Data stories and dashboard development: a case study of an aviation schedule and delay causes

Data stories and dashboard development: a case study of an aviation schedule and delay causes

S S Salleh[1,2]*, A S Shukri[1], N I Othman[1] and N S M Saad[1]

[1] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Cawangan Negeri Sembilan, Kampus Seremban, Negeri Sembilan, Malaysia
[2] Malaysia Institute of Transport (MITRANS), Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia

*ssalwa@uitm.edu.my

Abstract. In this case study, five key processes in modelling a data story of aviation data patterns during COVID-19 have been executed. It started with the collection of secondary data from relevant sources. Data inspection, transformation, and preparation activities, including data cleaning, filtering, and sampling, are all included in this work. Iterative exploratory data analysis (EDA) has been conducted to determine the pattern of each independent attribute, followed by an assessment after the data story is modelled and integrated on a dashboard. The questionnaire has been distributed and the visuals were assessed by giving respondents a few tasks to interpret stories based on their comprehension. The result shows that the data stories have been interpreted in a similar narrative by all the respondents. The overall mean score is 4.71, and this significantly shows that the respondents agree and strongly agree that the visual objects help in communicating patterns and stories. The overall process gives researchers experience and guidelines for future work. Overall, the objectives of the study have been met. Nevertheless, it gives researchers a lot of experience in interpreting data, cleansing and transformation, analysis, modelling the visualisation by selecting suitable charts, and integrating the objects together into a dashboard.

Keywords: Data visualisation. Data stories, aviation, dashboard, descriptive analysis

1. Introduction

Exploring data stories and presenting the outcomes in the form of a visual uses a lot of basic and advanced graphics as its main aim is to convey insights and tell a dataset's story [1]. Its goal is to help the audience understand the story by using a simplified form of two-dimensional or three-dimensional charts. Typically, these visuals are incorporated into a dashboard for operational analysis because they allow users to explore the story interactively. The dashboard gives users an overall perspective of the situation as it currently stands without going into too much detail [2]. In this context, a dashboard is generally defined as a straightforward front end for monitoring, analysing, and optimising important business processes by empowering people at all hierarchical levels to make better decisions [2, 3]. Ideally, the visuals must be understandable and intuitive for non-technical users in terms of data display to ensure that users gain meaningful knowledge to make timely decisions and other associated activities [3]. It can be in the form of single-view reporting displays or interactive interfaces with numerous views and uses [4]. However, the type of chart used in the dashboard may or may not facilitate comprehension, whereby choosing the incorrect chart type could confuse users or cause data misinterpretation. Thus, the whole process of developing a dashboard must be exercised.

In the context of the aviation industry, discovering data stories is one of the topics that has been researched as there is a fragment in the ecosystem, or a gap between the data scientists and the domain experts in the industry [5]. It is important for the industry to investigate and extract data stories of the aviation schedule, as well as identify the pattern of flight operations to help the operators understand the current situation. One of the most significant issues is flight delays, as it is a performance indicator of any transportation system. Notably, commercial aviation defines delay as the amount of time an aircraft is running late or delayed [6]. The U.S. Department of Transportation reported that flight problems were the highest category of the complaints received in June 2022, with 1,686 (28.8%) concerned with cancellations, delays, or other deviations from airlines' schedules [7]. This demonstrates how important it is to analyse the flight delay and its causes.

In his work, Vishwakarma et al. mentioned that delays cause effects leading from passengers' monetary value to dissatisfaction [8]. He analysed and stated that flight delays are caused by many diverse factors, ranging from adverse weather conditions, technical problems, restrictions in air-traffic, and many more. They used big data technologies like Hadoop MapReduce, Apache Hive, Apache Pig, Hadoop Distributed File System (HDFS), and MySQL in the analysis. Results show that the effect of bad weather has a major impact on flight delays. However, there was no data visualisation output shown in the study. Another study by Kumar et al. [9] proposed a method that addressed flight delay prediction that applied computational and machine learning approaches. Besides, they also presented a timeline of major works that depicts the relationship between the factors of the flight delay. His work used some visuals, but not in a dashboard interactive mode. Richard et al. [10] work on an expert system that acts as an intelligent agent to interrogate past aircraft occurrences using a Fuzzy Logic System (FLS). The expert system is presented graphically and interactively using AcciMaps. However, not much visual has been assessed and discussed in his work. Other works have been reviewed [11,12,13], but work on aviation datasets and dashboard development is extremely limited. Thus, this research represents an effort to investigate and present the process of analyzing, designing, and developing a dashboard with visuals related to aviation operations monitoring. To demonstrate the findings, this study proposes a prototype that has been developed. To meet that purpose, the study's objectives are as follows: (i) to identify, select, and prepare relevant attributes for aviation delay monitoring and causes; (ii) to design visual objects to facilitate the data stories of the attributes; and (iii) to assess the dashboard's formative functionality and usability of the data story.

This paper is organised and starts with an introduction (I) that covers the study background, problem's definition, objectives, and related work. In Section 2, methods and materials (M) are presented, followed by Section 3 that presents the result (R) and discussion (D). Concluding remarks are in the final section, where future work is proposed. Within the dataset, the objective of data stories that are intended to be obtained are: (i) the trend of scheduled flights and actual flights based on monthly; the highest flight distribution by states; and the busiest airline; (ii) the pattern of the on-time, delayed, and cancelled status of flights and the flight's punctuality based on the duration of delays during flight departure and flight arrival; and (iii) identify the most likely causes of flight delay and reasons for flight cancellation.

2. Methods and Material

This study consists of four steps. The following Figure 1 shows the overall steps. The first step was to obtain the data from relevant sources, where technical skills such as MySQL are necessary to analyse the data. The second step involved data examination, transformation, and preparation duties, where data scrubbing, filtering, and sampling have been done. Corresponding to that, scrubbing data also includes suggesting values for the missing data as well as deriving new attributes. The third stage comes afterwards by exploring the data to get an overview of each attribute pattern. Different data types, such as numerical data, categorical data, ordinal and nominal data, will require different treatments. An Explanatory Data Analysis (EDA) technique has been implemented iteratively in steps 2 and 3 as it helps to identify the pattern of data in each independent attribute. The fourth stage is the modelling of the data story in the form of visual representation and arranging each of the objects on a dashboard.

Here, the selection of significant types of data visualisation delivers the correct story of the data. In the final step, which is an assessment where the models and data stories will be interpreted by giving 13 respondents a few tasks to interpret the data stories based on their understanding. Findings from the assessment help with improving the data visualisation and dashboard design.

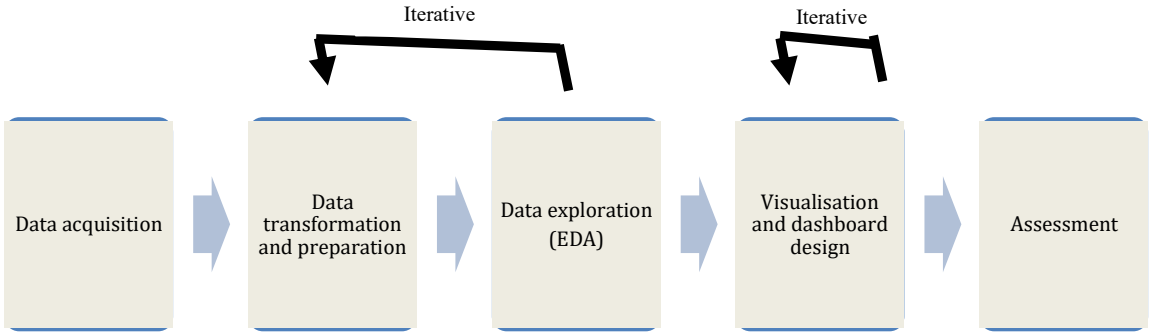

Figure 1. Steps involved in the study

The target users of the dashboard are airline R&D personnel. The dashboard will be useful for them to view and understand the patterns interactively and to understand the patterns of previous data stories, which helps in making concise decision-making that minimises the cost incurred due to rescheduling flight delays and cancellations. The dataset used in this study is "*On-Time: Reporting Carrier On-Time Performance (1987–present)*". It has been gathered and compiled by the Bureau of Transportation Statistics of the United States Department of Transportation and is accessible from [14]. The original datasets contain 4,688,354 records of one-year data from January 1st, 2020, to December 31st, 2020. Random data sampling had been done to cater to only 2% of the entire sample size, as it was reduced to only 68,567 records. Twenty attributes have been chosen as they are closely related to the data story objectives. To maintain the pattern in the sample, the actual pattern has been referred to from the original dataset. As mentioned above, the EDA has been performed on independent attributes as this is to explore data patterns, identify any anomalies, and verify assumptions using summary statistics and graphical representations. Figure 2 shows the visual of the attribute types of airlines: departure, cancellation status, reasons of cancellation, and causes of delay. During EDA, blank values for some of the attributes have been found. However, from analysis, these blank values have their own meanings. For instance, the blank values in the departure delay and arrival delay will indicate that the flights are being cancelled. The blank values for the reasons of the cancellation will indicate that the flights are not being cancelled. On the other hand, the blank values in the delay time for all the causes of delay have been tokenized as 0 to indicate that there are no delays for that flight corresponding to those causes. Apart from that, there are several outliers as shown in the boxplot for the departure delay and arrival delay attributes. To overcome this, all the charts that are related to both attributes will use the average delay time for each flight departure and flight arrival. This is to ensure that the outliers do not affect the overall data story. Data preparation involves a lot of tokenization tasks. For example, in the token for the cancellation status, 0 means the flight is not cancelled; 1 means the flight is cancelled. For the purposes of this cancellation, token A means carrier, B means weather, C means National Aviation System (NAS), and D means security.

Figure 2. Data visualisation of an independent attribute from an EDA exercise

To achieve the data story objectives, two new attributes have been derived, which are "Departure Status" and "Arrival Status". According to the Federal Aviation Administration (FAA), a flight is said to be delayed when it is late for more than 15 minutes of its scheduled time [15]. Figure 3 shows the snapshots of the derivative attributes.

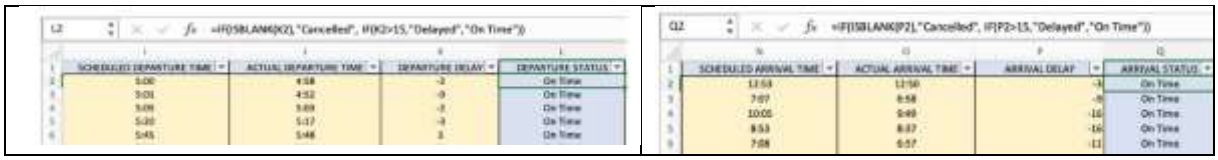

Figure 3. Snapshots of derivative attributes

In the design and development process, the arrangement of visual objects in the dashboard has been done in accordance with Gestalt theory [16]. While the chart was chosen based on the proposed study by [17], The study mentioned that it is important to use charts that are appropriate and capable of providing an in-depth interpretation and presentation of the data in a split second. To accomplish this, the design and implementation exercise went through three cycles of data story interpretation exercises and experimented with various types of charts before settling on the final one. Literally, it is a repetitive process where changes and modifications are made frequently until the dashboard is finally completed and satisfied by the respondents. The respondents of the exercise were three experts in data visualisation and ten graduate students who had taken a course on data visualization, and they were asked to rate the dashboard. Figure 4 shows the dashboard design and its reading flow have been indicated in five dotted lines. Charts constructed to achieve related objectives have been shown in the red box. Each of the objects and charts serves its own purposes as described in Table 1.

Dashboards provide flexibility for users to modify the visual representations of objects inside those objects or select the dimensions and measures to visualise [4]. During the assessment, the reading flow of the dashboard starts with page 1. In the first row, the user will first look at 4 filters, which are filters for the month, airline, state, and flight status. These filters will be used to navigate the data according to the user's preferences. Starting with the column labelled as 2, the user will look at the 5 cards that display the total number of scheduled flights, actual flights, cancelled flights, airlines, and states. Following that, the user will move to the next column at 3. In this column, the user will first look at the combination of bar chart and line chart on the number of flights by month, as well as a map chart on the distribution of flights by state. Lastly, for page 1, a bubble chart of the number of flights by airline is displayed. Continue to page 2, where the user will look at the fourth column. This column consists of a simple chart of the number of flights by status. Next to this chart, there will be two tornado charts that display the average departure and arrival delays by month, as well as the percentage of flights by departure and arrival punctuality. Lastly, the user will look at the last column, labelled as 5. In this column, there will be a spider chart and a 100% stacked bar chart. The respective charts are used for the percentage likelihood of the cause of delay and the reasons of cancellation by month. In addition, all the charts have been arranged according to the sequence of the data story objectives. Meanwhile, the second objective is achieved by the first 3 charts on page 2, indicated by label number 4. These charts explain the on-time, delayed, and cancelled status of flights and the flight's punctuality based on the duration of delays during flight departure and flight arrival. Lastly, the last 2 charts on page 2, indicated by label number 5, explain the most common causes of flight delays and the reasons for flight cancellations by month.

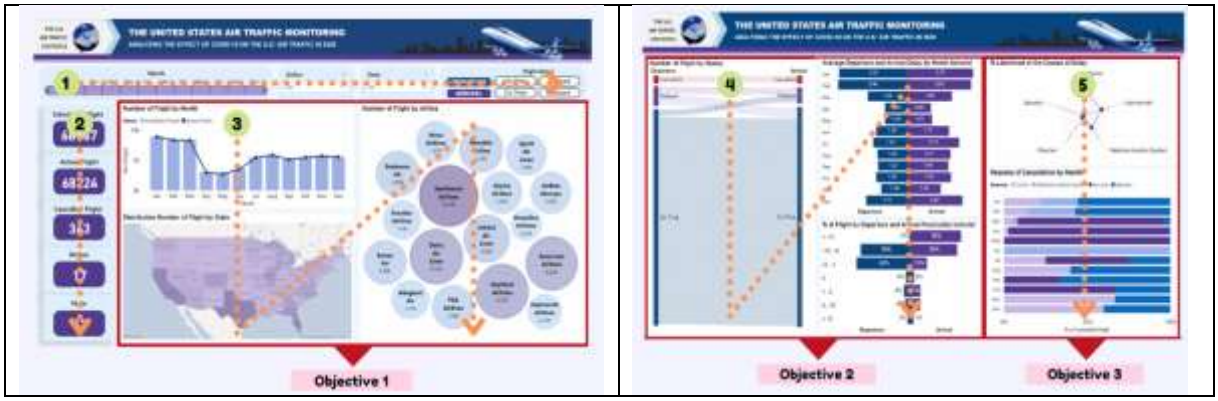

Figure 4. Screen shots of the dashboard shows its reading flow

Table 1. Visual objects and its purposes

No	Charts Title	Purposes
1	Number of Flight by Month	To analyse the number of flights that follow the schedule every month,
2	Distribution of Flight by State	To determine the busiest state based on the most flights scheduled
3	Number of Flight by Airline	To determine the busiest airline based on the most flights scheduled
4	Number of Flight by Status	To determine the status of the relationship between the flight departure and flight arrival.
5	Average Departure and Arrival Delay by Month (minute)	To compare and analyse the average flight delays between departure and arrival every month.
6	Percentage of Flight by Departure and Arrival Punctuality (minute)	To analyse the punctuality of the flights based on the percentage of delays during departure and arrival.
7	Percentage likelihood of the Causes of Delay	To investigate the overall main cause of flight delays.
8	Reasons of Cancellation by Month	To investigate the main reason for flight cancellations every month.

3. Results

The assessment followed and adopted a technique, namely "*Evaluating Communication through Visualization*" (CTV) [18,19]. The result was obtained based on the Likert scale used in the instrument on a 5-scale where 0 to 5 indicates strongly disagree, neutral, agree, and strongly agree, respectively. The respondents had been given a set of questions to answer, and in the last part, there were three writing tasks at random on storytelling narration. The scores for the narration were based on a 5-scale rubric of adequate, adequate but not comprehensive, partly comprehensive, comprehensive, and very comprehensive, respectively. The overall mean score is 4.71, and this significantly shows that the respondents strongly agree that the visual objects are able to communicate the stories significantly. The outcome of the story-telling narration tasks indicated that the users were able to comprehend the data stories.

Table 2. Result of CTV and narration tasks exercise

Respondents	Questions							Overall Mean	Std Dev
	I learn the data insights better and/or faster using the visualization tools	Can I learn the data insights better and/or faster using the visualisation objects	Each visual object is helpful in explaining and communicating the insights	I interact with each visualisations object using filters, sliders, and boxes	Can useful information be extracted from a casual information visualization? Task 1	Task 2	Task 3		
1	4	4	4	3	3	4	4	3.71	0.49
2	4	5	5	5	4	4	4	4.43	0.53
3	4	5	5	5	4	4	4	4.43	0.53
4	4	5	5	5	4	4	4	4.43	0.53
5	5	5	5	5	5	5	5	5.00	0.00
6	4	5	4	5	4	4	4	4.29	0.49
7	4	5	5	4	4	4	4	4.29	0.49
8	4	5	5	5	4	4	4	4.43	0.53
9	4	5	4	5	4	5	5	4.57	0.53
10	5	5	5	5	4	5	5	4.86	0.38
11	5	5	5	5	4	4	4	4.57	0.53
12	4	5	4	4	4	4	4	4.14	0.38
13	5	5	5	5	4	5	4	4.71	0.49
Average								4.45	0.45

4. Discussion

4.1. Data story

The objectives of the data stories have been successfully achieved, and respondents have consistently written similar narrations of stories in the exercises. The monthly flight schedule showed that most of the flights have accordingly followed the schedule throughout the months, except for the first and second quarter of the year, especially in March and April. Besides, the number of flights has also significantly dropped during these two months. As a matter of fact, the occurrence of COVID-19 during the early months of the year 2020 is evidence of its contributing to flight cancellation. It was also found that the busiest states that operated the highest number of flights were covered by Texas and California (a total of 21% of flights) on the West and South sides of the mainland of the United States, respectively. The busiest airline is Southwest Airlines, where they operate 20% of the flights throughout the year. Despite the impact that comes from the COVID-19 outbreak, most of the flights are on-time during both departure and arrival at more than 90%. Adding to that, both the monthly average departure delay and the monthly average arrival delay only range between 0.93 minutes and 2.88 minutes, which is considerably not too late. Most of the flights are also likely to be punctual upon departure and arrival, with more than 80% of the flights in total. Regardless, the most common cause of delay is due to carrier issues. To minimise the likelihood of the same cause occurring again, circumstances within the airline's control, such as maintenance or crew problems, aircraft cleaning, baggage loading, fuelling, need to be managed excessively from time to time. As for the common reason for cancellation, it happens to be

security issues, with more than 60% of cancelled flights for each month of occurrence. However, bad weather has the major occurrence of cancelled flights with a total occurrence of 10 out of 12 months as compared to security with only 6 months of occurrences. The analysis demonstrates that COVID-19 indeed has a clear impact on the United States' air traffic, where the number of flights has dropped by more than 40% on average. However, the evidence of an inconsistent rise and fall in the demand for flights throughout the third and fourth quarter of the year shows that airline companies have been slowly recovering but still hurting from the unforeseen consequences caused by the spread of COVID-19 variants. With that, it can be concluded that there are 17 airline companies, 343 flights cancelled, 5 causes of flight delays, and 4 reasons for the flight to be cancelled.

4.2. Lesson Learnt

Based on this study, especially through the EDA, it was found that preparing the dataset sample has a few problems:

a) The original dataset is huge and may contain many attributes that are not relevant to the study objectives. Thus, the selection of specific attributes needs to be conducted carefully to ensure that all relevant attributes are not overlooked. On the other hand, some of the attributes must be derived to fulfil the analysis aims set at the initial stage.
b) Values in attribute “delay” in this dataset can be caused by multiple issues at the same time. For instance, a delayed flight can be caused by both carrier issues and security issues at the same time. The causes of these were divided into multiple attributes in this study. These five causes of delay need a machine learning algorithm to group them together automatically. In our future work, further analysis will be done to handle this issue.
c) It has been discovered that some attributes contain blank values with valuable meanings. Since the values in each attribute have not been properly defined, mistakes in data interpretation could happen as researchers may assume that the data is not complete and not valuable. Fortunately, in this study, the researchers were able to realise the importance of these blank values as they have meaning. For example, the blank values in the departure delay and arrival delay indicate that the flights are being cancelled. On the other hand, the blank values for the attribute "reasons of the cancellation" indicate that the flights are not being cancelled.

5. Conclusion

In this study, five key processes in the modelling of a data story in the form of a visual were carried out, and the objects were integrated into a dashboard. It started with data collection from relevant sources. Data inspection, transformation, and preparation activities, including data cleaning, filtering, and sampling, are all included in this work. The dataset was then examined to gain a general understanding of each attribute pattern. An iterative exploratory data analysis (EDA) has been employed to determine anomalies and the data pattern in each independent attribute. An assessment has been conducted after the data story and the dashboard have been modelled and developed. The result shows that, overall, the data stories have been interpreted in a similar narrative by all the respondents. This study proves that collaborative and iterative efforts are required and helps in ensuring data representation is chosen in a way that tells the data's remarkable story. The overall process and findings help us in our future work to improve data visualisation and the dashboard design process. Since all pertinent attributes had been correctly selected and they had been prepared accordingly, the study's overall objectives were all achieved. Implicitly or not, this study has provided researchers with such comprehensive experience in conducting data analysis, data visualisation, and dashboard development. In the future, analytics based on machine learning will be added to the dashboard to make it work better.

Acknowledgements

The authors would like to thank the Malaysia Institute of Transport (MITRANS), UiTM for sponsoring this publication.

References

[1] Khalid A, Hassan N, Razak N and Baharuden A 2020 Business intelligence dashboard for driver performance in fleet management *Proceedings of the 2020 11th International Conference on E-Education, E-Business, E-Management, and E-Learning* 347-351
[2] Salleh S S, Mohamed N A, and Shah N A 2021 Simulating data stories of clients' credit card default *Application of Modelling and Simulation* **6** 184 – 190
[3] Carlos J Costa and Manuela A 2019 Supporting the decision on dashboard design charts *The IIER International Conference*, p10-15
[4] Sarikaya A, Correll M, Bartram L, Tory M and Fisher D 2019 what do we talk about when we talk about dashboards? *IEEE Transactions on Visualization and Computer Graphics* **25**(1) 682-692
[5] Paula Lopex 2017 Discovering hidden knowledge in aviation data, Retrieved on 28/9/2022 https://datascience.aero/discovering-hidden-knowledge-aviation-data/
[6] Carvalho L, Sternberg A, Maia Goncalves L, Beatriz Cruz A, Soares J, Brandao D, Carvalho D, and Ogasawara E 2020 On the relevance of data science for flight delay research: a systematic review *Transport Reviews* **41**(4) 499-528
[7] Air travel consumer report: consumer complaints up from may nearly 270 percent above pre-pandemic levels 2022 US Department of Transportation *Technical report* retrieved on 28/9/2022, https://www.transportation.gov/briefing-room/air-travel-consumer-report-consumer-complaints-may-nearly-270-percent-above-pre
[8] Prasant V 2019 Big data analytic on aviation trends in U.S. and determining possible enhancements for flight delays *Technical Report*
[9] Manjunatha K B, Achyutha P, Kalashetty J, Rekha V S and Nirmala G 2022 Business analysis and modelling of flight delays using artificial intelligence *International Journal of Health Sciences* **6**(S1) 7897–7908
[10] Ng C, Bil C, Sardina S and O'bree T 2022 Designing an expert system to support aviation occurrence investigations *Expert Systems with Applications* 117994 207
[11] Jiang Y, Li S, Huang J, and Scott N 2019 Worry and anger from flight delay: Antecedents and consequences *International Journal of Tourism Research* **22**(3) 289 – 302
[12] Sivaiah G V and Sheeja R 2020 Flight scheduling for airport throughput and flight delay optimization *Journal of Critical Reviews* **7**(14) 1211 – 1217
[13] Tian H, Presa-Reyes M, Tao Y, Wang T, Pouyanfar S, Miguel A, Luis S, Shyu M, Chen S, and Iyengar S 2021 Data analytics for air travel data: A survey and new Perspectives *ACM Computer. Survey* **54**(8) 1-35
[14] Bureau of Transportation Statistics, 2022 https://www.transtats.bts.gov/DL_SelectFields.asp?gnoyr_VQ=FGJ&QO_fu146_anzr=b0-gvzr, *On-Time: Reporting Carrier On-Time Performance (1987-present).* Retrieved on 30 Sept 2022
[15] Hajko J and Badánik B 2020 Airline on-time performance management *Transp. Res.* Procedia **51** 82–97, https://doi: 10.1016/j.trpro.2020.11.011
[16] Liang Y 2018 Application of Gestalt psychology in product human-machine Interface design. IOP Conference Series: Materials Science and Engineering, 392, 062054. https://doi.org/10.1088/1757- 899x/392/6/062054
[17] Tamara M 2015 *Visualization Analysis & Design*, CRC Press by Taylor & Francis Group, ISBN: 13: 978-1-4665-0893-4
[18] Bertini E, Lam H and Peter A 2011 Summaries: A special issue on evaluation for information visualisation *Information Visualization* **10**(3) 161-161
[19] Yeh N 2019 How to efficiently evaluate information visualization? https://medium.com/visumd/how-to-efficiently-evaluate-information-visualization-69bece7b30b1, Retrieved on 30 Sept 2022