WESTERN SYDNEY
UNIVERSITY

Visual Analytics - Assignment 2

Xuan Bach Tran - 22027453
Thanh Thong Tran - 22027485

School of Computer, Data and Mathematical Sciences,
Western Sydney University

DECLARATION

- I hold a copy of this assignment that I can produce if the original is lost or damaged.
- I hereby certify that no part of this assignment/product has been copied from any other student's.
- work or from any other source except where due acknowledgement is made in the assignment.
- No part of this assignment/product has been written/produced for me by another person except where such collaboration has been authorised by the subject lecturer/tutor concerned.

I. Introduction

The contemporary world of data is expanding at ever great speed, with millions of gigabytes worth of data being processed, analysed, and stored in a day. Furthermore, data doesn't come in just a simple frame with one or two variables, but many attributes and measurements are taken into consideration while performing data collection. Essentially, multidimensional datasets extend beyond the traditional data formats, such as the typically known two-dimensional data, offering more in-depth analysis and richer information mining. A study by Liu and co-authors emphasised their effectiveness in discovering patterns in climate data, fuelling better decision making at environmental level (Liu et al. 2016).

This report leverages the dataset from "Domestic airline on time performance", which is a monthly report issued by the Bureau of Infrastructure and Transport Research Economics, a governmental body that is responsible for monitoring the punctuality performance of key airlines operating on Australian airports. The ultimate purpose of the report is to analyse and visualise the assigned dataset to discover and evaluate any patterns, relationships, or identify irregularities that may seem significant for further examination.

II. Literature Review

Data does not come in just a simple frame with one or two variables, but many attributes and measurements are taken into consideration while performing data collection. Essentially, multidimensional datasets extend beyond the traditional data formats, such as the typically known two-dimensional data, offering more in-depth analysis and richer information mining. A study by Liu and co-authors emphasised their effectiveness in discovering patterns in climate data, fuelling better decision making at environmental level (Liu et al. 2016).

Transforming raw multidimensional data into presentable visualisation can be a challenge since extensive refinement and manipulation for such types of data are more demanding of knowledge and skills than one might expect. Some of the visualisation methods to be used in this report included bar charts, line charts, or a combination of both in one view, geographic maps, etc. The implementation of multiple visualisation techniques serves to explore various aspects of the dataset that one single graph or chart will not be able to convey, while aiding readers or users to get a good grasp of the whole picture; they will also help fuel more in-depth analysis and research on relationships, patterns found during different stages of the exploration.

2.1/ Symbolic display: Line chart, bar chart

In terms of universally recognised symbolic displays, line charts usually stand out as one of the most effective and simple methods of data visualisation. According to Tableau, line chart is described as a powerful tool for presenting data trends, patterns which fluctuate over time (Tableau); each data point is plotted based on its corresponding values on each axis of the graph, and from there, one can observe how y-axis changes in relation to x-axis by connecting these points. For example, one can examine the number of sales for a product over a specified time frame, or even forecast how the profit is going to look like in the next few months, based on the connected data point on the chart. While the line chart proves its capabilities in demonstrating overall trends and patterns, a data set with too many categories might render the method less effective due to lines overlapping; such problems can be somewhat mitigated by proper colour encoding (Wilkinson, 2005).

While line charts illustrate continuous values of a certain dimension, bar charts compare between various categorical variables. A typical bar chart also consists of the horizontal axis, which displays the categories subjected to visualisation, and the vertical axis wherein the length or size of the ruler demonstrate the corresponding values of those certain categories (Wilkinson, 2005). Bar charts can be stacked or grouped, allowing more effective visualisation of complex multidimensional data. Nevertheless, bar charts may as well not be effective if the multidimensional data is too large, which could result in difficulty comparison due to cluttered categories (Wilkinson, 2005). The Analysis section will take into consideration such factors.

2.2/ Graph visualisation: Treemap

The Treemap was first introduced by Ben Schneiderman in his research paper namely *Tree-maps: a space-filling approach to the visualisation of hierarchical information structures* in 1991 (Johnson & Schneiderman, 1991). This distinctive visualisation technique leverages the size, or the colour intensity of the nested rectangles to illustrate accordingly with the quantitative measures of the targeted categories or subcategories (Brühlmann & Keim, 2008). Treemap can be grouped into subgroups, enabling the readers to navigate through the hierarchical relationships and patterns. Despite the ability to make optimal use of display space and revealing patterns, precise comparison between categories is challenging as it does not offer clear separations as bar charts does (Shneiderman, 1998).

2.3/ GeoVisualisation: Symbol map

Although symbol maps are not as effective as visualising multidimensional data, they can be a useful tool for geographic display. According to Monroe and Qian, depiction of quantitative value distribution across a certain region can be achieved by utilising the density, size or colour of the symbol, larger symbol means higher intensity or more populated with data points (Monroe & Qian, 2004). Symbol maps allow the users and the readers to have better comprehension of the geographic distribution, while simultaneously getting a good grasp of the data characteristics through various attributes.

2.4/ Parallel coordinates plot

The parallel coordinates plot is exceptionally useful with hypervariate data as it leverages equally spaced vertical axes to represent each numerical variable. The data points are plotted with vertices connected across the axes by a polyline, and the position of the point corresponds to a specific value (Inselberg, 1986). By visualising the plot, one can observe the position of the vertex in relation to its quantitative measurements, as well as comparing between various dimensions. While parallel coordinates plot can be cluttered when dealing with too many dimensions, it still provides a unique way of observing and analysing the patterns or relationships that are difficult to recognise in other methods.

III. Analytical discussion

The given data consists of multiple tables, each detailing the performance of airlines across Australian domestic and international airports. Therefore, the tables were unified by taking advantage of the union function at the Data Source tab on Tableau, which allows the users to append values from one table to another. After completing the union, a new worksheet would appear with all the data now in one big cluster, simplifying the latter examination process (Figure 1).

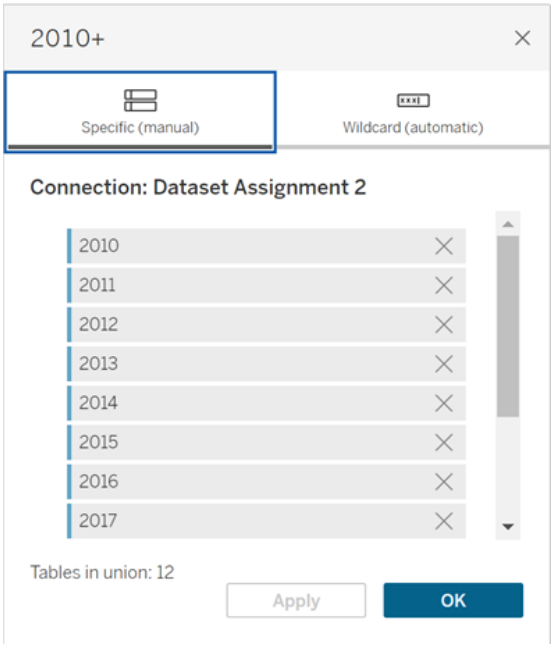

Figure 1. Union by connecting different tables in multiple years.

3.1/ General visualisations

a. Flights shares by airlines

Although there is no clear hierarchical structure in the dataset, the Treemap layout can be utilised to demonstrate the percentage that each airline takes up, based on their state of origin. For example, Virgin Australia is a commercial airline from Queensland, Jetstar is from Victoria, etc. (Figure 2). A grouped parameter was created called “Airline (group)”, consisting of present Australian states for grouping purposes, and filters for airlines are also applied to ensure that there is no irrelevant information being accounted into the visualisation.

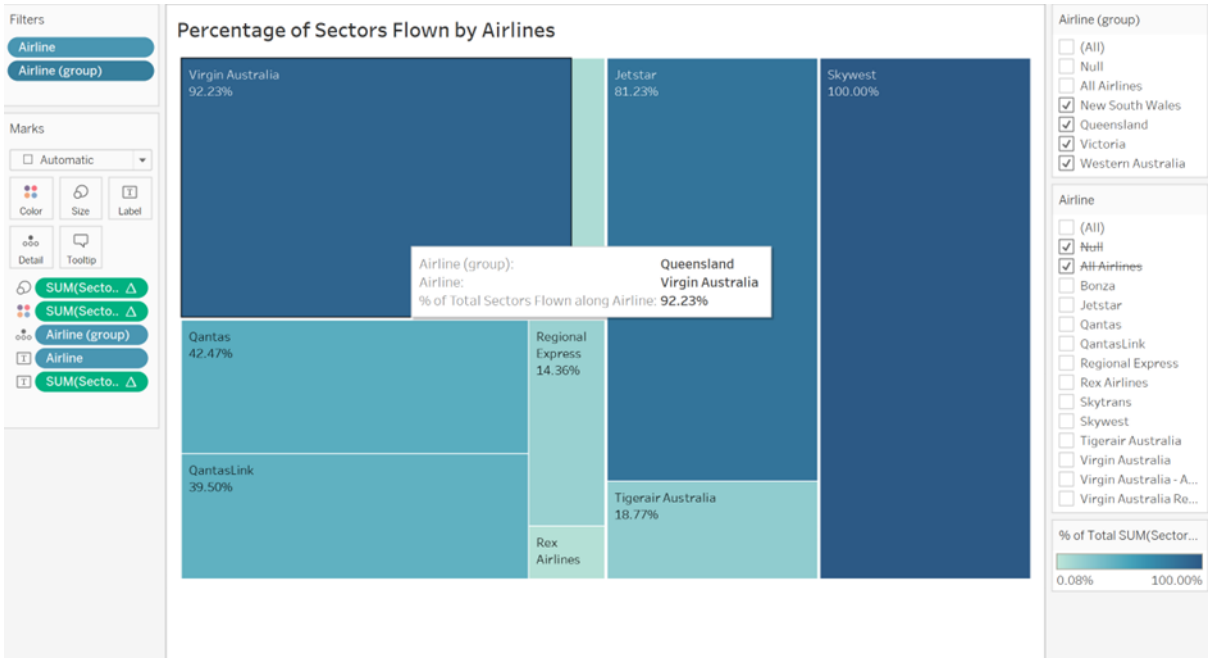

Figure 2. Percentage of Sectors Flown by airlines (grouped by states).

Illustration shows major airlines such as Virgin Australia, Qantas/QantasLink, Jetstar take the dominant proportion in terms of flight shares on the market, compared to the competitors within the same state. Furthermore, Skywest surprisingly appears to be the monopolistic player in Western Australia as it accounts for 100% of the market share.

b. Number of flights at arriving and departing ports

Potential approaches when dealing with figure comparison between distinct categories or subcategories includes a bar chart, as one may base off the length of the ruler to pre-attentively select zones of interest. In the following visualisation, the bar chart presents the number of flights at arriving ports. Zooming technique was also applied to view top eight airports with the highest sectors flown as the bar chart extends beyond the visible area of the horizontal axis, due to the large number of categories, and it would be time consuming to untick unwanted airports (Figure 3).

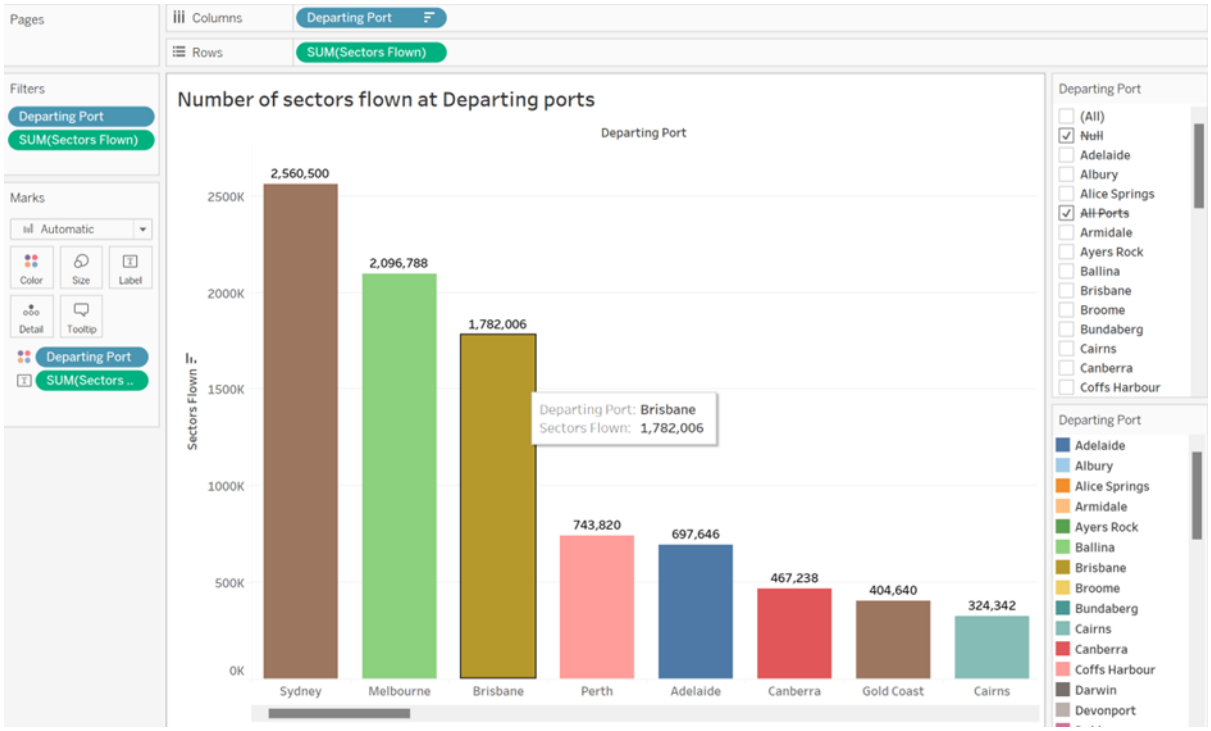

Figure 3. Bar chart for the number of sectors completed at departing ports.

Based on the bar chart above, it is evident that major airports such as Sydney reported the highest figure for the number of departing flights, with 2.5 million flights in a span of 14 years, followed by Melbourne with nearly 2.1 million flights and Brisbane with roughly 1.8 million. The other five airports witnessed an average of 528.000 flights at the same time (Figure 4).

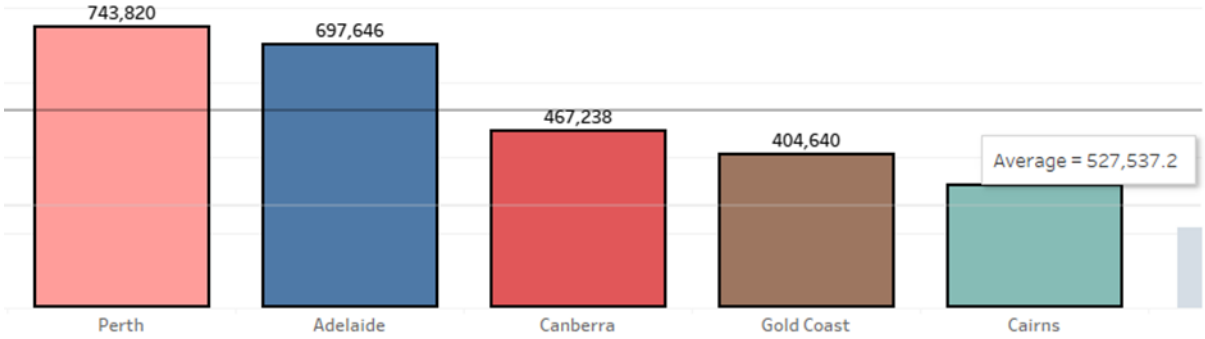

Figure 4. Number of sectors flown for the rest of the top eight.

3.2/ Analysis of airports performance with Treemap and geographic visualisation

a. Treemap

In the first stage, it is necessary to determine which airports account for significant domestic flights to obtain an objective view of the Australian aviation industry. In order words, information on the number of sectors flown retrieved from airports would be visualised, which is appropriate with the treemap method. Both departing and arriving airports are aggregated by the sector flown. The larger the box and the darker the colour, the higher the proportion of domestic flights in Australia that the airport accounts

for. Hence, it can be seen from Figure 5 that the five major airports in terms of both departing and arriving are Sydney, Melbourne, Brisbane, Perth, and Adelaide.

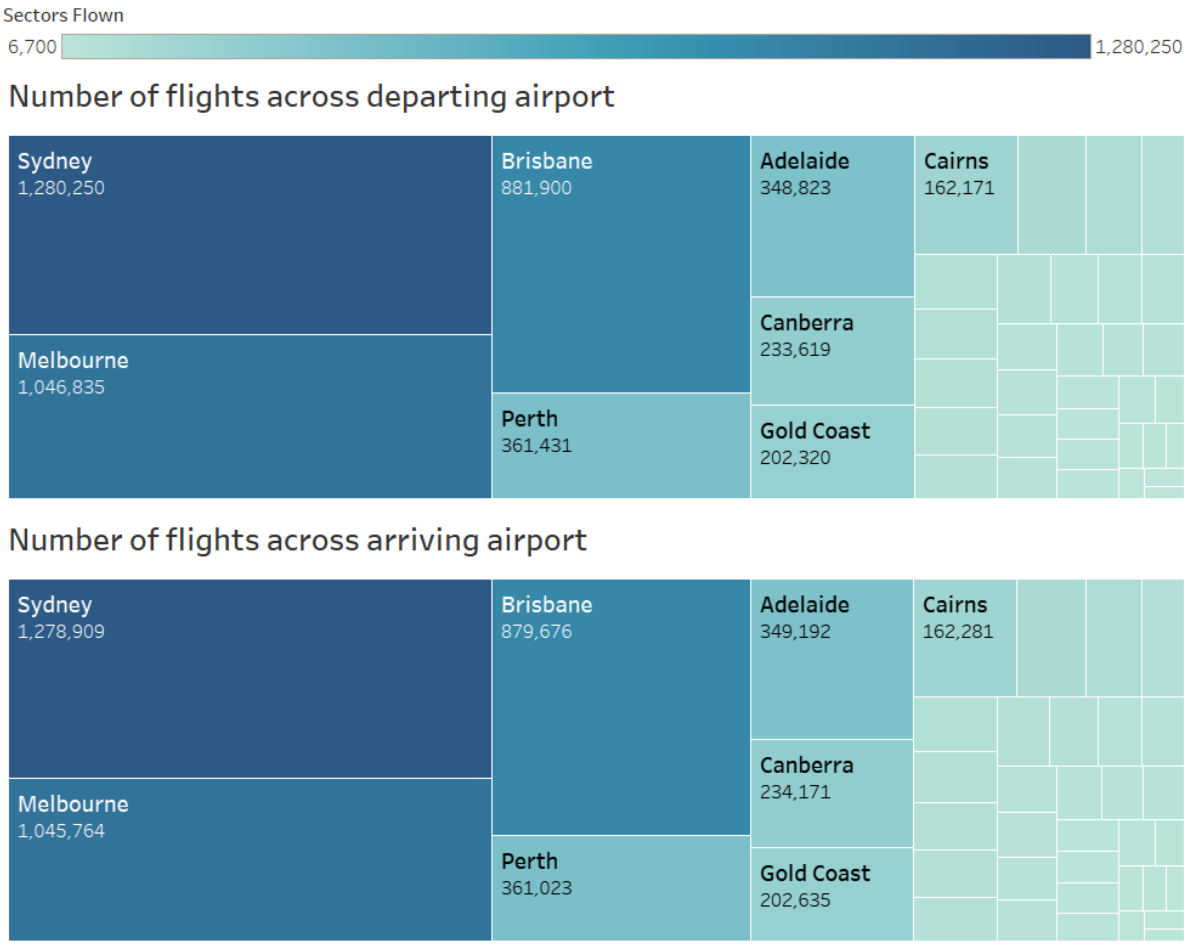

Figure 5. Number of sectors flown across departing and arriving airports.

b. GeoVisualisation: Symbol map

The performance of departing and arriving airports is shown by leveraging geographic visualisations. Specifically, departure and arrival airports are shown on two maps colour-sorted according to the average percentage of on-time flights. Red which is below average represents low performance and conversely blue represents high performance in terms of departure or arrival on time. The circle size of each airport expresses how many sectors flown that airport has, which is also used to identify the five major airports. It is figured out that these major airports perform well on departure flights but not on arrival flights (as shown in Figure 6). This pattern raises a point for operators to improve their effectiveness in arriving flights.

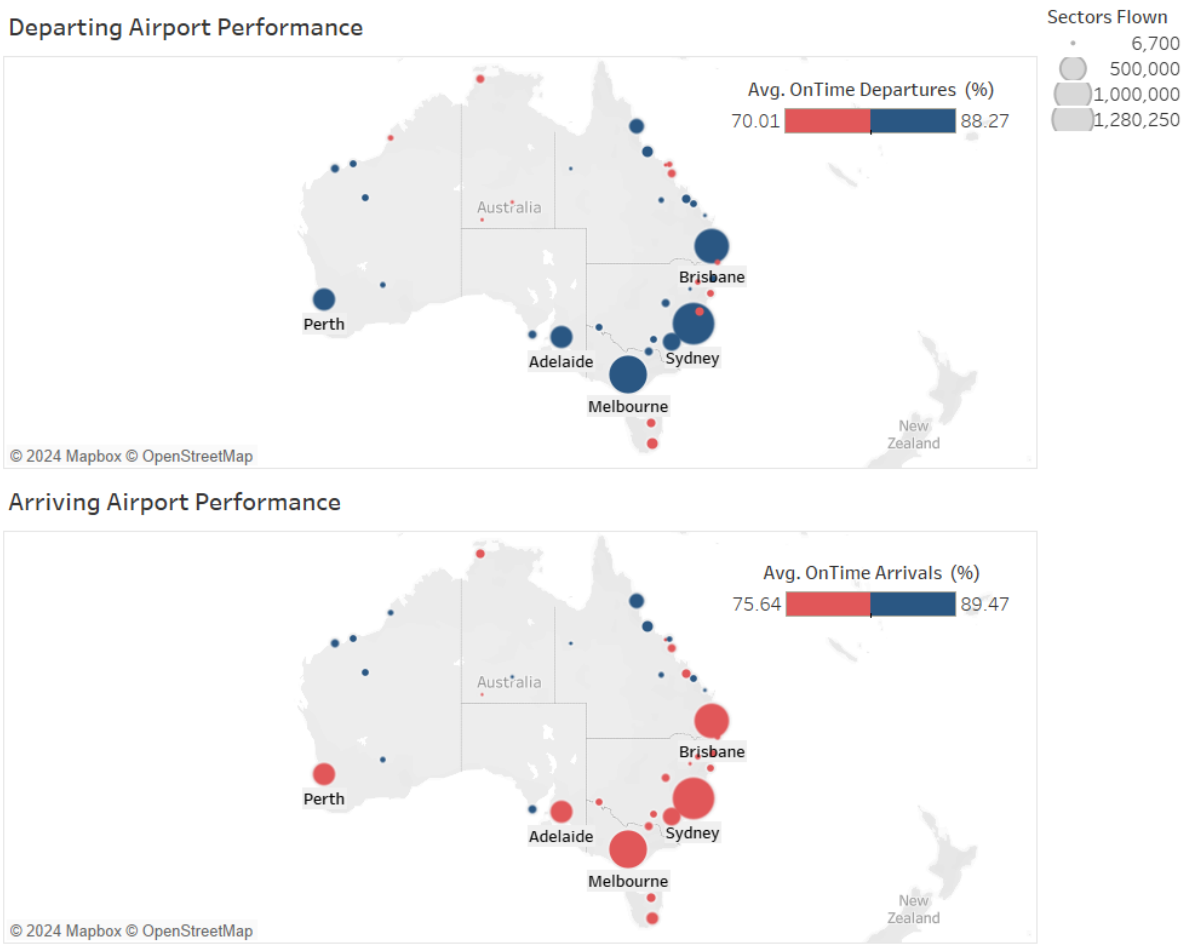

Figure 6. Departing and Arriving airport performance.

3.3/ Airports' performance through the years via time series visualisations

a. Line chart

Acknowledging there may be changes over time in terms of the number of sectors flown by airlines, the line chart is chosen for its capability to illustrate and reveal insightful information on continuous variables (Figure 7). A few airlines that don't serve any analytical purposes have been left out.

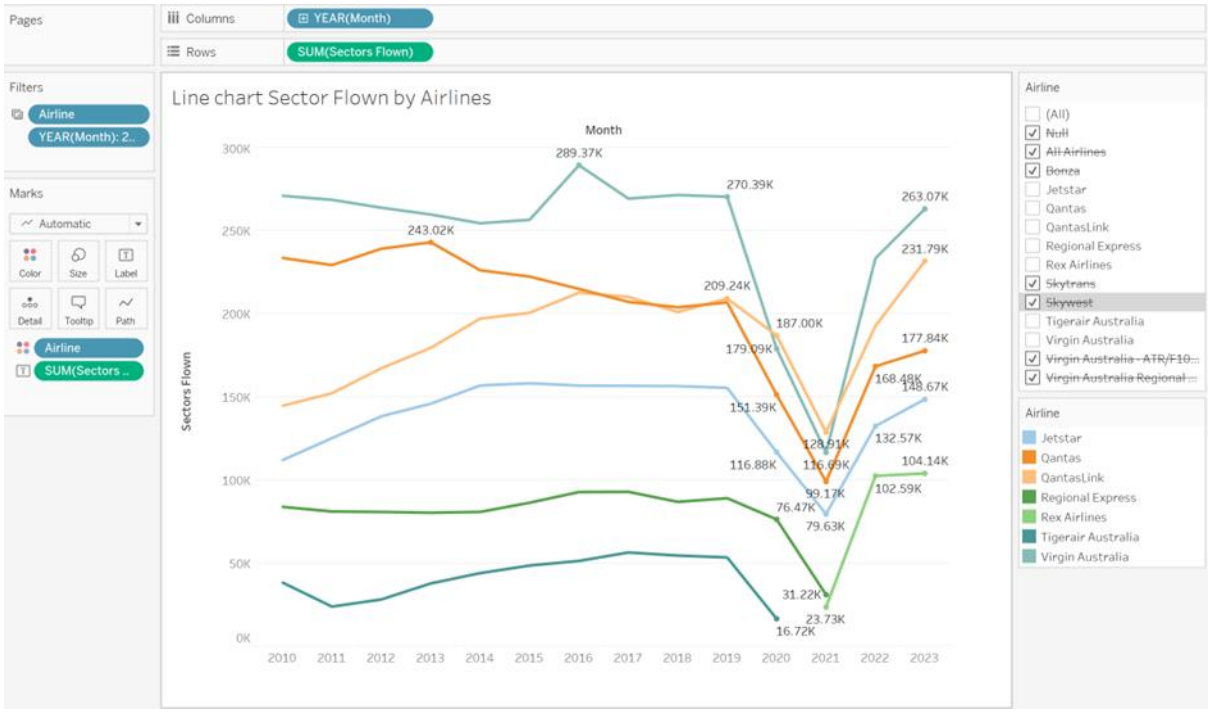

Figure 7. Line chart illustrating sectors flown by airlines in 14 years.

Observation shows the majority of airlines' number of sectors remained relatively steady from 2010 to 2019, with a peak in 2016 for Virgin Australia. During the time frame, it is worth taking into consideration the most significant event that drastically disrupted the airline industry, which was COVID-19. During the pandemic, all Australian airlines experienced a sharp decline in the number of flights completed, Jetstar is shown to have the largest drop from 270.000 to 116.000 sectors flown between 2019 and 2021 (Figure 7). A few airlines such as Tigerair Australia, or Regional Express appeared to have ceased operation, as there was no further data recorded since 2021.

b. Dual-axis (line chart and bar chart combination)

It was hypothesised that a decline in the number of sectors flown could possibly mean an increase in the level of flight cancellations. Hence, a combination of bar chart and line chart was implemented to see if there was such correlation between the two variables. The combination is plotted using dual axis, visualising the number of sectors flown on the left y-axis, and the cancellations on the right (Figure 8).

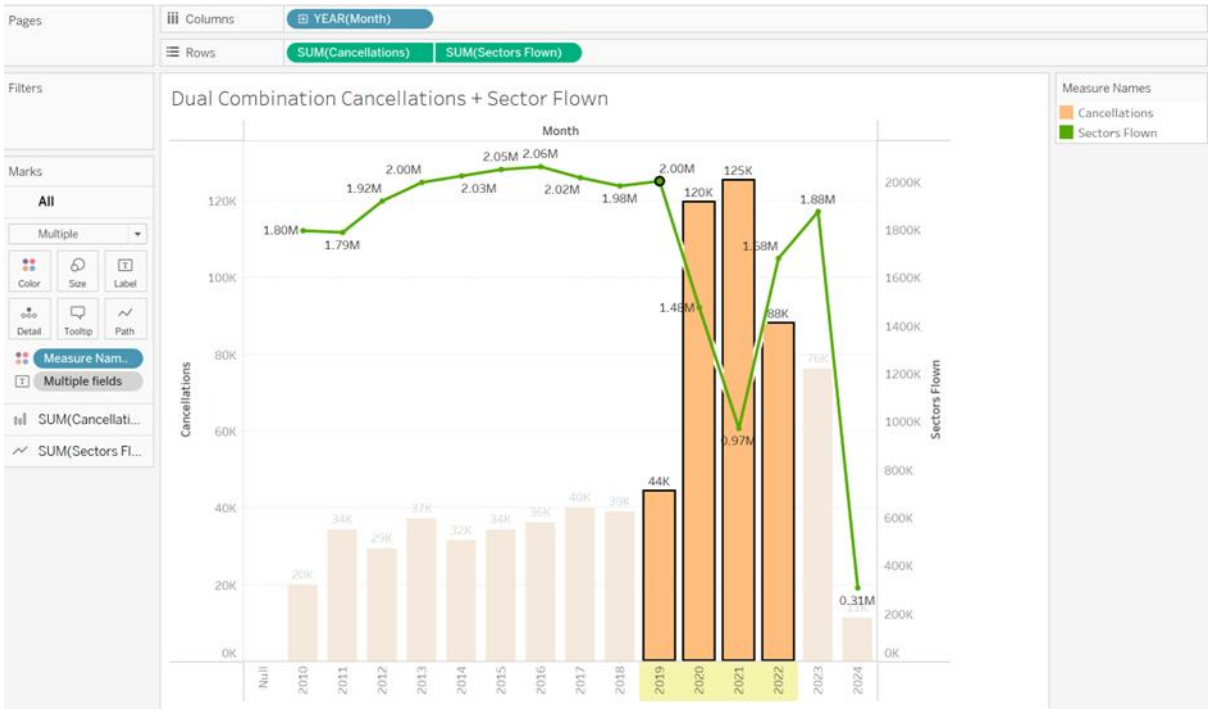

Figure 8. Dual-axis visualisation for sectors flown and cancellations.

It is clear that while the cancellation rate skyrocketed during the pandemic, a significant decline in the number of sectors flown can also be recognised. From 2010 to 2019, cancellations stayed at an average of 35.000, while its counterpart was at around 1.5 million to 2 million flights per year. The figure for cancellations then shot up to 120.000 in 2020 and 125.000 in 2021, dragging the total sectors flown down to only 970.000 in 2021. According to the Australian National Audit Office, the pandemic resulted in 95% decrease in both domestic and international passenger activity, and negatively impacted the country's aviation industry (ANAO, 2022).

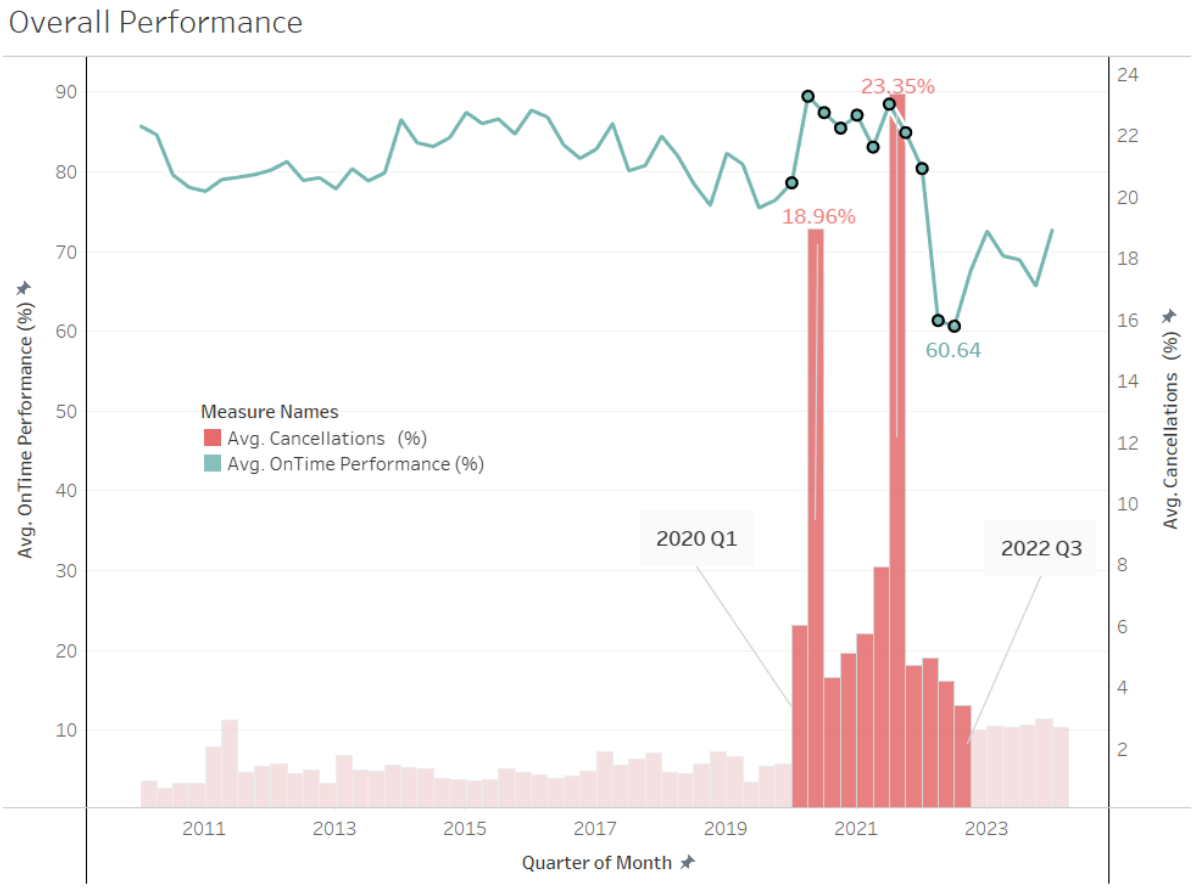

Figure 9. Overall performance over time.

Figure 9 above also shows a significant increase in the percentage of flight cancellations recorded in the period of approximately more than two years from the first quarter of 2020 to the third quarter of 2022. The two peak periods in the second quarter of 2020 and the third quarter of 2021 recorded nearly 20% and 23.5% of cancelled flights, respectively. This may also be the reason for the record low % on time performance remarkably in the third quarter of 2022, only 60.64%. This notable period also provides useful information as this is the period when the COVID-19 epidemic broke out and raged, causing many challenges for the aviation industry.

3.4/ Airports' performance through 12 months of a selected year

It can be seen that in 2023, after the COVID-19 epidemic, the aviation industry gradually recovered and operated normally, which is also the time chosen to visualise the performance of airports. By visualising data over a period of time, a line graph and a bar chart would be the most appropriate, which represents the average percentage of on-time performance and cancellations. A geographic map is also integrated to illustrate the attributes of different airports. Figure 10 shows that the proportion of on-time flights tends to possess a downward trend in 2023, from about 74% in January to nearly 63% in the last month of the year. In addition, another pattern also found in Figure 10 is that the beginning of the second and third quarters, which are April and July in 2023, recorded a sudden increase in the percentage of cancelled flights of more than 4%. In addition, among the five major airports, only Melbourne Airport exhibits low performance in ensuring flights depart or arrive on time.

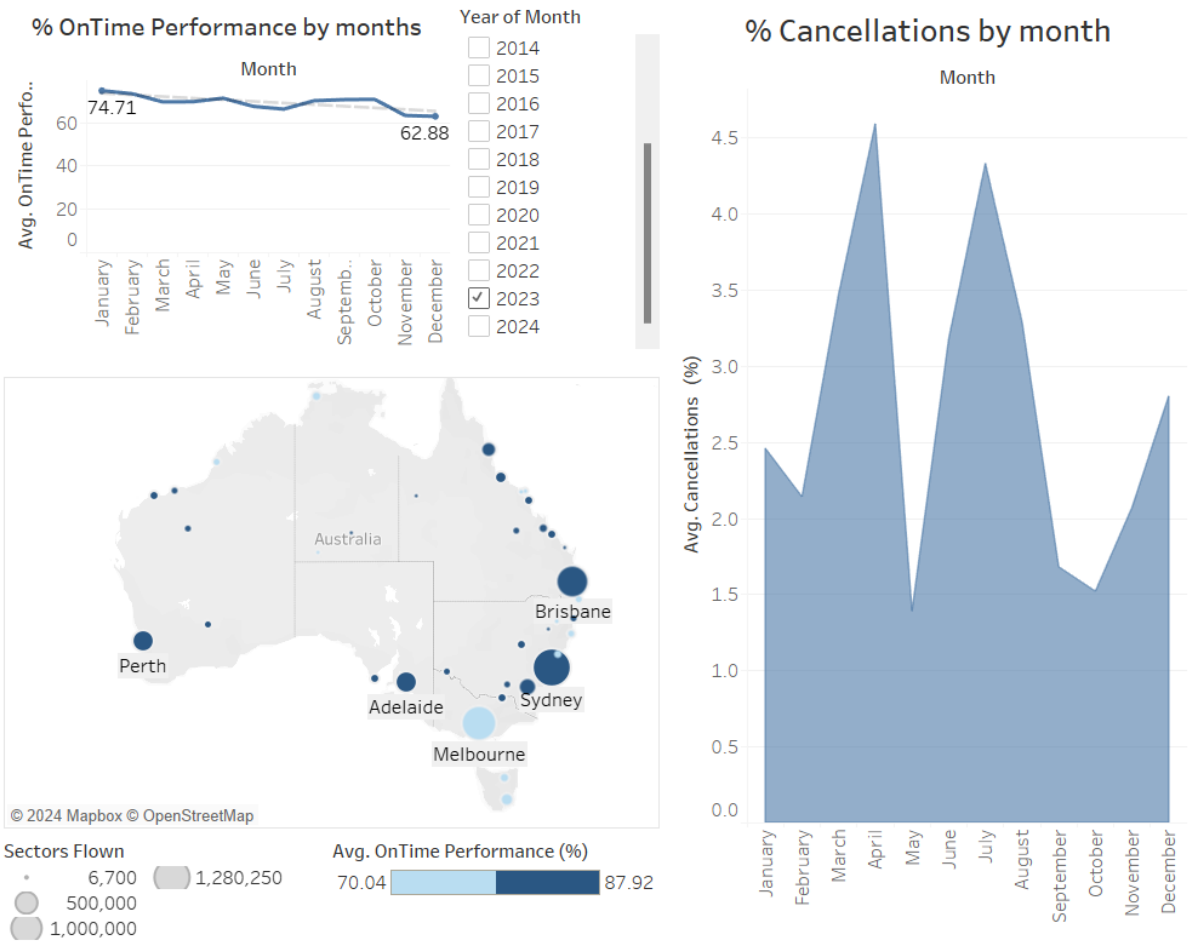

Figure 10. Airport performance through 12 months.

3.5/ Airlines' performance via Parallel Coordinates Plot

Normally, the examination of various dimensions is least likely to be effectively visualised using simple charts or graphs; therefore, parallel coordinate plot layout was selected as the primary method for hypervariate visualisation to explore the performance of airports. Multiple steps were taken to create the plot, including automated display of airport statistics such as Arrival on time, Arrival delayed, Departure on time, etc. when the cursor is hovered on the polylines. The visualisation is shown in Figure 11 below.

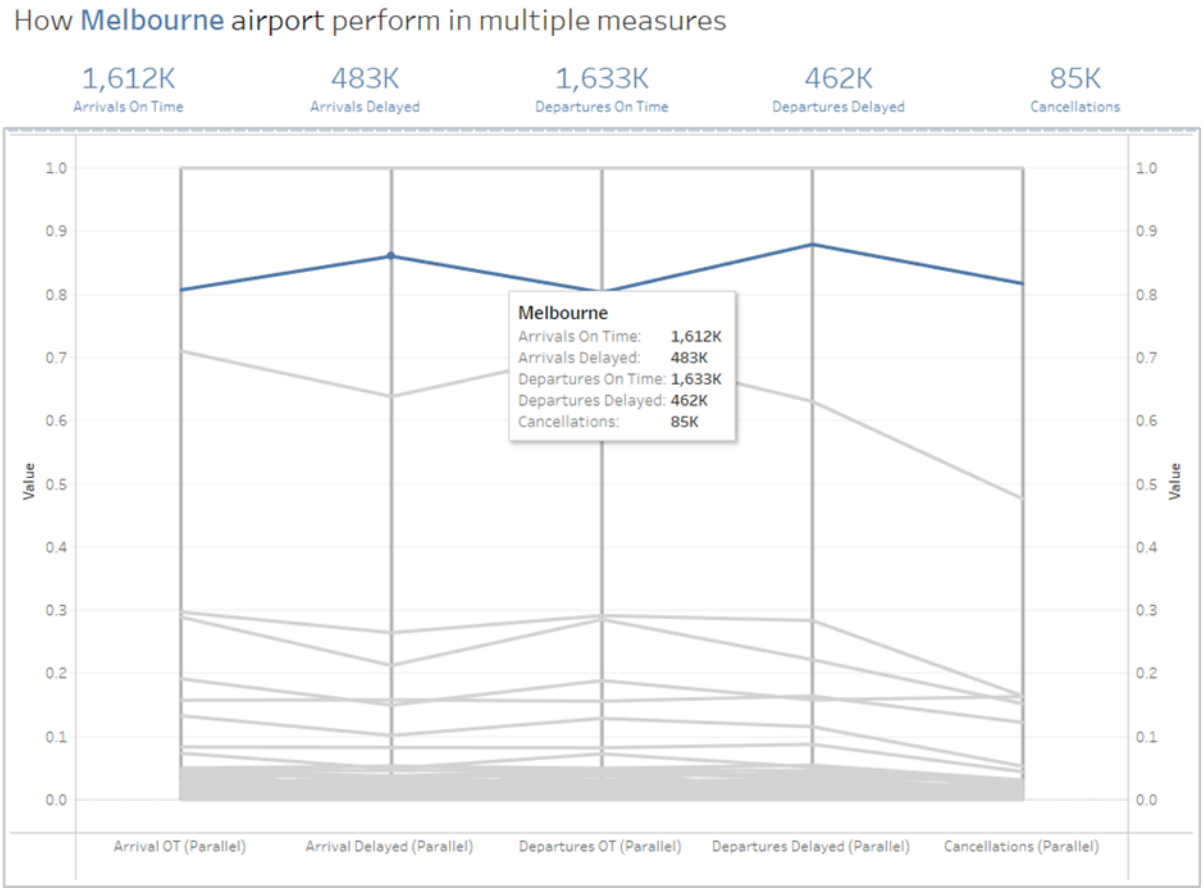

Figure 11. Parallel Coordinates Plot for the overall performance of airports.

Apart from Sydney, Melbourne, and Brisbane airports being the top performers, there appears to be a lot of overlapping between polylines that represent smaller airports. This circumstance is often caused by the relatively skewed distribution of flights when major key airports are in place. Therefore, the positions of airports remained largely similar to what was demonstrated in the bar chart. However, the clutter and overlapping of polylines appear densely in the lower end of the plot since the performance of smaller airports had minimal differences, given that the number of sectors flown in these airports are somewhat fewer, compared to major ones.

3.6/ Flight efficiency of the routes and the airlines

Multiple views are leveraged in Tableau by combining several worksheets in a dashboard. A horizontal bar chart is used to depict the average percentage of on-time performance of routes corresponding to a flight map. Accordingly, packed bubbles help to identify which airline accounted for the most significant sectors flown with chosen routes. Based on Figure 12, the five most effective routes are classified by sorting the percentage of on-time performance in descending order. QantasLink and Regional Express airlines are identified as accounting for a significant proportion of the total number of flights with this good performance level. The bar chart's colour property also demonstrates that routes with high on-time performance only account for a low proportion of flights and vice versa. This irregularity needs to be exploited in further research.

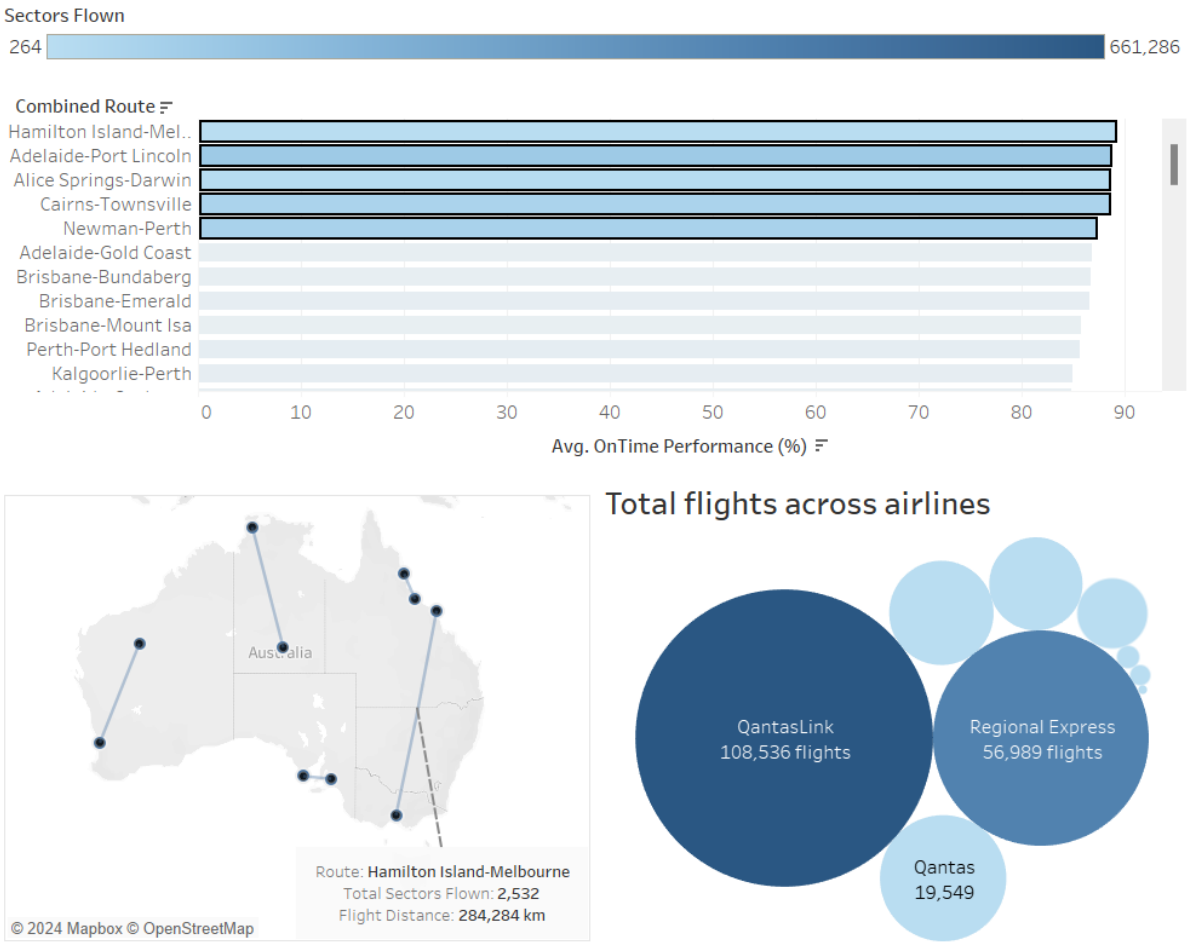

Figure 12. Route performance dashboard.

IV. Conclusion

This paper referred to numerous previously peer-reviewed research papers for qualitative literature review and applied various visualisation methods to deal with a given multidimensional data set by leveraging a powerful visual analytics platform, Tableau. Creation of parameters and calculated fields also unlocked more aspects of the dataset for exploration and examination. Understanding the structure of the data set helps implement several visualisation methods effectively and interpret them in dashboards comprehensively. Applying filters, highlight techniques, and colour, size, label attributes to optimise visualising large data sets also brings a meaningful visualisation. Based on that, patterns and interesting properties are found, serving as a strong premise for further research in the field of the aviation industry.

V. Reference list

1. ANAO 2022, *COVID-19 Support to the Aviation Sector*, Australian National Audit Office. https://www.anao.gov.au/work/performance-audit/covid-19-support-to-the-aviation-sector#:~:text=Following%20the%20implementation%20of%20travel,cent%20decrease%20in%20passenger%20activity.

2. Brühlmann, M, & Keim, D. A 2008 *Monitoring multidimensional data streams using visual hotlines*, In Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 812-820). https://dl.acm.org/

3. Bureau of Infrastructure and Transport Research Economics 2024, *Airline On Time Performance Monthly Reports*, Australian Government. https://www.bitre.gov.au/publications/ongoing/airline_on_time_monthly

4. Few, S 2009, *Visual Thinking for Design*, Journal of Design Research, vol. 8, no. 2, 2009, pp. 147-167.

5. Green, Tamara. D 2019, *Why Is Data Visualization Important? What Is Important in Data Visualization?*, Harvard Data Science Review. hdsr.mitpress.mit.edu/pub/issue_articles/why-is-data-visualization-important-what-is-important-in-data-visualization

6. Inselberg, Alfred, *Parallel coordinates: A tool for visualizing multivariate relational data*, Journal of the American Statistical Association 81.393 (1986): 661-678.

7. Liu, Y, Wu, F, & Wang, Y 2016, *Multidimensional analysis of climate extremes in China using a gridded dataset*, Theoretical and Applied Climatology, 127(3-4), 809-820. (https://www.sciencedirect.com/science/article/pii/S0921818123003028)

8. Monroe, M. C, & Qian, W 2007, *Digital mapping for geo-spatial research*. In I. N. Levine (Ed.), Geography and technology (pp. 147-168). Thousand Oaks, CA: SAGE Publications Ltd.

9. Munzner, T 2014, *Visualization Analysis and Design*, AK Peters/CRC Press.

10. Shneiderman, B 1998, *Designing the user interface for information visualization*, Addison-Wesley Longman Publishing Co., Inc.

11. Tableau, *Building Line Charts*, Tableau Help. https://help.tableau.com/current/pro/desktop/en-us/buildexamples_line.htm

12. Wilkinson, L 2005, *The Grammar of Graphics*. Springer-Verlag New York, Inc.