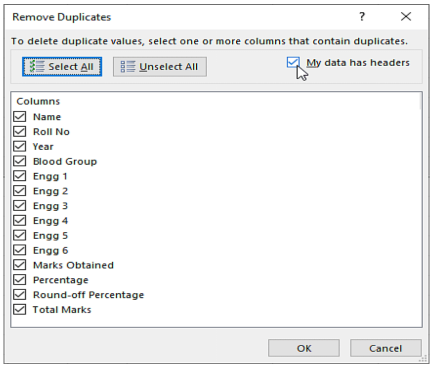
Remove Duplicates

Delete duplicate rows from a sheet.

You can pick which columns should be checked for duplicate information.

Fig. 7. Remove duplicates feature in excel

To access this feature, you can go to Data option placed in tool bar, there you can see "Remove Duplicate" feature as Fig. 8. Once you click on it, it will show you the duplicates and you can remove them if you want.



Remove Duplicates

To delete duplicate values, select one or more columns that contain duplicates.

Select All Unselect All ☑ My data has headers

Columns
☑ Name
☑ Roll No
☑ Year
☑ Blood Group
☑ Engg 1
☑ Engg 2
☑ Engg 3
☑ Engg 4
☑ Engg 5
☑ Engg 6
☑ Marks Obtained
☑ Percentage
☑ Round-off Percentage
☑ Total Marks

Fig. 8. Remove duplicates window

Once you are satisfied with columns checking you can hit ok and excel will remove duplicate column successfully.

3) Spell check: MS Excel has another useful feature as shown in Fig. 9 which helps you check for spelling mistakes in a data. This feature can come very handy in pointing out typos. For this feature to work, you can select your desired column from database and then click on "Review" option from tool bar. There you can see spell check option.

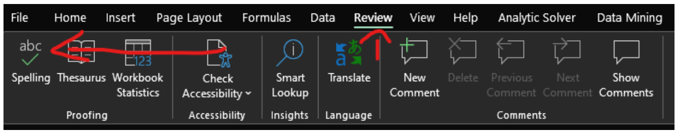
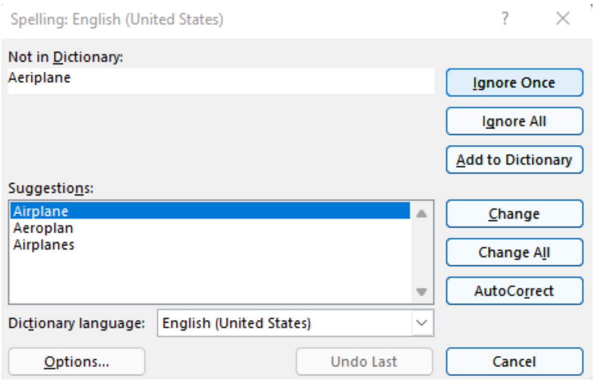


Fig. 9. Spelling check in excel

Once you click on it, Excel will automatically scan for any spelling mistakes in the data and will show as per the Fig. 10.


Spelling: English (United States)

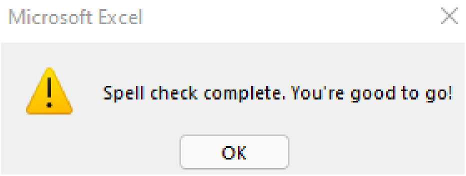
Not in Dictionary:
Aeriplane

Suggestions:
Airplane
Aeroplan
Airplanes

Dictionary language: English (United States)

Fig. 10. Spelling check configuration

You can see this Pop-up, from here you can choose which spelling is suitable for you. If you think the word spell is ok, you can click on "Add to Dictionary" or "Ignore All/once". If you want to change the word with the given suggestions, you can highlight preferred suggestion and click on "Change /Change All". If you want the system to automatically correct, you can click on "Auto Correct". If there are no spelling errors the following pop up will appear as shown in Fig. 11.

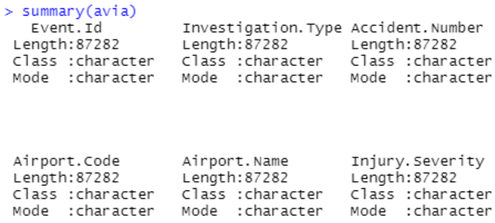

Microsoft Excel

Spell check complete. You're good to go!

Fig. 11. Successful configuration of spelling check

RStudio is an open-source IDE (integrated development environment) for R programming that has been widely utilised by the community for statistical and computing purposes. The RStudio was utilised in this project since various tasks required the use of the R language to aid in data visualisation and pre-processing. Due to the ease of access and pre-experiences of the program being used, the authors decided to use this program as the pre-processing tool due to the familiarity of library tools usage from the past and the flexibility of data processing it provides. Fig. 12 demonstrates the techniques that were used in R Studio to perform data cleaning and preparation:

4) Display summary of a data frame: summary(df)


```
> summary(avia)
   Event.Id         Investigation.Type Accident.Number   
 Length:87282       Length:87282       Length:87282      
 Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character  



 Airport.Code       Airport.Name       Injury.Severity   
 Length:87282       Length:87282       Length:87282      
 Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character  
```

Fig. 12. Display summary of non-numerical attributes of a data frame

Using this function, we will get a summary of every attribute in a data frame as shown in Fig. 12. For each attribute, it will display the number of records and the attribute type.

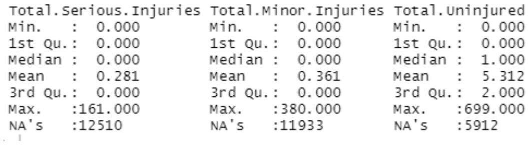

```
 Total.Serious.Injuries Total.Minor.Injuries Total.Uninjured 
 Min.   :  0.000        Min.   :  0.000      Min.   :  0.000 
 1st Qu.:  0.000        1st Qu.:  0.000      1st Qu.:  0.000 
 Median :  0.000        Median :  0.000      Median :  1.000 
 Mean   :  0.281        Mean   :  0.361      Mean   :  5.312 
 3rd Qu.:  0.000        3rd Qu.:  0.000      3rd Qu.:  2.000 
 Max.   :161.000        Max.   :380.000      Max.   :699.000 
 NA's   :12510          NA's   :11933        NA's   :5912    
```

Fig. 13. Display summary of numerical attributes of a data frame

For numerical attributes, R will be able to display even more details of the attribute, like the first quarter value, third quarter value, minimum and maximum value found, number of missing values etc as shown in Fig. 13. Using this function, we will be able to check the type of every attribute and also the number of missing values. However, as this function is unable to display more information about non-numerical attributes, we have to use other functions to learn more about our dataset.

5) Display number of missing values or empty columns for each attribute:

colSums(is.na(df)| df == "")

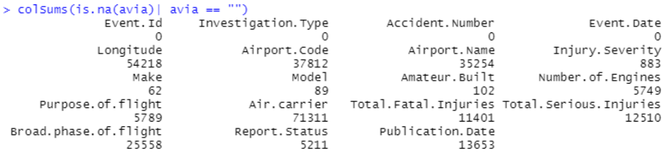

```
> colSums(is.na(avia)| avia == "")
              Event.Id     Investigation.Type        Accident.Number             Event.Date 
                     0                      0                      0                      0 
             Longitude           Airport.Code           Airport.Name        Injury.Severity 
                 54218                  37812                  35254                    883 
                  Make                  Model          Amateur.Built      Number.of.Engines 
                    62                     89                    102                   5749 
     Purpose.of.flight            Air.carrier   Total.Fatal.Injuries Total.Serious.Injuries 
                  5789                  71311                  11401                  12510 
 Broad.phase.of.flight          Report.Status       Publication.Date 
                 25558                   5211                  13653 
```

Fig. 14. Display number of missing values or empty columns for each attribute

Through this function, we are able to clearly see the number of missing values and empty columns every attribute has as shown in Fig. 14. Any missing values or empty columns are bad for making visualizations through BI dashboards. Thus, we have to figure out what to deal with attributes that have an overwhelming number of missing values or empty columns.

6) Remove columns from a data frame:

df = subset(df, select = -c(attribute1, attribute2, attribute3))

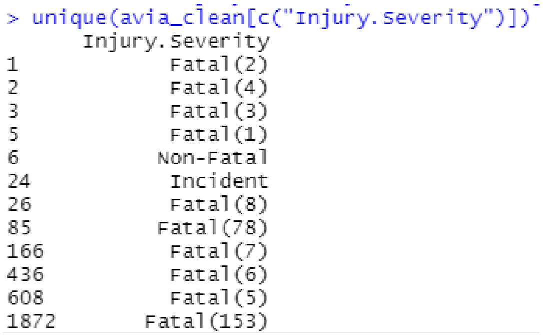
With this function, we are subsetting attributes to create a new data frame. However, as we are specifying the “-c” argument, we are actually creating a new data frame that excludes the attributes mentioned in the command. We used this function to remove any attributes that have way too many missing values or empty columns to be used in our BI dashboard visualizations.

7) Remove rows with missing values:
df <- na.omit(df)

We used this function to remove the rows that had only missing values in our data frame.

8) Format values in an attribute: In our data set, the attribute that mentions the injury severity of aviation accidents also mention the number of fatal injuries in Fig. 15 (i.e. Fatal(4) meaning four meaning suffered fatal injuries). As the data set already has an attribute that stores the number of fatal injuries, we can make this attribute to only mention whether the injury severity is fatal, non-fatal or incidental. To achieve this, we used two functions:

unique(df[c("attribute1")])

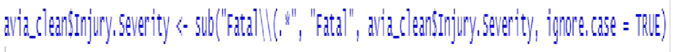


```
> unique(avia_clean[c("Injury.Severity")])
     Injury.Severity
1           Fatal(2)
2           Fatal(4)
3           Fatal(3)
5           Fatal(1)
6          Non-Fatal
24          Incident
26          Fatal(8)
85         Fatal(78)
166         Fatal(7)
436         Fatal(6)
608         Fatal(5)
1872      Fatal(153)
```

Fig. 15. Display unique values in an attribute

Fig. 15 shows us the unique values in an attribute. Using this command, we first checked whether the attribute had a big number of unique attributes, which is bad for a categorical attribute.

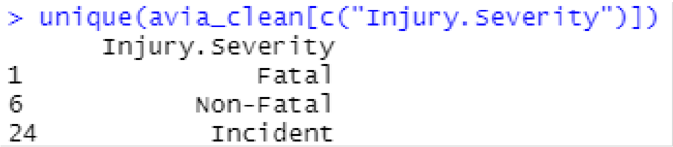
df$attribute1 <- sub("Searched sting", "Substitute string", df$attribute1, ignore.case = TRUE)



```
avia_clean$Injury.Severity <- sub("Fatal\\(.*", "Fatal", avia_clean$Injury.Severity, ignore.case = TRUE)
```

Fig. 16. Command to replace strings in an attribute

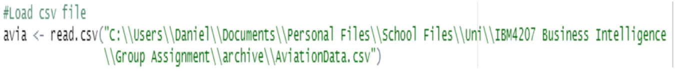
As shown in Fig. 16, we managed to automatically clear out the number of fatal injuries in this attribute, leaving only the word “Fatal”. Displaying the unique values again as per the Fig. 17, the command was successful in clearing out the number of fatal injuries in the attribute.



```
> unique(avia_clean[c("Injury.Severity")])
     Injury.Severity
1              Fatal
6          Non-Fatal
24          Incident
```

Fig. 17. Display unique values after replacing strings

9) Read and write files:
df <- read.csv("File location path")

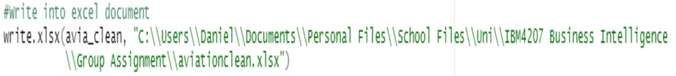


```
#Load csv file
avia <- read.csv("C:\\Users\\Daniel\\Documents\\Personal Files\\School Files\\Uni\\IBM4207 Business Intelligence
                 \\Group Assignment\\archive\\AviationData.csv")
```

Fig. 18. Command to read CSV file into a data frame

Fig. 18 depicts that we could read the csv format file that our data set came in. A csv file is a delimited text file which uses commas to separate between different values, and using this function, we imported the file into R Studio as a data frame to perform data cleaning and preparation.

write.xlsx(df, "File location path")



```
#Write into excel document
write.xlsx(avia_clean, "C:\\Users\\Daniel\\Documents\\Personal Files\\School Files\\Uni\\IBM4207 Business Intelligence
           \\Group Assignment\\aviationclean.xlsx")
```

Fig. 19. Command to write data frame into an excel document

After our data cleaning and preparation is completed, we want to export our data frame into an Excel file so that it can be used for our BI dashboards. We used this function, which is part of the “openxlsx” R library to write our clean data frame into an Excel file as shown in Fig. 19.

Due to challenges during data recording and such, several attributes of our dataset had a large number of missing values/empty columns. For example, the air carrier column had 71,311 missing values, longitude and latitude had 54,218 and 54,209 missing values each. As our dataset has 87,282 records in total, including attributes like these with a large number of missing values will be problematic. We also cannot try to fill in the missing values with techniques like using the mean as it will distort the meaning of the data. Because of this issue, we decided to outright remove the attributes that has a large number of missing values.

C. Data Modeling

The process of generating a visual representation of a comprehensive information system or specific components of that system in order to express the relationships that exist between different data points and organizational structures is known as data modelling [4]. The purpose of data modelling is to offer illustrations of the many kinds of data that are utilized and stored inside the system, as well as representations of their connections, potential groupings and organizational structures, format types, and attribute values.

In data mining, an algorithm is a set of related computations and heuristics that are used to construct a model based on a collection of data. Before designing a model, the algorithm does an initial analysis on the data that people provide, looking for different sorts of patterns or trends. The results of this study are used by the algorithm, which goes through a number of rounds to determine which parameters will provide the most accurate mining model [12]. After that, the whole of the data gathering is analyzed using these criteria in order to derive actionable patterns and comprehensive statistics. In this section, the algorithms that were employed are the decision tree method, the Support Vector Machine (SVM) algorithm, and the Naïve Bayes algorithm.

1) Naïve bayes algorithm: The Naïve Bayes algorithm is a classification approach that is based on the Bayes Theorem and operates on the assumption that predictors are independent of one another [3]. To put it simply, a Naïve Bayes classifier is one that operates on the assumption that the existence of one feature in a class has absolutely no bearing on the presence of any other feature. The Fig. 20 illustrates an example of the formula for Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Fig. 20. Calculation for bayes' theorem

- The possibility that a certain hypothesis (A) will in fact come to pass is quantified by the posterior probability, abbreviated as P(A|B).
- Likelihood Probability, abbreviated as P(B|A), is a measurement that determines how probable it is, based on the data that is now available, that a particular hypothesis is right.
- Priority probability, sometimes abbreviated as P(A), refers to the possibility that a hypothesis exists prior to seeing the data.
- The likelihood of evidence is referred to as P(B), which stands for marginal probability.

2) Decision tree: A decision tree is a technique of supervised learning that may be used with both discrete and continuous variables in its analysis [6]. The most essential aspect of the dataset is used to determine the subgroups that are created from the dataset. The algorithms are what decide how this characteristic is categorized by the decision tree and how the divisions are made between the categories. The terminal or leaf nodes, which do not further split, are created when the root node, which represents the most significant predictor, divides into decision nodes, which are sub-nodes as shown in Fig. 21 [7].

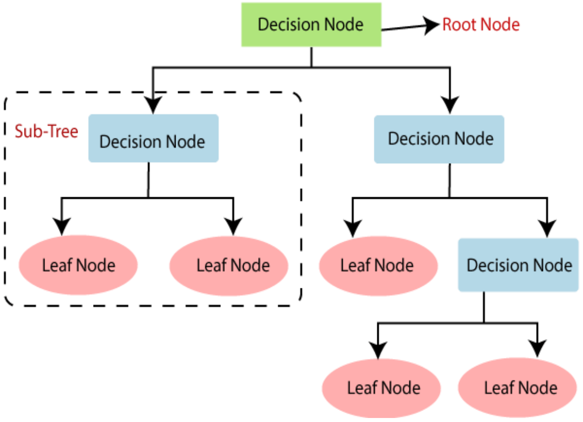


Fig. 21. Decision tree structure

3) Support vector machine (SVM): The objective of the Support Vector Machine technique is to locate, in a space of N dimensions (where N represents the number of features), a hyperplane that categorizes the data points in a way that is unambiguous [8]. The two distinct groups of data points may be partitioned using any one of several hyperplanes that are available to choose from [9]. Finding a plane that has the highest margin, or the greatest separation between the data points of both classes, is the goal here. As per Fig. 22,When the margin distance is increased to its maximum, more support is added, which in turn increases the level of confidence with which subsequent data points may be categorized.

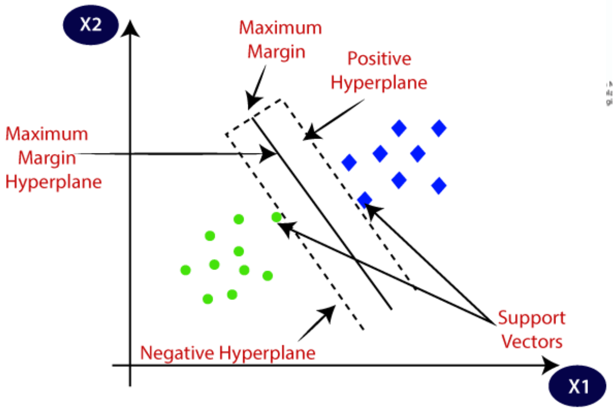


Fig. 22. SVM algorithm

IV. Business Intelligence Architecture

The framework that a business establishes in order to run applications for business intelligence and analytics is referred to as a business intelligence architecture. It discusses the information technology platforms and software tools that are used to collect, integrate, store, and analyze business intelligence data before providing it to high-level executives and other corporate users as information on operational processes and statistics. Implementing an efficient BI program that allows data analysis and reporting to assist an organisation in monitoring business performance, optimizing business processes, recognizing new revenue opportunities, working on improving strategic planning, and providing better decisions overall requires a critical component known as the underlying BI architecture.

The BI architecture design shows how all the data and processes are connected to each other to support Valve business. The diagram shows the seven main components and the flow in Valve when dealing with data through business intelligence [2]. The seven main components consist of data sources, ETL (extract, transform and loading), data repository, business intelligence tools, interfaces, data management and security management as shown in Fig. 23.

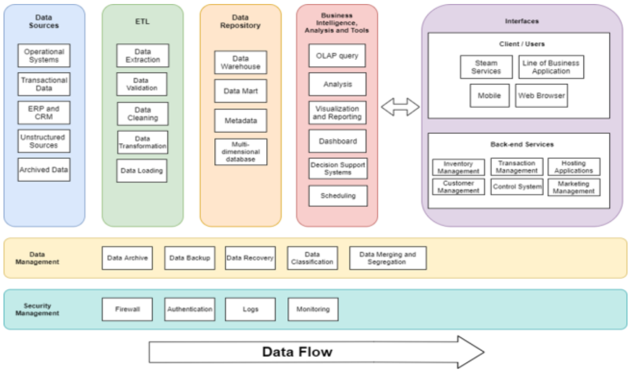


Fig. 23. Proposed BI architecture

A. *Data Source*

Data sources mainly refer to where data is obtained. Any business organisation would get data differently depending on their mode of business, and their operations done. In Valve, data can come from multiple sources. For instance, transactions or purchases on games and the accessories can be categorised into transactional data, whereas daily processes and routines would contribute to data obtained from operational systems. This architecture design suggests that data sources may also involve data from unstructured forms, archives, and other systems such as ERP and CRM. All these data will then be staged into the ETL stage.

B. *ETL (Extract, Transform and Load)*

This stage involves three major steps in dealing with data. First, data identified from the various data sources are extracted. There are multiple methods to extract data. It is crucial to discover which extraction method and platform would benefit the business the most. Usually when data are extracted from various data sources, the data vary from each other in terms of formatting. For instance, two data sources might store "date" data differently from each other. This is where the second step, transformation comes in. Transformation relates to changing the data into the wanted and appropriate format. This step also involves data cleaning and attributes derivations. This step is significant to ensure data integrity and quality, while making sure that the data would reach a common dimension and storage method. The data will then be loaded into the repository once all the data has been transformed.

C. *Data Repository*

Data repository refers to where data is stored after being processed. Valve stores over ten million of players' data, and management side. Therefore, Valve will need numerous data repositories and in different forms, such as data marts and data warehouses. For example, data that are generated to players and require high availability would be stored in data marts, and data that are not frequently used can be placed into data warehouses. Valve can store summarised and transformed data in data marts for respective departments to ensure productivity, while metadata that describes data and the relationships, together with raw data can be stored in a data warehouse. Multidimensional databases can help integrate data warehouses and OLAP applications to generate complex data to users [11].

D. *Business Intelligence, Analytics and Tools*

This section comprises of all the business intelligence tools used in Valve. All of these tools assist the organisation in making better decisions and supporting the interfaces on both clients and back-end services. The tools include OLAP operations, analysis, visualisation, reporting, dashboard, and decision support applications.

E. *Interfaces*

The interfaces involve how input and output would result through different platforms such as mobile, web applications, line of business application and steam application. Interfaces comprise of back-end services from within the company and externally from third-party services. These services extend to aid in the management side in Valve, such as inventory management, transaction management, customer management and more. Interface is the result where the client and users can view on the screen. The interfaces will communicate with the business intelligence part in order to output the wanted information.

F. *Data Management*

It is important for every organisation to have a sustainable data management plan to ensure that data can be kept up to date and saved. This will help to maintain data quality and availability. It is also vital to archive and backup data in a set time interval to avoid any losses due to data loss or breach of security to protect the business operations. Data management is usually done by the administrators and certain technical employees throughout the data processes up to the interfaces.

G. *Security Management*

Data is a crucial component in any organisation. Hence, they need to be protected and maintained at all times. A well-secured organisation would only allow authorized and recognized personnel to access the operational part of the organisation. Hence, firewalls and authentication are highly required in order to protect the data used by Valve. It is also crucial to monitor the activities of the personnel involved in the business operations in order to find out the culprit when there is a breach of security or data loss through logs. This is because Valve is a big company, comprising hundreds of employees and, which suggests that anything could happen.

V. DATAWAREHOUSE AND OLAP MODEL

The dimensional model serves as the foundation for all analytics and reporting. It is used throughout the entire science of dimensional modelling, not only in the construction of the model but also in the execution of reports and queries, the development of extract, transform, and load tools, and the use of business intelligence tools [13]. Even though the business people in the organisation might never get to see the actual data models that are developed, they would become very acquainted with the reports and dashboards they help produce. Unless those models are capable of generating transparent, effective reporting and analysis, they will be ineffective in assisting the company in viewing data and using it to make informed choices that have an impact on operations. The ability to generate this vital business information is dependent on the use of dimensional modelling.

The dimensional model provides a logical data model for the presentation level of a Data Warehousing and Business Intelligence (DWBI) application, through which the end-user dashboards would take data. It contains a list of the entities and attributes required for the envisioned dashboards. The entities that provide measures are referred to as facts [12][13]. It is termed dimensions when they provide qualifiers that allow facts to be disaggregated, filtered, and organised in various ways. The Steam Sales Schema's data warehouse architecture is centred on the classic star model. Compared to the Snowflake and Starflake schema models, the star schema design has been the most useful when dealing with historical information since it has the lowest query complexity. This allows them to operate at peak performance in data warehouses, data marts, business intelligence applications, and online analytical processing (OLAP).

The fact data can be organised in a fact table at the model's centre, whereas the dimensional data is organised in dimension tables and surrounds the fact table. The fact table is the integration component at the heart of the data warehouse's star structure. They allow machine learning algorithms to analyse the data in its totality and concurrent access to the data by other subsystems. The star schema's dimension tables have no foreign keys. This allows star schema databases to be optimised for lookup and query performance using actual dimensions. They may also be customised to give the best performance and the specific parameters that are deemed most important or often queried by the organisation in question. Data may be inputted transactionally as it is received, or it can be imported in bulk and subsequently validated and correctly denormalized.

The fact table of Boeing's dimensional model is the accident fact table, which holds details of the accident like the accident date, publication date, latitude, longitude, number of injuries by type, weather condition when the accident happened, flight phase during the accident and the report's status. From the fact table, it is connected to three dimension tables, namely, the flight dimension, state dimension and the airport dimension.

In the flight dimension, it has an ID number to identify the flight, along with the flight's schedule, purpose and number of passengers. It connects to the aircraft dimension, which records the aircraft's damage, category and identifies whether it is amateurly built. From the aircraft dimension, it connects to two dimensions. The first one is the airline dimension, which identifies the airline owning the aircraft, along with their name, country, region, email and contact number. The second dimension is the model dimension, which identifies the model of the aircraft, with the make, number of engines, engine type and description provided by the Federal Aviation Regulations (FAR). The airline dimension is connected to the manufacturer dimension, which identifies the manufacturer's details and contact methods.

The airport dimension holds information about the flight's departing airport, with details like the airport code, name, country, region, email and contact number. On the other hand, the state dimension recalls where the accident happened, along with the state name. It is connected to the country dimension, which holds information like the country name and the country's region.

These nine tables make up the dimensional model designed for The Boeing Company. It enables the company to keep track of aviation accidents with paper reports and BI dashboards, while updating it with new records regularly.

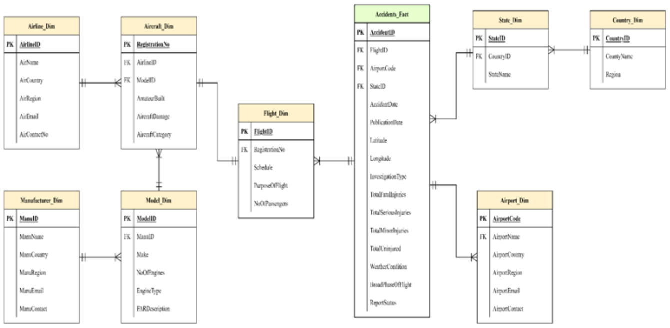


Fig. 24. Diagram of data warehouse schema

The dimensional model that was developed for The Boeing Company uses the snowflake schema as depicted in Fig. 24. Due to the complexity of records to be kept for an aviation accident, the dimensional model needs to use the snowflake schema in order to avoid data redundancy and ensure relevant data is stored in their respective tables. This is possible through the normalization of the dimensional structure, which causes the tables to form this snowflake-like shape, where the name comes from.

VI. Dashboard Visualization

This project calls for the development of five different dashboards, each of which will be based on an analysis of either aircraft MAKE as shown in Fig. 25, purpose of flight, injury, weather condition or aircraft engine, respectively. Through the use of maps and graphs, data visualization enables us to better comprehend the significance of the information by putting it in a visual context. Due to the fact that the information is simpler to comprehend by the human mind, it is thus far less difficult to recognize trends, patterns, and outliers in massive data sets.

A. Dashboard for Aircraft MAKE Analysis

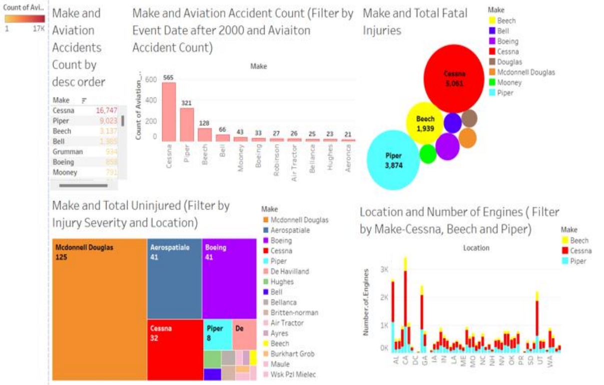
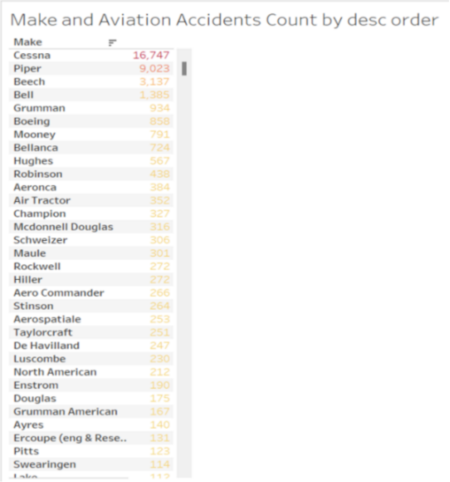


Fig. 25. Dashboard for aircraft MAKE analysis

The term “dashboard” refers to an electronic tracking tool that is used by companies to show and summarize the data that they have collected. The figure above demonstrates the overview of the dashboard showing factors with the “Make” column. This column represents the brand type of the aircraft. As evident from the above dashboard, the Make data is compared with aviation accidents total count, and another similar case where it is also filtered by event date of accidents that occurred after the year 2000. It is also visually represented along with total fatal injuries, denoting only the major ones. Total uninjured is also compared where it is filtered by severity of the accident and location. Final set of comparison has been done taking the top 3 makes having the most fatal injuries and representing them in a bar graph based on location and the number of engines. The dashboard gives an overview of the idea that can be gained by analyzing the representations.

On comparison of make with the count of aviation accidents, it can be found that Cessna had the highest count of accidents being 16,747, which is followed by Piper (9,023) and Beech (3,137) shown in Fig. 26. Evidently, when make is compared with total fatal injuries, Cessna has the highest rate with it being 5,061 and Piper and Beech following the same previous pattern having 3,874 and 1,939 fatal injuries, respectively.



Make and Aviation Accidents Count by desc order

Make	
Cessna	16,747
Piper	9,023
Beech	3,137
Bell	1,385
Grumman	934
Boeing	858
Mooney	791
Bellanca	724
Hughes	567
Robinson	438
Aeronca	384
Air Tractor	352
Champion	327
Mcdonnell Douglas	316
Schweizer	306
Maule	301
Rockwell	272
Hiller	272
Aero Commander	266
Stinson	264
Aerospatiale	253
Taylorcraft	251
De Havilland	247
Luscombe	230
North American	212
Enstrom	190
Douglas	175
Grumman American	167
Ayres	140
Ercoupe (eng & Rese..	131
Pitts	123
Swearingen	114
Lake	112

Fig. 26. Make and aviation accidents count by descending order

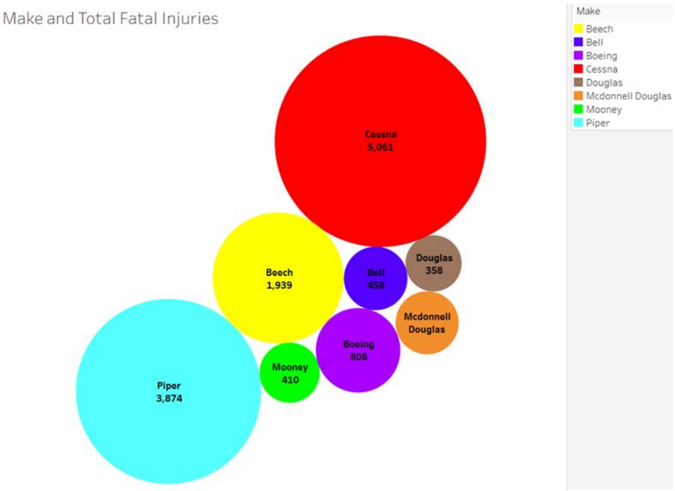


Fig. 27. Make and total fatal injuries

The two visual represents go hand in hand with each other and can aid the company in identifying the makes that have the highest rate of accidents and thus dig deeper into finding the reasons considering other real-life factors into account. It can also be noted that in the first representation, the number is arranged in the decreasing order with the color gradient being darker for the highest number and fading as it lowers. These tools are great for having a more toned representation of data. In the comparison with total fatal injuries, the packed bubble representation gives out a very clear and easy to understand pictorial representation with different makes being color-coded and the higher number having a larger bubble with the count represented as label in Fig. 27.

Based on the insights gained by the previous comparisons, the data can be filtered according to the needs of the user to display certain factors. For example, the tri-colored bar chart has specifically been filtered by the top 3 makes involved in the highest accidents (Cessna, Piper and Beech) and compared with the location of the accident and the number of engines present in the aircraft as shown in Fig. 28.

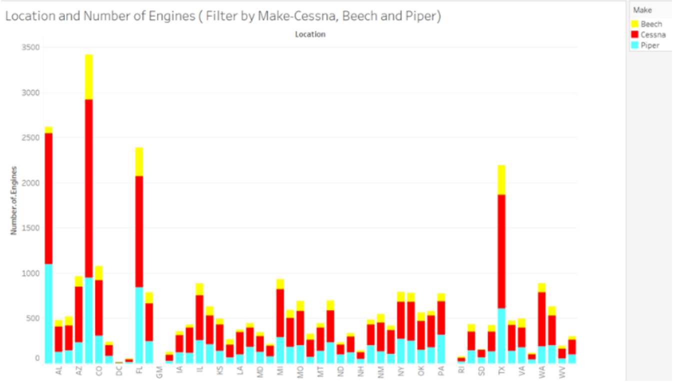


Fig. 28. Location and number of engines, filter by make

Each color represents one of the makes and exhibits a very organized pattern as information. By making use of previously identified information and viewing such representation based on specific requirement, more focused insights can be obtained. For example, it is proof that Arizona (AZ) has the highest number of accidents happening with the three makes.

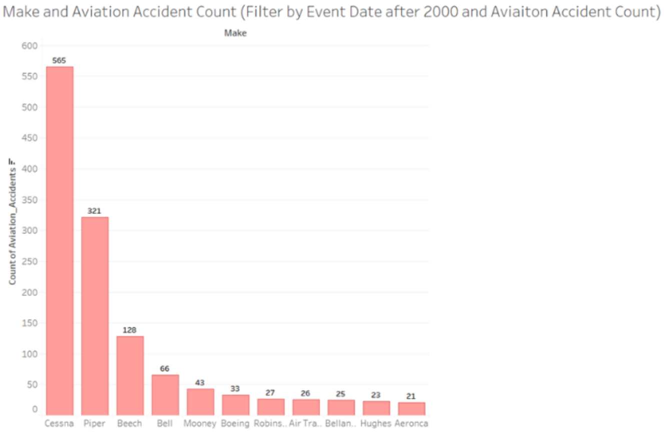


Fig. 29. Make and aviation accident count, filter by event date

Another example can be the bar chart in Fig. 29 depicting the make with the aviation count which is filtered by data to only show those that happened after the year 2000 and by already knowing the major values from the text tables, it has been filtered to show only those with count 21 and above as these alone give a large contrast between the highest values. This allows viewing specific data as per requirement and constraint. In this case, it can be seen that Cessna takes the lead of accidents even in this decade.

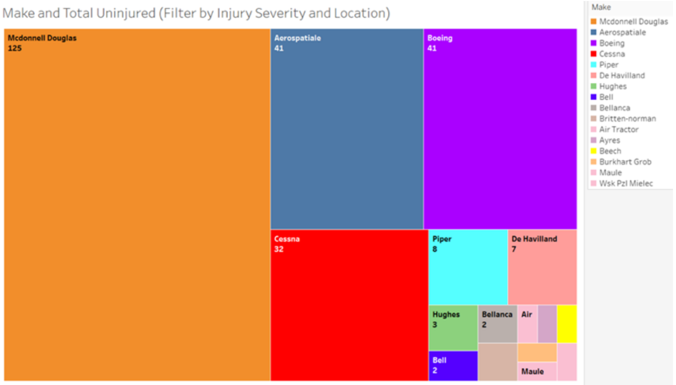
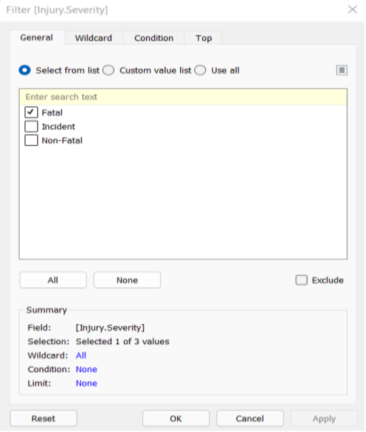


Fig. 30. Make and total uninjured, filter by severity and location

Finally, the Fig. 30 tree-maps present a colorful view of the make and total uninjured in each case that has been filtered by injury severity and location. In this example, two different categories have been allocated as filters with each allocation allowing the flexibility to choose the type and amount of data to be filtered with.


Filter [Injury.Severity]

General | Wildcard | Condition | Top

● Select from list ○ Custom value list ○ Use all

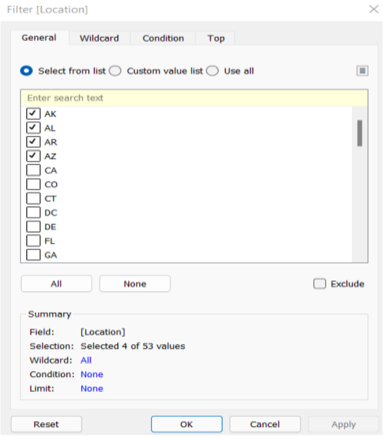
Enter search text
☑ Fatal
☐ Incident
☐ Non-Fatal

All | None | ☐ Exclude

Summary
Field: [Injury.Severity]
Selection: Selected 1 of 3 values
Wildcard: All
Condition: None
Limit: None

Reset | OK | Cancel | Apply

Fig. 31. Filter severity


Filter [Location]

General | Wildcard | Condition | Top

● Select from list ○ Custom value list ○ Use all

Enter search text
☑ AK
☑ AL
☑ AR
☑ AZ
☐ CA
☐ CO
☐ CT
☐ DC
☐ DE
☐ FL
☐ GA

All | None | ☐ Exclude

Summary
Field: [Location]
Selection: Selected 4 of 53 values
Wildcard: All
Condition: None
Limit: None

Reset | OK | Cancel | Apply

Fig. 32. Filter location

In this case, the severity has been filtered by selecting 'fatal' and the location has been restricted to Alaska (AK), Alabama (AL), Arkansas (AR), Arizona (AZ) as shown in Fig. 31 and Fig. 32. Using only these constraints, the output has been presented. This is usually effective when deep analysis has to be made for a specific area or with a unique set of categories.

B. *Dashboard for Purpose of Flight Analysis*

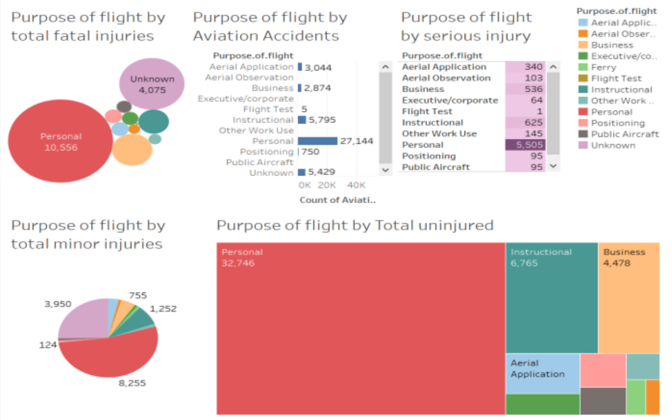


Fig. 33. Dashboard view for purpose of flight analysis

The Fig. 33 represents the dashboard that compares the purpose of flight with multiple factors present in the dataset. The dashboard is useful in providing an overall view of all the insights that have been gathered which can be further used to improve the current situation and identify relations between different factors. In this case, the factor has been compared with total fatal injuries, aviation accidents, total serious injury, total minor injuries and total uninjured.

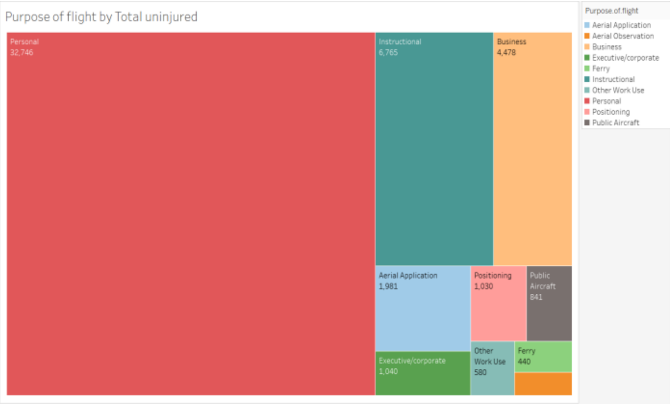


Fig. 34. Tree map for purpose of flight by total uninjured

The Fig. 34 tree-maps represent, the purpose of flight with total uninjured. It can be seen that majority of the flights were due to personal reasons and they also have the highest number of total uninjured people (32,746). With this visual presentation of data, some important information can be collected. In this case, the major reason for travel can be identified, along with the least factor that causes people to travel which is aerial observation.

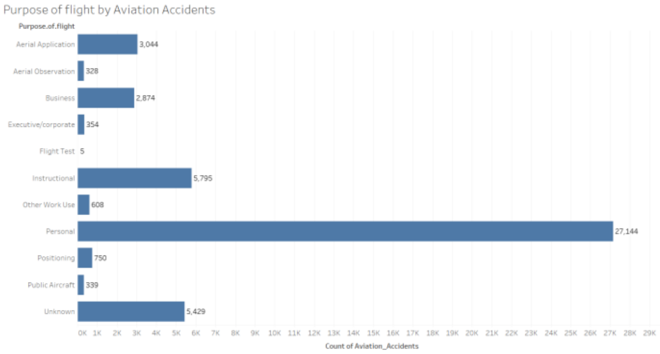


Fig. 35. Bar graph for purpose of flight by aviation accidents

The Fig. 35 bar chart presents the purpose of flight with aviation accidents count. It can be noted that the highest number of accidents is 27,344 and those were people travelling for personal reasons. This is very much related to the previous comparison as both cases denote majority of the people involved in the accidents, travelled for personal use and therefore, they also contribute to a higher number of people uninjured. Bar chart gives a clear and defined view of the data in a very cleansed manner.

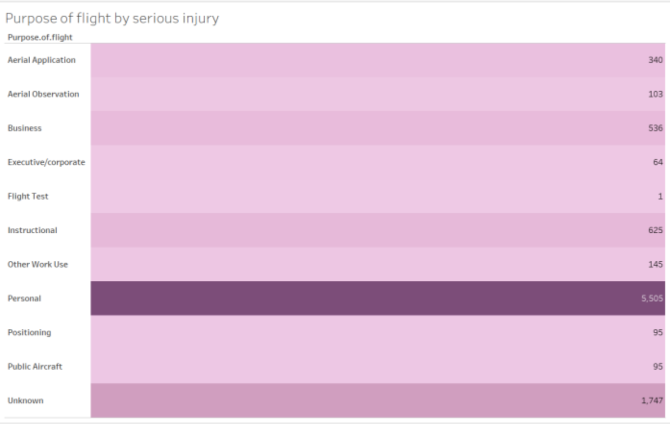


Fig. 36. Bullet graph for purpose of flight by serious injury

The Fig. 36 bullet graph shows the relation between purpose of flight and the number of serious injuries. The highest number of serious injuries is 5,505 which is among people travelling for personal use. As the insight gathered before, we can here specifically identify injury level among passengers travelling with different purposes.

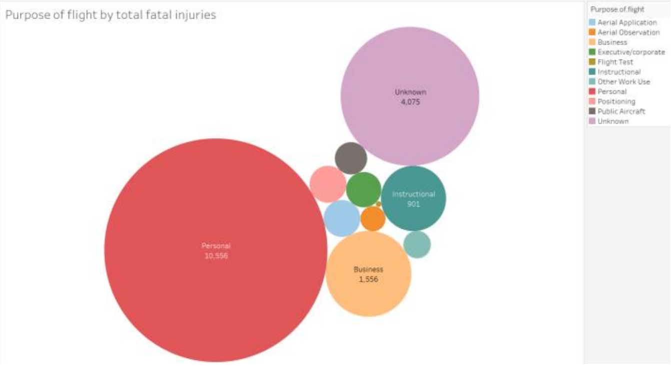


Fig. 37. Packed bubble graph for purpose of flight by total fatal injuries

The Fig. 37 bubble chart shows the amount of people who have been fatally injured due to the accidents and the reasons they travelled. Out of the many categories, many people who travelled for personal and business reasons had fatal injuries. A total of 10,556 injured people were personal travelling passengers while there were around 1,556 business passengers. In this case, the size of the bubble presents the bubble represents the fatal injury count whereas the color represents the different categories of passengers. It shows a very precise and easy to understand view of the data for insight gathering.

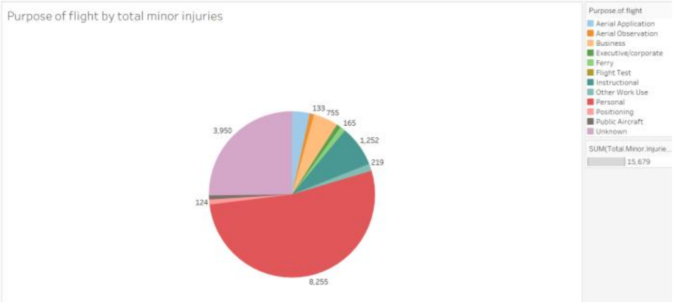


Fig. 38. Pie chart for purpose of flight by total minor injuries

The Fig. 38 pie chart represents the purpose of flight compared with total minor injuries. The different colours depict different categories of passengers while the space occupied by a colour represents the quantity of the people injured. As evident from the previous visual representations, many numbers of people were traveling for personal purposes and therefore, people from that category had the most numbers of injured and uninjured people. In this case, there was a total of 8,255 people from the personal travel category who were had minor injury. The above depictions show the identification of the highest category of people travelling and a deeper insight into the accident and injury data for those different categories.

C. *Dashboard for Injury Analysis*

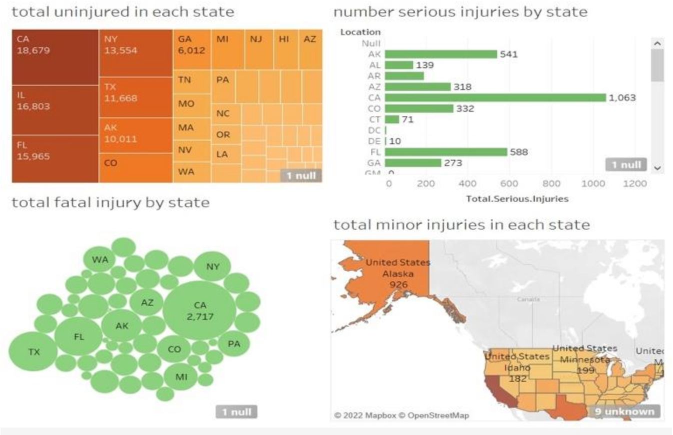


Fig. 39. Dashboard for injury analysis

The Fig. 39 illustrates an overview of the dashboard showing multiple factors from "Location" column. The location column shows the factors according to each state. In this case I used 4 factors named 'total uninjured, total serious, total fatal and total minor'. Each of which is represented using four different kinds of figures. The dashboard talks all about injuries and deaths. This will help the airline companies to improve their security measures in each state according to the data. Fact Bubbles easily identify where the factor is in high value. Maps are easy to identify geographically. Tree maps can easily give scores for each state and is easy to identify using colour scheme too. Bar charts give accurate statistics of the factors. Following is the explanation of each:

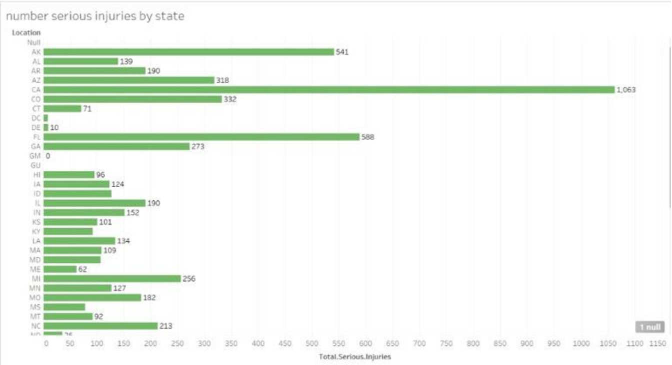


Fig. 40. Bar graph for number of serious injuries by state

According to this Fig. 40, we can analyze that in US, most of the serious injuries are in state CA (California) which are almost close to 1000. After that second highest injuries are in FL and AK. Meanwhile GM and GU have zero serious injuries.

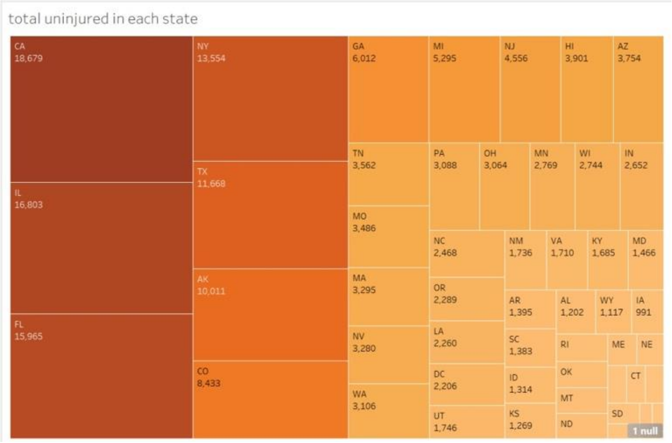


Fig. 41. Tree map for total uninjured in each state

The Fig. 41 relates to uninjured people in each state. We have seen that California had most serious injuries in previous figure but on the other hand it also has most injured people among other states too. IL, FL, and NY are on 2nd, 3rd and 4th ranks respectively, each having more than 10,000 people uninjured.

Now coming to the injuries which almost led to death, California is again ranked 1st with 2,717 fatal injuries. This might be because of excess number of travellers in CA. TX again is in 2nd position with more than 1200 fatal injuries and FL is ranked 3rd. Other states have very less reported fatal injuries due to less travellers as shown in Fig. 42.

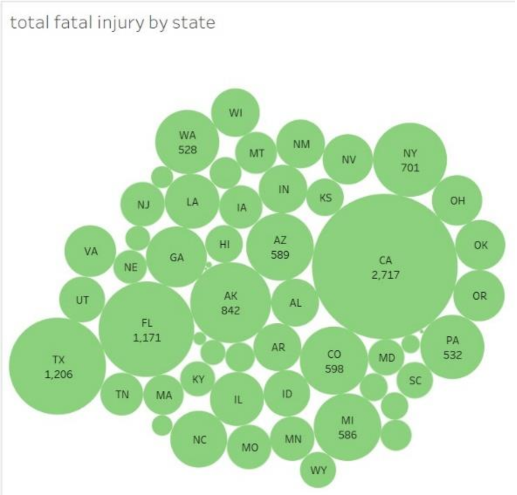


Fig. 42. Packed bubble graph for total fatal injury by state

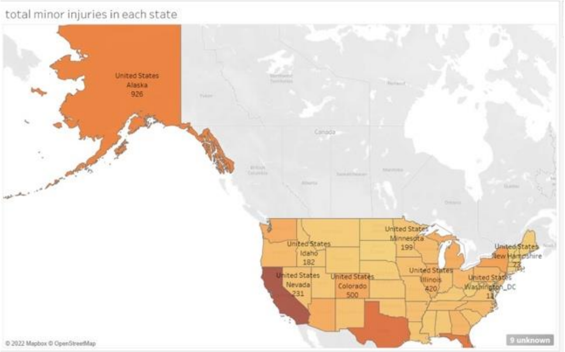
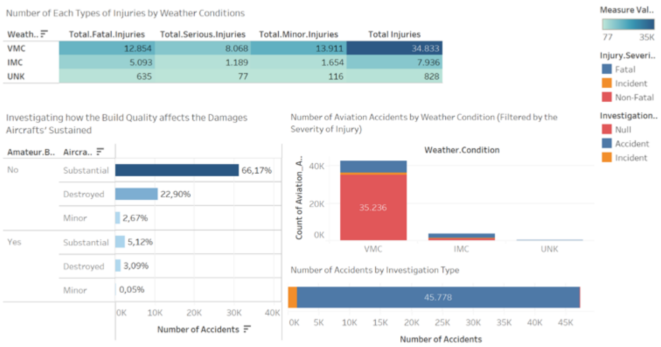


Fig. 43. Map for total minor injuries in each state

Coming to the minor injuries in each state, we used maps to locate them out. The region with most minor injuries is Alaska, with over 900 minor injuries confirmed. Second is Colorado with exactly 500 minor injuries and then ranks Illinois with 420 minor injuries as shown in Fig. 43.

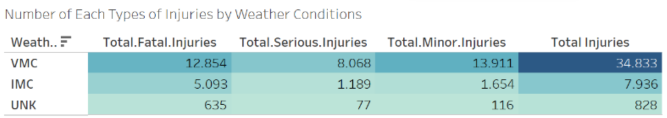
D. Dashboard for Weather Condition Analysis


Number of Each Types of Injuries by Weather Conditions

Weath..	Total.Fatal.Injuries	Total.Serious.Injuries	Total.Minor.Injuries	Total Injuries
VMC	12.854	8.068	13.911	34.833
IMC	5.093	1.189	1.654	7.936
UNK	635	77	116	828

Fig. 44. Dashboard for weather condition analysis

The Fig. 44 displays a dashboard that compares data based on the chosen criteria. In this dashboard, five pieces of information are shown. Three pieces of information used weather conditions as the factor to be compared to another criteria. One other displayed information of damage sustained based on the quality of the builder, which can be seen in the bottom left. The last one, investigation type is compared with the number of accidents that happened. One of the benefits of a dashboard is that it provides us with the means to view and analyze the key data that has been constructed. Hence, a conclusion can be drawn, and plans can be made accordingly.


Number of Each Types of Injuries by Weather Conditions

Weath..	Total.Fatal.Injuries	Total.Serious.Injuries	Total.Minor.Injuries	Total Injuries
VMC	12.854	8.068	13.911	34.833
IMC	5.093	1.189	1.654	7.936
UNK	635	77	116	828

Fig. 45. Table for number of each types of injuries by weather conditions

The Fig. 45 represents the number of injuries based on the weather conditions. It is displayed in the form of a highlighted text table. The table is constructed to enable the engineers to understand the weather situation that leads to accidents and what types of injuries it leads to. The total injuries caused by a particular weather condition can be analyzed too. Example, the above table shows that VMC generates a higher number of injuries compared to IMC on every level. From there, it can be concluded that humans tend to make more errors even with the situation is favorable. This is because IMC technically describe poor conditions, so machines are more heavily used. Therefore, the engineers can build parts that can improve the performance of humans (pilot and co-pilot).

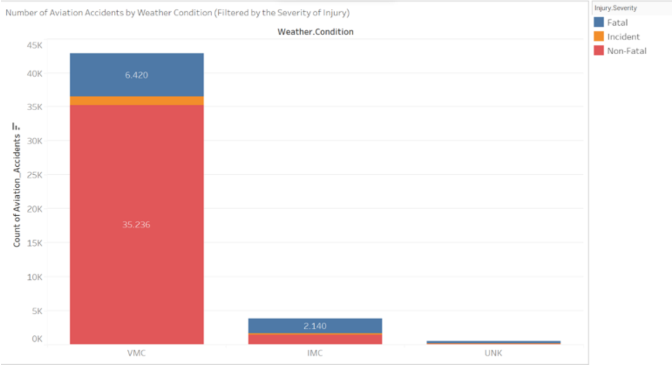


Fig. 46. Stacked bar graph for number of aviation accidents by weather condition

The Fig. 46 bar chart shows the number of accidents caused by weather conditions, where the bar chart is stacked with information on the severity of injury. The severity of injury inside the bar chart is separated with colour marks so the data can be easier to understood. In addition, the number of accidents per severity of injury is also displayed. The data visualization has a similar purpose to the previous one. The difference is that in this data, it describes a more detailed representation of the injury.

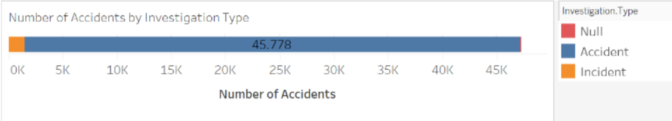


Fig. 47. Stacked bar graph for number of accidents by investigation type

The Fig. 47 stacked bar chart shows the type of accidents. It illustrates the differences in number of each accident type by separating it inside the bar. To differentiate between the investigation types, colour marks are used. Moreover, the number of accidents for each accident type is also displayed.

There are two reasons why stacked bar is used. The first is to fit the dashboard. The second is because we are dealing with small data where the factor is only one. For efficiency purposes, the number of accidents attribute is put in the column side. This data can be used by engineers to understand the fatalities of the events. An accident is described as an unexpected event that leads to damage and injury. While incident is described as an unexpected event that leads to some minor injury.

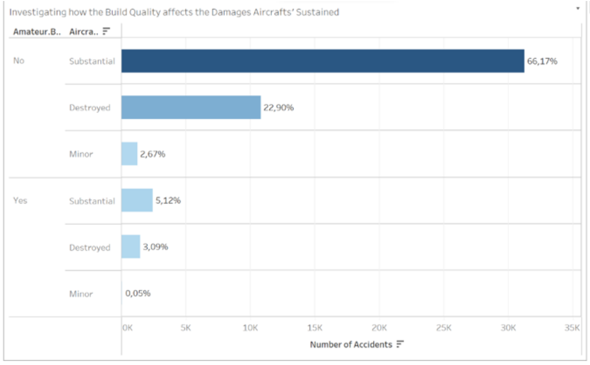


Fig. 48. Graph for investigating how the build quality affects the damages aircrafts' sustained

The data above is visualized using a bar chart in Fig. 48 that displays the damage aircraft sustained because of the build quality. The graph is constructed through two dimensions; the build quality and aircraft damage, and compare it by the number of accidents. The build quality is divided into two parts that compromise the components of aircraft damage. Colour is included to indicate the volume of the data in each criterion. The percentage in the chart above is over the total accidents.

The purpose of the above chart is not to compare directly the number of aircraft damages between aircraft that are maturely built or not. This is because the data distribution between the two criteria is too far. Hence, the above can be used to show the damage distribution on the aircraft. In each build quality, the aircraft mostly sustained substantial damage. It can also be noted that rarely aircraft suffered minor damage during aviation accidents.

E. Dashboard for Aircraft Engine Analysis

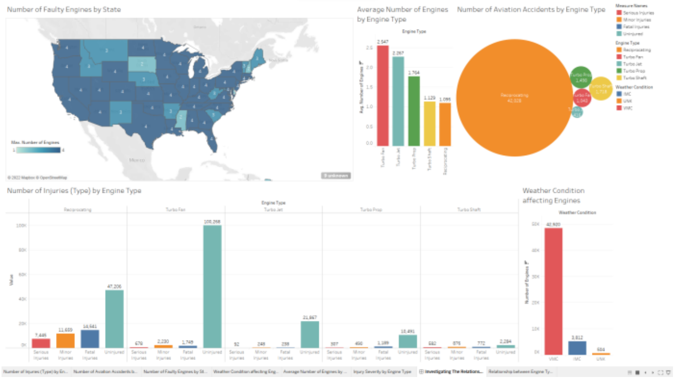

Number of Faulty Engines by State
Average Number of Engines by Engine Type
Number of Aviation Accidents by Engine Type
Number of Injuries (Type) by Engine Type
Weather Condition affecting Engines

Fig. 49. Dashboard for aircraft engine analysis

Fig. 49 shows a dashboard, it uses information about the aircraft's engines to perform visualizations. In here, there are five different graphs, all designed to help engineers inside of The Boeing Company to get statistics about their aircraft's engines. All these five graphs are there to help engineers and analysists in The Boeing Company to identify faults in their engine's designs or implementation through the performance and safety data. With this information, improvements or changes can be made to the aircrafts produced by Boeing in an effort to increase the safety of the flights and ensure the security of their flight crew and passengers.

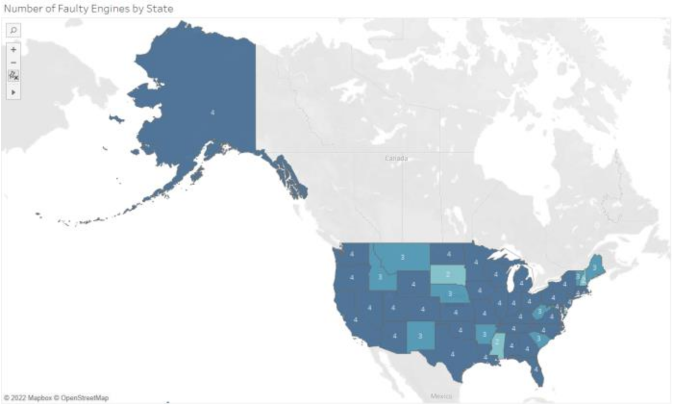
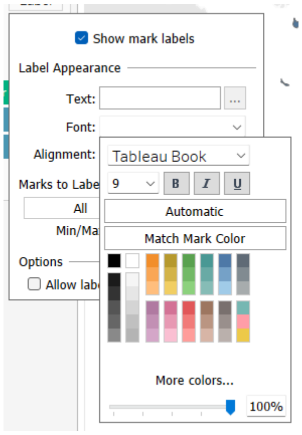

Number of Faulty Engines by State
Canada
Mexico
© 2022 Mapbox © OpenStreetMap

Fig. 50. Map visualization for the number of faulty engines by state

The first graph in the dashboard is a map that Fig. 50 illustrates the number of faulty engines for aviation accidents from each state in the US. It is done by using the state data, along with the maximum number of engines in the accident records for that particular state.



☑ Show mark labels

Label Appearance

Text:

Font:

Alignment:

Marks to Labe

All

Min/Ma

Options

□ Allow lab

Tableau Book

9 B I U

Automatic

Match Mark Color

More colors...

100%

Fig. 51. Settings for label on map

At first, the map only provides an illustration of the number of faulty engines through the shade of the blue colour, being the darker the blue, the more engines that are faulty for that state. However, we thought that this is unclear and would not be enough for the analysts and engineers inside Boeing. Thus, the "Show mark labels" option was enabled as shown Fig. 51, with the label's text colour adjusted to white, as a black font might be hard to read with dark blues.

This map graph is created so the engineers in Boeing can get to know the maximum number of engines that turned out to be faulty for each of their aircrafts. It is supported with the map to differentiate the data by state, so that in the future, if the departure point of the flight is available for Boing's BI dashboard, the engineers can measure the distance between the departing point and point of accident, while investigating the effects from the number of engines used.

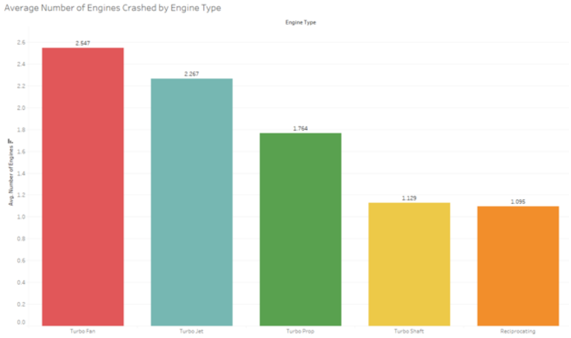


Fig. 52. Bar graph of the average number of engines crashed by engine type

For the second visualization, Fig. 52 it is a bar graph which shows the average number of engines in aircrafts that crashed, separated by its engine type. The different types of engines are shown with the colour mark, which allows for easy reading of the visualization and data. The average number of engines by engine type is also shown, as a label, on top of each bar.

The bar graph would help the aircraft engineers understand the average number of engines installed for each engine type which was involved in aviation accidents. With this data, the engineers can perform a more thorough investigation on the relationship between the number of engines installed, the installed engine type, its performance and reliability.

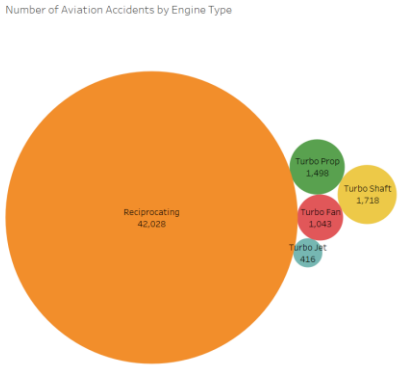


Fig. 53. Packed bubble visualization for the number of aviation accidents by engine type

The third visualization for the engine dashboard is the packed bubble graph as shown in Fig. 53, which shows the number of aviation accidents by its engine type. The packed bubble graph illustrates the difference in accidents between each engine type by differentiating each bubble's size. The bubbles are also identified with the colour marking option, to make it more easily readable.

The purpose for comparing the number of aviation accidents caused by each engine type through this packed bubble graph is to allow for an easier to understand visualization of the relationship between these two data. The data analysts and engineers can look at this graph and get to know the number of accidents involved by each different engine type. For example, the packed bubble graph shows that the reciprocating engine type caused the greatest number of aviation accidents. Thus, they can investigate into the reason behind the engine type's performance, to determine whether improvements should be made, or a different engine should be used.

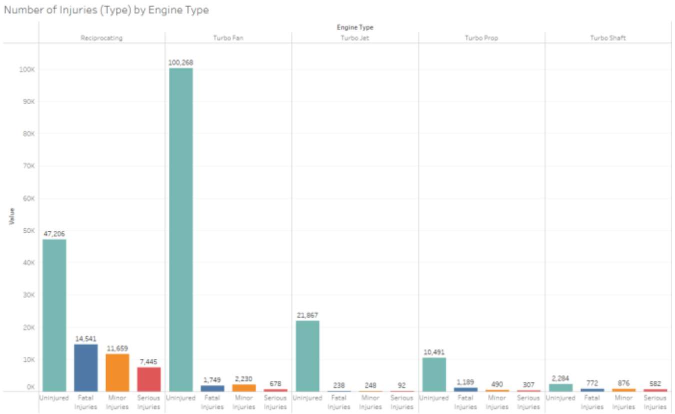


Fig. 54. Bar graph of the number of injuries (type) by engine type

The fourth visualization is shown in Fig. 54 bar graph, which shows the difference in number of each injury type, separated by the engine type. This graph is possible by putting the total number of fatal injuries, serious injuries, minor injuries and uninjured into one measure values group, then comparing it with the engine type. Each engine type is divided into one section, with five sections in total. While the number of injuries by type are shown for each engine type using different colours for each injury type.

This bar graph can be considered as the one that ties all of the previous visualization together. This is because it shows the total for each type of injury, compared among different engine types. Previously, we had a bar graph which showed the average number of engines for each engine type and a packed bubble graph to illustrate the number of aviation accidents by engine type. From these two graphs, the engineer or the analyst inside Boeing may conclude that the low number of average engines is one of the reasons that reciprocating engine types were involved in most aviation accidents. While the turbo fan engine type used, on average, the most engines per aircraft, at the same time having the second least aviation accidents.

This bar graph goes further to help the analysis by showing that, while luckily most of the passengers involved were uninjured, the reciprocating engine type had the greatest number of injuries for every injury type. On the other hand, the high average number of engines used for turbo fan aircrafts attained the greatest number of uninjured passengers. Judging from the data on the dashboard, the engineers may try to incorporate characteristics or improvements of turbo fan engines into reciprocating engines, in hopes to improve the safety and reliability of aircrafts with the engine type. Otherwise, the engineers can phase out reciprocating engine aircrafts in favour for safer engine types.

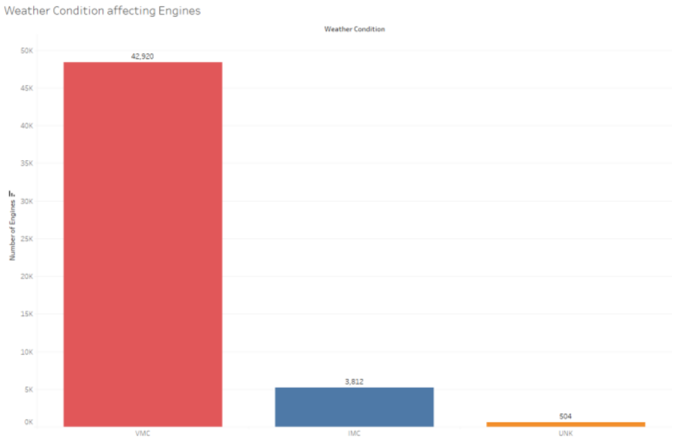


Fig. 55. Bar graph for the weather condition affecting engines

The last visualization for this engine dashboard is shown in Fig. 55 the bar graph which shows the how the different weather conditions affect the aircraft's engines. It is done by color coding the different weather conditions and comparing them with the accident numbers with the engine types.

In aviation, VMC means that the weather condition is good enough that the pilot can attain visuals of the skies, nearby possible aircrafts and the nearby terrain. Meanwhile, IMC describes when the visibility of the aircraft's surroundings is limited by bad weather conditions or when flying through thick clouds. In a situation like this, trained pilots will have to rely on the hearing equipment available in the cockpit to remain updated about the aircraft's current whereabouts and surroundings.

This current state of the graph eliminates one common suspicion of engineers, which is that bad weather conditions are the main reasons that aviation accidents happen. However according to this bar graph, it is not the case as most of the accidents recorded actually happened when the weather condition is satisfactory for flight. While bad weather conditions do also play considerably small part in affecting the chance of an aviation accident happening, pilots nowadays have access to more modern and advanced monitoring equipment that allows them to keep track of the plane's status, location, elevation and such. This lends into the argument that weather conditions do not necessarily cause aviation accidents, but instead fault in the aircraft's design or the pilot's inexperience or mistake had caused the accident to happen.

VII. Conclusion

BI is put into action using programming or devices which assemble tremendous measures of information and change them into valuable data which can be investigated to likewise acquire bits of knowledge and decide. Utilizing this methodology, the association will actually want to recognize past patterns and information that untruths concealed inside verifiable information gathered throughout the long term.

Boeing being the main organization in the flying area has numerous obligations and should stand against all the challenges it is facing right now. Most of the challenges include unexpected government regulations, market demands; technological advancements market demands, global coordination risk and preapproval tests. Boeing can combat all these problems easily by working on the market opportunities. For instance, it can increase its seating capacity and produce engines with more fuel efficiency as compared to competitors. Moreover, the Boeing team can analyze past incidents and an effective design change considering past accidents.

Finally, it can improve its technological side by doing some advancement in exchange of digital information, weather prediction, connected cabin systems and cameras to provide passengers a view of outside plane. Coming to the solutions, Boeing can implement prediction of encountering into a crash using BI and then use countermeasures to prevent it from happening. Another application where BI can be used is aircraft maintenance scheduling using predictive maintenance. Business Intelligence is also extremely beneficial in identifying the cause of aircraft crash. The autopilot system also learns from such situations and prevents the craft from being going into deadly situations too.

To sum this up BI is an essential tool in modern day businesses and organizations. It helps companies falling from mishaps and increases their performance along with user security. The data was clearly assessed in this assignment, and we got valuable statistics from it which can also help the company improve its insights.

REFERENCES

[1] Boeing (2019). *Boeing: Data Analytics*. [online] www.boeing.com. Available at: https://www.boeing.com/services/government/data-analytics.page.

[2] Calzon, B. (2022). *Discover Data Warehouse & Business Intelligence Architecture*. [online] BI Blog | Data Visualization & Analytics Blog | datapine. Available at: https://www.datapine.com/blog/data-warehousing-and-business-intelligence-architecture/.

[3] HT (2019). *Naive Bayes Algorithm*. [online] Medium. Available at: https://medium.com/@hackares/naive-bayes-algorithm-e565daa89eb7.

[4] IBM Cloud Education (2020). *What is Data Modeling?* [online] www.ibm.com. Available at: https://www.ibm.com/cloud/learn/data-modeling.

[5] Kaggle (2022). *Aviation Accident Database & Synopses*. [online] kaggle.com. Available at: https://www.kaggle.com/khsamaha/aviation-accident-database-synopses.

[6] KDnuggets (2020). *Decision Tree Algorithm, Explained*. [online] KDnuggets. Available at: https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html.

[7] Mark Logic (2022). *Boeing: Querying Model Based Systems Definition Data | MarkLogic World*. [online] MarkLogic. Available at: https://www.marklogic.com/resources/boeing-querying-model-based-systems-definition-data/

[8] Mihai, A. (2015). Airline Applications of Business Intelligence Systems. *INCAS BULLETIN*, 7(3), pp.153–160. doi:10.13111/2066-8201.2015.7.3.14.

[9] Rushikesh Pupale (2018). *Medium*. [online] Medium. Available at: https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989.

[10] Saud, D. (2021). *How can business intelligence revolutionize the airline industry*. [online] Folio3 Dynamics Blog. Available at: https://dynamics.folio3.com/blog/business-intelligence-revolutionize-airline-industry/.

[11] Segal, E. (2022). *Boeing Faces New Challenges To Image, Reputation And Credibility*. [online] Forbes. Available at: https://www.forbes.com/sites/edwardsegal/2021/10/15/boeing-faces-new-challenges-to-image-reputation-and-credibility/?sh=2d1b3fbb6bff

[12] Sherman, R. (2020). *7 Data Modeling Techniques and Concepts for Business*. [online] SearchDataManagement. Available at: https://www.techtarget.com/searchdatamanagement/tip/7-data-modeling-techniques-and-concepts-for-business.

[13] Taylor, D. (2022). *What is Dimensional Modeling in Data Warehouse? Learn Types*. [online] www.guru99.com. Available at: https://www.guru99.com/dimensional-model-data-warehouse.html.

[14] Tripathi, A., Bagga, T. and Aggarwal, R.K. (2020). Strategic Impact of Business Intelligence : A Review of Literature. *Prabandhan: Indian Journal of Management*, 13(3), p.35. doi:10.17010/pijom/2020/v13i3/151175.