BIRTH STATISTICS AND SCHOOL FUNDING: A CASE STUDY

THE BIG 'WHY'

Understanding the factors leading to healthy births is of paramount importance to the healthcare industry. Birth statistics are rigorously recorded and studied to produce the most desirable outcomes possible. But what if we can also determine if any seemingly unrelated factors from one's childhood can shape *future* birth statistics? Could we then use this information to indirectly guide the paths of mothers in years to come?

FUNDING: FIRM FOOTING OR FLIMSY FLIM-FLAM?

In this project, I explored one such possible factor: public school funding. The reasoning that led to the question went like this:

- Are birth weights affected by the age of the mother?
- If yes, is the age of the mother influenced by the level of education attained?
- If yes, is the level of education attained influenced by the amount of public school funding received in grade school?

Put simply: **Does school funding influence the birth weights of the following generation?**

TOOLS AND DATA

The analysis was conducted over the course of three weeks with Python through Jupyter Notebook. The analysis included pandas, NumPy, OS, seaborn, matplotlib, folium, json, pylab, and sci-kit learn libraries and employed geographic visualization, linear regression, and clustering. Additional visualization was done in Tableau Public.

The data for this project comes from the merging of two data sets from Kaggle. The first data set, "US Births by Year, State, and Education Level" was originally sourced from the CDC's Natality page using the WONDER retrieval tool. It includes detailed information about recorded births from 2016 to 2021. The second data set, "U.S. Educational Finances," was originally sourced from the United States Census Bureau's annual surveys of grade schools. It contains information about the revenues and expenditures for the years between 1992 and 2016.

THE DEEP DIVE

The first big step to any project--besides obtaining and understanding the data--is making it useful. In order to answer my questions, I had to make some early choices that defined how the whole project would progress.

Determining School Years. When attempting to determine what amount of funding the mothers received in public school, the only relevant data I had contained *average* ages for the mothers, the level of education received, and the years they gave birth. I used this information to create what I considered to be a reasonable approach:

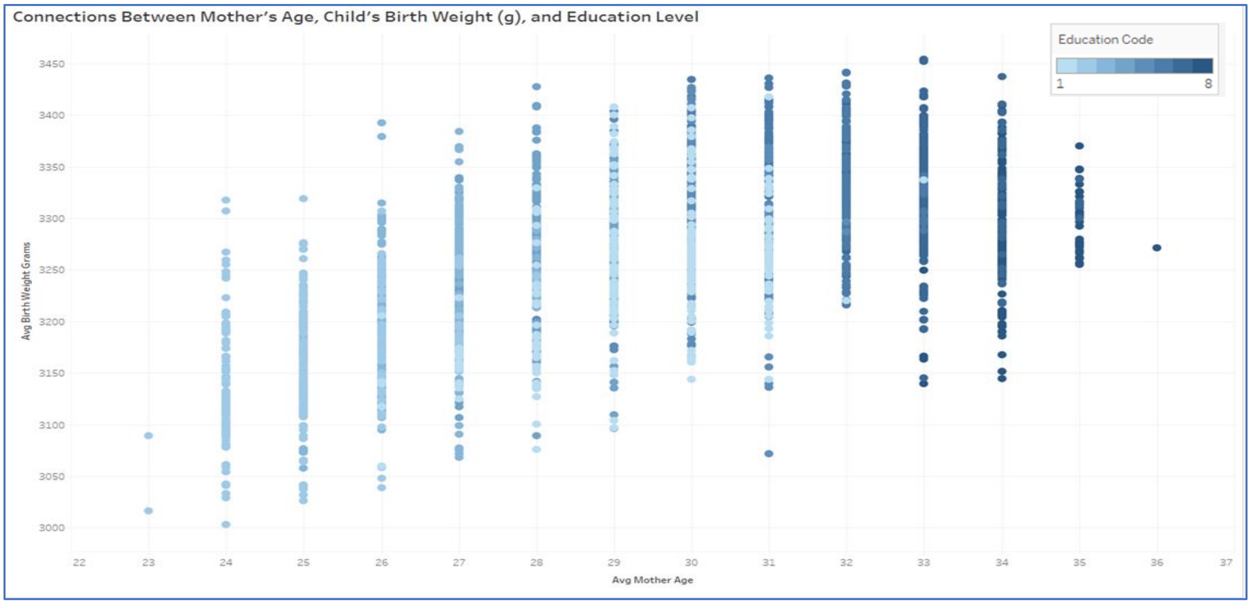
(Year – Average Mother's Age) + Assumed School Age = Reference Year

where Assumed School Age is 13 for mothers who attained 8th grade or lower, 15 for some high school with no diploma, and 16 for high school graduates and above. Unknown education levels were assigned a value of 14 years.

Once a reference year was obtained, I calculated the average school expenditure for that year and the three years before and after it, creating a 7-year average total expenditure per student. This was done to reflect (a) the funding received over several years of education, and (b) that the reference year is only based on an average age, so this would hopefully blend and blur the lines a little from that specific year to something a bit more general and applicable to more mothers in the category.

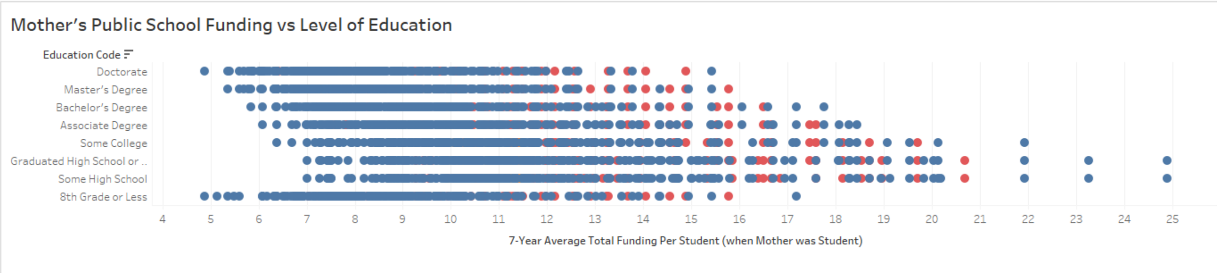
```
# Add '7yr_avg_total_per_student' column

finances['7yr_avg_total_per_student'] =
finances.groupby(['state'])['total_expenditure_per_student'].rolling(7, min_periods =
1, center = True).mean().values
```

Charts, Charts, and More Charts. Each variable being tested needed to be compared with each other to establish relationships. Using seaborn, I made scatterplots with regression model fits of *average birth weight (g) vs. average age of mother*, *education level of mother vs average age of mother*, and *average birth weight vs. education level of mother*. In each of these cases, there was indeed a positive correlation. Checking these with a correlation heatmap yielded values of 0.52, 0.38, and 0.62, respectively. A chart relating all three was later created in Tableau and is shown here.



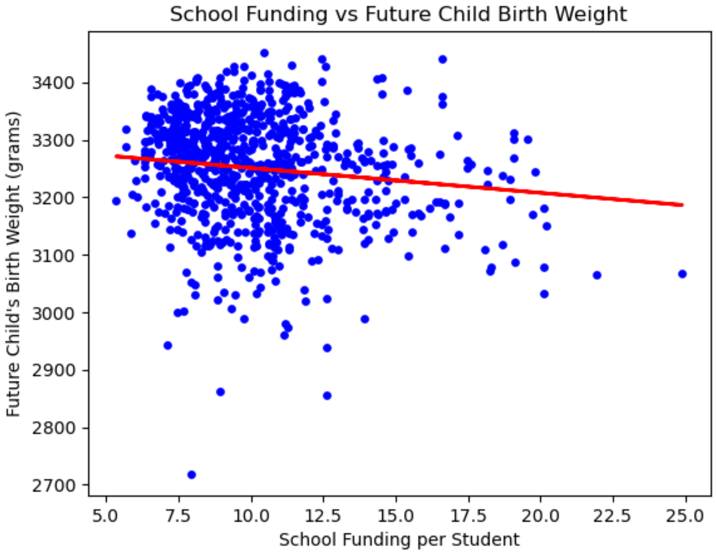
On the *x*-axis, we have the average age of the mother, and on the *y* we have the average birth weight of their child in grams. The educational level attained by the mother is indicated by the shade of blue: the darker the shade, the higher the level of education. Here, we verify that as the mother's age goes up, so too does the birth weight until it levels off around the age of 32, and the education level likewise increases. Clearly, time pursuing education causes many women to delay having a child.

A Bad Connection. Now that we know these three variables are connected, I had to establish the final connection between educational level attained and public school funding. At first, the data didn't look promising. A quick glance at the data seemed to indicate that greater funding led to *lower* achievement.



That didn't seem right. It felt like something was wrong, but I couldn't put my finger on the problem. I resolved to continue further analysis elsewhere and return to this problem a little later.

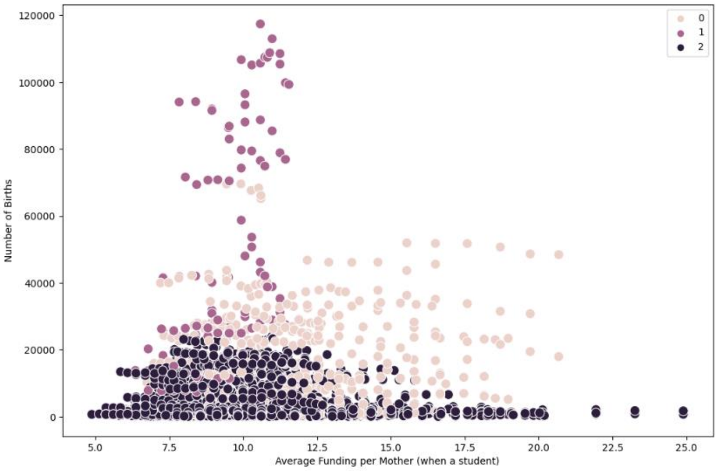
Regression. By plotting a regression line of school funding and birth weights, we ended up with...a whole lot of nothing.



The relationship was a poor one. The r^2 value was only a mere 0.036. It seemed there was a bad link somewhere, so I pressed on and planned to return to this relationship after a bit more digging.

I performed a cluster analysis with a k-means algorithm to try to uncover any further meaningful insights. At first, the results looked no clearer. I tried reviewing several other variables to make sense of the clustering pattern, and one stood out to me.

This plot of *number of births vs average funding per mother* had the clusters clearly defined. So why would there be such a high number of births for one cluster but not the others? I had three possible ideas for division: income (in case there's a connection between income and number of children in a family), geography (populations), or a mix of both. I did not have income data, but I did have geographic data.



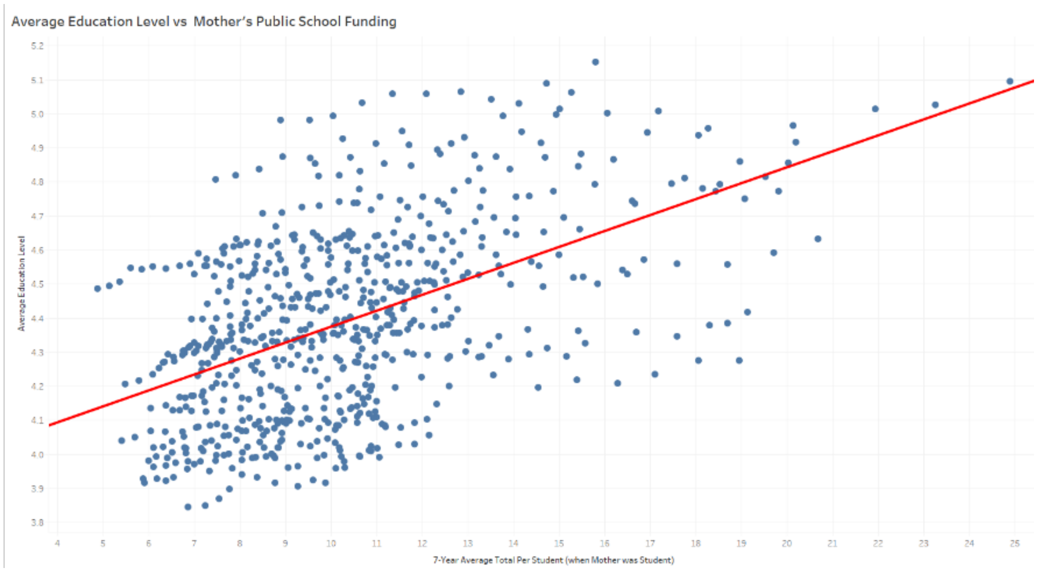
Using the filtering functionality of Tableau, I was able to confirm that the cluster with the highest number of births was limited to California and Texas—the two most populous states in the U.S. The middle cluster included values from the next ten most populous states, and the 3rd cluster contained the rest.

Connection Repaired. The geographic analysis sparked an epiphany: the problem with the earlier graphic of funding and educational level ignored the *change of educational level in a state with increased funding*. Before, all the states were lumped together, so an actual *progression* based on funding was not readily viewable. This brought me back into the data to standardize the education level.

The data itself already utilized an education code, where "1" denoted someone who didn't make it past 8th grade, "2" denoted some high school without a diploma, "3" denoted a high school graduate or GED recipient, and so on all the way up to "8" for a doctorate degree. I used these values as the basis of an *education ratio*. The education code was multiplied by the number of births for that level (and thus number of mothers) within each state and year. The sum of those values was then divided by the total number of births that year. The result was a number indicating the average education code for each state and year. To put it in mathematical terms, for each state-year combination,

$$\text{education ratio} = \frac{\sum_{e=1}^{8} e n_e}{\sum_{e=1}^{8} n_e}$$

where e is the education code and n_e is the number of births where the mother attained said education code. This education ratio now allowed me to directly tie state funding to the state's achievements. Lo and behold, once I plotted this new data as average education level and school funding, a moderate positive relationship emerged.



CAVEATS

There were some limitations in the analysis that need to be addressed.

1. While we have the cost of expenditures for each state in each year, what we do not know is the effect of inflation on these expenditures. It may be that what we have recorded as an increase in funding is simply an increase in cost to maintain the same level of support.

2. Much of the data was presented as averages. This leads to scenarios where margin cases are getting ignored and errors stack up. For example, none of the birth weights recorded are medically low nor high—all fall into a healthy range. This may obfuscate some of the potential for the analysis's impact.

3. Some data for Maine, Montana, and Nebraska were suppressed by the CDC and left out of the set.

4. State-level analysis may not be granular enough to identify geographic or economical impact. More interesting information may be gleaned from county- or school-district level data.

SO WHAT DID WE LEARN?

- Birth weights increase as the age of the mother increases
- The age of the mother increases as the level of education attained increases
- The level of education attained increases as the amount of public school funding increases
- Almost paradoxically, an increase in the amount of public school funding is **NOT** a significant factor in birth weights!

This project was my first use of python machine learning techniques and they proved powerful. I enjoyed the journey that the data took me on. It was like a good book: I was carried along and validated in my reasoning, variables continuing to verify a connection I only had in my head, until a plot twist came along and surprised me. I expected to have a definitive answer at the end of this project, and instead I ended up with evidence that some things are just more complicated than they seem. And that, of course, is why data analysis exists at all.

To see more of this project, please access the GitHub repository here.

Analysis Performed by Sebastian Lombardo

Case Study Written 07/01/2023