2024

Fundamentals of Digital Visualisation

FINAL REPORT

LOVELY FERNANDEZ | FHNW

Contents

Abstract

This report documents my understanding of the five key learning objectives of Data Visualisation (GDS), where each learning objective is discussed in its own 2-page essay. *Data visualisations of the given theories are included within the report* which are also included in a Jupyter Notebook attached to this assignment.

My aim is to display a clear and engaging visual representation to support my understanding of data visualisation: using perceptual tools and processing raw, unstructured data with the aim of conveying complex information of the real world.

Introduction

What is data visualisation? Data visualisation is a tool used to communicate specific observations of a dataset to support the decision-making process. Since there is often an abundance of raw data, it is vital to filter the dataset to focus the reader's attention on the purpose of the visual.

According to Steven Franconeri's conference, *'How to Make Perfect Pancakes'* (Franconeri, 2019), it is important to consider the idea of different perceptions and how a graph could be interpreted differently by individuals. He emphasised the complexity of visual perception - how the brain perceives visual stimuli in multiple ways - by demonstrating the ambiguity of the duck-rabbit illusion.

As Franconeri suggested, there are three primary visual tools the brain uses to understand visual data: object recognition, which is the quick identification of objects based on familiarity and patterns; feature distribution analysis, which involves the assessment of statistical properties such as size and colour distributions in a visual; and comparative analysis, which is the process of comparing different visual elements.

Using *'How to Make Perfect Pancakes'* underlines the importance of using data visualisation as it allows different interpretations of the same data based on how one manipulates perceptual tools.

Application of Basic Visualisation Principles

Learning Objective 01

In this section, I will use the dataset sourced from Kaggle, 'Student Mental Health' (Islam, n.d.)*, which focuses on students and their mental health. My aim is to create a visual comparison across each subject's average performance (CGPA) results between male and female students.*

To achieve an impactful and effective representation of a dataset, it is important to define two main points: (1) *"What is the purpose of the visual?"* and (2) *"How do I direct the reader's attention to that purpose?"*

As highlighted in the introduction, the human brain relies on three primary visual tools to understand visual data. By integrating these perceptual tools with the techniques recommended in *'Better Data Visualisations'* (Schwabish, 2022), we can refine our methodology for analysing and presenting complex datasets - guiding one's interpretation of a data visual.

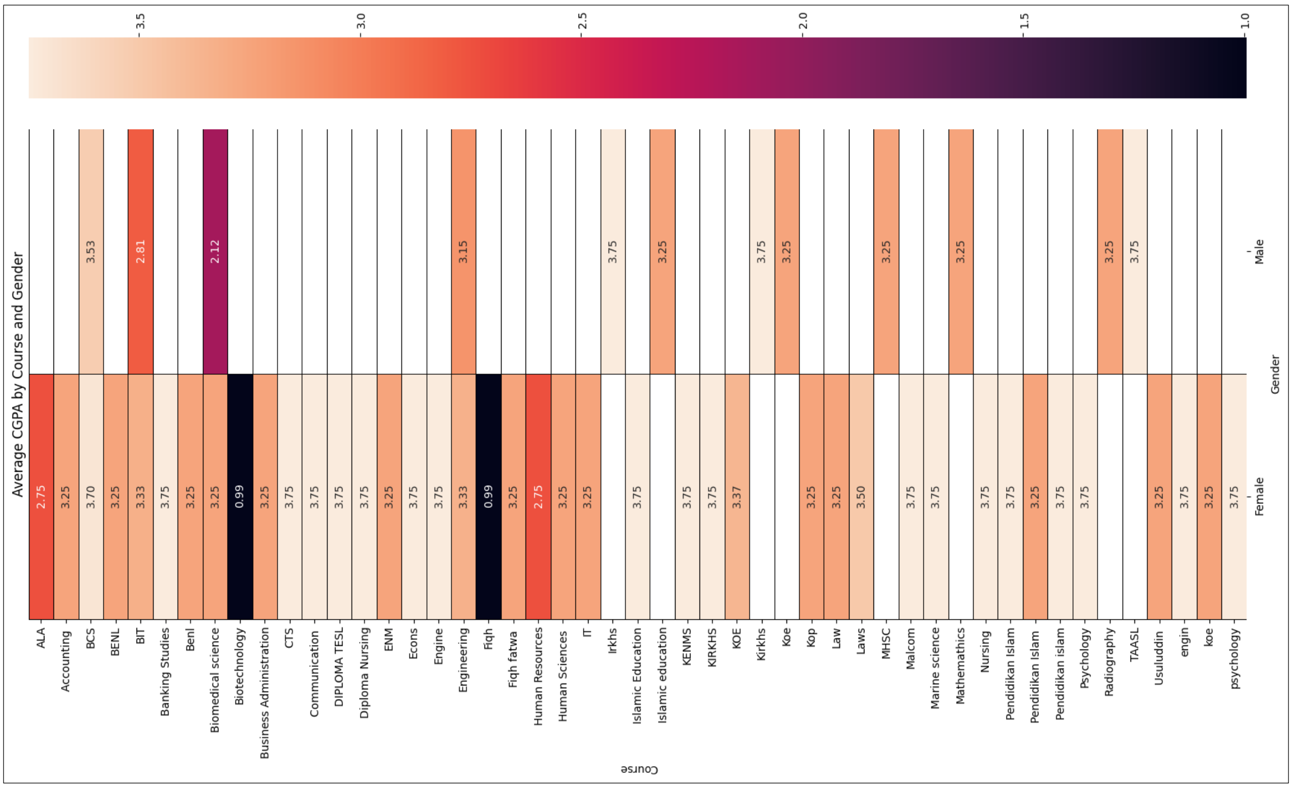

Figure 01. Table Format of Average CGPA by Course and Gender

Using a table to represent data is a common yet often an inefficient approach if working with numerous variables. They can easily become cluttered and ineligible, thus going against the idea of *object recognition*. While *'Better Data Visualisations'* provides standards for improving tables (Better Data Visualisations, p. 344), applying a more interactive visual representation can attract the reader's attention more effectively.

Comparative analysis is a method used to compare differences and relationships. By employing this method, I will present the differences and the total average of the CGPA results for both male and female students in each subject, offering insights that may aid in understanding the relationship between them. Using a Pyramid Chart is an optimal choice, as they are commonly employed to compare distributions between two variables.

'The advantage of the pyramid chart is that we can assess the overall shape of the distributions because both groups sit on the same vertical baseline.' - (Better Data Visualisations, p. 185)

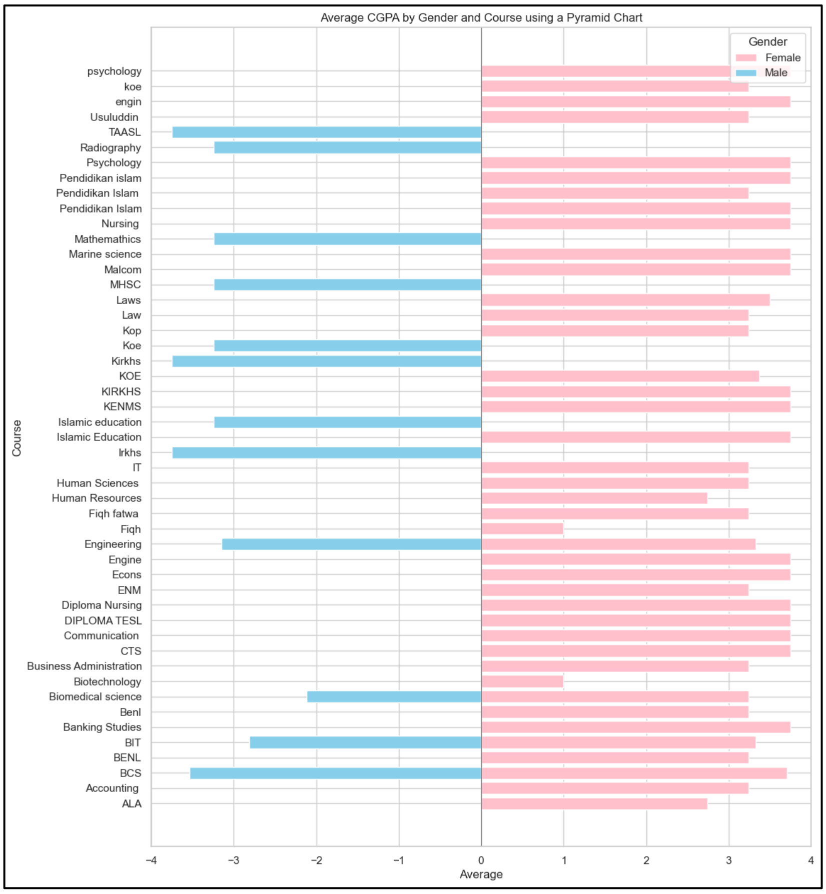

Figure 02. Pyramid Chart of the Differences of CGPA

Taking inspiration from the concept of *feature distribution analysis*, integrating characteristics of a Lollipop Chart (Better Data Visualisations, p. 80) into the Pyramid Chart may enhance the visual appeal of the data. This hybrid approach allows more variables to be added without cluttering the chart, making it easier to read and interpret. Additionally, we can order the subjects based on the average size of the CGPA results and highlight each subject with its own distinct colour (Better Data Visualisations, pp. 69 - 70, 80).

As the dataset comprises various types of subjects to be integrated into the graph, it is more convenient to reverse the chart and employ the Y-axis for long axis labels (Schwabish, 2022, p. 79). This approach ensures enough space for subject labels with varying lengths without cluttering the graphs.

With numerous subject labels on the Y-axis, adding grade labels from 0 to 4 on the X-axis could also crowd the graph and make it challenging to read. Simplifying the graph by removing the X-axis will allow the reader to solely focus on the subjects. Instead of labelling CGPA scores across the X-axis, they can be incorporated into each subject's lollipop. This enhances readability and reduces graph complexity (Schwabish, 2022, pp. 78 - 79).

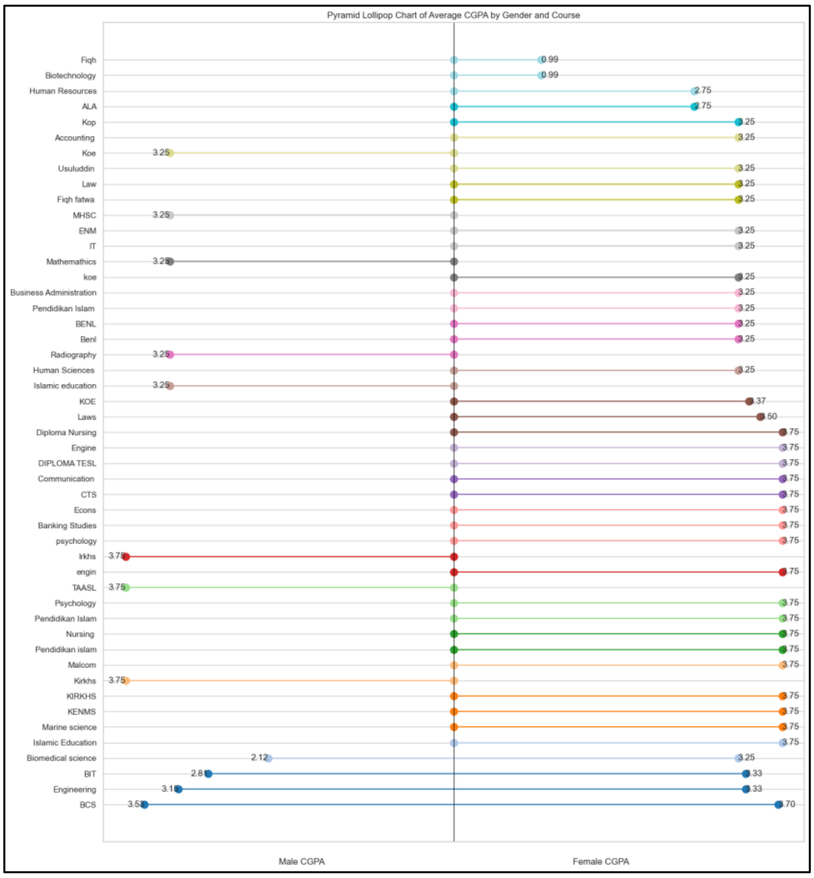

Figure 03. Hybrid Chart between Lollipop and Pyramid Chart

The main objective of this visualisation is to display the differences in CGPA results along with their overall average scores. In complement to the pyramid chart, a separate lollipop chart can be introduced to exclusively highlight these differences. Following the same principles and structure of the pyramid chart, this additional chart will solely display the differences of the CGPA results between male and female students.

Figure 04. Lollipop Chart focused on Differences

As the fundamental aim of visualisation is to make data easier to read, eliminating the need for readers to manually compare and calculate the differences themselves streamlines the process. Therefore, the addition of a lollipop chart supports the reader and removes additional tasks (Schwabish, 2022, p. 85).

Visual Perception

Learning Objective 02

For this learning objective, I will reference a dataset sourced from Kaggle titled 'Internet Usage' (Pavan Kalyan, n.d.). *The dataset contains information about various age groups using different types of media.*

Visual perception is a basic element that shapes our understanding and interaction with visual data. Understanding how individuals perceive visual stimuli and how to apply perceptual tools with this in mind is essential for effective and clear visualisation.

According to David McCandless, in his TED Talk *'The Beauty of Data Visualization,'* he claimed that visual information is processed unconsciously and highlights the eye's sensitivity to patterns, colours, and shapes (McCandless, 2012).

Larking and Simon explain the advantages of using diagrammatic representations, such as charts or graphs, and why it is more optimal than textual descriptions. This idea applies perfectly to visualisation systems as they involve three essential cognitive processes: looking for information, noticing patterns or shapes, and interpretation.

"From a formal point of view, Larking & Simon [497] give a rationale about advantages of using diagrammatic representations instead of textual ones. It can be applied to the successful use of a visualization system to the effect that the three reasoning processes—searching, recognition and inference—are also present in the use of visualization systems." - (Human-Centered Visualization Environments, p. 15)

Understanding these processes and adjusting visuals to match the reader's unconscious processing capabilities improves the effectiveness of visuals at successfully presenting the data and its purpose.

The *Gestalt Theory* or *Gestalt Principles* further discusses how the human brain process and organises information. It also implies that individuals perceive complete images or visuals rather than seeing separate parts. Recognising patterns and connecting these helps understand complex stimuli. There are six main organisational principles the Gestalt Theory discuses on how the human brain connects these patterns: proximity, similarity, enclosure, closer, continuity, and connection (Schwabish, 2022, p. 22).

Implementing this theory, I will illustrate the different types of medias and pinpoint the primary media within each age demographic: young adults (18-24), adults (25-34), middle-age (35-49), and seniors (50+ years).

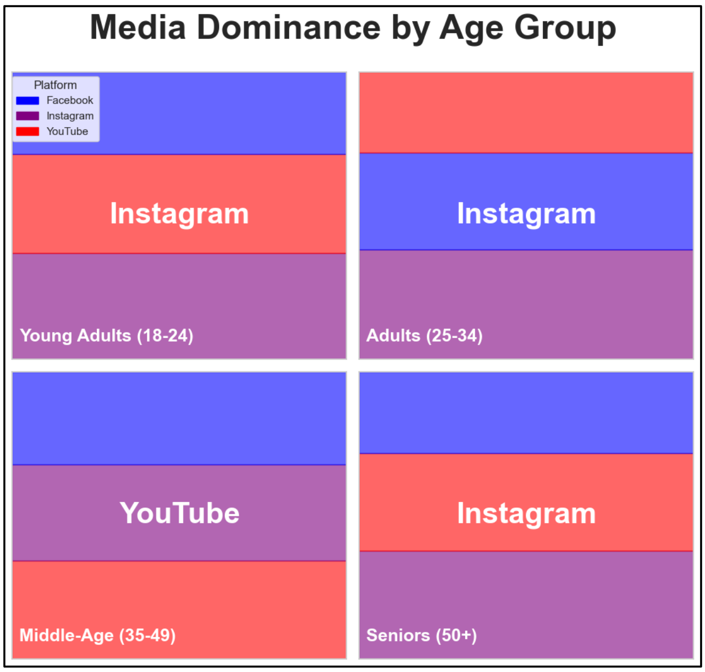

Figure 05. Quadrant Chart representing Media Dominance by Age Group

In Chapter 2 on '*Better Data Visualizations'*, it provides five primary guidelines aimed to improve the effectiveness and clarity of visualisations: prioritise main data, reduce clutter, integrate graphics and text, avoid the spaghetti chart, and start with a grey scale.

"Your reader can only grasp your point, argument, or story if they see the data. This doesn't mean that all the data must be shown, but it does mean that you should highlight the values that are important to your argument. As chart creators, our challenge is deciding how much data to show and the best way to show it." - (Schwabish, 2022, pp. 29 - 30)

The 'Media Dominance by Age Group' graph aims to highlight the primary media platform preferred by each age demographic, focusing on age groups and their corresponding media choices. Initially, I proposed a scatter plot graph to display individual media preferences, but it resulted in a cluttered and noisy visualisation. To improve clarity, I updated the graph implementing the part-to-whole concept from 'Better Data Visualisations, Chapter 5' (Schwabish, 2022, p. 290).

The comparison between *Figure 05* and *Figure 06* illustrates the reality of Gestalt Theory, where one understands better when connecting segments as one visual than focusing on separate parts. The scatterplot depicted in *Figure 06* is a loud and cluttered visual, making it difficult to follow. In contrast, *Figure 05* offers a unified presentation of media platforms within each age group, providing a simpler yet easy-to-understand visualisation. In this graph, the dominant group can be quickly identified by the size of the media segments per age group, as well as aided by labels directly on the graph (Schwabish, 2022, p. 34).

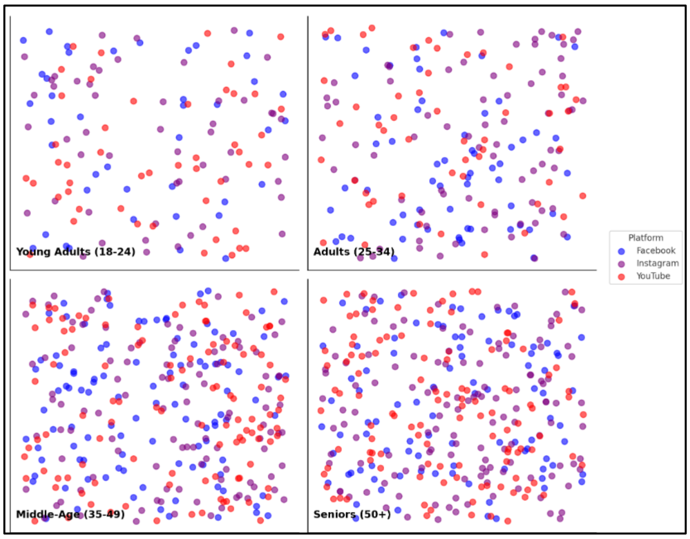

Figure 06. Scatter Plot Chart representing Media Dominance by Age Group

Following Bertin's theory of familiarisation with colours, I strategically chose a palette that mirrors the primary colours of the selected media platforms: purple for Instagram, red for YouTube, and blue for Facebook. By using these recognisable colours, readers can immediately recognise and understand the graph.

The role of visualisation is to promote clarity, transparency, and creativity, to support individuals on navigating through complex information of the real-world. Understanding the fundamental aspect of visual perception, we can reveal hidden patterns and connections, hence, transforming raw data into meaningful information.

Design Principles vs. Data

Learning Objective 03

To support learning objective 03, I will reference a dataset sourced from Kaggle titled 'Internet Usage' (Pavan Kalyan, n.d.). *The dataset contains information about females and their preferred types of posts on Instagram.*

Data preprocessing explains the process of transforming unstructured, raw data into a usable format for visualisation, allowing one to make informed decisions. These practices are vital in a data scientist's day-to-day tasks as without correctly prepared data, visualising complex data would be impossible. The relationship between data preprocessing and design principles is the foundation for effectively designing a visual.

The *'IBM Developer Tutorial'* by M. Jones (Working with Messy Data, 2017) addresses the common challenges involved in cleansing and validating data for better quality and usability. The data processing pipeline includes data cleansing, machine learning analysis, and finally, data visualisation.

The tutorial further discusses the three common types of data issues and its characteristics: such as missing values or incomplete sections within the dataset; incorrect formatting, where the dataset may contain errors or inconsistencies in its structure; and in general, inconsistencies across datasets. An unstructured dataset marks the data unusable for machine learning processing algorithms, thus, underlining the importance of data cleansing to enhance quality, providing accurate results and facilitating informed decisions.

"Incorrect or inconsistent data leads to false conclusions. And so, how well you clean and understand the data has a high impact on the quality of the results." - (Elgabry, 2019)

Data cleansing techniques may involve parsing data, identifying and rectifying errors, normalising data from various sources, and ensuring consistency across datasets. Other methods may also include defining schemas to check data types and validate field entries, employing rules for data transformation, and removing duplicates.

Once the data is cleansed, profiling helps analyse the data to identify any remaining inconsistencies, outliers, or errors. Reporting how robust data is provides transparency and accountability in the data cleaning process, ensuring that data is consistent and reliable. With a structured, high-quality dataset, an accurate visualisation of real-world problems can be demonstrated, enabling individuals to make informed decisions.

In the following demonstration, I've illustrated the primary techniques of data preprocessing using the *'Internet Usage'* dataset sourced from Kaggle. Specifically, I've focused on female preferences for browsing content on Instagram. I've crafted a bar chart to visually represent these interests, applying the principles outlined in Chapter 4 of *'Better Data Visualisation'* (Schwabish, 2022, p. 67).

	age	gender	time_spent	platform	interests	location	demographics	profession	income	indebt	isHomeOwner	Owns_Car
0	64	non-binary	7	Instagram	Lifestlye	Australia	Sub_Urban	Software Engineer	12658	True	True	False
1	64	male	1	Facebook	Travel	United States	Urban	Marketer Manager	10501	False	False	False
2	64	non-binary	4	Instagram	Lifestlye	United States	Rural	Marketer Manager	18880	True	True	True
3	64	male	4	Instagram	Travel	United Kingdom	Urban	Marketer Manager	12823	False	False	True
4	64	male	6	Instagram	Lifestlye	United States	Rural	Student	14760	False	False	False

Figure 07. Raw Unstructured Data of Internet Usage Dataset

In *Figure 07*, I imported the CSV file and displayed its contents as a data frame. The dataset contains several columns: *age, gender, time spent, platform, interests, location demographics, profession, income, debt, home ownership, and car ownership*. Given my focus on female preferences for Instagram content browsing, I filtered the attributes gender, interests, and platform and dropped the other columns. Additionally, I selected females exclusively from the gender data, as the dataset comprises multiple gender types.

I updated the data frame with the relevant columns and variables, then aggregated the data to count the number of females interested in each type of content on Instagram – *lifestyle, travel, or sport*. The results are demonstrated in *Figure 08* and *Figure 09*.

	gender	interests	platform
0	non-binary	Lifestlye	Instagram
1	male	Travel	Facebook
2	non-binary	Lifestlye	Instagram
3	male	Travel	Instagram
4	male	Lifestlye	Instagram

Figure 08. Filtering Dataset to Gender, Interests, and Platform

Female interests on Instagram:

	Interests	Count
0	Lifestlye	53
1	Sports	48
2	Travel	34

Figure 09. Filtering by Gender and

Following the principles outlined in Chapter 4 of *'Better Data Visualisation'* (Schwabish, 2022, p. 78) and learning objective 01, I created a bar chart to accurately and clearly illustrate *Figure 09*.

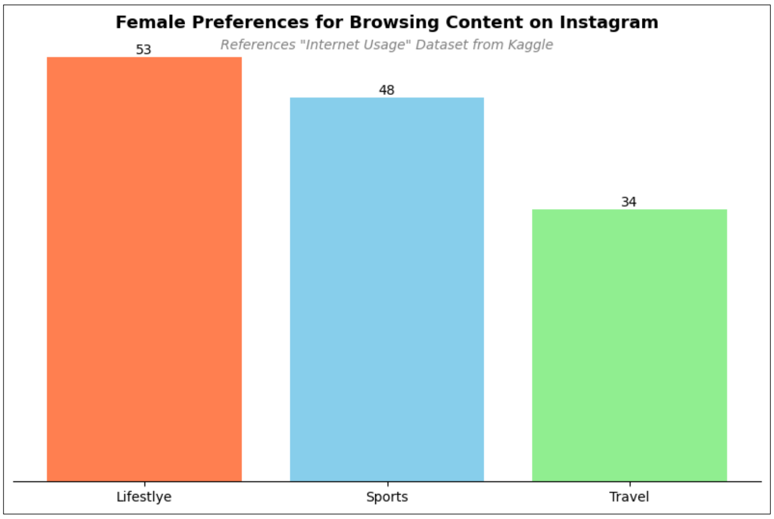

Figure 10. Bar Chart to demonstrate Female Preferences to Browse

Analysing the visual, *Figure 10*, I can presume that females are more interested in lifestyle content, in comparison with travel content, which appears to be the least preferred type. Considering this, when working with media algorithms for *content filtering*, also known as *content recommendation*, prioritising lifestyle-focused posts may be more beneficial for female users.

Grammar of Graphics Tools

Learning Objective 04

For this learning objective, I will use the 'League of Legends Champions Dataset' referenced from Kaggle (Patel, n.d.). *This dataset contains the statistics of the characters' metrics and attributes of the game.*

The *Grammar of Graphics* framework is a set of guidelines to create, deconstruct, and interpret data visualisation. It is a systemic approach on fundamental visualisation concepts that aims to enhance visual perception and implementation. This framework was initially proposed by Leland Wilkinson on his book, *'The Grammar of Graphics'* (Wilkinson, 2005) where he discusses when and how to use visual variables to support decision-making.

Wilkinson covers major aspects essential for effective data visualisation to ensure they are both visually effective and practically implementable. This framework includes practical tips such as: identifying key variables and relationships, thoroughly understanding the dataset, mapping aesthetics to consider how different styles in colour, shape, and size impact readability and interpretation, ensuring scales accurately represent the range and distribution of data, selecting the correct type of geometric object to match the nature of the data, applying statistical summaries to uncover key patterns and insights, using facets to break down complex data into smaller manageable subplots, and selecting the appropriate coordinate system that best fits the data (Wickham, 2010, p. 8).

Hadley Wickham further discusses these principles through the *ggplot2 package* in R, although the concepts remain applicable across various programming languages and tools that support data visualisation. Wickham's grammar builds on Wilkinson's style guide, which focuses on enhancing the *clarity, flexibility, and efficiency.* This approach is widely adopted in data analysis and visualisation workflows due to its systematic and scalable nature.

"…a comprehensive data visualisation style guide breaks down the parts of graphs, charts, and tables to demonstrate best practices and strategies to design and style your charts. Elements like font and colour, the widths of lines and style gridlines, and the use of tick marks are all choices that determine whether a graph is clear, engaging, and consistent – or whether it isn't." - (Schwabish, 2022, p. 349)

The *visualisation pipeline* guides the creation of visual data representations. It starts with data analysis, preparing data through filters and corrections. Mapping then transforms translating data into points or lines with attributes like colour and size which later is rendered into visual images. This structured process ensures effective and expressive visualisations for data analysis and understanding (Visualisation Pipelines, 2007).

Developing a data visualisation style guide serves three main purposes: Firstly, it provides detailed style guidelines and expectations to keep all team members on the same page. Secondly, individuals unfamiliar with the personalised styling and branding guidelines the organization uses may face challenges in creating graphs, which can be time-consuming.

A style guide streamlines this process and automates the application of graph styles. Lastly, as important as other branding materials, a style guide sets the tone and expectations - establishing standards that structure the *style, appearance,* and *details* of data visualisation. (Schwabish, 2022, pp. 349 - 350). As a result, the styling of graphs remains consistent, and the streamlined process creates a more efficient environment. Additionally, with a style guide, individuals can develop their own unique brand and identity.

"Colour has unmistakable power in our visualisations, it may be the first thing people notice about our graphs. Colour can evoke and draw attention. As Vincent van Gogh

wrote to his brother in 1885, 'Colour expresses something in itself. One can't do without it; one must make use of it." - (Schwabish, 2022, p. 358)

There 5 primary colour schemes, in which can improve a graph's appearance and clarity: *binary, sequential, diverging, categorical, highlighting, and transparency.* It is advisable to adhere to colour theory principles and consider how colours complement or contrast each other, as this can significantly influence one's perception of the graph. According to 'Better Data Visualisations,' it is recommended to avoid using a rainbow palette because it does not logically correspond to the *number system* in the data. (Schwabish, 2022, p. 361).

"Finally, be mindful that colours can reinforce stereotypes or hold different meanings in different cultures...how different cultures use and perceive different colours." (Schwabish, 2022, p. 362)

Working with the 'League of Legends Champions' dataset, I will display the champions with the highest and lowest base health and base mana. In this gaming culture, health is traditionally represented in red, while mana is depicted in blue. I plan to primarily use these two colours to highlight these attributes to emphasise the distinctions between attributes, as recommended in the *'NCI Colour Palette'* (Schwabish, 2022, p. 358).

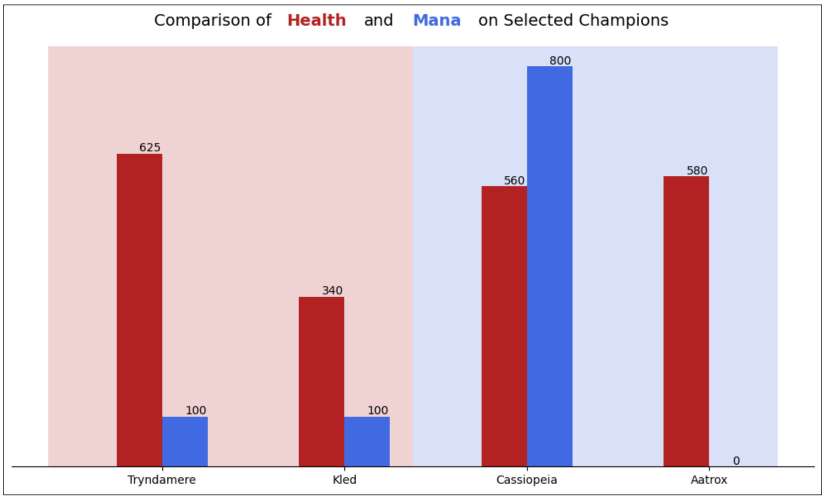

Figure 11. Comparison on League of Legends Champions Health and Mana

In *Figure 11*, I illustrated the champions with the highest and lowest values of health (left) and mana (right). I have added a translucent overlay of their respective colours to highlight each attribute. Additionally, I removed the legend and clarified the purpose of each colour by incorporating them into the title. Directly labelling wherever possible minimises disconnection between the audience and the information presented on the graph. (Schwabish, 2022, pp. 34 - 35).

By limiting my palette to two, I avoid colour clutter on the graph. Adjusting the opacity of the background colour also prevents it from overpowering the bars but instead, complements them. By repurposing the stereotypical identities of red and blue - representing health and mana, respectively, based on the game's concept - I enable the target audience to quickly grasp the intended message.

Effective colour palette manipulation enhances the reader's ability to perceive and understand graphs more effectively than plain text alone. Understanding the audience and their perception is important to determine the most effective way to communicate information.

"Colour speak louder than words" - (Rao, 2021)

Evaluation

Learning Objective 05

In this section, I will evaluate and refine Figure 02, the 'Pyramid Chart of the Differences in CGPA', aiming to create a clearer and more understandable graph for the audience.

What determines if a data visualisation is good? A good data visualisation is one that allows the reader to immediately understand the information presented without difficulty. *How can we establish if readers understand the graph?* To ensure a visualisation effectively communicates the intended information, conducting usability tests is the key. *Usability testing* help in refining data visualisations for the audience by understanding their behaviours and preferences, as well as identifying any issues missed during development from an external user perspective.

"We all need people who will give us feedback. That's how we improve..." – Bill Gates, 2013

Finalising an evaluation involves four main areas. Firstly, establishing clear objectives for what needs to be evaluated and identifying specific criteria to focus on. Secondly, select the appropriate method that align with these objectives. Thirdly, prepare and execute usability tests based on the chosen methods. Lastly, review data analysis and findings and identify common usability issues. Additionally, proposing design improvements and documenting the results in a report are important steps to effectively conclude the evaluation process.

There are two main types of evaluation methods: *qualitative* or *quantitative*. Qualitative methods involve focus groups and structured interviews aimed at gathering insights into how people use a product or service, making them ideal for understanding user experience. On the other hand, quantitative methods involve control groups and surveys, which focus on collecting benchmarks for analysis (Moran, 2019). As we are seeking user feedback based on their experience with graphs, the best choice is to opt for a qualitative method.

Referencing *Figure 02*, the *'Pyramid Chart of the Differences in CGPA'*, I provided a brief review, identifying issues with the graph and proposing solutions to enhance its clarity and effectiveness as a data visualisation.

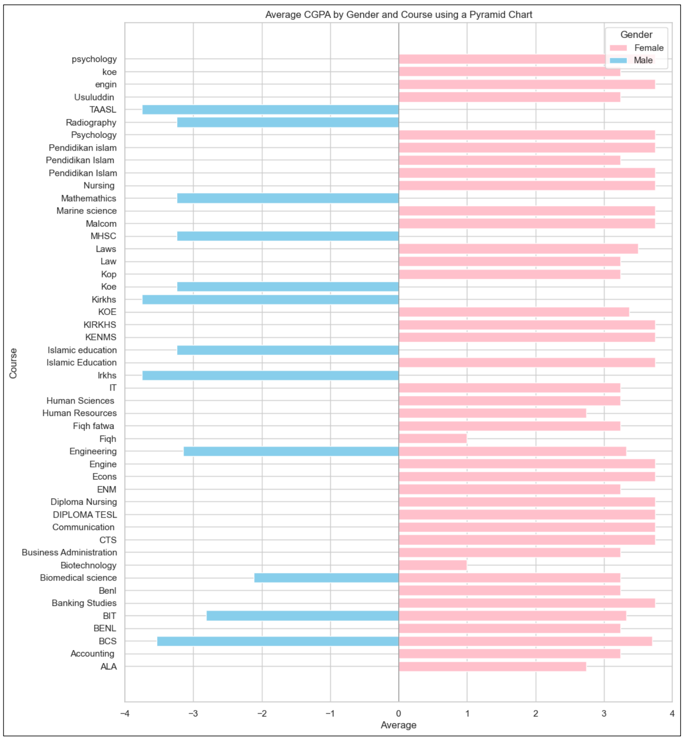

Figure 02. Pyramid Chart of the Differences in CGPA

Without reading the title or context yet, I already feel overwhelmed by the number of bar charts in this graph. While the graph displays all subjects, its purpose is to compare the differences in CGPA between male and female students. Therefore, focusing solely on the subjects that both genders are studying would be more appropriate.

"It's not about showing the least amount of data, it's about showing the data that matter most" - (Schwabish, 2022, p. 31)

Focusing solely on the differences in CGPA, there is no need for two separate charts plotting each gender. This approach would require readers to manually calculate the differences between each CGPA results, adding unnecessary work. Instead, this process can be streamlined by simplifying the pyramid chart into a bar chart that illustrates the differences.

Additionally, avoiding the initial colour palette may be appropriate, as it may be perceived differently due to social stereotypes that vary across cultures and time periods. Instead, I will choose a neutral colour to visualise the differences.

"For many years, pink and blue colours were used to differentiate data values for women and men. But in modern-day western cultures, these colours come with gendered stereotype baggage, pink suggest weakness and blue suggests strengths...up

until about the mid-twentieth century, it was the opposite...'Pink is, after all, just faded red, which in the era of scarlet-jacketed soldiers and red-robed cardinals was the most masculine colour, while blue was the signature hue of the Virgin Mary."- (Schwabish, 2022, p. 362)

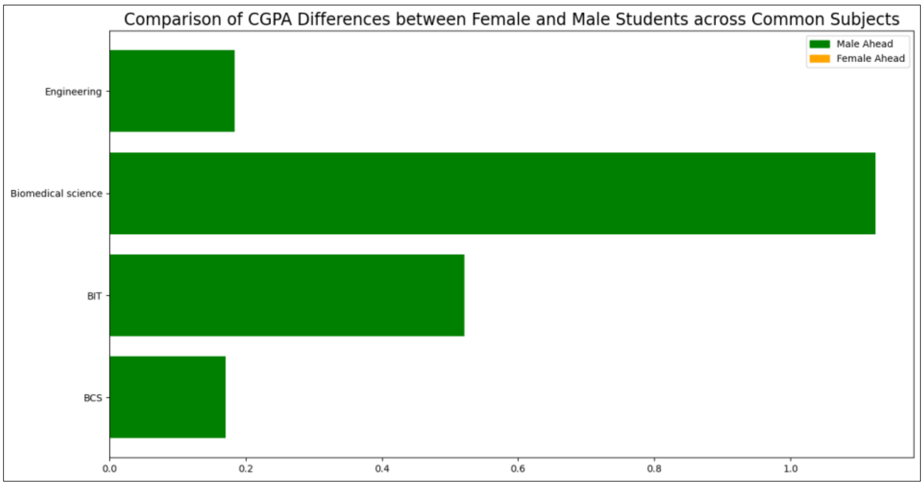

Figure 12. Updated Version of Figure 02

In *Figure 12*, significant improvements are evident when compared to *Figure 02*. The visualisation has been simplified, enhancing clarity and focusing more sharply on the primary objective of depicting CGPA differences between genders across common subjects. By transitioning from a dual-chart pyramid structure to a single bar chart displaying only the differences in CGPA, the information presented is now more accessible and straightforward for interpretation.

Conclusion

"...the best way to improve and refine your data visualisation technique is to create visualisations. Exploring different data sets, visualisation types, and tools can help you refine your aesthetic, and play with different techniques and forms." - (Schwabish, 2022, p. 408)

Throughout this assignment, I have delved into various types of datasets and experimented with different methods of visualising data. As I continue to work with more data in the future, my aim is to refine my skills further and develop my own aesthetic for data visualisation.

In addition, through this assignment, I have gained a foundational understanding of data visualisation, including its purposes and the ability to personalise graphs based on audience preferences - manipulating visual elements to enhance the perception and clarity of the data presented.

References

- Andreas Kerren, A. E. (2006). Human-Centered Visualization Environments. Revised Lectures.
- Elgabry, O. (2019, February 28). *The Ultimate Guide to Data Cleaning.* Retrieved from https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4.
- Franconeri, S. (2019). *How to Make Perfect Pancakes*. Retrieved from Youtube: https://www.youtube.com/watch?v=Jq2Rc0WlYTE&t=8s
- Gates, B. (2013). Bill Gates: Good Feedback Is the Key to Improvement. *Inc*.
- Islam, M. S. (n.d.). *Student Mental health*. Retrieved from Kaggle: https://www.kaggle.com/datasets/shariful07/student-mental-health
- Jones, M. (2017, December 13). *Working with Messy Data.* Retrieved from IBM: https://developer.ibm.com/tutorials/ba-cleanse-process-visualize-data-set-1/
- McCandless, D. (2012). *The Beauty of Data Visualisation*. Retrieved from Youtube: https://www.youtube.com/watch?v=5Zg-C8AAIGg
- Moran, K. (2019). *Usability Testing 101*. Retrieved from NNg: https://www.nngroup.com/articles/usability-testing-101/
- Patel, A. (n.d.). *League of Legends Champions Dataset*. Retrieved from Kaggle: https://www.kaggle.com/datasets/dem0nking/league-of-legends-champions-dataset?resource=download
- Pavan Kalyan. (n.d.). *Internet Usage*. Retrieved from Kaggle: https://www.kaggle.com/datasets/pavan9065/internet-usage
- Rao, S. K. (2021). *Colours Speak Louder Than Words*. Retrieved from EPL: https://www.eplglobal.com/colours-speak-louder-than-words/
- Schwabish, J. (2022). Better Data Visualisations. In J. Schwabish, *Better Data Visualisations.*
- *Visualisation Pipelines.* (2007). Retrieved from Wiki: https://infovis-wiki.net/wiki/Visualization_Pipeline
- Wickham, H. (2010). *A layered grammar of graphics.* Retrieved from https://vita.had.co.nz/papers/layered-grammar.html
- Wilkinson, L. (2005). *The Grammar of Graphics.*