

HR DATA APP

Data Visualization

Group 15:

- Jiayun Liu
- Yuxiao Xiong
- Jinglei Xu

Index

1. Dataset	2
2. Problem 1. How does the position, department, personal context (Gender, Hispanic Latino, US Citizen), recruitment source and engagement survey influence their income, satisfaction score, or performance score?.....	2
3. Problem 2. In which areas of company(states) the employees usually have more salary than in others?	4
4. Problem 3. How to find the best candidate for each job area? / What kind of profiles suits best for each job area?	6
5. Instructions on hands-on files and how to run the tool	9
6. Conclusions and limitations	11

1. Dataset

The dataset that we use in this project is retrieved from Kaggle¹. In real world, HR data can be hard to come by, and HR professionals generally lag behind with respect to analytics and data visualization competency. Thus, this assignment is all about visualizing and analyzing human resources to help HR professionals to know the situation of employees to see how to improve their satisfaction and working performance and to find out new ideal candidates.

In order to facilitate the future analysis, the dataset is pre-processed, for example, some data has leading and trailing white spaces which could cause problems when comparing strings. Additionally, values of some columns are case ignored. This process of data cleaning is done with OpenRefine.

The dataset obtained to do the visual analytics is a table which is structured by items(rows) and attributes(columns). The dataset contains 36 attributes, some of them are categorical and others are ordered.

2. Problem 1. How does the position, department, personal context (Gender, Hispanic Latino, US Citizen), recruitment source and engagement survey influence their income, satisfaction score, or performance score?

The initial idea to answer this question is through plots which show the distribution of the influence factor and target attribute. As there are both categorical and quantitative variables, we are going to use a boxplot or a violin plot for categorical-quantitative visualization and a mosaic plot for categorical-categorical visualization.

Data abstraction

- Attributes: Position, RecruitmentSource, Department, Sex, HispanicLatino, CitizenDesc, Salary, PerformanceScore, EmpSatisfaction
- Categorical attributes: Position, RecruitmentSource, Department, Sex, HispanicLatino, CitizenDesc, PerformanceScore, EmpSatisfaction
- Ordered attributes: Salary (Quantitative)

For this task, we selected the variables that are in the question that we wanted to answer and

¹ Human Resources Data Set, <https://www.kaggle.com/datasets/rhuebner/human-resources-data-set?resource=download>

some more that are similar and interesting like RecruitmentSource.

Task abstraction:

Action **Consume**:

- Present: The distribution of influence factors and response variables is presented to the user to identify patterns and trends to help the equality of gender, race and nationalities in future recruitment process.

Action **Search**:

- Lookup: The user knows what (variables that the user wants to analyze) and where (the violin plot or mosaic plot) to look.

Action **Query**:

- Compare: Multiple variables are compared using the violin plot or mosaic plot.

Target:

Identify trends, patterns and outliers.

Design choices:

- Encode: Boxplot and violin-plots express values with aligned vertical position, separate key(response) variables with horizontal position, show major part of values with area. Mosaic plots shows the values with area-coded subcomponents for each category of key attribute 1 and hue for key attribute 2.
- Manipulate: we use the library plotly to draw the plots so the plots are interactive where the user can change the aspect of view, select elements in the view and navigate to change the point of view.
- Reduce: the filters personal information and working situation indicators.

Visualization + Interaction

The user needs to select a personal information, which represents the explanatory variable, and a working situation indicator, which represents a response variable. In this case, the response variables are Salary, PerformanceScore, and EmpSatisfaction.

Depending on the combination of the two variables:

- Categorical-Quantitative: A violin plot is used, where the shape of the violin plot will provide information about the distribution of the response variable, which is Salary in case of Figure 1, for different levels of explanatory variables.
- Categorical-Categorical: A mosaic plot is used, which is similar to the violin plot, it shows the proportion of observation in each combination of categories by using rectangles.

Both are common to compare the distribution of multiple groups or variables. Therefore, we can provide a comprehensive view of the relationship between the variables and help the user to identify patterns or trends.

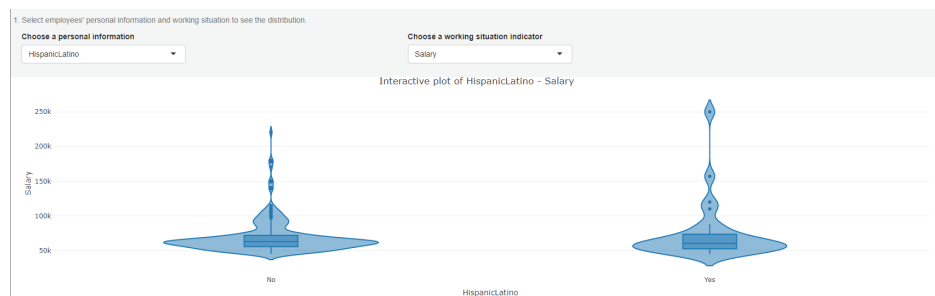


Figure 1. Violin plot HispanicLatino-Salary

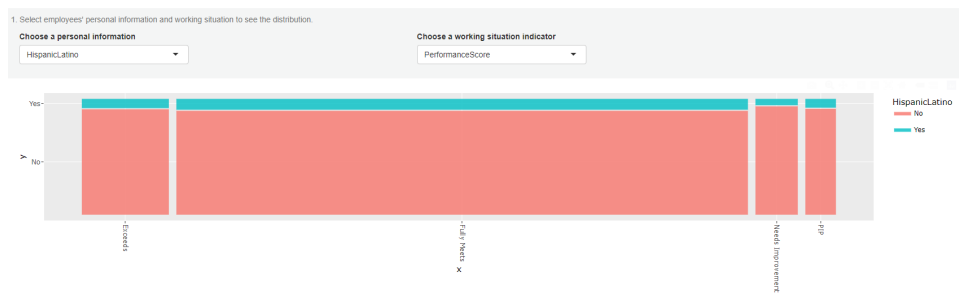


Figure 2. Mosaic plot HispanicLatino-PerformanceScore

3. Problem 2. In which areas of company(states) the employees usually have more salary than in others?

As the problem is asking about different states, a choropleth map will be used to solve the issue, on which information on the distribution of wages for different positions in different areas.

The figure below is our initial thoughts on this issue:

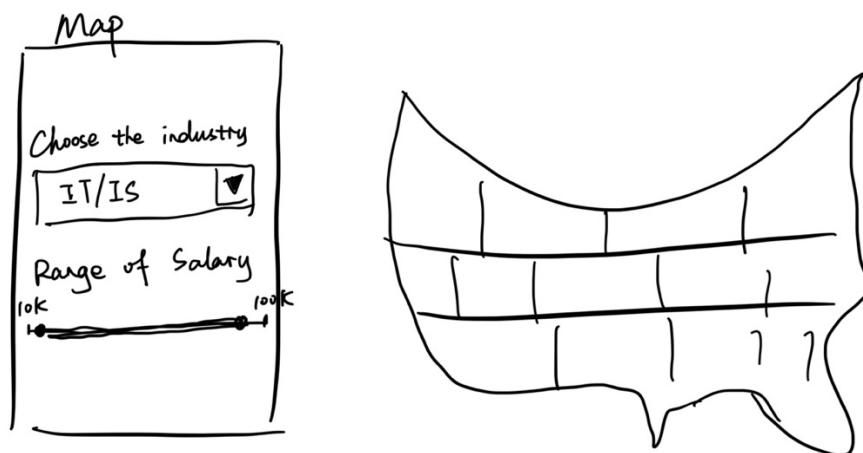


Figure 3: The initial idea of the Map

Data abstraction:

- Attributes: Salary, State, Department
- Ordered attributes: Salary (Quantitative)
- Categorical attributes: State, Department

A choropleth map is usually designed for geographic data which mark the boundaries of each area; however, the value of "State" attribute is the initials of the names of the states in United States. This problem is solved thanks to the R libraries `maps` and `mapproj` which help us to draw the map of states with states names. Additionally, to draw the map, we need to establish a relationship between salary and region. In order to more easily see the salary gap between employees in different job industries, we also choose the attribute department.

Tasks abstraction:

Action **Consume**:

- Present: a map of salary of different job industries is illustrated.
- Produce (annotate): graphical information is associated when drawing the map.

Action **Search**:

- Locate: The user tries to locate the salary in map of states.

Action **Query**:

- Summarize: The visualization summarizes the salary of different job industries in different states.

Target:

- Dependencies: Salary in different states depending on the industry.

Design choices:

- Encode: Map with area and saturation.
- Manipulate: select the range of salary to see different elements
- Reduce: the filter of job industries

Visualization + Interaction:

We are going to use a map to represent the salary in different states, the user will be able to manipulate the map by changing the job industry using a filter. A range of the amount of salary is also provided to the user to view the states with salary in that range. When the user selects a salary range, the content in the legend will also change accordingly, and its content display depends on the salary range selected by the user. In Figure 3 and Figure 4, we can see that we have selected the IT_IS department and moved the salary range, which is determined to be

between \$101814-\$250000, and as the salary value increases, the color of the map also changes from dark to light, which means IT_IS departments in those areas don't have more salary.

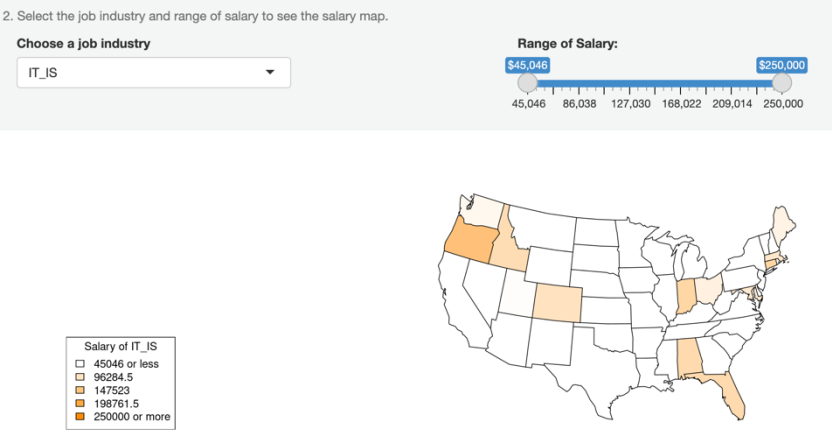


Figure 4: Solution of problem 2

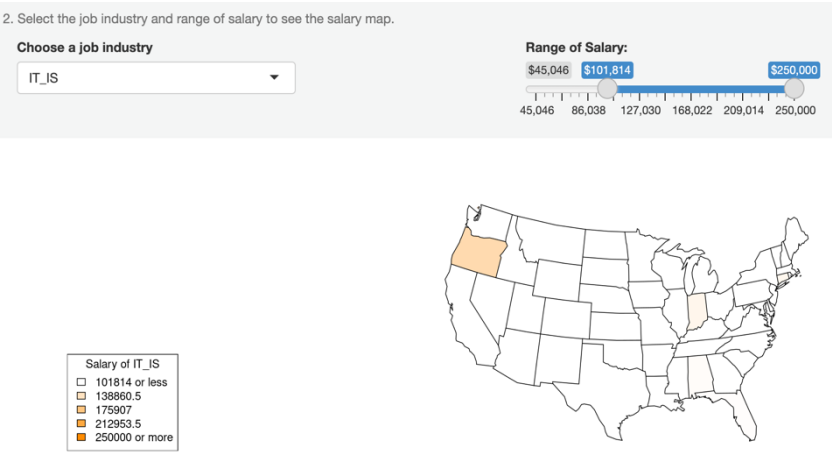


Figure 5: Different range of Salary of IT_IS

4. Problem 3. How to find the best candidate for each job area? / What kind of profiles suits best for each job area?

To solve this problem, we come up with an idea of a parallel coordinates plot where every y-axis could be the information interested by a HR expert who wants to hire new employees. Figure 6 shows an initial idea of the solution but some filters that may destroy the interactivity of the plot are removed in final application.

Choose the features to take into consideration: (Multiple)

- | | |
|--|--|
| <input type="checkbox"/> Performance Score | <input type="checkbox"/> Marital Status |
| <input type="checkbox"/> Job position | <input type="checkbox"/> University |
| <input type="checkbox"/> Salary | <input type="checkbox"/> State |
| | <input type="checkbox"/> Hispanic/Latino |

DONE

• Select the requirements of each feature and fill in the importance of each (0-100):

Performance Score:



Salary of previous job



Previous job position related? ☐ Yes ☐ No **40**

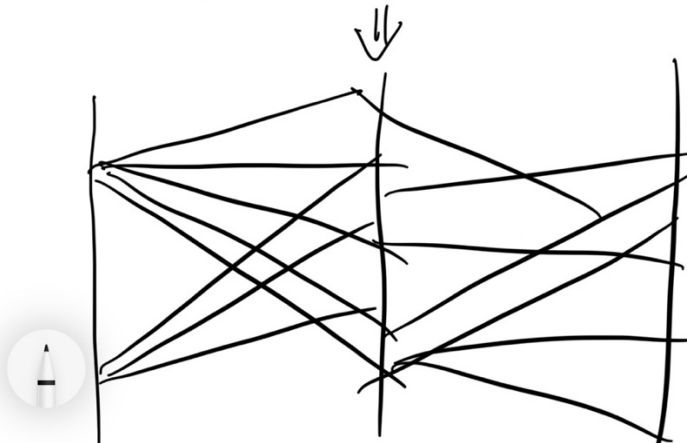


Figure 6. Initial idea of problem 3

Data abstraction

- Attributes: Department, PerformanceScore, EmploymentStatus, Sex, Salary, job_related
- Categorical: Department, PerformanceScore, EmploymentStatus, Sex, job_related
- Ordered: Salary (Quantitative)

These variables are chosen because we think that when a HR expert wants to hire someone, he/she could be interested in the situation of the previous jobs of the candidates, such as salary, job industry and working performance. The attribute job_related is a new categorical attribute generated from Department which depends on the user's choice that is explained in the design choice part.

Task abstraction:

Action **Consume**:

- Present: We will present the working performance and salary of employees and show if the previous job is related to the job area the user chooses
- Produce (Derive): a new attribute "job_related" is generated from Department.

Action **Search**:

- Browse: The user knows to look at y-axis and the colored lines of the parallel coordinates.

Action **Query**:

- Compare: Comparing the salary, performance score and previous job relations between different candidates.

Target:

Identify trends, features and correlations.

Design choices:

- Encode: Horizontal spatial position used to separate axes (PerformanceScore, Salary and job_related), vertical spatial position to express values along each aligned axis with connection line marks as segment between them.
- Manipulate: select the range of salary to see different elements.
- Reduce: the filter of job areas, the value of this filter affects to the categorical job_related.

For example, if the user selects Production of the filter, the value of job_related of all employees with working experience in Production Department will be "Yes" and the value will be "No" for employees with working experience in other departments.

Visualization + Interaction

The user needs to choose the range of salary, if he/she minds that the candidate was fired by personal reasons such as poor attendance and bad working performance and the area of job title. The parallel coordinates plot will show the performance score, salary and previous job related of those employees that satisfy the conditions. The color of the lines of the plot shows if the employee is female or male (Figure 7). The visualization itself is also interactive in the way that the user can choose the range of y-axis. For example, if the user only wants employees whose performance score is "Exceeds", he/she can draw a line through the "Exceeds" part of the "Performance Score" axis, so that the color of the lines that do not connect to this part will change to grey (Figure 8).



Figure 7. Solution of problem 3

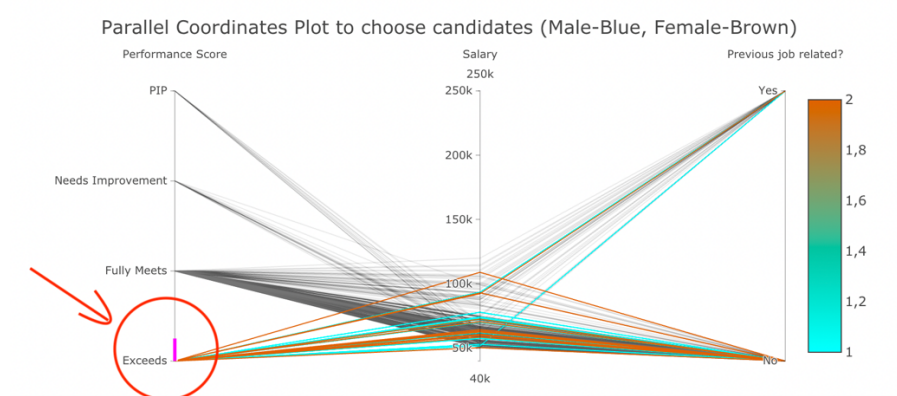


Figure 8. Interactive part of the parallel coordinates plot

5. Instructions on hands-on files and how to run the tool

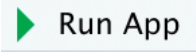
Hands-on files

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
- helpers.R: auxiliary R file to help to draw the map
- Data/data.csv: dataset to use
- rsconnect/shinyapps.io/estrellaxyx/HumanResourcesDataApp.dcf: file that contains the information that connects local app to shinyapps.io
- report.pdf: this report

Run the app locally

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

- Open the app.R file in RStudio and change the working directory to where app.R is located.
- Install if the working dependencies do not have the following packages: shiny, maps, mapproj, ggmosaic, ggplot2 and plotly.
- Introduce `runApp('app.R')` in the R console or click on the  icon of the window.

Instructions of how to use the app

The app contains three visualizations that can be used to solve three problems described in the report.

1. To interact with the first visualization:
 - The first filter is about some personal information which contains Department, if the employee is a Hispanic Latino(HispanicLatino) and a US-citizen(CitizenDesc), the marital status(MaritalDesc), race(RaceDesc), recruitment source(RecruitmentSource) and gender(Sex) of the employee.
 - The second filter is about the working situation indicators, which include the salary(Salary), performance score of previous job(PerformanceScore) and satisfaction level of job company(EmpSatisfaction) of the employee.
 - To see the exact value of somewhere in the plot, place the cursor nearby. There is also a navbar of plotly on top-right corner to change axes and zoom in or out.
2. To interact with the second visualization:
 - The first filter is about job industry, if the employee belongs to Production, IT/IS, Software Engineering, Sales, Admin offices or executive offices.
 - The second filter is to choose the range of salary for situations like, the user wants to know the map of Production industry but only those with salary below \$80K.
3. To interact with the third visualization:
 - The first filter is about job area, if the employee has experience with Production, IT/IS, Software Engineering, Sales, Admin offices or executive offices.
 - The second filter is to choose the range of salary expected to offer to the employees.
 - The third filter is to ask the user(HR expert) if he/she minds that the employee was fired due to reasons like poor attendance, bad performance on work, not showing up, not answering calls, etc.
 - The vertical axes of the visualization is interactive, in a way that the user can choose the range of values. For example, if the user only wants employees whose performance score is "Exceeds", he/she can draw a line through the "Exceeds" part of the "Performance Score" axis, so that the color of the lines that do not connect to this part will change to grey.

Web app

The data visualization application is also published in ShinyApps. The link to access it is <https://estrellaxyx.shinyapps.io/HumanResourcesDataApp/>. Figure 9 also shows a screenshot of the entire application.

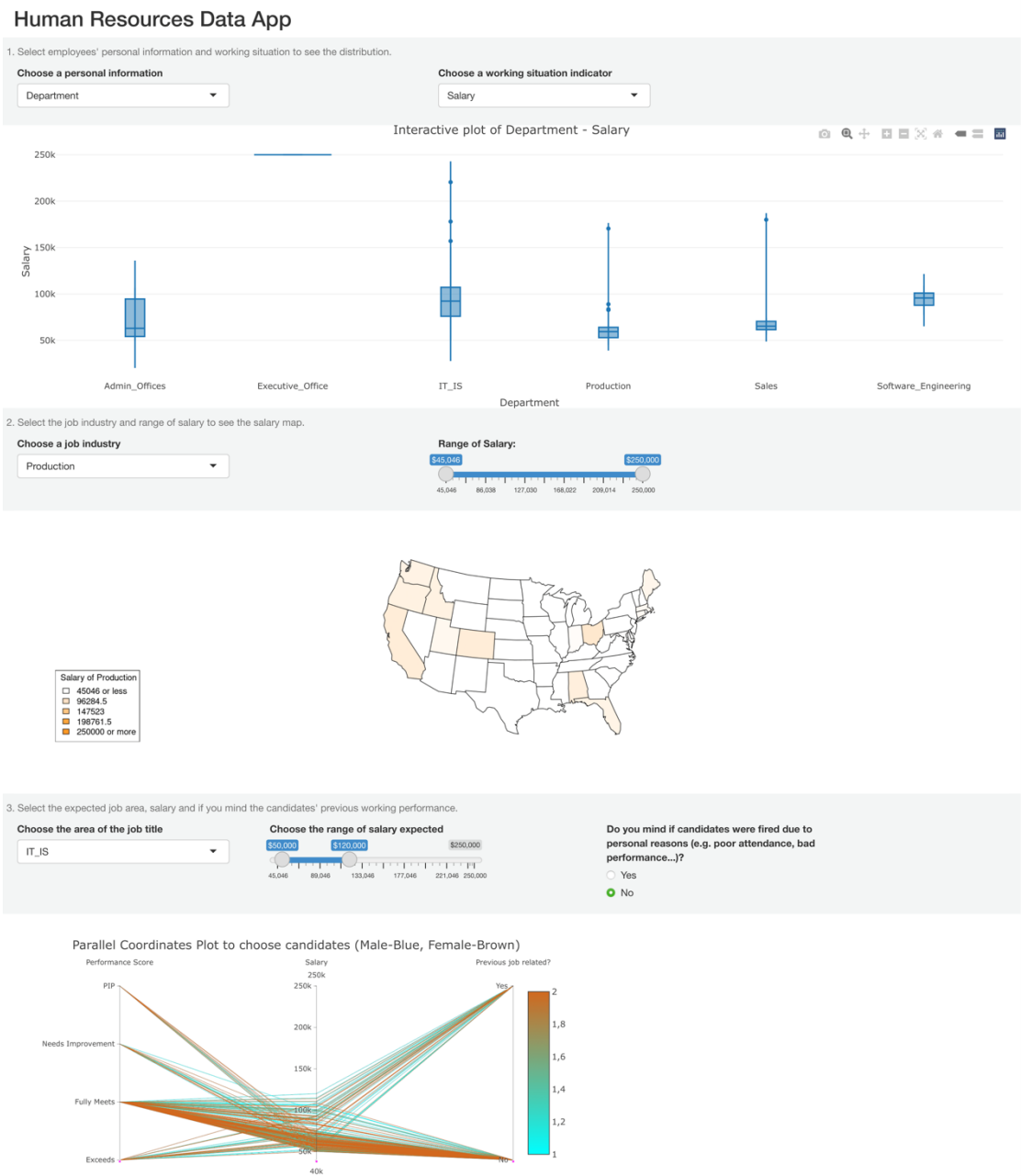


Figure 9. Human Resources Data App

6. Conclusions and limitations

In conclusion, the three problems listed are solved with distinct interactive visualizations illustrated in a Shiny App. These three visualizations which show the distributions of

information that indicate the situation of employees can help HR experts to make better decisions or plans to hire new employees. However, there are some limitations and improvements. On the one hand, the visualizations are not completely interactive. For example, we may add interactions on the map so that when the user places the cursor on some area, the visualization will show the details of salary of the state. On the other hand, the dataset only contains about 300 observations from a few numbers of job industries and states. To get more general results, we may need more data sources and employ some data science techniques (such as clustering, classification, etc.) to enrich our visualizations. These improvements will remain as a future work for the application.