Homework 3: Data Visualization

Shashwat Shashi Mehta (827008698)

Discussions and Challenges

The question I ask in this exercise is - **Is college worth it?** I devise a simple metric to approximate the need for a college degree within some major and then ask questions about how having a college degree affects one's employment status, and also their salary.

The hypotheses I had before starting was that, with a college degree, one should be able to find a good job. This would mean higher salaries and lower unemployment rates. After the analysis, my hypotheses turned out to be true, although not by any significant measure.

Before settling on this question, I put a lot of thought into what question I can ask of the limited data we have. I initially tried to investigate the gender pay gap between different majors. However, due to the lack of granularity in the data provided, I couldn't think of any way around it.

The limited data granularity was definitely a challenge. We know that the 1000 people, 400 of whom are women, are paid on average 80K, but we don't know the same for just the women. If instead, we had individual survey responses with us rather than these aggregates, I could picture several different insightful investigations.

Tools Used

1. Bokeh
2. Pandas
3. Photoshop
4. Canva

Design Rationale

The charts I've drawn are pretty easy on the eyes. They use bright, meaningful colors, and show only the relevant information to the user.

1. **College Ratio by Major Category Bar Graph**

   The purpose of this graph is to validate my hypothesis. After plotting the graph, we can visually inspect what majors need degrees mostly, and which ones don't. Then we can justify the answers using real-world examples.

   Regarding the design, the graph uses **colors** to separate different major categories. Please note that the **same colors** are used in the subsequent graphs as well. The x-axis **labels** the majors by name; and the y-axis just as our ratio. I also ordered the bars in **descending order** to make comparisons easier. The graph also has a **grid** behind it to easily lookup the actual values from the y-axis.

2. **Degree Vs Salary**

   I compare the correlation between our metric and the median salary of the major. The **colors** of the scatter points indicate the category of the major. I've also added relevant labels beside the dots themselves in the same color. I've used the design principals of **association by color** and **association by proximity** to make sure the labels are easily identified. We observe interesting relations here, which I've discussed in the poster itself. Finally, on the top right corner, we can see a helpful **correlation score**.

3. **Degree Vs Unemployment**

   This is a very simple scatter plot. As before, the colors of the scatter points show the major categories. We also have the correlation score, the same as before. The **legend** is to the bottom right where it doesn't block many points from view. Lastly, the **faint background grid** makes it easy for the viewer to orient their scales so they can compare and contrast easily.

Pros and Cons of Design

Pros:

1. **Easy on your eyes** -- bright, beautiful colors, proper alignments and paddings.
2. **Easy to find relevant information** -- legend, labels, scores, etc, are placed wherever it makes the most UX sense.
3. **Easy to follow** -- Tells a story properly in a linear fashion. Everything is to the point and contributes to the overall story.
4. **Separation of concerns** -- Like paragraphs, disjoint events are separated.

5. **Attention to detail** -- Every subtle color, background, border, whitespace, has been a deliberate decision.

Cons:

1. **Some of the texts may be too small** for some people -- I've tried my best to make everything legible. However, you may have to zoom and pan sometimes.
2. **Non-interactive format** -- Bokeh creates everything as an interactive web component. Select, highlight and tooltip hovers are really useful features that are obviously missing in this 2D rendering.
3. **Bokeh isn't the most pretty** -- We were constrained to using python libraries. As such, bokeh was my choice. However, javascript has much prettier libraries for data visualization.
4. **Doesn't show ALL labels** -- It's neither possible, nor a good idea to show labels for all scatter points. However, the viewer may find interest in some particular point on the graph. This, of course, wouldn't be possible with this non-interactive format.
5. **Not optimized for color-blindness** -- I could use symbols instead of colors to solve this, but that wouldn't be very pretty. However, if I was implementing this on a website, I would just put a toggle to turn the color blindness mode on, which would change the scatter points accordingly.