Graduate School Dataset Visualization

Aishwarya Shastry, Nivedita Prasad, Riya Maan

December 22, 2017

1 Introduction

Data visualization is used to leverage the concept of analysis of data for better understanding. Visualizing charts and graphs along with interaction helps data analyst/user to understand the data better and make informed decisions. Information Visualization is the area where interactive projections of the visualized charts annotate human thought process to arrive at intelligent decisions. We have chosen Graduate School dataset that has information about the salaries that the graduates receive after graduating. The dataset considers salaries in accordance with different fields of major, regions and types of colleges. We have developed a website to visualize the same.

In this information visualization report, we have explained about different visualization techniques used. In Section 2, we have discussed about our Graduate School dataset and how we have preprocessed it for the analysis. In Section 3, we have explained in detail about the visualizations techniques and the justification of why we chose them. This section also includes the details about dashboard and the user interaction with the charts and map. Next, section 4 puts forward the analysis done and trends observed on the dataset through our dashboard. Lastly, section 5 states limitations of our information visualization project.

2 Grad School Dataset

The dataset chosen and used for our information visualization project is from Kaggle website [Kag17]. This dataset was first obtained by kaggle from the Wall Street Journal based on data from Payscale, Inc. We have further cleansed the data set using python script. The dataset is distributed over three comma separated(.csv) files with 320 entries in "salaries by region" containing the school name and region, 269 entries in "salaries by college type" containing school name and school type and 50 entries in "degrees that pay back" containing undergraduate majors and their salary estimates. All three files are accompanied by corresponding starting average salary and average, tenth, twenty-fifth, seventy-fifth and ninetieth percentile of mid career salary. Each of the visualization that we have created uses one of the three files.The types of schools considered are State, Liberal arts, Engineering, Party and Ivy League.

We had to preprocess the dataset before using it. First of all, we removed "$" symbol from all columns containing salaries; because we had to use ascending/ descending order of salaries for axes of plots, so we needed salaries in numeric format. Also, there were some NaN entries in dataset, we had to set them to zero to prevent error in reading file.

The average salaries earned at the beginning of the student's career after graduation showed a noticeable increase over time. The salaries varied based on the region where the colleges are situated in USA. Colleges in some regions have a substantial higher salary than those in others. Also, salary trends vary according to different college types and degree being pursued by student. Students pay hefty prices to attend colleges in USA. Hence, it is crucial to check salary trends of students upon graduation, to get an idea if it is worth to invest in that school or should the

Salaries by College Type	
School Name	Massachusetts Institute of Technology (MIT)
School Type	Engineering
Starting Median Salary	72000
Mid-Career Median Salary	126000
Mid-Career 10th Percentile Salary	768000
Mid-Career 25th Percentile Salary	99200
Mid-Career 75th Percentile Salary	168000
Mid-Career 90th Percentile Salary	220000

Degrees that Pay Back	
Undergraduate Major	Accounting
Starting Median Salary	46000
Mid-Career Median Salary	77100
Percent change from Starting to Mi	67.6
Mid-Career 10th Percentile Salary	42200
Mid-Career 25th Percentile Salary	56100
Mid-Career 75th Percentile Salary	108000
Mid-Career 90th Percentile Salary	152000

Salaries by Region	
School Name	Stanford University
Region	California
Starting Median Salary	70400
Mid-Career Median Salary	129000
Mid-Career 10th Percentile Salary	68400
Mid-Career 25th Percentile Salary	93100
Mid-Career 75th Percentile Salary	184000
Mid-Career 90th Percentile Salary	257000

Figure 1: Examples from dataset

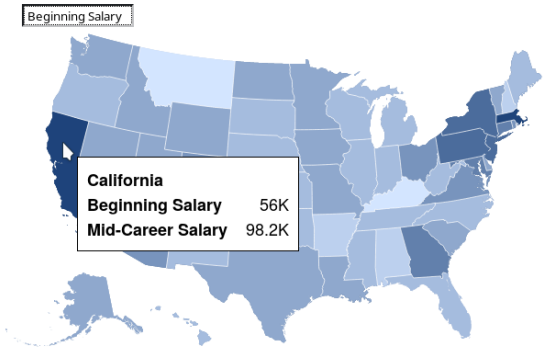

Figure 2: Choropleth map depicting state-wise salary distribution

student consider other school. We want to analyze the salary trends across school types, regions in the USA, and the degrees that really pays back after graduation.

No additional processing of data was required for horizontal box plots, parallel coordinates and heat map. While for choropleth map, state-wise data was not available, hence state was added to the table containing "salaries by region" and then it was plotted on map. This enabled us to show state wise trends for salaries.

3 Visualization Tasks

We have presented four visualization tasks through our dashboard we have designed. The tasks are categorized into analysis task and presentation task. Our presentation task shows distribution of salaries across different states in United States as "Starting Career Salary", "Mid Career Salary" and "End Career Salary". This allows users to look for states with lowest and highest salaries and give an idea about the Economic distribution within the country.

There are three analysis tasks. First of all, we wanted to analyze the the salary distribution by different fields of study. We have achieved this by using box and whiskers plot. This plot gives an insight of the salary distribution by the area of study during undergraduate degree. Through this we want to gain an understanding that which undergraduate major results in a high mid career salary. Further, we tried to analyze salary trends by college type using parallel coordinates plot. Or we can say that we wanted to know which college type results in better salary after graduation. This is helpful as we can decide which college type is best to invest in, so as to get maximum return (salary). Lastly, we wanted to analyze that trends between starting and mid-career salary of college graduates. We have used different colors to plot different regions of colleges. As a hypothesis, we state that students from colleges lying near Silicon Valley have highest starting and mid-career salary averages upon graduation.

We believe that these visualizations would be really helpful for prospective students to decide upon their college and form an idea of what to salaries to expect post graduation. They can form a better decision regarding, which college is worthy enough to invest in. No one wants to pay exorbitantly high amount of tuition fees for the college and land up in low paying job. These visualizations can help students in forming better decision.

4 Design Choices

4.1 Choropleth Map

Our presentation task was to show the distribution of economy across the different states in United States. We achieved this by plotting the Salary distributions on the Choropleth Map. A Choropleth Map is a method for graphically plotting areas in which the areas are shaded or colored in proportion to the measurement of statistical variable displayed on the map.

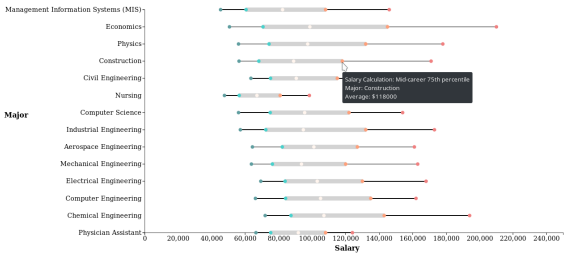

Figure 3: Horizontal box and whiskers plot for different fields of study

4.1.1 Design Decisions

Choropleth Maps are an efficient representation of spatial data. Javascripts like GeoJSON provide simple geographical features. TopoJSON, an extension of GeoJSON encodes topology. Our visualization task included showing salary distribution across different states in US. We used the map to show Beginning and Mid-Career Salary across all the states. We calculated the average salaries across regions and took it as our dataset. We obtained data for 52 states and for each state average beginning salary and 50th percentile of mid-career salary was recorded.

We have used the color coding technique such that the state with maximum Beginning and Mid Career Salary is shaded with the darkest shade of blue and the lowest Salary with the lightest shade.

4.1.2 User Interaction

We have developed the Choropleth map in such a way that it allows users to interact easily and thus give them a better understanding of the data. On hovering mouse over the region, information about the plot is shown in the tooltip as shown in the figure 2.

4.2 Box and Whiskers Plot

Box plot is a method for graphically depicting groups of numerical data through percentile range. Lines extending from box (whiskers) indicate variability outside lower and upper percentile. Median is depicted by a line inside the box.

4.2.1 Design Decisions

One of our visualization task included showing salary trends for different fields of study. Since box and whiskers plot provides an efficient way of depicting data distribution based on percentile ranges, we decided to use it to depict salary trends in mid-career of college graduates. We had data of 50 different fields of study. For each field, 10th, 25th, 50th, 75th and 90th percentile of mid-career salary was recorded. This plot provides the best medium to compare these distributions.

We have used horizontal box and whiskers plot. It uses different fields of study on y-axis and salary on x-axis. The scale on y-axis depicts 50 different fields of study, while the scale on x-axis depicts salary range from $0 to $250,000. All the salaries lie within this range. The box is colored in gray color, while whiskers are colored in black. For plotting values of percentiles, we have superimposed circles on box and whiskers. These circles have different colors corresponding to different percentile value. For lower ranges, shades of blue are used; while for higher ranges shades of red are used. Darker shades represent extremes, while lighter shades represent intermediate values. We prefer higher salaries, hence it is colored in red, because it's closer to us, while lower salaries are away , hence shown in blue. To represent median, we have used "sea shell" shade. Figure 2 shows the salary trends for different fields of major.

4.2.2 User Interaction

The horizontal box and whiskers plot developed by us allows user to interact with it so as to help him/her get better understanding of data. On hovering mouse over the circles and box and whiskers, information about the plot is shown in tooltip. Over box and whiskers, name of the corresponding major is displayed. While on circles, the corresponding percentile, major and salary

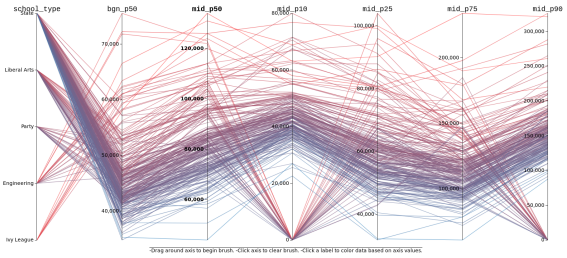

Figure 4: Parallel coordinates plot showing salary trends by college-type

is shown. This can be seen in figure 3. Tooltip background is chosen such that it is clearly distinguishable from surrounding and readable at same time.

4.3 Parallel Coordinates Plot

Parallel coordinate plots help to represent data in higher dimensions as 2D by adding parallel axes and plotting a polyline.The reordering of axes in parallel coordinates plot is important for data analysis.

4.3.1 Design Decisions

We had to analyze salary trends for different college type. For that, we had beginning career salary and mid-career salary for different percentiles. For each college, we had 5 columns each. So, to visualize them in single chart, parallel coordinates is one good option. The first axis represents school type and the rest of them have salaries. Initially, all lines are ordered according to third column, colored red for higher salaries and blue for lower salaries[BI07] after much research , as seen in figure 4. The ordinal scale "School Type" displays five distinct types of schools selected for analysis.

4.3.2 User Interaction

We are leveraging the power of visualization for analysis during user interaction by the use of interactions. When the user is clicks on an axis, the highest salary for that selected column or axis will be displayed in red and gradually fading into blue for decreasing salary. This can be done for all seven axes displayed in the parallel coordinate plot. The user can select the mid or starting career axes which they want to view and lines will get colored according to that. Additionally, the brush mode has been implemented to specifically view a specific school for over the complete career lifespan chosen for analysis. This brush mode can be used to highlight salary range for all the axes for specific scrutiny of the salary trends after graduation of the college graduates from a particular type of school.

For much clear understanding and interaction, instructions such as "Drag around axis to begin brush","Click axis to clear brush","Click label to color data based on axis value" has been stated below the parallel coordinates for guidance.

4.4 Scatter Plot

A scatter plot is a mathematical plot using cartesian coordinates to display values for two variables of data. We have used the Scatter Plot to Plot Beginning and Mid Career Salaries trends for different schools.

4.4.1 Design Decisions

The most efficient way to represent a relationship between two variables is Scatter Plot. Thus, we decided to analyze the trends between Beginning and Mid Career Salary in different regions of United States by Scatter Plot. We used different colors to represent different regions.

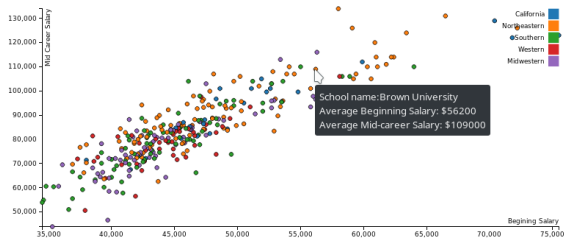

Figure 5: Scatter plot for beginning career salary vs mid-career salary

4.4.2 User Interaction

We have provided a tooltip such that when the mouse hovers over the bubbles, it shows the information about the School name, Beginning and Mid Career Salary. This can be seen in the figure 5.

5 Analysis

5.1 Chloropleth Map

The choropleth map shows the distribution of salary in the United States, Thus giving us an insight into the richest states and poorest states in terms of Income of one's career. Looking at this distribution one can decide, where they want to work at the beginning of their career and at the mid of the career. They can decide if they want to move to a different state to get better salaries and see which state provides the most increase in salaries and thus decide to move in there.

5.2 Horizontal Box-plots

On plotting horizontal boxplot for different fields of study, as shown in figure 3, we can clearly see that economics is the highest paying degree followed by finance and chemical engineering. Also, least salary for civil engineering is more than other fields, but highest salary is not that high. Salary distribution is over a small range. While for economics, salary range is wide. So, we can not get an estimate of salary after graduation. For some fields like, nursing and nutrition, salary range is quite small; hence, we can easily estimate mid-career salary.

5.3 Parallel Coordinates Plot

When a student wants to decide on school type for their career, they can drag and select one or more school types on the plotline to get an idea of how much salary to expect after graduating. For example, the Ivy League school type plot in figure 6 lines depict: the starting average salary is high, around sixty to seventy thousand dollars, and then it rises to around hundred and thirty thousand dollars in their mid career range. From this we can conclude that Ivy League universities can be a better choice to study and experience booming career life for those students who are aiming at pursuing technical career. The engineering and liberal arts school type shows a stagnancy in the salary after some time in career life.The party school graduates earn less in the beginning of the career but shows a significant increase in the salary of the graduates which is almost equal to ninetieth percentile salary of the engineering or Ivy school students.The state school type start at really low salaries but the salaries escalate swiftly and become stagnant at $100-200k. But the Ivy League colleges' starting salary is high, as expected, when compared with other schools' starting salary. Hence, we can say that parallel coordinate plot shows a relation between salary distribution and different school types.

5.4 Scatter Plot

From the Scatter Plot, we can analyse that colleges in the Bay area start with the highest salary and have a significant increase over the years. While, The Northeastern region starts with a lower salary but has higher increase than most of the colleges in Bay area over the years. We can infer

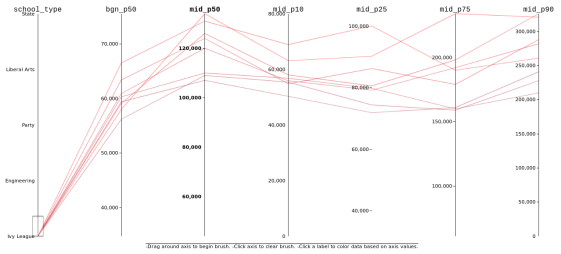

Figure 6: Parallel coordinates plot showing salary trends for Ivy League colleges

from the plot that the Western region has the lowest increase in Salary trends over Beginning and Mid career Salary. Most of the salaries are concentrated over the Values (40K,80K) and (50K,100k) in the plot, giving us an insight into the average salaries in the United States.

5.5 Limitations and Challenges

For data that were missing we filled the NaN values with zero, which lead to the plotting of salaries of zero in the parallel coordinate plots.Even though the existence of zero salaries was not possible, this could not be avoided.This was one of the limitations of the dataset. Also, we were unable to show region wise salary variation on US map, using choropleth map.

References

[BI07] David Borland and Russell M. Taylor Ii. Rainbow color map (still) considered harmful. *IEEE*, 2007.

[Kag17] Kaggle.com. Exploration of college salaries by major | kaggle. *Robotics and Autonomous Systems*, 2017.

A Individual Reports

A.1 Aishwarya Shastry

Our visualization task was to provide a presentation task and do three analysis task. We were supposed to show the salary distribution by region and analyze the Beginning and Mid Career Salary trends by region,school type and undergraduate major. My task was to show the salary distribution by us states. I achieved this by using javascript libraries Topojson which has topological features to plot the map. I color coded the states such that the highest salary has the darkest shade of blue and lowest being the lightest. I added a drop-down menu which changes the plotting of the choropleth by Beginning Salary or by Mid-Career Salary. This plot shows the economic distribution across the different states. The second task we were trying to achieve to show the infer a relationship between the Beginning Salary and Mid-Career Salary. For this, I plotted a Scatter Plot which had Beginning Salary trends on the X axis and Mid-Career Salary trends on the Y axis. I color coded the circles such that Schools from the same region are plotted with the same color. We can infer from the Scatter plot that Northeastern and California has the highest salaries both at the beginning and after years of working. Most of the salaries are concentrated over the Values (40K,80K) and (50K,100k) in the plot, giving us an insight into the average salaries in the United States.As a group we tried to use our combined ideas and efforts to this project and learned new javascript libraries. Infoviz project taught us team work,cooperation and how combined ideas can be put together for a better cause.

A.2 Nivedita Prasad

Our Infoviz project had the preprocessing as the first stage and my task was to clean all the 3 datasets obtained from Kaggle into data which contained numeric column values with $ symbols, commas and NaN values removed. I achieved this using python script with pandas library. My second task was to show the trends in salary versus School type for the graduate school types .I achieved this by using d3.js and parallel coordinate libraries .I have additionally added the user instructions for guidelines. The brush feature is provided , the axes can be dragged and moved beside any of the desired axes for clear understanding of the trends along with ordinal lines containing names of the school types. I have contributed for the introduction and dataset, parallel coordinate plot's corresponding design choices, analysis,user interaction. I contributed to the video by screen recording the whole video process and the voice recording was contributed by my team mate Riya and ideas by Aishwarya. I really enjoyed this Infoviz Project and we as a team learnt a lot from this project.

A.3 Riya Maan

We chose to analyze graduate school dataset using one presentation and three analysis tasks. For these, we used choropleth map, horizontal box and whiskers plot, parallel coordinates plot and scatter plot. I worked on the code of horizontal box and whiskers plot, to depict salary distribution for different fields of study. I used different colors to present lower and higher percentile of salaries. This plot gave me a better understanding of how salaries vary for different major. For some major, salary varies over a smaller range, while others have huge variations in salary. I think that this visualization would prove to be very helpful for prospective students to decide their college. Working in group with Aishwarya and Nivedita was a great experience. We all worked together to complete this project on time. This project gave me an opportunity to learn d3.js. This tool made developing visualization easy. We can use this tool in future to analyze data. Overall, it was a learning experience.