Final Project: Data Analysis

James Parker Tonra

12/07/22

CS 1030

After looking over the different options for sourcing data, FiveThirtyEight seemed like a good place to start since the site featured many different types of data. The data set I ended up choosing to analyze contained information pertaining to graduates from various college majors. Sourced with data obtained from a 2010 American Community Survey, the data set included a category for each major as well as some useful statistics. Most notable of the statistics were the total amount, median incomes, and unemployment rates for each major. I thought it would be interesting to find out which majors/categories of majors had the best/worst unemployment rates. I downloaded the .CSV and imported it into Microsoft Excel (Data > Get & Transform Data > From Text/CSV). This led me to raise a few more questions:

> (1) Which major has the highest unemployment rate?
>
> (2) Which category of majors has the highest average unemployment rate?
>
> (3) Which major has the lowest unemployment rate?
>
> (4) Is there a correlation between total in major & unemployment rate?

To find the answer to questions **(1)** & **(3)**, I sorted the data by unemployment rate (UR). This was a super simple task since the .CSV was neatly formatted. I clicked the dropdown arrow next to the “Unemployment Rate” column & selected “Sort Largest to Smallest”. “Miscellaneous Fine Arts” had the highest UR by a significant amount, about 50% higher than its runner up (Clinical Psychology). “Educational Administration and Supervision” & “Geological and Geophysical Engineering” had the lowest URs: both were tied at zero **(Figures a, b)**

Major_code	Major	Major_category	Total	Unemployment_rate
6099	MISCELLANEOUS FINE ARTS	Arts	8511	0.156147487
5202	CLINICAL PSYCHOLOGY	Psychology & Social Work	7638	0.102712161
3801	MILITARY TECHNOLOGIES	Industrial Arts & Consumer Services	4315	0.101796407
2303	SCHOOL STUDENT COUNSELING	Education	2396	0.101745936
3501	LIBRARY SCIENCE	Education	16193	0.094842992
6003	VISUAL AND PERFORMING ARTS	Arts	55141	0.094658002
2101	COMPUTER PROGRAMMING AND DATA PROCESSING	Computers & Mathematics	29317	0.090264217
5206	SOCIAL PSYCHOLOGY	Psychology & Social Work	10871	0.087336245

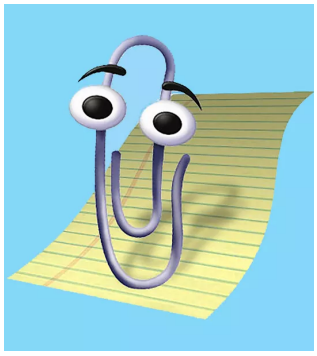
Figure a: Majors with the 8 highest unemployment rates (irrelevant columns hidden)

6107	NURSING	Health	1769892	0.026796818
6109	TREATMENT THERAPY PROFESSIONS	Health	252138	0.026291603
1100	GENERAL AGRICULTURE	Agriculture & Natural Resources	128148	0.026147106
4005	MATHEMATICS AND COMPUTER SCIENCE	Computers & Mathematics	7184	0.024900398
5008	MATERIALS SCIENCE	Engineering	7208	0.022333333
3607	PHARMACOLOGY	Biology & Life Science	5015	0.016110797
2301	EDUCATIONAL ADMINISTRATION AND SUPERVISION	Education	4037	0
2411	GEOLOGICAL AND GEOPHYSICAL ENGINEERING	Engineering	6264	0

Figure b: Majors with the 8 lowest unemployment rates (irrelevant columns hidden)

To determine the answer to question **(2)**, I made a pivot table from the data. I did so by selecting all the cells in the table and clicking “PivotTable” (Insert > Tables > PivotTable). From Microsoft’s website:

> A PivotTable is *a powerful tool to calculate, summarize, and analyze data that lets you see comparisons, patterns, and trends in your data*.



Creating a pivot table allowed me to zoom in on certain parts of the data while leaving others out. Additionally, I now had the option to add columns with the average or sum of values from any category. I formatted the unemployment rate to display as a 2 decimal point percentage (Format Cells > Number > Percentage) to make the UR column a bit easier to read. Excel performed all of the calculations for me. I then sorted the categories by decreasing UR once again to determine that "Arts" had a higher UR than any other category **(Figure c)**.

Row Labels	Sum of Unemployed	Average of Unemployment_rate
⊞ Arts	104125	8.76%
⊞ Psychology & Social Work	104206	7.79%
⊞ Interdisciplinary	2990	7.73%
⊞ Humanities & Liberal Arts	179136	6.94%
⊞ Communications & Journalism	101199	6.91%
⊞ Law & Public Policy	43049	6.79%
⊞ Social Science	132150	6.57%
⊞ Computers & Mathematics	79974	5.94%
⊞ Industrial Arts & Consumer Services	40360	5.85%
⊞ Physical Sciences	38221	5.45%
⊞ Business	434397	5.45%
⊞ Engineering	146389	5.06%
⊞ Biology & Life Science	57335	4.99%
⊞ Health	75013	4.72%
⊞ Education	125336	4.68%
⊞ Agriculture & Natural Resources	18551	3.96%
Grand Total	1682431	5.74%

Figure c: Pivot table with major categories and their unemployment rates.

To answer question **(4)**, I made a scatterplot. I was curious to see if there was a correlation between lots of people being in a major and its unemployment rate. I inserted a scatter plot (Insert > Charts > Scatter) and selected the values in the "Total" column as the x-axis & values in the "UR" column as the y-axis. I then added a trendline (Chart elements > Trendline) to determine that there was a weak negative correlation between "Total" and "UR" values within the data **(Figure d)**:

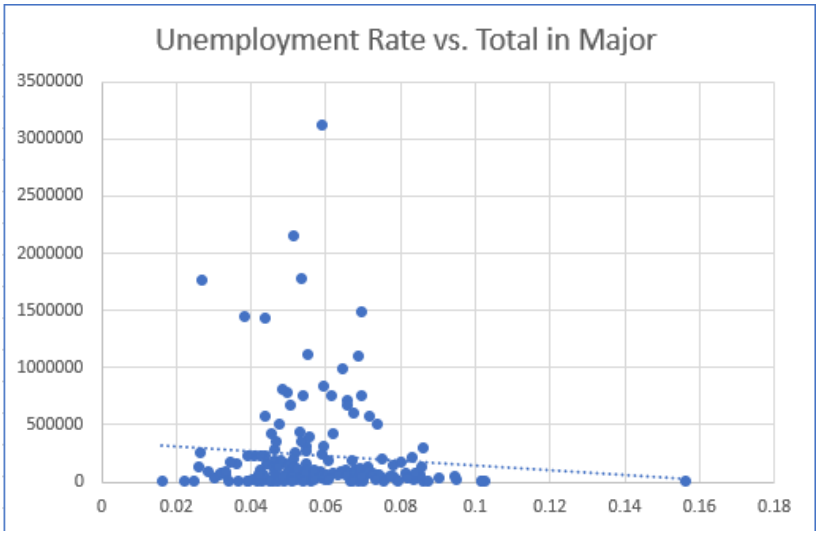


Figure d: Scatter plot with total number of graduates in major vs unemployment rate

Conclusion:

As a musician, I can't say that I was incredibly surprised to learn that "miscellaneous fine arts" majors were the most likely to be unemployed after graduating. Conversely, I was quite surprised to learn that there was a negative correlation between total number graduates & unemployment rate. I figured that a more graduates within a major would inevitably result in more of them being unemployed. By looking at business management, it's easy to see where I was wrong. Despite having the greatest total number of graduates (over three million), business management's UR was only 5.89% (pretty close to the median). Even upon removing extreme outliers (top/bottom 3 UR values), a weak negative correlation was still observed **(Figure e)**. Overall, my experience with the data analysis project was quite rewarding. Excel/spreadsheets in general have a history of intimidating me, but once I got into a groove I found the workflow quite enjoyable.

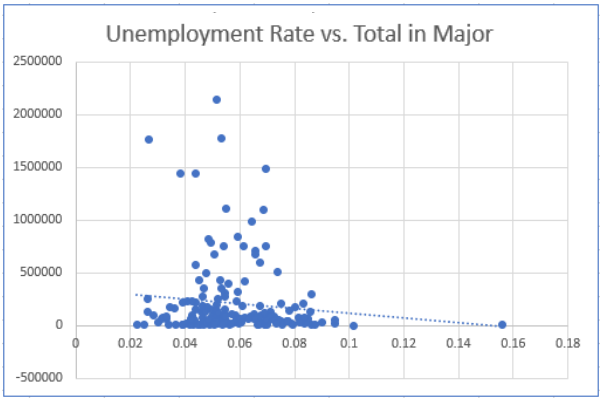


Figure e: Scatter plot with outliers removed.