Introduction

The demand for data scientist role has been sky-rocketed since 2012, along with the huge hike in their salary compared to other professionals. This makes data scientists the modern unicorn in the 21st century. However, the salary varies with many factors. Therefore, data visualization is needed to analyze complex data, identify patterns, extract information, and lastly present the data visually for the audience to understand quickly, thus making decisions accurately (Kesav Kalluri, 2020).

To conduct the study, various graphs, charts, and plots of variables in relation to salary will be used and evaluated as a part of EDA process. Suitable diagrams will be included in the storyboards which catered to diverse groups of audience: Job seekers, recruiters, and human resource. The storyboards will be designed with several design principles to appeal to the viewers.

Problem Statement

In the dataset containing salary of data scientists, the absence of benchmark and non-standardised salary caused a significant dissimilarity in the pay of the same skills sets or titles. The relationship between variables and salaries are unclear to determine a fixed amount for the roles.

Objective

Study the dataset by creating charts or plotting graphs to visualize the relationship between variables in univariate, bivariate and multivariate analysis. Also, apply streamgraphs to benchmark salary for the required factors throughout the years. Next, create customized storyboards for identified audience groups with suitable diagrams created earlier on. To ensure the audience's understanding, apply Gestalt's principle and contrast colours in the storyboard to grab their attention, guide them in reading and understanding the information presented.

Lastly, the storyboards created with the respective diagrams along with salary benchmark to give better visibility to viewers in determining the worth of their skill sets in different parts of the world.

Dataset

The dataset named “2023 Data Scientists Salary” contains 11 rows (variables) and 3755 rows (records).

Variables	Data types	Meaning
work_year	Integer	The year when salary was paid
experience_level	String	Job experience
employment_type	String	Employment type
job_title	String	Role title, also represents their skill sets
salary	Integer	Annual salary
salary_currency	String	Currency of the salary
salary_in_usd	Integer	Amount of salary in USD
employee_residence	String	Where employee lives
remote_ratio	Integer	Amount of work done remotely
company_location	String	Where the company located
company_size	String	Size of the company

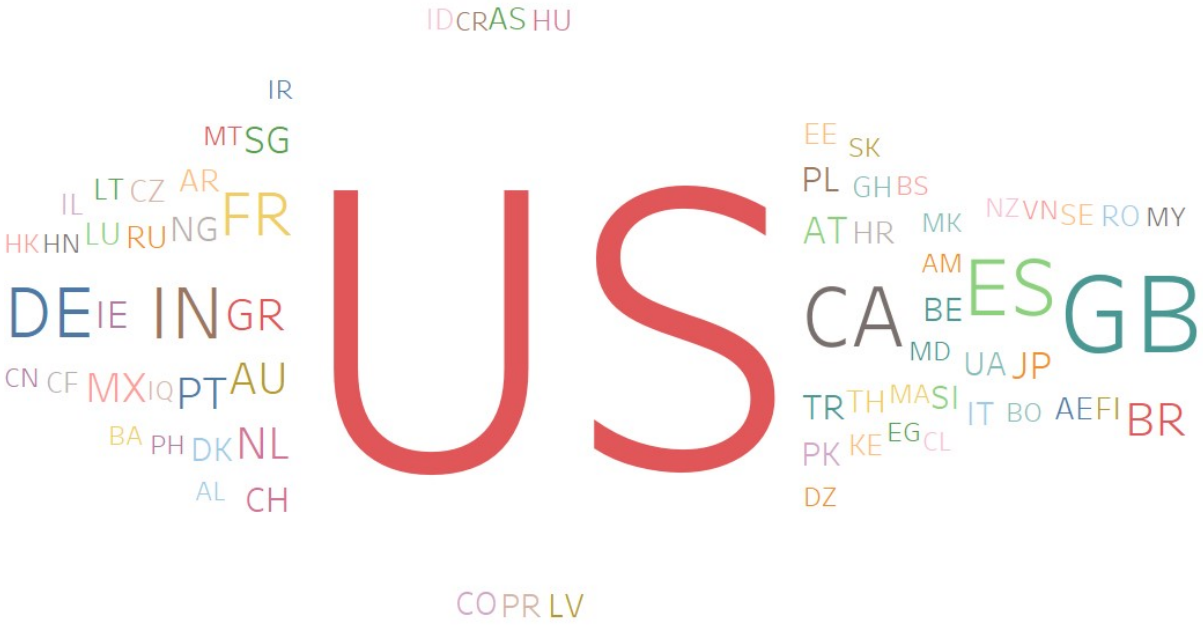
Identify Audience

The dataset is beneficial to several groups of audience due to the nature of their role and respective focus.

Audience	Role description	Focus
General public	• High school graduate, or someone who is checking the future career opportunities while surveying for the courses to further study. • Looking for a quick summary of role and salary information	• Top popular roles • Job market by location • Display a summary of average salary by levels of experience
Talent Acquisitions	• Recruiters or talent acquisitions from agencies • Company human resource	• Analyse the salary by experience and job titles, employment type, company size, company location etc. • Details of the salary distribution by minimum, maximum, median, range. • Market rate, baseline, percentile of salary by the year trend.
Job seeker	• Employees • Someone already in the work force	• Display popularity of job title and company location • Display the salary by experience, company size, company location and job title • Chance of working remotely

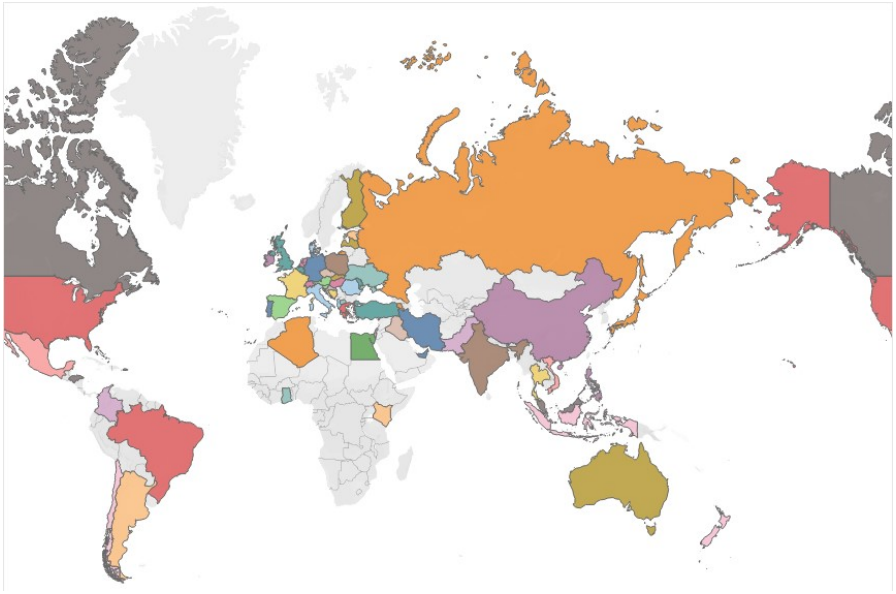
Visualization

Word Cloud



Word Cloud represents the count of company locations in the dataset. Each country is presented by their code, where the size of text represents the count in the dataset, the bigger the text, the higher the count. In the diagram, United State has the most companies with data science related records. This method is a straightforward way to tell the audience how popular the role in various places, the colours and text size are the pre-attentive attribute that catch their attention instantly.

Map



The map displays the company locations in the world, where the countries present in the dataset are colored and highlighted, while the dull light grey regions are places not captured in the dataset. This map gives a visible view of job opportunities around the world.

From the map, the company locations are all over the world, covered the entire Europe and Asia region including some remote islands like Hawaii and New Zealand, but scarce in Africa and Central Asia, none in Greenland.

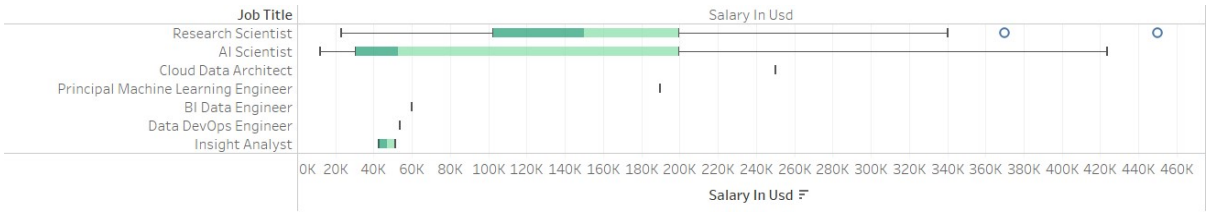
However, viewers might confuse with the countries positioned on the map are sharing the same colours due to limited colour palette, for instance, red colour is used for US and Brazil, orange for Russia and Algeria; light blue in Italy, Denmark, Albania, and Romania. Second issue happen when a similar colour is used side-by-side, such as purple (China) and light purple (Pakistan), light green (Ukraine) and light blue (Romania). These low contrast colours cause the viewers to misinterpret that the countries of similar colours are one country.

Packed Bubble Chart

A Packed Bubble Chart is an alternative to a bar chart for small sample size with at least one dimension and measure to visualize information in a compact way. Otherwise, the chart will be overwhelmed by the number of bubbles, especially when some of the bubbles are of equivalent size to judge the values or too small to see (*Understanding and Using Bubble Charts | Tableau*, 2019).

The chart above displayed the top 5 data jobs in which the popularity is determined by the size and colour of the bubble, where bigger size and darker colour suggest higher popularity, and vice versa. The most popular role is Data Engineer, followed by Data Scientist, Data Analyst, Machine Learning Engineer and lastly Analytics Engineer.

Box plot

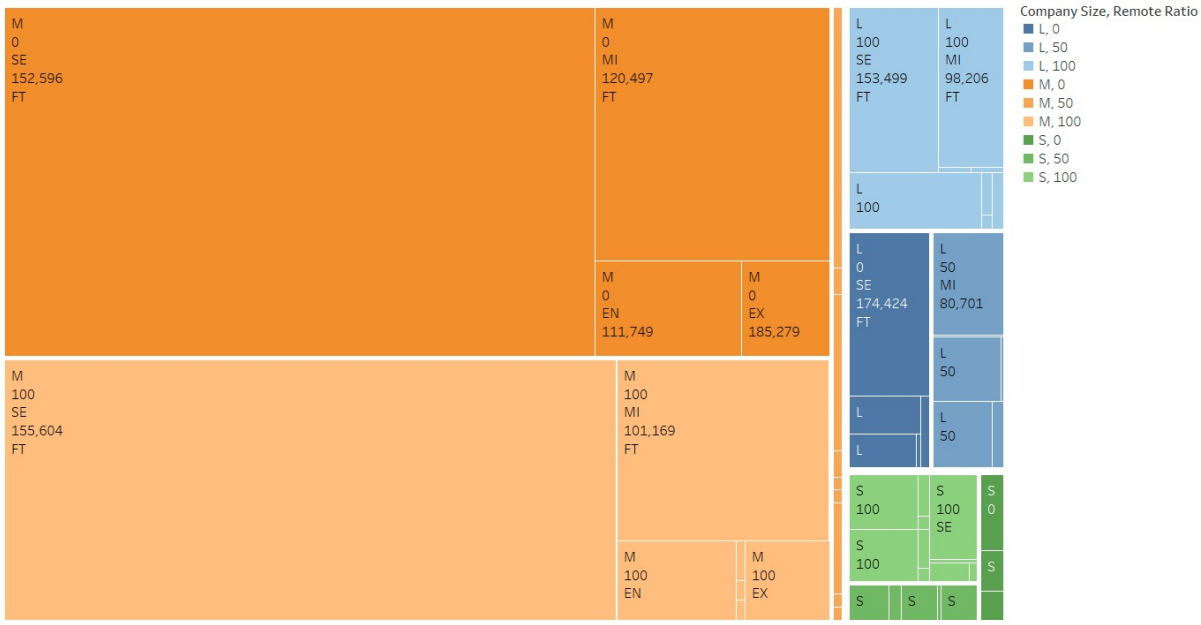


The box plot provides a clear view for salary distribution of each job title, by displaying the outliers, lower and upper whiskers, quartiles and median. This graphical display allows the large data summary to be viewed effectively. When the plot is sorted by descending order, Research Scientist and AI Scientist are placed at the first and second places

with the highest pay. The salary range of Research Scientist is between $23,000 to $340,000, while the range for AI Scientist is $12,000 to $423,834. But AI Scientist is placed second, because Research Scientist has salary outlier of $450,000 that exceeds the range of AI Scientists.

However, the plot is not suitable for small dataset (Staff, 2024). For example, there are a few jobs that have little or only 1 record, which are Cloud Data Architect, Principal Machine Learning Engineer, BI Data Engineer etc. This causes the absence of whiskers or box in the plot, along with odd display of data when comparing the jobs such as Insight Analyst (2 records) and Data DevOps Engineer (1 record). The reading process can be challenging for analyzing a mass number categorical and numerical attributes due to the enormous amount of information to investigate.

Tree maps



The chart is an alternative to word cloud, it gives a hierarchical view of data and make it easy to spot patterns (*Create a Treemap Chart in Office - Microsoft Support*, 2021). From the map, the employment type, experience level and average salary in each company size and remote ration are shown in detail.

By looking at the colours, medium-sized companies have half of the work done on-site (0), half is done remotely (100), and the remaining small portion is through hybrid (50) mode. Meanwhile in large companies, on-site occupied half of the total work done, that is a combination of both remote and hybrid; most of the work done in small company was through remote. Therefore, the remote mode is playing a major part in all company sizes.

Also, the map shows the proportions of records among the 3 companies, in which small sized companies have the least records, followed by generous size, and lastly medium size has the most records.

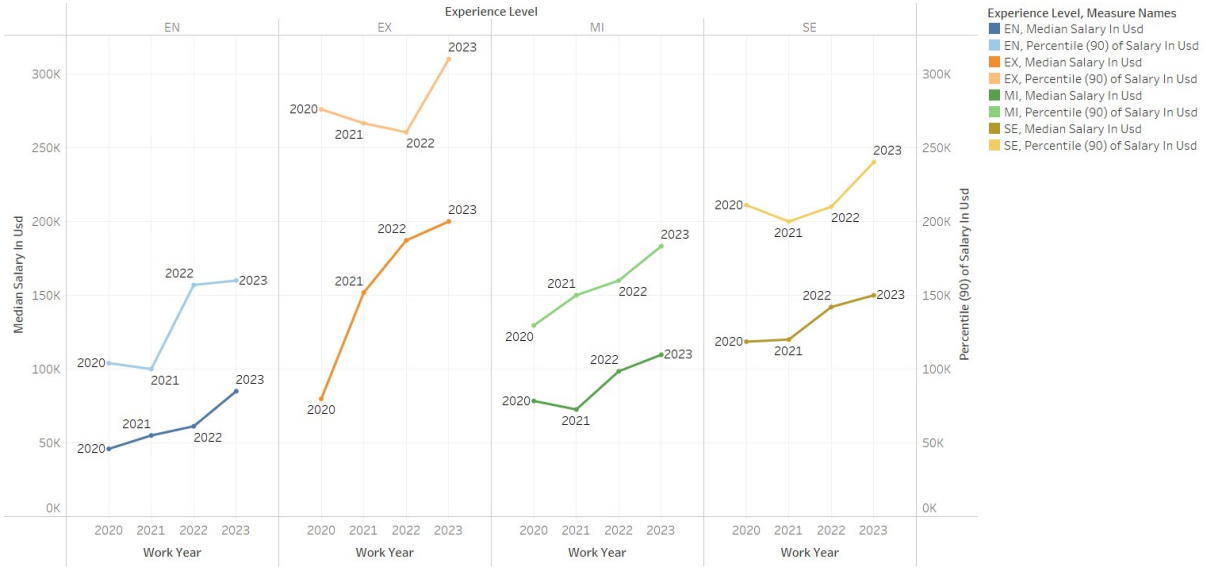
In terms of employment type and experience level, senior or mid-level full-timers did most of the work regardless of the company size. In other words, they build up most staff in those companies.

Bar chart

The bar chart is one of the straightforward ways to summarize a dataset based on categorical variables such as Experience Level, Employment Type, Job Title, Salary Currency, Employee Residence. The bars will be sorted in descending order and a darker colour shade is used for a bigger number to give a better visual effect to viewers. The categorical variables are placed at the y-axis, while numerical variables are at the x-axis for pleasurable reading.

In the Experience level vs Salary chart above, average salary amount will be shown at the end of each bar so the viewers can refer to it quickly and easily. The executive has the highest salary, followed by senior, mid, and entry level. The bar chart will be varied for different viewers, this chart displays basic information needed by the public or potential labour. More details will be displayed for recruiters and current workers in the field.

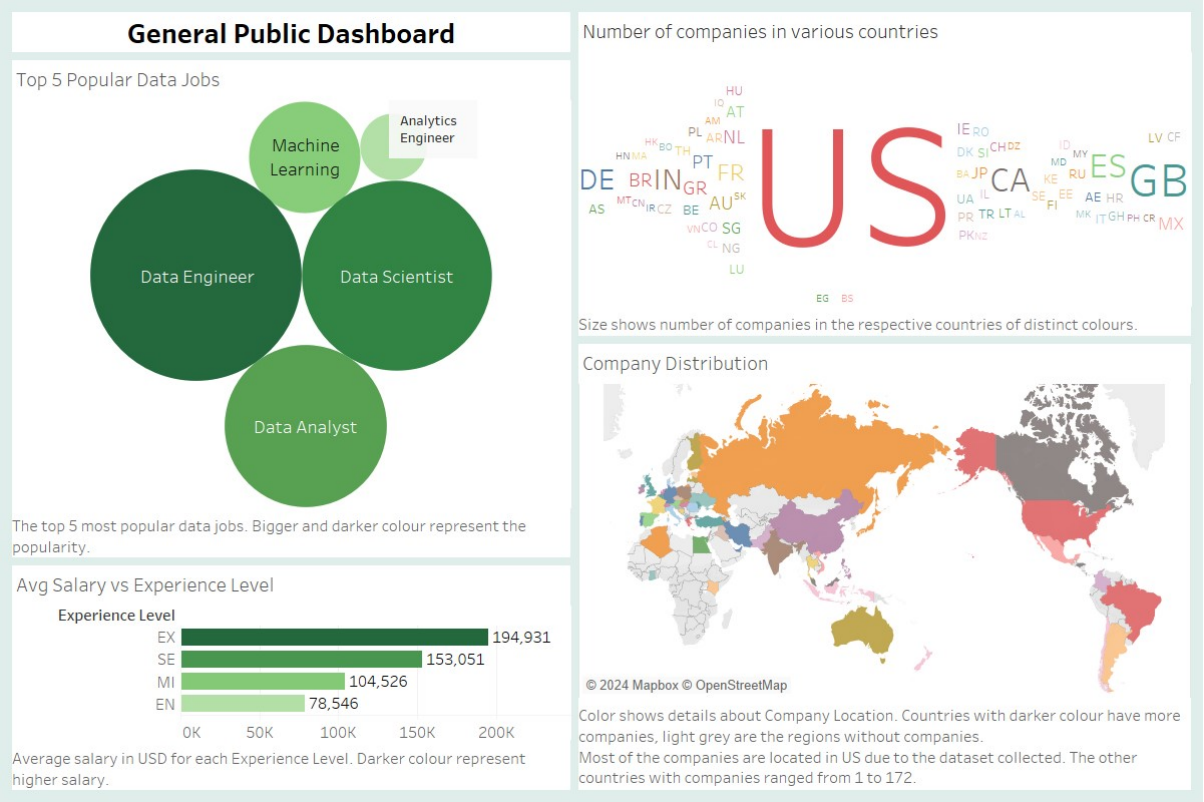
Line graph



The graph shows a general salary comparison of median and 90 percentiles from 2020 to 2023 for unique experience level. The graph shows the same pattern for all the experience levels where median salary is below 90 percentiles. However, the difference between 2 measurements can be unexpectedly big and have non-matching trends. For instance, EX in 2020 and 2021, where the salary gap between median and 90 percentile is wide in 2020; from 2020 to 2022, the trend of 90 percentile went downwards while the median salary increased exponentially.

Storyboards

Storyboard #1: General Public



The dashboard for the public shows some simple diagrams or charts that give the necessary information, easy to digest and understand through. The theme of the project is green; therefore, the dashboard background is lighter shades of green to avoid distraction. To ensure the flow of the dashboard, the audience's journey should start from the left, the bubble chart, to colourful maps, then to bar chart.

<u>Title: General Public Dashboard</u>

Title's font is large and bold as a visual cue for viewers to identify the main idea of content to be read in a section.

<u>Packed bubble chart</u>

This is an attention-grabbing tool to get viewers' eyes to focus on the large bubble of the data roles. The chart is straightforward to read and process, viewers would first land on the largest bubble and go through other bubbles quickly as there are only 5 of them. The Analytics Engineer is annotated due to its long text which does not show up in the small bubble.

Word map: Number of companies in various countries

The red “US” is an eye-catchy word that directs viewers’ attention to stay on the map for a few seconds to understand the popularity of the jobs in the US market compared to others.

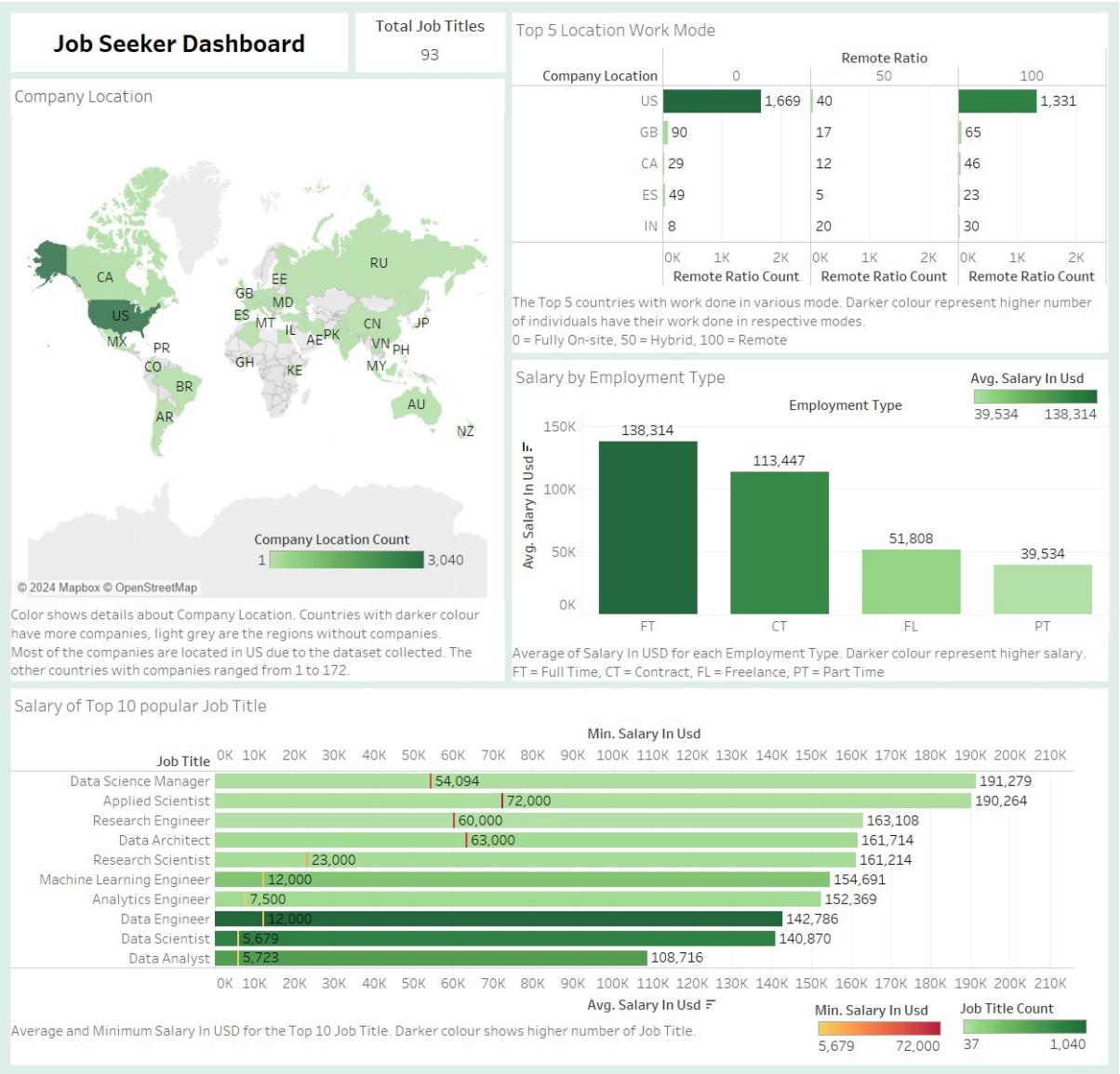
World map: Company distribution

The map shows the location of companies in the world, where the user would get a glance to know how famous the job is in the global market before moving to the next part.

Bar chart: Average Salary vs Experience Level

The chart reveals the salary at distinct stages, which is the answer to the audience’s curiosity.

Storyboard #2: Job seeker



The dashboard contains detailed information to feed the appetite of those who are already in the work force. The employee's emphasis is on the salary amounts and other factors such as overseas job opportunities and work mode.

Total Job Titles

In the Job Seeker dashboard, the audience would know there are 93 different titles in the data science field.

Company location

The map highlighted US has the most companies involved in data science compared to the rest. Also, data science jobs are famous in the world as green regions are more than grey.

Top 5 location work mode

Hybrid work mode is significantly lesser than on-site and remote for all countries. Most of the work is done either on-site or remote.

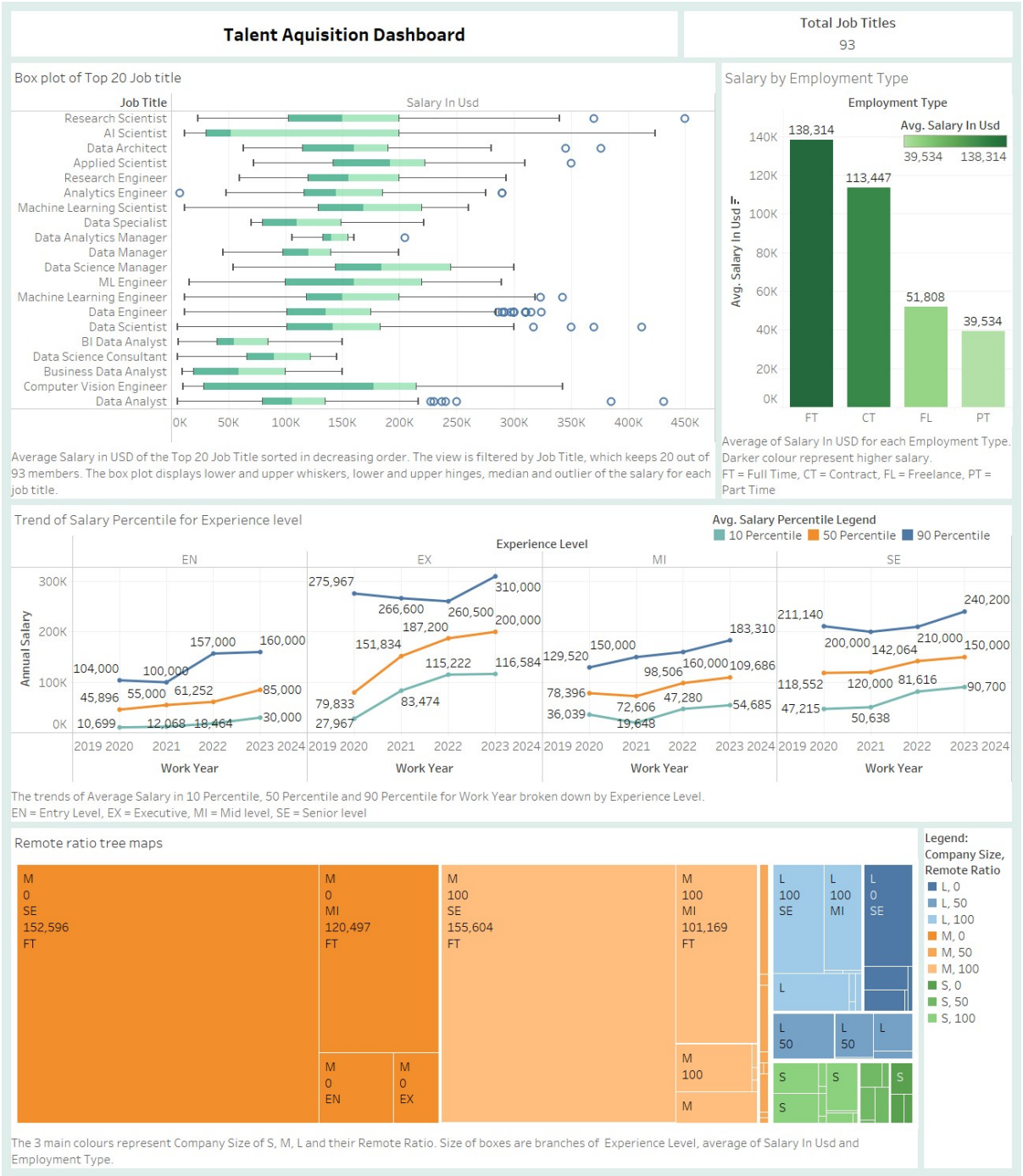
Salary by Employment Type

The average salary of a full-timer is the highest among all. Contractors' salary is closest to full-timer, while a sharp drop is observed in the salary of freelancer and part-timer.

Salary of Top 10 Popular Job Title

Data Science Manager and Applied Scientist have the highest salary although they are not as popular as Data Engineer, Data Scientist and Data Analyst. The minimum salary of the roles is also shown in Gantt bar with a contrast gradient colour of red and orange for easy spotting. The minimum salary allows the audience to check if they are not underpaid.

Storyboard #3: Talent Acquisitions



The dashboard for recruiters has more features and precise information for them to make complex analysis for gaining accurate insights. From the visualization, the information studied can be used to make a complete salary guide that benefits all parties.

Total job titles

Recruiters must know the 93 job titles in this data science ecosystem. Each of them has unique job scopes and skills sets, therefore a different salary range.

Box plot of Top 20 Job Title

The top 20 roles are focused for the analysis to avoid the dashboard from congestion. The box plot shows the range of salary and the outliers, which are helpful information for studying the salary market of data jobs and hiring for the specific roles.

Salary by Employment Type

Another piece of information needed for hiring diverse types of employment contract.

Trend of Salary Percentile for Experience Level

Useful for comparing the market trend throughout the years for people with unique experience. The 10, 50 and 90 percentile helps in drawing the salary baseline and standards. This also allows predictions to be made for future salaries.

Remote Tree Maps

The map is a direct way to visualize the population of company size. The addition of more variables contributes to detailing the map by sorting out specific information, for instance, the average salary of a full-time senior level remote employee in a medium sized company is $155,604.

Conclusion

Data visualization simplifies complex data, helps in decision making and finds areas that need improvement or attention (Team Atlan, 2023). The dashboards can be customized to tell stories and information from different perspectives for the audience, this allows them to understand the data efficiently. However, there are some limitations in the current data in making the visualization, in which the remote ratio values can be confusing. Therefore, the dataset shall go through data cleaning to ensure the quality, and enrichment process to discover more insights before the visualization process.

Word Count: 2345 words

References

Kesav Kalluri. (2020, August 14). *Importance, Purpose, and Benefit of Data Visualization Tools!* Business Analytics Platform with Pre-Built Insights | SplashBI. https://splashbi.com/importance-purpose-benefit-of-data-visualization-tools/#importance

Staff, C. (2024, March 21). *What Is a Box Plot?* Coursera. https://www.coursera.org/articles/what-is-a-box-plot

Create a treemap chart in Office - Microsoft Support. (2021). Microsoft.com. https://support.microsoft.com/en-us/office/create-a-treemap-chart-in-office-dfe86d28-a610-4ef5-9b30-362d5c624b68#:~:text=A%20treemap%20chart%20provides%20a,shown%20as%20a%20smaller%20rectangle.

Understanding and Using Bubble Charts | Tableau. (2019). Tableau. https://www.tableau.com/data-insights/reference-library/visual-analytics/charts/packed-bubbles

Team Atlan. (2023, July 21). *11 Benefits of Data Visualization You Can't Ignore in 2024*. Atlan.com; Atlan. https://atlan.com/benefits-of-data-visualization/