College Majors, Earnings, and Employability: A Visual Investigation

David Kwan
dkwan33@yorku.ca
York University
Toronto, Ontario

ABSTRACT

I present a novel series of visualizations created in Tableau for the purpose of investigating the earnings, employability, and gender wage gap between nearly all possible college majors and fields. Core principles of visualization design are adhered to, and question-targeted interactive dashboards give the user the flexibility to control and investigate whatever relationships they desire.

1 INTRODUCTION

For many students, perhaps almost every student, a major concern after graduation is employment. A college degree is by no means a guarantee of employment, nor is it a guarantee of economic safety, but it most certainly can help. The choice of major is a one of earliest decisions in a student's career, it can open many doors and offer many opportunities, but to many, the "wrong" choice may seem like a waste of time. After all, an unemployable or "useless" degree still takes just as much time as other degrees. There are a plethora of questions that prospective students, current students, or even recent graduates could be interested in:

- What kind of jobs are there?
- What subjects are more popular?
- What are students studying?
- What do students specialize in?
- Who makes more money?
- Who is more employable?
- Does the degree even matter?

These are all valuable questions and are only a fraction of the possible questions regarding choice of college major. To that end, this investigation aims to answer all of the above and more using data visualization techniques built in the Tableau software platform.

, ,

2 DATA SET

In 2014, Ben Casselman wrote an article entitled *The Economic Guide To Picking A College Major* published on the political and economic blog *FiveThirtyEight* associated with ABC News [1]. This article used a full data set from the 2014 United States Census courtesy of the United States Census Bureau Public-Use Microdata. However, using this data set, Casselman only "ranked" the college majors in terms of earnings. Furthermore, Casselman did not create any data visualization. This is unfortunate as the data set itself holds much more potential. *FiveThirtyEight* has made this data set public on Github (although the source was public to begin with, it is difficult to work with due to sheer breadth), along with the R script used to pull all the relevant data from the Census Bureau [2].

The data itself is comprised of a few tables. The two most detailed contain information pertaining to recent graduates, being those of age 28 or less, as well as data on all ages. Each row corresponds to a college major (e.g. Biomedical Engineering), and the first and foremost data column contains the broader major category (e.g. Engineering). Both of these are categorical data fields. The rest of the columns, of which there are many, are all quantitative. These data fields include the total number of people with that major, the total employed, the total unemployed, median salary of that major, gender, and much more. Data transformation was performed when required.

3 VISUALIZATION OVERVIEW

Using this data set, I have created a series of charts and dashboards. My focus is on recent graduates, so much of the data is from the recent graduates table, although data from the all-ages table is used as well whenever the comparison is worthwhile. Details on this will follow in section 4.

The charts created include bar charts, box plots, scatterplots, and a map chart. Both the expressiveness and effectiveness principles of visualization design [3] were adhered to as best as possible. Other available options such as packed-bubble charts were not chosen because it provided no clear advantage, even when placed on a dashboard with other types of charts. The map chart is a good example of a chart that, although does not strictly follow the expressiveness

Figure 1: Dashboard 1: Total Counts for each Major Category for both recent grads and all ages

principle, had value especially when placed on a dashboard with interactivity.

Multiple dashboards were created with the goal of answering one or more specific questions that the user may have. All these dashboards provide value with interactive features, allowing the user to choose what they want to focus on. The interactivity cannot be easily shown on paper, nor is it particularly valuable to show screenshots of individual filters, but both filtering and highlighting can be done based on either major category or major specialization in all dashboards. Any filtering or highlighting done on one chart of the dashboard updates all other charts on the dashboard. This was immensely important for this data set considering that there were 172 different majors to choose from. It is reasonable to expect that most users would only be interested in a few, or even one major specialization, and the corresponding major field, and thus would want to highlight or filter for the major that they wanted to investigate.

Furthermore, the dashboards provide juxtaposition between related graphs for comparison. One of the most prominent benefits of juxtaposition in this visualization is to show a comparison between recent graduates and all ages along the same metrics. Juxtaposition also allows the user to more easily compare between certain majors on multiple metrics.

To avoid confusion and to provide clarification, "major" will be commonly referred to as "specialization", while "major category" will be commonly referred to as "field".

4 CHARTS, DASHBOARDS, QUESTIONS, AND ANSWERS

In this section I will discuss each Dashboard one at a time. The user is meant to traverse through each if they wish to follow a logical progression or story. Of course, the user can also view dashboards in whatever order they wish or pick and choose to investigate whatever question(s) they may have. The motivation and target for each dashboard will be discussed, as well as the visual encoding, any major drawbacks, and any interesting observations.

Dashboard 1: Fields and their Popularity

Dashboard 1 shown in Fig. 1 introduces the user to the data set, as it only shows the major categories and not the major specializations. Following the expressiveness and effectiveness principles, the bar charts here represent the quantitative variable (Total) as height, while the categorical variable (Major category) is represented with colour hue. It is important to note here that the colours used for the major categories

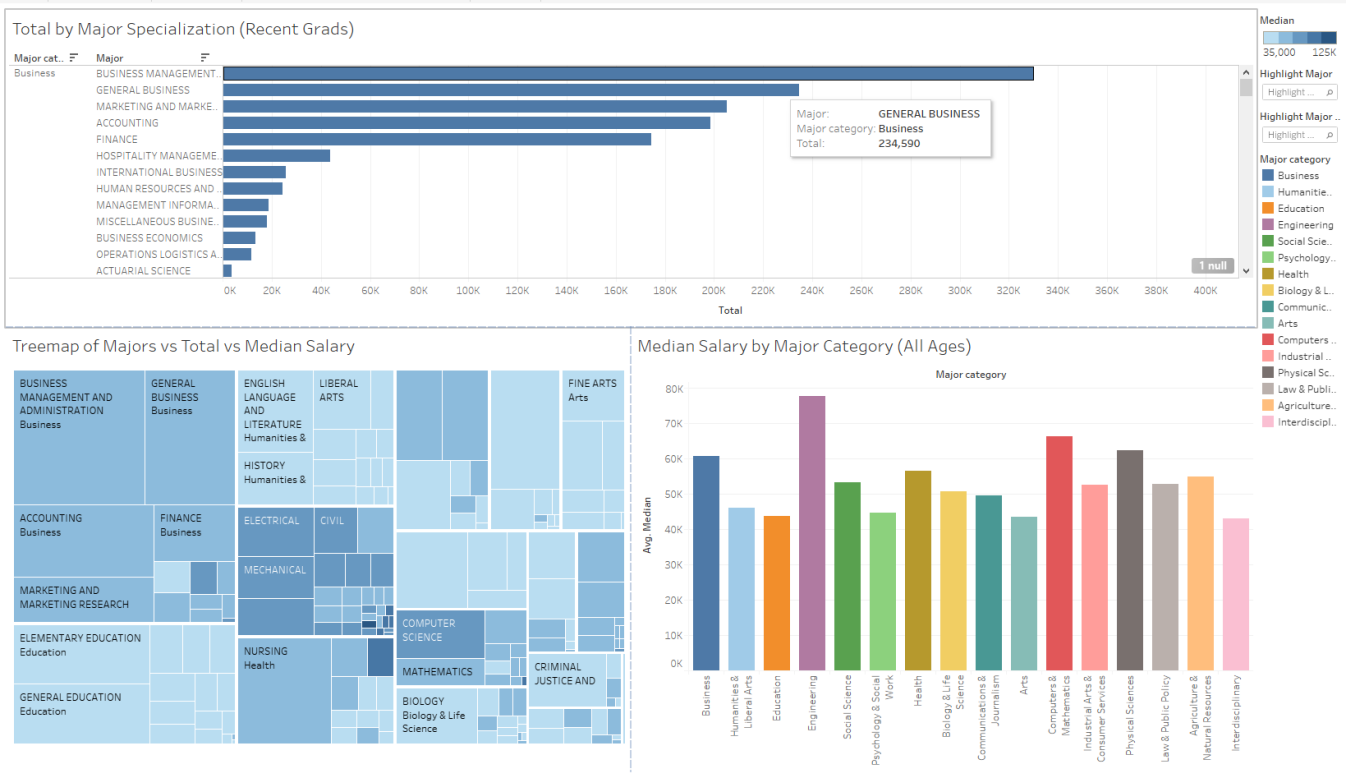

Figure 2: Dashboard 2: Total Counts for each Major Specialization for recent grads juxtaposed with Treemap

here are used throughout all the visualizations. This is important of course, for consistency. Furthermore, the large number of categories here means many colour hues, 16 to be precise, had to be used, which can confuse the user and be hard to distinguish from one another. This is an unfortunate drawback to my visualization solution, but cannot be avoided, as the major categories cannot be further grouped, nor would it be fair to eliminate any fields as part of a base filter.

This dashboard answers the question "What kind of jobs are there?". For now, we can assume what people are studying is representative of the jobs that exist, although Dashboard 5 shows why this may not be the case when we look at underutilization rate. It may be technically more accurate to say "What are people studying?" but that seems less pertinent to the intended reader. Clearly business is the dominant field in terms of popularity by a large margin. Physical sciences, Law, Agriculture are relatively unpopular. The juxtaposition of recent graduates vs all ages on this dashboard lets the reader infer if there are fields that are losing or gaining popularity. The most significant change seems to be a large increase in the current popularity of Biology and Life Science.

Dashboard 2: Specializations, Popularity, and an Introduction to Earnings

The purpose of Dashboard 2 shown in Fig. 2 is to, firstly, give the user an introduction to major specializations. Rather than showing the broader categories as in Fig. 1, we now show each field broken down into its individual specializations, and the total count for each specialization is given as well. Secondly, this dashboard serves to give the user a broad overview of earnings by field, as well as allow the user to at-a-glance look at earnings of interesting specializations. This dashboard was designed to answer the following questions:

- Which specializations are more popular?
- Which specializations make less money than the rest of the field? Which make more?
- Which fields make more money?

The first chart on Dashboard 2 is a grouped bar chart. This bar chart, similar to the charts on Dashboard 1, shows the total counts for each major specialization. Since there are 172 specializations, with approximately a dozen specializations per field, it is unreasonable to use a stacked bar chart. Furthermore, a grouped bar chart is more useful for comparing differences between the specializations themselves, as

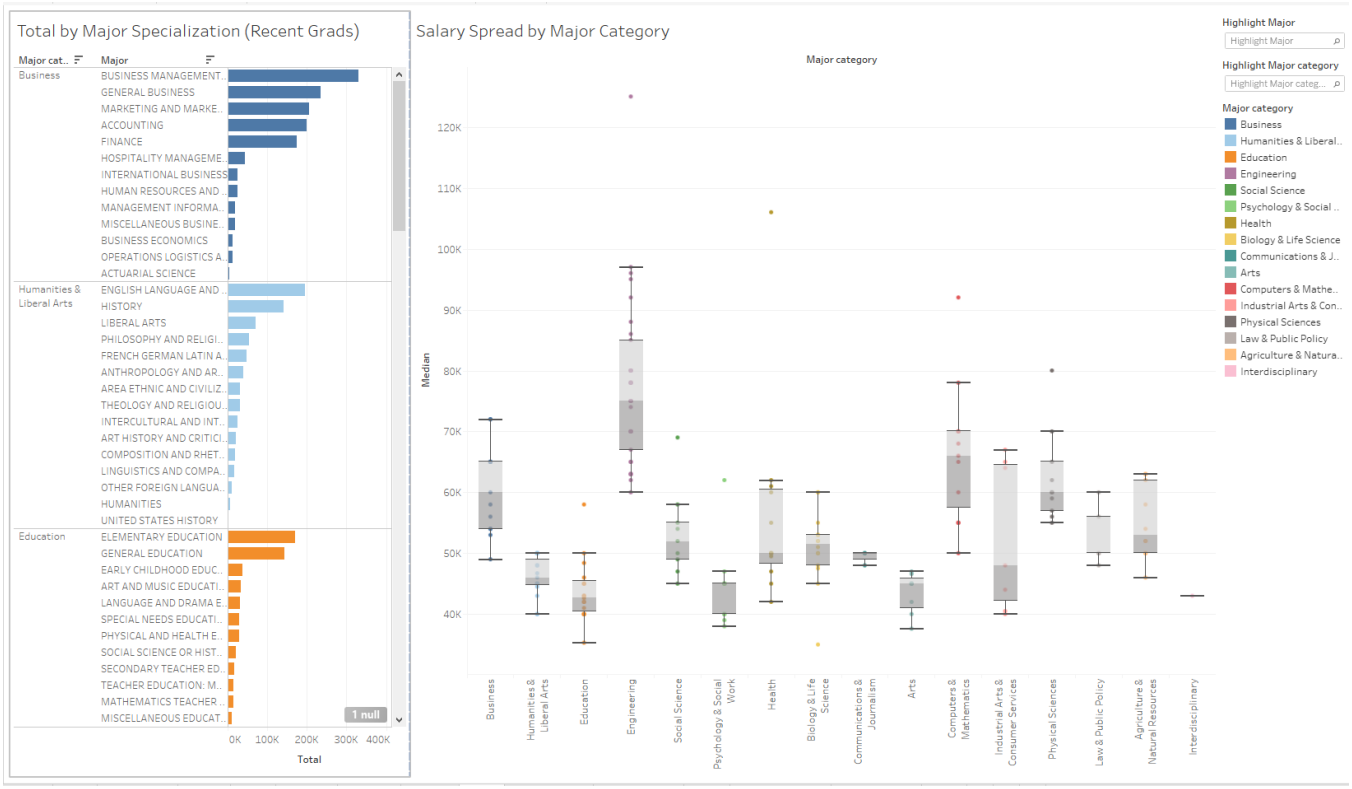

Figure 3: Dashboard 3: Major Categories and a Box Plot of all Earnings for Majors and Major Categories

the user would be interested in comparing individual majors to other majors, such as Computer Science vs Computer and Information Systems. The grouped bar chart is allocated enough space on the dashboard to show one field at a time. The user can filter by major or major category to view whatever they are interested in. There is most definitely a problem with scalability here, but in this way, the problem of scale is addressed with filtering and highlighting.

The second chart is map chart showing fields, their specializations, the total count encoded by area, and the salary encoded by colour saturation. Each field is the parent and each specialization is a node. Area and saturation do not strictly follow the expressiveness principle, but the map chart has many beneficial trade-offs. The map chart deals with scale significantly better than other charts, as it can fit all 172 majors on very little real-estate. It provides a good broad overview of the fields and can at a glance, due to pre-attentive processing [3], show any significant outliers within the region. It is easy for the user to identify which fields have higher earnings simply by how dark the region is. Any specializations within the field that are significantly lighter or significantly darker can be seen quickly by the user, and with a mouse-over tooltip, that specialization can be revealed. For example, the Pharmaceutical specialization within the Health field is significantly darker and thus has much higher earnings than the rest of the field. Similarly, Hospitality Management within the Business field is significantly lighter and thus has much lower earnings. The encoding of salary to colour hue does not allow for easy perception of exact numeric differences. A tooltip showing the median salary for a specialization can be brought up with a mouse-over, but if the user is interested in numerical differences rather than trends, they are better-off moving onto Dashboard 3.

The third chart is bar chart showing the average earnings of each field. This is the same information shown by the saturation of the region in the map chart, but since saturation is a relatively poor encoding compared to height, this bar chart was included to give users a more clear picture of the earnings per field.

It is important to note that all portrayals of salary on this dashboard use the salary information for all ages rather than for recent graduates. This was done deliberately because the user, when looking at general salary information, is mostly interested in how much they *can* make rather than how much they will make right out of school. Furthermore, the salary levels of recent graduates are very similar, and would have resulted in a map chart with mostly homogeneous colour. This extends to Dashboard 3 as well, but not to Dashboard 4,

Figure 4: Dashboard 4: Unemployment Rate vs Median Salary for recent grads and all ages

where the salary is clearly defined as either recent graduate or all ages, since the comparison between the two is what is important on that dashboard.

Dashboard 3: An In-Depth Look at Earnings

Dashboard 3 shown in Fig. 3 shows a much more in-depth overview of earnings. The primary focus here is the box plot of earnings for each field, with points (that can be highlighted or filtered) for each specialization. This dashboard is particularly useful for outlining the salary info of an individual specialization (using highlighting) and how it relates to other specializations, both within and outside the field. It is also useful for comparing with the median and quartiles of the field, to see where that specialization lies within the field itself. With this dashboard, the user can answer the following questions:

- Which fields earn more? Which specializations?
- What are the salary differences between fields? Between specializations?
- What about between specializations within the same subject?
- Does salary within a particular field vary much?
- Are there specializations that make much more or much less than the field?

A user can highlight a specific specialization they are interested in to bring up a tooltip with relevant information as well as to see where that point lies within the overall distribution. The box plot on this dashboard is effective at showing range and outliers. The range is valuable for seeing the spread of salaries within the field, and the outliers let you immediately see if there are specializations that make significantly more or less than the rest of the field. The map chart in dashboard 2 also showed this information, but that was encoded with colour saturation, and was thus not as effective as the height encoding here.

There are 7 positive outliers, some of the more notable ones include Petroleum Engineering in the field of Engineering, Pharmaceutical Sciences in the field of Health, and Economics in the field of Social Science. There is one negative outlier, quite a surprising one: Neuroscience in the field of Biology and Life Science.

The box plot on this dashboard has not been ordered based on ascending or descending salary. Instead, the same ordering from Dashboard 1 has been kept. This consistent ordering helps combat the colour encoding issue where we have too many colour hues for too many fields. In fact, this ordering is particularly important on the box plot because colour hue

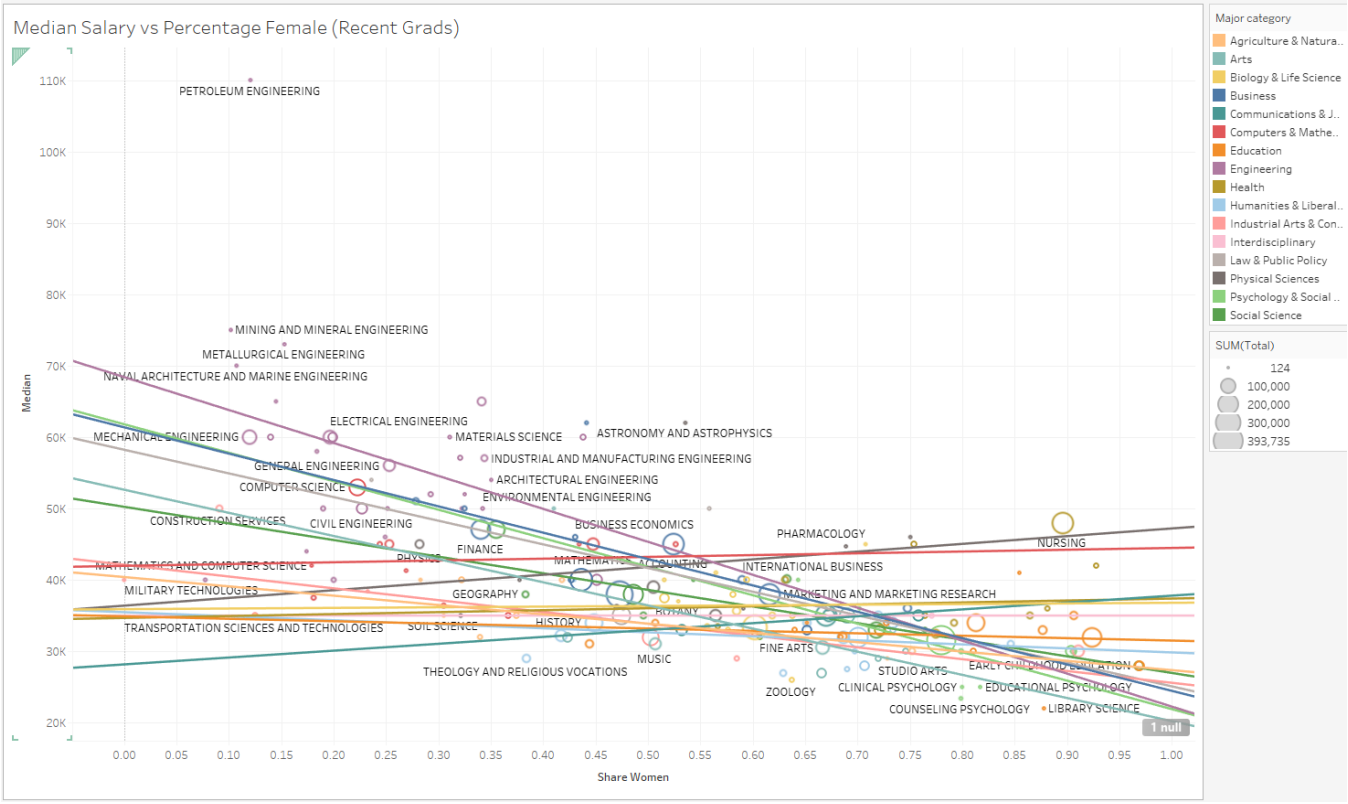

Figure 5: Median Salary vs Percentage Female

is hard to distinguish on the plot due to the small size of the marks.

Dashboard 4: Earnings and Employability

Dashboard 4 shown in Fig. 4 allows the user to investigate employability. Furthermore, it specifically aims to investigate the difference between recent graduates and all ages. From this dashboard, the following questions can be answered:

- Which specializations are more employable? How does this relate to earnings?
- Does the employability for a specialization improve over time?
- How does the employability for a specialization compare to other specializations?
- Are there fields that are more employable or less employable?

The dashboard shows two scatter plots. Each scatter plot shows the Unemployment Rate vs Median Salary for each specialization on the X and Y axes respectively. Colour hue, as-always, represents the field, and the size of each mark indicates the total count for the specialization.

The mean lines for each quantitative variable are also shown as grey lines running through the plot to give the user a frame of reference for the centre. More importantly, it divides each plot into usable quadrants, from which the user quick inferences. By looking at any mark, the user can at a glance see which quadrant the mark lies in, and thus draw conclusions. Marks that lie in the upper-left quadrant are both high-earning and in demand. This quadrant is dominated by the field of Engineering (Purple), and to a lesser extent, the fields of Computers and Mathematics (Red) as well as Business (Dark Blue). Marks that lie in the bottom-right quadrant are both low-earning and in low demand. Here you can find mostly Humanities and Liberal Arts (Light Blue), as well as Psychology and Social Work (Light Green). The bottom-left and top-right quadrants are highly interesting. The bottom-left quadrant represents majors that are low-earning yet also in high demand. This quadrant is very clearly represented by the field of Education (Orange), and to a lesser extent, the field of Biology and Life Sciences (Yellow).

The upper-right quadrant does not contain very many marks compared to the other three. Majors that lie here are both high-paying, yet also in low demand, which seems counter-intuitive at first. This is where the juxtaposition of the all ages graph is highly useful. As we can see, most of the

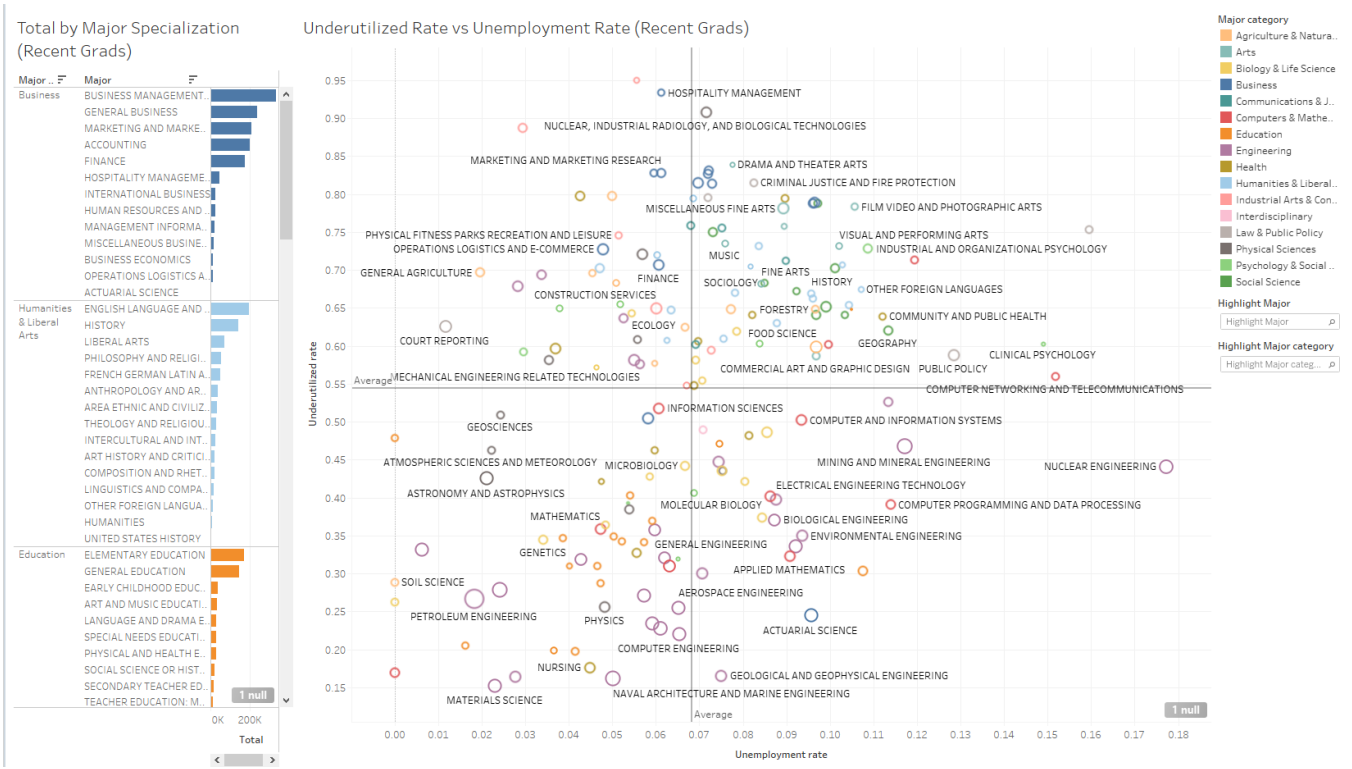

Figure 6: Dashboard 5: Underutilized Rate vs Unemployment Rate

marks that appear in this quadrant, such as Mining and Mineral Engineering, Nuclear Engineering, and Economics, have much lower unemployment than recent graduates. From this, we can infer that these specializations do not commonly employ recent graduates, and instead, prefer those that have experience. This makes a lot of sense especially when you consider the nature of the work of Nuclear Engineers or Economists.

Earnings vs. Gender

Figure 5 displays a scatter plot showing salary vs the percentage of women with that major. This was not added to any dashboard as it is not particularly useful when juxtaposed with other graphs, since it already displays the two metrics that users are interested in when discussing this topic: the gender wage gap. The purpose of this chart is to investigate the following:

- How does gender relate to salary?
- Do different fields have different gender-wage gaps compared to others?

The main focus of this scatter plot is the trend lines. Unlike the previous scatter plots on Fig. 4, the overall trend is infact, the primary thing the user wishes to see. For the most part, all the trend lines are either negatively correlated, or neutral. From this we can infer that the in general, a higher percentage of women is correlated with a lower salary, but this is different depending on the field. We can see that some fields have very steep negative correlations, such as Engineering and Business, while others, such as Education or Biology and Life Sciences are relatively neutral. Interestingly enough, and perhaps surprisingly, the field of Computers and Mathematics has a positive correlation.

Dashboard 5: An In-Depth Look at Employability

So far, many of our observations have been based on the assumption that the total count of those with a specific college major represents the number of jobs for the major. This is true for most, but for many, may not be an accurate assumption at all. To investigate this, I created a transformed data field that I have called "Underutilized Rate". The equation for the underutilized rate is as follows:

$$\frac{(NonCollegeJobs + LowWageJobs)}{(NonCollegeJobs + LowWageJobs + CollegeJobs)}$$

Effectively, the underutilized rate shows the number of those who are working jobs that do not actually utilize a college degree. It is important to note that due to limitations in the data set, it is not the under-utilization of that specific degree since census data does not exist for whether or not a job

requires their college degree specifically; only whether or not they have a degree at all. However, it is likely safe to assume that it representative nonetheless, after all, school is simply the first building block in the career.

Dashboard 5 shown in Fig. 6 shows a scatter plot of the underutilized rate vs unemployment rate for recent graduates. The grouped bar chart showing major specializations is present as well, to make it easier for the user to filer for their field or specialization of interest. It is important to note here that, unlike all the previous scatter plots, the size encoding does not encode total count but rather salary. This was done because the grouped bar chart already shows total count quite effectively, and so it would have been redundant to show it again as a size encoding on the scatter plot. Instead we show salary information since it is not present anywhere else on the dashboard, and, following the effectiveness principle, it is encoded with size since it is not the primary or even secondary metric that we are interested in.

This dashboard allows the user to answer the following questions:

- Are degrees useful?
- Which degrees are more useful than others?
- Is a field or specialization possibly over-saturated?
- Are there salary differences due to over-saturation?

Similar to Dashboard 4, mean lines on the scatter plot provide a measure of centre and also split the plot into easy-to-discern quadrants. Marks that lie in the lower-left quadrant represent specializations that are both employed and utilized. Conversely, marks that lie in the upper-right quadrant represent specializations that have both low employment and are underutilized. Marks that lie in the bottom-right quadrant are utilized, but not employed. The specializations that are found here, such as Nuclear Engineering, seem to be similar to those found in the upper-right quadrant of Dashboard 4: indicating fields that are highly-specialized and require experience. Perhaps many of those with these degrees are still pursuing higher education such as doctorate degrees. Marks that lie in the upper-left quadrant are perhaps the most interesting on this dashboard. The specializations found here are highly employed but are underutilized. Specializations such as "Medical Assisting Services" and "Operations Logistics and E-commerce" are perhaps, at least according to this analysis, not particularly useful.

5 FUTURE WORK AND CONCLUSION

In the future I would like improve this visualization by adding animation, especially if time-scale data could be made available and integrated with the current data. Dashboard 4, where the changes in both earnings and employability can be seen, is in my opinion the most valuable dashboard and yet relatively cluttered. It is easy to miss certain interesting finds if you do not specifically filter for it. For example, the Pharmaceutical Sciences specialization makes a massive jump where it "shoots" from the lower left quadrant on the recent graduates chart all the way up to the second highest earning specialization on the all ages chart.

I have presented a series of interactive visualizations intended for prospective students, current students, and recent graduates to make informed decisions about their future. Many common questions regarding the earnings and employability of 172 college majors can be answered from these dashboards, and I have made it clear which dashboards are effective at answering which questions. A countless number of conclusions can be drawn from this data, and I have made a few of the most interesting observations within section 4. Further observations can be investigated at the users' pleasure, and I am sure students would benefit greatly from utilizing this visualization regarding the earnings and employability of various fields and specializations.

REFERENCES

[1] Ben Casselman. 2014. The Economic Guide To Picking A College Major. https://fivethirtyeight.com/features/the-economic-guide-to-picking-a-college-major/

[2] Fivethirtyeight. 2019. fivethirtyeight/data. https://github.com/fivethirtyeight/data

[3] Tamara Munzner and Eamonn Maguire. 2015. *Visualization analysis & design.* CRC Press Taylor & Francis Group.