Challenges of Data Visualization

Data visualization is a graphical method to present information and data in such a way that is straightforward and effective by simplifying data with charts, graphs, info-graphics, and so on (https://www.facebook.com/kdnuggets, 2024). This method allows analysts to identify patterns and aids the audience understand better, hence it is a persuasive tool in decision making.

However, there are challenges in data visualization. Firstly, the veracity of data must be ensured to avoid redundant or inaccurate results. Secondly, data investigation shall be carried out to choose the correct chart type and seek audience feedback in design phase, this could keep away from misinterpretation and helps in decision making (Maya, 2024). Next, the dashboard containing charts must be simplified to avoid information overload which could potentially lead to confusion instead of clarity. Therefore, the data must be segmented into smaller parts during visualization rather than a piece of giant graphic (Solutions, 2024).

Furthermore, the alignment, contrast and white space need to be utilized to improve user experience. High colour contrast could affect viewers' perception of the value disparity to be greater than they are (Bowers, 2020). Adequate white space helps viewers focus their attention and increases readability. Lastly, alignment plays a role in organizing the design and improving readability. Unfinished and cluttered look is caused by poor alignment in design (*Design Dictionary | Venngage | Alignment [Design Principle Definition]*, 2019). Meanwhile, a good alignment gives a tidy and clearer impact to the viewers.

Dataset

The “2023 Data Scientists Salary” dataset is obtained from Kaggle, where the dataset contains factors that affect the salaries.

work_year	experience_level	employmer	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
2023	SE	FT	Principal Data Scientist	80000	EUR	85847	ES	100	ES	L
2023	MI	CT	ML Engineer	30000	USD	30000	US	100	US	S
2023	MI	CT	ML Engineer	25500	USD	25500	US	100	US	S
2023	SE	FT	Data Scientist	175000	USD	175000	CA	100	CA	M
2023	SE	FT	Data Scientist	120000	USD	120000	CA	100	CA	M
2023	SE	FT	Applied Scientist	222200	USD	222200	US	0	US	L
2023	SE	FT	Applied Scientist	136000	USD	136000	US	0	US	L
2023	SE	FT	Data Scientist	219000	USD	219000	CA	0	CA	M
2023	SE	FT	Data Scientist	141000	USD	141000	CA	0	CA	M
2023	SE	FT	Data Scientist	147100	USD	147100	US	0	US	M
2023	SE	FT	Data Scientist	90700	USD	90700	US	0	US	M
2023	SE	FT	Data Analyst	130000	USD	130000	US	100	US	M
2023	SE	FT	Data Analyst	100000	USD	100000	US	100	US	M
2023	EN	FT	Applied Scientist	213660	USD	213660	US	0	US	L
2023	EN	FT	Applied Scientist	130760	USD	130760	US	0	US	L
2023	SE	FT	Data Modeler	147100	USD	147100	US	0	US	M
2023	SE	FT	Data Modeler	90700	USD	90700	US	0	US	M
2023	SE	FT	Data Scientist	170000	USD	170000	US	0	US	M
2023	SE	FT	Data Scientist	150000	USD	150000	US	0	US	M
2023	MI	FT	Data Analyst	150000	USD	150000	US	100	US	M
2023	MI	FT	Data Analyst	110000	USD	110000	US	100	US	M

(Shan, 2023)

Variables	Data types	Meaning	Values
work_year	Integer	The year when salary was paid	2020, 2021, 2022, 2023
experience_level	String	Job experience in the year: • EN: Entry-level / Junior • MI: Mid-level / Intermediate • SE: Senior-level / Expert • EX: Executive / Director	EN, MI, SE, EX
employment_type	String	Employment type: • PT: Part-time • FT: Full-time • CT: Contract • FL: Freelance	PT, FT, CT, FL
job_title	String	Role title, represent their skill sets	Principal Data Scientist, Machine Learning Engineer , etc.

salary	Integer	Annual salary paid.	Ranged between 6,000 to 3,040,000
salary_currency	String	The currency of the salary paid followed ISO 4217 currency code. i.e: JPY stands for Japanese Yen	EUR, USD, JPY, CAD, etc.
salary_in_usd	Integer	Salary amount currency in USD	Ranged from 5,132 to 450,000
employee_residence	String	Employee's country of residence followed ISO 3166 country code. i.e: AU is Australia	ES, US, CA, etc.
remote_ratio	Integer	Amount of work done remotely: • 0 No remote • 50 Partially remote • 100 Fully remote	100, 50, 0
company_location	String	The country where the company's head quarter is located, followed ISO 3166 country code	ES, US, CA, etc.
company_size	String	Number of people that worked for the company: • S: less than 50 • M: 50 to 250 • L: more than 250	S, M, L

Problem Statement

The data scientist's salary dataset contains an array of details regarding the salaries, whereby those salaries are not standardised, and absence of benchmark that leads to a pay difference for the same skill sets. Benchmarking is important for organisations to make comparison with other businesses and acknowledge of the pay differences in their own sector, also ensure a fair package after considering every aspect of remuneration (*Why Is It Important to Salary Benchmark?* 2024). This could also promote pay transparency which contributes to pay equity, motivate performance, and builds trust between employees and employers (Pay *Transparency: Understanding What It Is and Why It Is Important*, 2024).

To tackle the problem, the dataset will be studied, and a dashboard needs to be created along with a salary benchmark to give better visibility to viewers in determining the worth of numerous skill sets in different parts of the world.

Research Questions

Who are the audience, and what to do with the dashboard?
How to grab the attention of audience and direct their eyes for a right flow in the dashboard?
What are the charts to be created for univariate, bivariate and multivariate analysis to display the relationship among variables?
What are the salary benchmarks for various job titles specific to their region, experience, remote ratio, company size and location across the years?

Objectives

To present the customised dashboards with different charts , graphs, and action elements for identified audience groups.
To apply sharp contrast colours for the text and some graphs to gain audiences' attention, then use Gestalt's Principles in designing the dashboards.
To visualize the relationship between the variables by using charts and graphs in univariate, bivariate and multivariate analysis.
Apply streamgraphs to benchmark salary for the required factors throughout the years.

Proposed Solutions / Outcomes

A good data visualization has four characteristics, also known as ACES: accurate, clear, empowering, and succinct (David, 2020). The contents ought to be simple and digestible for audience consumptions for audience to quickly understand information from a high-level view (Top *10 Proven Data Visualization Best Practices, 2023)*. The aim of studying the dataset of data scientists' salary is to determine and benchmark salary based on skill set in different regions through dashboards created.

The dataset contains both quantitative and categorical variables, hence the Exploratory Data Analysis (EDA) is applied to discover insights of patterns, trends, and relationships between variables with the aid of statistics and data visualization tools (Shah, 2021). There are 3 methods in the EDA process: univariate analysis, where the single categorical variables will be visualized by using pie or bar chart; bivariate and multivariate analysis will involve line graphs, histograms, and bubble charts to study interaction between multiple categorical variables and salary. Box and whisker plot will be used to show a summary of data through its quartiles (*17 Important Data Visualization Techniques | HBS Online*, 2019), whereby salary and experience levels would be the variables. After the analysis, the streamgraphs can be applied to display evolution of salaries of different job titles throughout the 4 years period. The stacked area chart can be used to display the values of multiple groups (Holtz, 2015), where the progression of salaries, and other factors can be checked within the same figure.

In the presentation of the dashboard, several design principles will be applied to ensure audience engagement while keeping the relevancy for target audiences. Therefore, dashboards will be customized to different viewers and minimize elements or excessive details (Bernardita Calzon, 2023). Meanwhile, the board itself will follow Gestalt's Principles, utilize the white space and alignment to look tidy and less clutter; eye-catchy colours will be applied to some text and graphs to grab attention and lead the audience to follow the flow.

Summary

Data scientists is a new role with high salary and exponential growth in demand in the past 10 years. To understand the salary distribution, the dataset of data scientists' salary would be used for study and analysis.

A data visualization dashboard will then be used to display the correlation of salary and several factors in charts and diagrams in the EDA process. Then, a salary benchmark will be created for a better presentation of its evolution to determine the worth of various skill sets in different regions.

Lastly, the dashboard will be designed in such a way that is comfortable for viewers and customized the contents to match the expectation of respective audience groups.

(980 words excluding introduction, table, and reference)

References

1. *Data Scientist: The Sexiest Job of the 21st Century*. (2012, October). Harvard Business Review. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century
2. *Data Scientist Still the Sexiest Job of the 21st Century?* (2022, July 15). Harvard Business Review. https://hbr.org/2022/07/is-data-scientist-still-the-sexiest-job-of-the-21st-century
3. Shan, H. (2023). *2023 Data Scientists Salary*. Kaggle.com. https://www.kaggle.com/datasets/henryshan/2023-data-scientists-salary
4. https://www.facebook.com/kdnuggets. (2024). *Data Visualization: Presenting Complex Information Effectively - KDnuggets*. KDnuggets. https://www.kdnuggets.com/data-visualization-presenting-complex-information-effectively#:~:text=The%20purpose%20of%20data%20visualization,an%20audience%20toward%20a%20conclusion.
5. Maya. (2024, January 2). *10 Common Challenges of Data Visualization & Their Solutions*. Synodus; Synodus. https://synodus.com/blog/big-data/challenges-of-data-visualization/
6. Solutions, P. 3. (2024, April). *Top 7 challenges in data Visualization & ways to overcome*. Platform 3 Solutions | Application Decommissioning, Data Migration &

Archival Solutions. https://platform3solutions.com/top-7-challenges-in-data-visualization-and-how-to-overcome-them/

7. Bowers, M. (2020). *Numbers Shouldn't Lie: An Overview of Common Data Visualization Mistakes*. Toptal Design Blog; Toptal. https://www.toptal.com/designers/ux/data-visualization-mistakes#:~:text=One%20of%20the%20challenges%20in,rendered%20in%20blue%20and%20green.%E2%80%8C
8. *Design Dictionary | Venngage | Alignment [Design Principle Definition]*. (2019). Venngage.com. https://venngage.com/design-dictionary/alignment-design-principle-definition/#:~:text=Alignment%20is%20a%20design%20principle,will%20look%20cluttered%20and%20unfinished.
9. Herskowitz, Z. (2024, February 14). *Pay Transparency in 2024: Everything You Need to Know*. Onward Search. https://www.onwardsearch.com/blog/2024/02/pay-transparency/#:~:text=Salary%20transparency%20helps%20foster%20trust,the%20pay%20is%20the%20same.*Pay Transparency: Understanding What It Is and Why It Is Important*. (2024). Betterup.com. https://www.betterup.com/blog/pay-transparency#:~:text=Transparency%20in%20pay%20can%20build,to%20feel%20valued%20and%20engaged.
10. *Why is it important to salary benchmark?* (2024). Robertwalters.co.uk. https://www.robertwalters.co.uk/insights/hiring-advice/blog/why-is-it-important-to-salary-benchmark.html
11. David, M. (2020). *5 Data Visualization Best Practices: The Secrets Behind Easily Digestible Visualizations*. Chartio; Chartio. https://chartio.com/learn/business-intelligence/5-data-visualization-best-practices/#:~:text=Accurate%3A%20The%20visualization%20should%20accurately,t%20take%20long%20to%20resonate.
12. *Top 10 Proven Data Visualization Best Practices*. (2023, November 2). GoodData. https://www.gooddata.com/blog/5-data-visualization-best-practices/
13. Holtz, Y. (2015). Stacked Area Graph. Data-To-Viz.com. https://www.data-to-viz.com/graph/stackedarea.html
14. Bernardita Calzon. (2023, April 5). *25 Dashboard Design Principles, Best Practices & How To's*. BI Blog | Data Visualization & Analytics Blog | Datapine. https://www.datapine.com/blog/dashboard-design-principles-and-best-practices/

15. Shah, K. (2021, April 19). *Exploratory Analysis: Using Univariate, Bivariate, & Multivariate Analysis Techniques*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/04/exploratory-analysis-using-univariate-bivariate-and-multivariate-analysis-techniques/
16. Kabacoff, R. (2024). *Chapter 4 Univariate Graphs | Modern Data Visualization with R*. Github.io. https://rkabacoff.github.io/datavis/Univariate.html#quantitative