Escuela Técnica Superior de
Ingenieros Informáticos
Universidad Politécnica de Madrid
antes Facultad de Informática

Second Exercise: Design of a New Interactive Data Analysis Tool for Data Science Salaries

Master in Data Science
Data Visualization

POLITÉCNICA

Enrique Martínez Martel

Carlos Almodóvar Román

Contents

1. Introduction

In the modern era of complex datasets, data visualization is an extremely powerful tool for analyzing them. The design of interactive data analysis tools plays a crucial role in aiding analysts and end-users to obtain meaningful insights.

Such tools serve as an important instrument for individuals seeking to explore, interpret, and draw conclusions from datasets. As the volume and complexity of data continue to grow, traditional static methods of analysis fall short in providing a comprehensive understanding. Interactive tools aid analysts to explore data dynamically and visualize patterns in real-time, providing a deeper engagement with the information at hand. These tools not only enhance the speed and accuracy of analysis but also promote a more iterative and exploratory approach, allowing users to uncover hidden relationships and trends.

2. Problem Characterization

The challenge presented in this exercise centers around crafting a new interactive data analysis tool. This tool aims to provide information and insights to job seekers, employers, HR professionals or even the general public about data scientist salaries. In this report the problem will be defined within a specific application domain, selecting an approved dataset and framing questions suitable for visualization. In this case, the topic is about Data Science Job Salaries.

In the field of data science job salaries, understanding the intricate patterns and influencing factors is crucial for employers and professionals alike. The complex nature of data needs sophisticated analytical tools such as Shiny. Shiny facilitates the exploration and visualization of salary distributions, trends, and correlations, offering valuable insights for strategic decision-making.

The questions the tool will be answering are:

- How do salaries vary depending on the company location or employee residence in the world?
  This question will be answered using a "Map Chart" and the variables that will vary will be Company Location and Employee Residence.

- How does salary depend on company size, experience level, remote ratio and employment type?
  This question will be answered using a "Bar Chart" and the variables that will vary will be Company Size, Experience Level, Remote Ratio and Employment Type.

- How mean salary and job opportunities have fluctuated for each type of company and experience level per year?

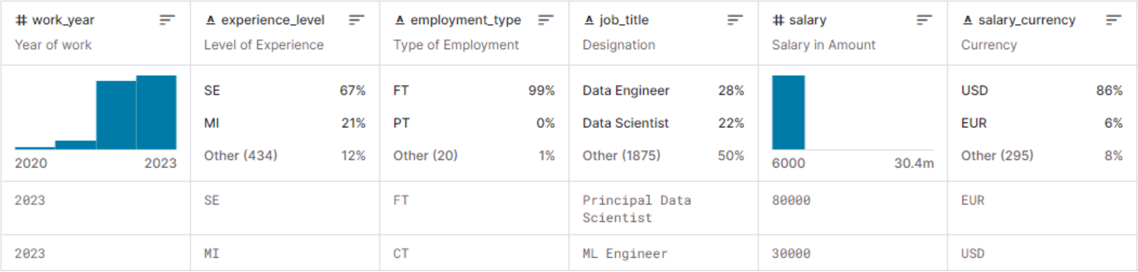
This question will be answered using a "Stream Graph" and the variables that will vary will be Job Opportunities and Salaries.

3. Data and Task Abstraction

In the following lines a data and task abstraction will be presented to provide a clear understanding of the goal of this assignment. The data science job salaries dataset chosen comes from *aijobs.net* contains 11 columns and 3755 rows. Each attribute contains information about each individual's salary:

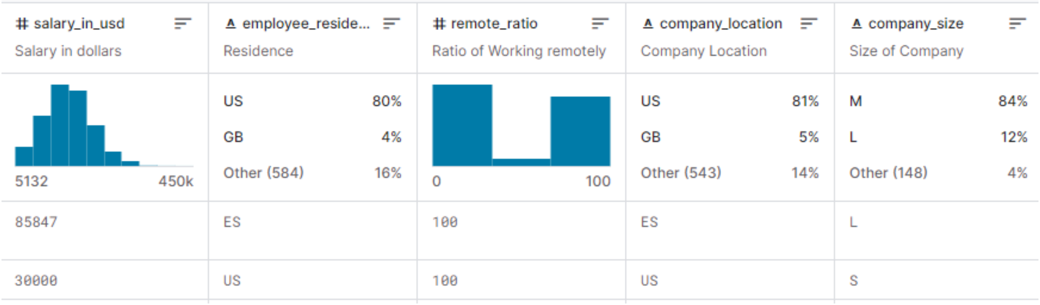
1. **work_year**: The year the salary was paid.
2. **experience_level**: The experience level in the job during the year
3. **employment_type**: The type of employment for the role
4. **job_title**: The role worked during the year.
5. **salary**: The total gross salary amount paid.
6. **salary_currency**: The currency of the salary paid as an ISO 4217 currency code.
7. **salaryinusd**: The salary in USD
8. **employee_residence**: Employee's primary country of residence during the work year as an ISO 3166 country code.
9. **remote_ratio**: The overall amount of work done remotely
10. **company_location**: The country of the employer's main office or contracting branch
11. **company_size**: The median number of people that worked for the company during the year

A small sample of the dataset can be seen on *Figure 1* and *Figure 2*.



# work_year Year of work	∆ experience_level Level of Experience	∆ employment_type Type of Employment	∆ job_title Designation	# salary Salary in Amount	∆ salary_currency Currency
2020 2023	SE 67% MI 21% Other (434) 12%	FT 99% PT 0% Other (20) 1%	Data Engineer 28% Data Scientist 22% Other (1875) 50%	6000 30.4m	USD 86% EUR 6% Other (295) 8%
2023	SE	FT	Principal Data Scientist	80000	EUR
2023	MI	CT	ML Engineer	30000	USD

Figure 1: First sample of dataset



# salary_in_usd	employee_reside...	# remote_ratio	company_location	company_size
Salary in dollars	Residence	Ratio of Working remotely	Company Location	Size of Company
5132 450k	US 80% GB 4% Other (584) 16%	0 100	US 81% GB 5% Other (543) 14%	M 84% L 12% Other (148) 4%
85847	ES	100	ES	L
30000	US	100	US	S

Figure 2: Second sample of dataset

The task abstraction involves exploring and interpreting intricate patterns within the data, using interactive tools to visualize the salary distributions, trends, and correlations. The selected questions include the variations in salaries based on global company location, employee residence, dependencies of salary on company size, experience level, remote ratio, and employment type, and the fluctuation of mean salary and job opportunities for each type of company and experience level per year.

4. Interaction and Visual Encoding

The interactive data analysis tool incorporates three main visualizations: Map Chart, Bar Chart, and Stream Graph. These visualizations enable users to explore and analyze the dataset interactively.

- Map Chart: Utilizing a custom function, the Map Chart dynamically illustrates how salaries vary across different countries based on the chosen variable (Company Location or Employee Residence). Users can interact with the Shiny app by selecting a variable and specifying a salary range, allowing for real-time exploration and analysis of global salary distributions.

- Bar Chart: The Bar Chart, implemented using the Plotly library, provides an interactive visualization of salary distributions based on categorical variables such as Company Size, Employment Type, Remote Ratio, and Experience Level. Users can select the X-axis variable from a dropdown menu, facilitating comparisons and analysis of salary variations across different categories.

- Stream Graph: The Stream Graph addresses questions related to the fluctuation of mean salary and number of job opportunities over each year and company size. Users can choose between two visualization types: "Ridge" or "Proportional." The Stream Graph is constructed using *ggplot2* and *ggstream* libraries, allowing users to dynamically explore patterns and trends in mean salary and job opportunities.

5. Implementation

5.1 Interface

Before moving on to the explanation of the backend of the code we must mention the frontend of our code, for the visualization of our website we have done it thanks to the Shiny package. On the same page we have included a title and a navigation bar that, by clicking on the three different options it contains, will show one of the three different diagrams. Each option contains the name of the diagram, the question it aims to answer and different options to select the variable that we will use to answer the question.

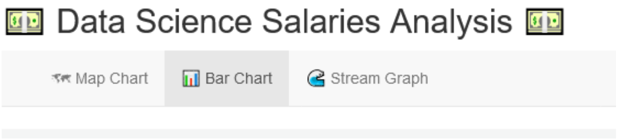


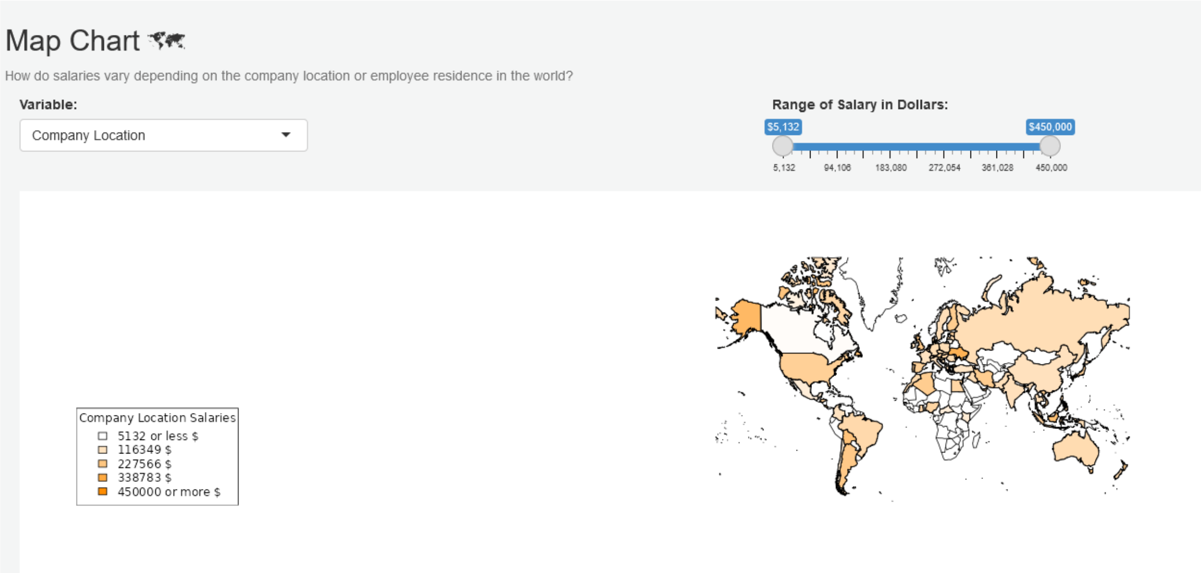
Figure 3: Navigation Bar from the app

5.2 Map

The map is implemented thanks to a custom function for creating the map. It takes in the selected countries variable and the salary range parameters. It plots the world map with specified fill colors, country regions and projection.

The different salaries in the world are generated thanks to a vector of fill colors for the map using a color ramp palette that could be green or orange depending on the variable. Furthermore, the function adjusts the salary values to fall within the specified range. So, based on the percentiles of the salary range it assigns colors, using the previously color ramp palette. Finally, It has a legend at the bottom-left of the map, indicating salary ranges with corresponding colors.

Users interact with the Shiny app by selecting a variable ("Company Location" or "Employee Residence") and specifying a salary range using the slider. So the map updates dynamically based on user inputs, illustrating how salaries vary across different countries for the chosen variable, allowing users to explore and analyze the data interactively in a visual representation of global salary distributions.



Map Chart

How do salaries vary depending on the company location or employee residence in the world?

Variable: Company Location

Range of Salary in Dollars: $5,132 – $450,000

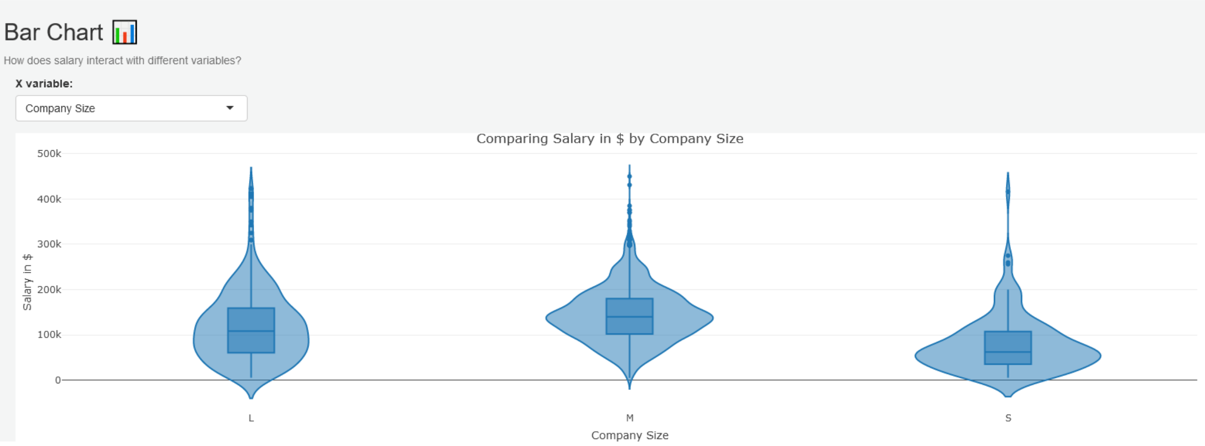
Figure 4: Map Chart from the app

5.3 Bar Chart

This R code is part of a Shiny app that generates an interactive bar chart using the Plotly library. It offers a dropdown menu for selecting the X-axis variable, which could be "Company Size," "Employment Type," "Remote Ratio," or "Experience Level."

This is a switch statement that dynamically selects the appropriate column from the dataset based on the chosen X-axis variable and the salary always depends on the Y-axis. Users interact with the Shiny app by selecting an X-axis variable (e.g., "Company Size") from the dropdown menu. The Plotly bar chart updates dynamically, displaying a violin plot that visualizes the distribution of salaries based on the selected variable.

The chart provides insights into how salaries vary across different categories of the chosen variable, facilitating comparisons and analysis.So, it is designed for exploring the relationship between salary and various categorical variables through an interactive and visually appealing bar chart.



Bar Chart 📊

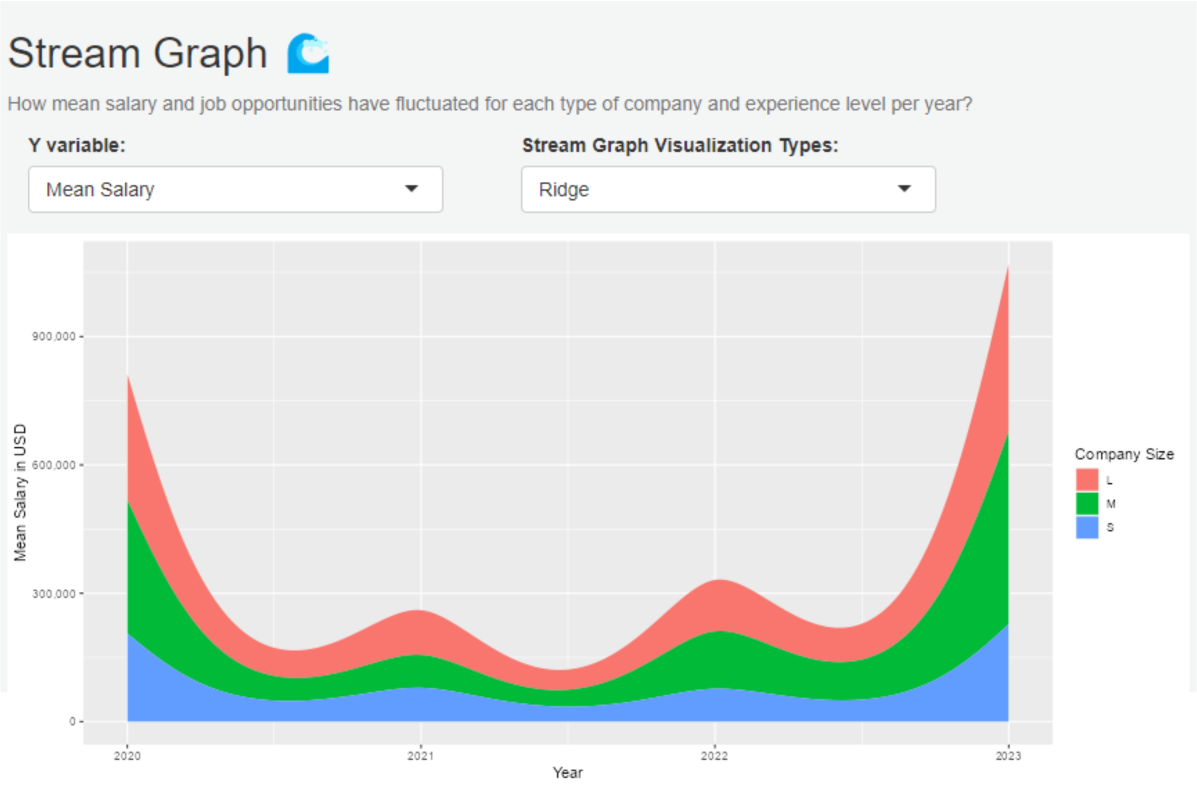
How does salary interact with different variables?

X variable: Company Size

Figure 5: Bar Chart from the app

5.4 Stream Graph

The implementation of the stream graph in the Shiny app involves conditional logic to differentiate between two scenarios: mean salary and job opportunities. For each case, the app aggregates data by grouping entries based on the work year and company size. The resulting summary statistics are then used to construct the stream graph using the ggplot2 and ggstream libraries. Users can further customize their visualization experience by choosing between two types of stream graph visualizations: "Ridge" or "Proportional." These options offer distinct ways to emphasize the patterns within the stream graph, providing flexibility for users to tailor the representation to their analytical needs.



Stream Graph

How mean salary and job opportunities have fluctuated for each type of company and experience level per year?

Y variable: Mean Salary

Stream Graph Visualization Types: Ridge

Figure 6: Stream Graph from the app

6. Instructions

6.1 Hands-on files

The zip file contains the following folders and files:

- **authors.txt:** name and id number of the authors.
- **app.R:** code of the main app.
- **Data/ds_salaries.csv:** data science salaries dataset.

- **rsconnect\shinyapps.io\datavisualization-group13\practical_work_assignment_2_-_r_visualization.dcf:** file that contains the information that connects local app to shinyapps.io.
- **report.pdf:** the report.

6.2 Run the app locally

In order to run the tool locally, please follow these instructions:

1. **Open the app.R** file in RStudio and change the working directory to where app.R is located.

2. **Install** if the working dependencies do not have the following packages:
    - mapproj
    - shiny
    - ggplot2
    - ggstream
    - tidyverse
    - plotly

3. **Click on** the icon of *RunApp* from RStudio.

6.3 Access the app via web

To access via web click on the following link: https://datavisualization-group13.shinyapps.io/practical_work_assignment_2_-_r_visualization/

7. Conclusion

In conclusion, the design and implementation of the interactive data analysis tool is really helpful for exploring and understanding complex data patterns within the field of data science job salaries. The three visualizations: Map Chart, Bar Chart, and Stream Graph, have served as an introduction to the field of data visualization and also have effectively addressed the questions regarding the data scientist salaries.

8. References

[1] Shiny. (n.d.). Welcome to Shiny. https://shiny.posit.co/r/getstarted/shiny-basics/lesson1/index.html

[2] An introduction to R. (n.d.). https://cran.r-project.org/doc/manuals/r-release/R-intro.html

[3] Data Science Salaries 2023 💸. (2023, April 13). Kaggle. https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023/data