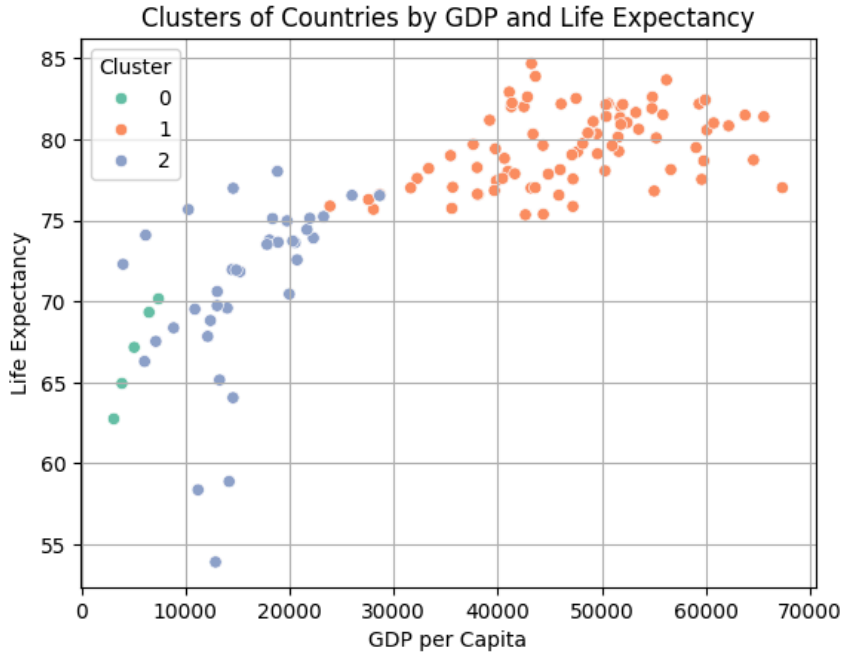
Achieng Kuir
CS/MATH-215
Data Manifesto
May 11, 2025.

When I first started this class, data felt kind of distant, just numbers in csv files and spreadsheets. But as the semester went on, especially when I was working on in-class activities and groups and individual projects, I started to see things differently. I realized data isn't just about stats or code it's about people, choices, and stories.

Readings like Giorgia Lupi's "Data Humanism," *Data Feminism* by Catherine D'Ignazio and Lauren Klein, and *Everybody Lies* (Chapter :Data Reimagined) by Seth Stephens-Davidowitz made me think a lot more deeply about what data really is and how we use it. I've come to believe that good data science is thoughtful, human centered, and always aware of the bigger picture. Here's what I now believe defines my approach to working with data.

Principle 1: Data Is About People

Lupi writes that we should see data as "a medium for empathy," and that stuck with me. When I worked with the life expectancy dataset in my final project, I wasn't just looking at numbers I was looking at lives. One plot showed how countries with similar GDPs could have very different life expectancies. That made me ask: what's really going on?



Data Feminism reminded me that data isn't neutral, it always reflects choices about what to collect, how to collect it, and who gets to interpret it. That made me think about who gets counted and who doesn't. That invisibility matters. In another class project, I looked at timestamp data (Project 7), and even though it was personal, it made me think about how data reflects behavior and life rhythms.

Principle 2: Patterns Are Not the Whole Story

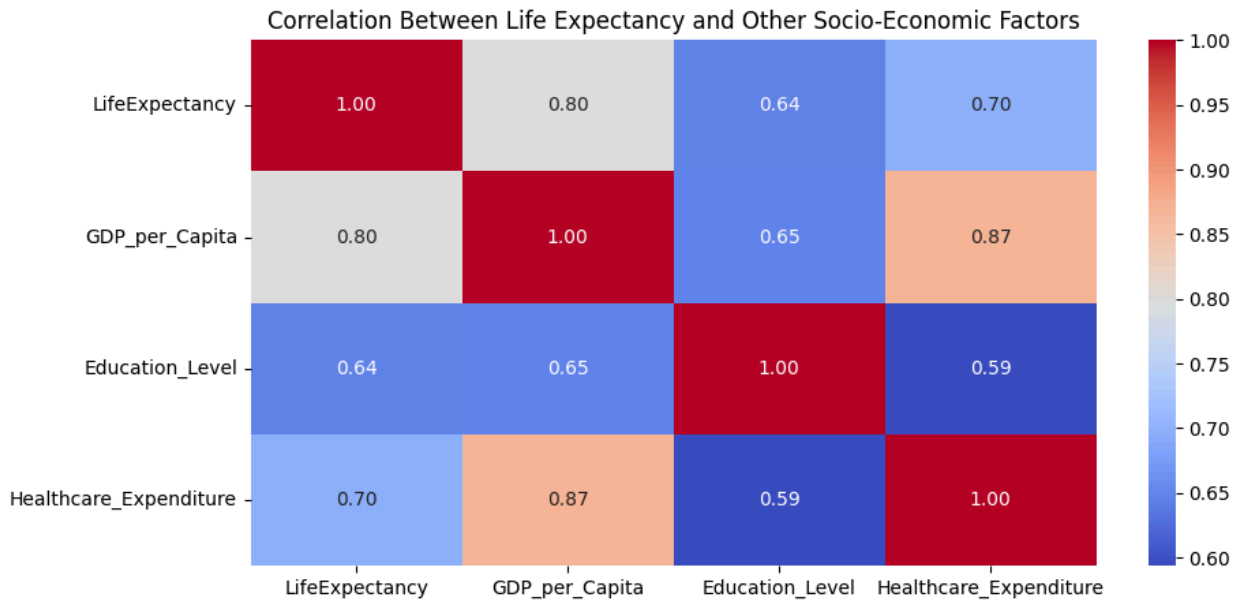
In my final project, I built a linear regression model, and analyzed corelation to see how education, healthcare, and GDP per capita might relate to life expectancy. The models gave me a general idea of correlation, but it didn't explain everything. Some countries didn't follow the trend, and I realized that just spotting a pattern doesn't mean we understand the reason behind it.

Regression Equation

$$\text{Life Expectancy} = 52.20 + 0.0002 \cdot \text{GDP\_per\_Capita} + 0.1746 \cdot \text{Education\_Level} + 0.0110 \cdot \text{Healthcare\_Expenditure}$$

The linear regression model estimates life expectancy based on the GDP per capita, education level, and healthcare expenditure. Each coefficient shows how much life expectancy is expected when that variable increases by one unit, holding all the other constant.

- Holding everything else constant, a $1 increase in GDP per capita is associated with a 0.0002 year increase in life expectancy.
- Holding everything else constant, a 1% point increase in education level is associated with 0.1746 year increase in life expectancy.
- Holding everything else constant, a 1% point increase in healthcare spending increases life expectancy by 0.0110 years.
- The intercept is 52.2 years, which serves as the baseline of the model.



This reminded me of our hierarchical clustering in-class activity (Week 6), where we used the U.S. College Scorecard dataset. The dataset included columns like INSTNM (institution name), Admission_Rate, Cost, Percent_PellGrants, and median_earnings_10yrs. We grouped colleges based on similarities in these attributes to form clusters. But just seeing which colleges were similar on paper didn't mean we fully understood what made those schools different in real life. Without knowing a college's history, mission, or student population, some of the clusters didn't tell the whole story. That taught me that data science should go beyond grouping or modeling. One example is when we made a scatterplot of college cost vs. median earnings after 10 years. Even though the chart showed a general trend that higher costs often relate to higher earnings, there were plenty of exceptions. That made me think harder about what might explain those outliers and reminded me that visualizations and patterns are only the beginning of a deeper analysis.

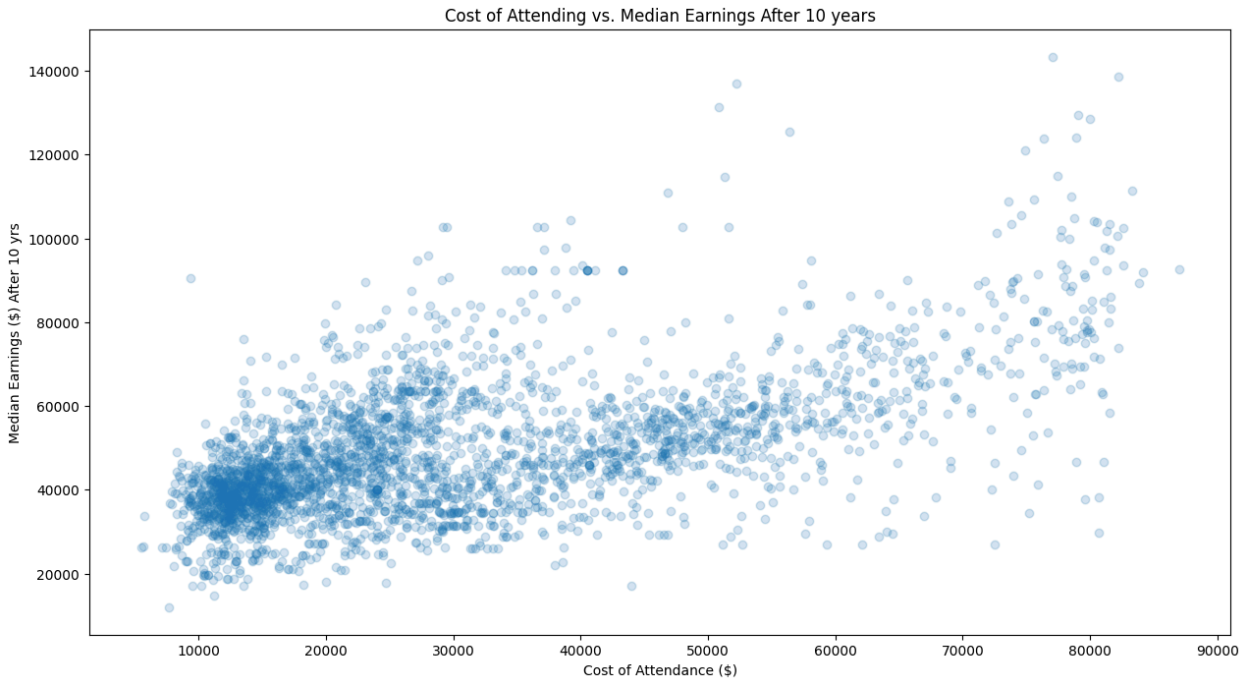
```
plt.figure(figsize=(15,8))

plt.title('Cost of Attending vs. Median Earnings After 10 years')

plt.xlabel('Cost of Attendance ($)')

plt.ylabel('Median Earnings ($) After 10 yrs')

plt.scatter(df['Cost'], df['median_earnings_10yrs'], alpha=0.2)
```



Principle 3: Stay Curious and Question Everything

One thing I learned from *Everybody Lies* was how unusual data (like Google searches for Spider Solitaire) can tell us a lot. In my API project (Project 5), I didn't always know what I was looking for at first, but asking questions and following weird patterns helped me discover surprising connections.

In the life expectancy project, I searched for cases where healthcare spending went up, but life expectancy dropped. It didn't fit the model and that's what made it so interesting. The project taught me to lean into the unknown and stay curious.

Principle 4: Data Science Isn't Neutral, and Neither Am I

The story of Christine Darden in *Data Feminism* showed me how people, especially women and people of color often have to use data to prove their worth. It reminded me that data science can reinforce power structures unless we actively try to question and challenge them.

I realized that I bring my own values and questions into my data work. Choosing to work on life expectancy (instead of something like consumer data) was my way of saying that human wellbeing matters. In my SQL project (baseball stats), I analyzed home runs and batting averages, and while that was fun, it also made me think about the difference between entertainment data and data that impacts lives.

Applying My Principles to Future Projects

If I were to study access to clean water across regions or any other project, I'd follow the same steps: start with curiosity, examine who's included and who's not, and question the assumptions built into the dataset. I'd visualize the data, but I'd also tell the stories behind the numbers. I'd bring in local context, qualitative insights, and remember that no dataset is the full picture.

Conclusion: What Data Science Means to Me Now

This class taught me more than just how to use Python or SQL. It taught me to think about what data represents and what stories it hides. Data is not just facts, it's part of a bigger system. readings like *Data Feminism* taught me that justice and equity should guide our work.

So if someone asked me what data science is, I'd say: it's not just analysis it's responsibility. It's a way of understanding the world, with care. As Giorgia Lupi said, "The revolution will be visualized." I want my work to visualize empathy, complexity, and people, not just patterns.