Exploratory Visual Analysis

1. Introduction

This document will take you through my journey of exploratory visual analysis of the World Development Indicators dataset provided by the World Bank. My approach involves employing Tableau Prep to meticulously clean the data and Tableau Desktop to delve deep into the dataset in the pursuit of a narrative. The project begins with data profiling, proceed through iterative data exploration via question or hypothesis and finally concludes with an effective, thoughtful visualization resulting from the exploration. Accordingly, this document has different sections for each of these phases.

2. Data Profile

To begin with, I downloaded the dataset "World Development Indicators" provided online by The World Bank. The World Development Indicators is a compilation of relevant, high-quality, and internationally comparable statistics about global development and the fight against poverty. The dataset is classified as Public under the Access to Information Classification Policy. So, users inside and outside the Bank can access this dataset. Since this dataset is licensed under Creative Commons Attribution 4.0, users are allowed to copy, modify, and distribute data in any format for any purpose, including commercial use but are obligated to give appropriate credit and indicate any changes they made. The downloaded dataset is a zip file (261.3 MB) with 6 individual csv files - WDICountry.csv, WDIcountry-series.csv, WDICSV.csv, WDISeries.csv, WDIseries-time.csv and WDIfootnote.csv.

To grasp the data structure, I loaded all six files into Tableau Prep and tried to get statistics and context of data through metadata. Subsequently, I determined that three out of the six CSV files would be most pertinent for my exploration and analysis. At this stage, I made the decision to disregard the remaining files and to focus exclusively on 'WDICSV.csv,' 'WDICountry.csv,' and 'WDISeries.csv.'

The primary data file, WDICSV, encompasses a wide array of 1477 distinct world development indicators derived from various sectors such as healthcare, energy, tourism, economy etc. accessible for 266 countries, spanning a 62-year timeline from 1960 to 2022. This comprehensive file is of 188.5 MB, comprising 392,883 rows and 67 columns. I observed that, for the earlier years, many indicators contained empty values, and each of these indicators utilized varying scales and had different statistical interpretations, encompassing aspects like weighted averages, sums, ratios, percentages of the total, and more. Given this complexity, I recognized the need for heightened caution when visualizing them collectively.

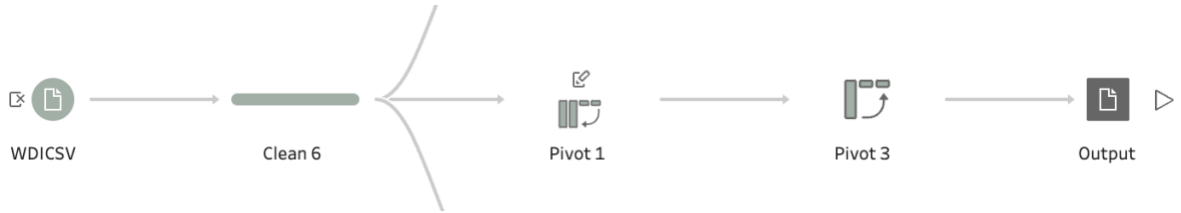


Figure 2.1 - Data manipulation using Tableau Prep.

For data preparation, I executed several essential cleaning tasks (figure 2.1) such as double pivoting and the removal of indicator codes, ensuring that each indicator value was extracted as distinct quantitative data while transforming the year into ordinal data. Additionally, I categorized both the country code and country name as nominal data.

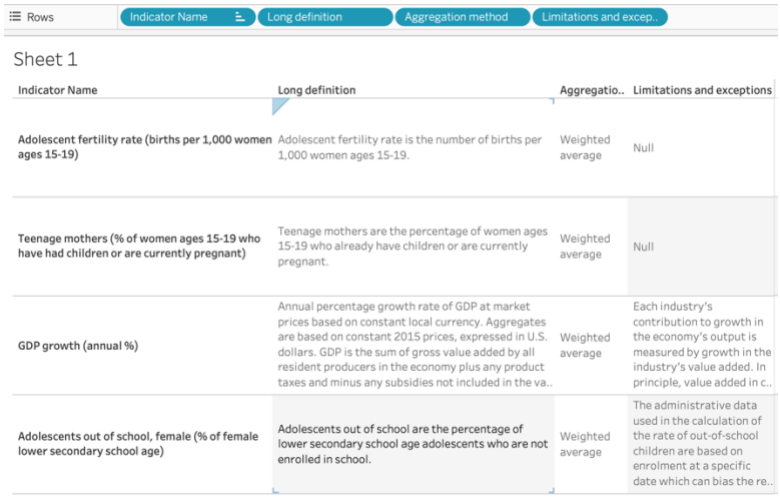
The WDISeries.csv is a concise 3.9 MB master file that offers additional information about each indicator. It contains 1477 records and 20 columns. Lastly, the WDICountry.csv, a relatively compact 152 KB file, providing distinctive characteristics of each country. This file features 265 records and 31 columns. These additional attributes about country and indicator series are all nominal data.

3. Hypothesis Exploration.

While analyzing, I encountered a set of indicators that piqued my interest, particularly those related to women's statistics. Among them, the indicator that stood out prominently was the adolescent fertility rate. Coming from India, I know one of the most common problems many developing countries are facing is early marriage and teen pregnancies. I believe the lack of education can be a contributing factor in elevating the risk of early adolescent pregnancy, primarily because these individuals may find themselves outside the protective school environment. Conversely, in cases where a school-going girl becomes pregnant, her education may be prematurely interrupted, potentially affecting her future job prospects. Better infrastructure and economy play pivotal role in shaping up future of citizens at this adolescent age.

Hypothesis: I believe there is a correlation between adolescent pregnancy, their education accessibility and the overall country's economic status.

I explored data from WDISeries by creating a dummy matrix as seen in Figure 3.1, in Tableau desktop with few attributes that would help me understand the meaning, aggregation methos and any limitation to be considered while using the indicators. I decided to use this matrix throughout my exploration to understand the context of indicators while relating to each other or while adding new ones to my exploration.



Indicator Name	Long definition	Aggregatio..	Limitations and exceptions
Adolescent fertility rate (births per 1,000 women ages 15-19)	Adolescent fertility rate is the number of births per 1,000 women ages 15-19.	Weighted average	Null
Teenage mothers (% of women ages 15-19 who have had children or are currently pregnant)	Teenage mothers are the percentage of women ages 15-19 who already have children or are currently pregnant.	Weighted average	Null
GDP growth (annual %)	Annual percentage growth rate of GDP at market prices based on constant local currency. Aggregates are based on constant 2015 prices, expressed in U.S. dollars. GDP is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the va..	Weighted average	Each industry's contribution to growth in the economy's output is measured by growth in the industry's value added. In principle, value added in c..
Adolescents out of school, female (% of female lower secondary school age)	Adolescents out of school are the percentage of lower secondary school age adolescents who are not enrolled in school.	Weighted average	The administrative data used in the calculation of the rate of out-of-school children are based on enrolment at a specific date which can bias the re..

Figure 3.1: Dummy Intermediate table to explore indicators as and when needed.

To work towards mt hypothesis, initially, I attempted to visually encode the indicator 'Teenage mothers (% of women ages 15-19 who have had children or are currently pregnant)' using shapes and colors through standard map, given the availability of geospatial data. However, I encountered significant data sparsity within this column. Consequently, I opted to make a second attempt, this time focusing on a similar indicator, the 'Adolescent fertility rate (births per 1,000 women ages 15-19)'. The data proved to be more favorable for this visualization. The difference in data availability can be seen in figure 3.2 using choropleth maps.

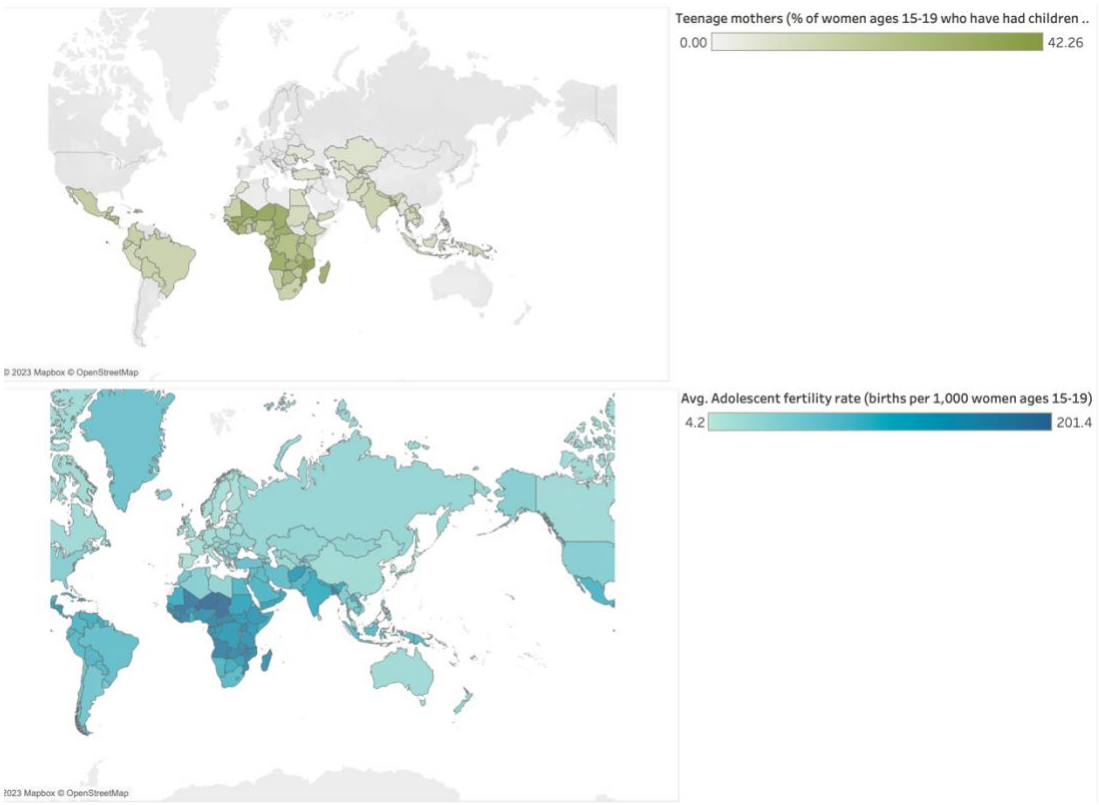


Figure 3.2: Map showing data sparsity among indicators.

Subsequently, I incorporated the percentage of female adolescents not currently enrolled in school, as illustrated in Figure 3.3. However, certain data points could not be plotted due to the logical grouping of countries. For instance, valid quantitative data existed under categories such as "European Union" or "Caribbean small states," representing specific geographical regions, which are not directly recognized by standard maps. In total, there were 49 instances of such null values that I had to filter out to generate the visualization depicted in Figure 3.3. Given that the primary objective was to substantiate the hypothesis by establishing correlations between the indicators, and considering our learning from class that filtering out valid data could potentially skew correlation trends, I made the decision to modify the visual representation.

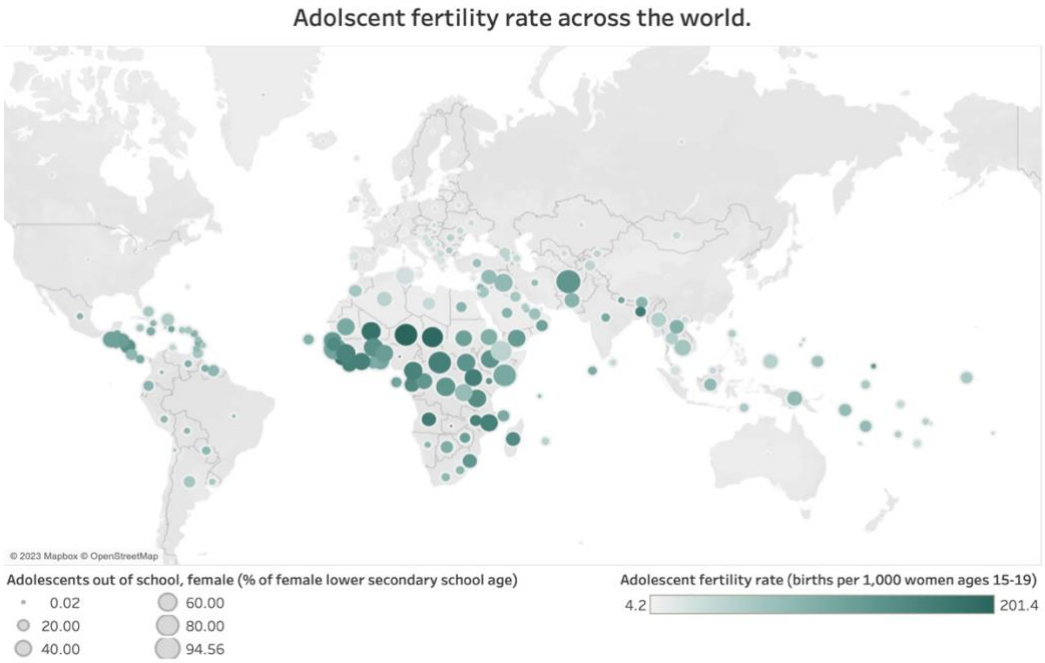


Figure 3.3: Map showing the Adolescent fertility rate vs adolescents out of school.

I successfully included all the necessary indicators for my hypothesis in a single line graph with dual axis (figure 3.4) due to variations in the scales of these indicators. This ensured I am not missing any valid data that could not be encoded geographically. While this visual representation effectively displayed the trends over time, it did not adequately illustrate the correlations between these factors. Also, the GDP indicator did not indicate any visible relation to other 2 variables. As a result, I had to revise the encoding once again.

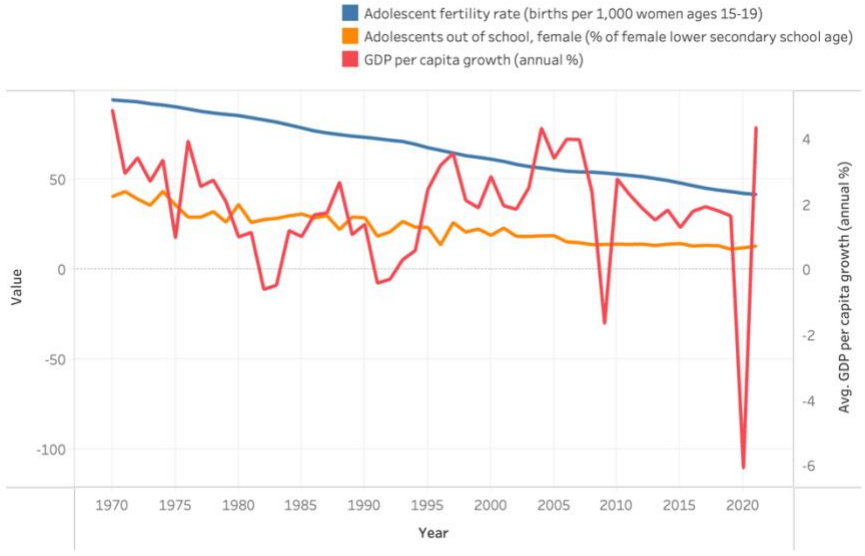


Figure 3.4: Line Graph showing the trends of Adolescent fertility rate, GDP per capita growth and Adolescent out of school.

From the lecture notes, I decided that best way to exhibit the multivariate correlation is via a scatter plot, and I decided to encode the variable one by one. To begin with, I encoded the two variables as shown. While the visual representation seemed to align with one part of my hypothesis, the sheer volume of data points proved challenging for human cognition, as it encompassed data from multiple countries across various years. This is seen in the figure 3.5. Parallelly, I now had to look for another economy related indicator as I dropped 'GDP per capita growth'.

Figure 3.5: Graph showing the positive correlation between adolescent pregnancy and adolescent out of school.

I needed to devise a strategy for grouping or categorizing the countries. Upon reviewing the WDICountries dataset, I determined that utilizing "Region" or "Income group" might be an effective way to segment the data. However, given that teenage pregnancy is primarily linked to socio-economic conditions rather than geographical location and the second part of my hypothesis connects to the country's economy, I opted to use the "income group" as the basis for categorizing the countries as it directly relates to economy and provides a wholistic economic status of a country. The resulting visualization appeared cluttered (Figure 3.6) and may not have been intellectually stimulating as intended. However, it still aligns with the hypothesis as the income negatively correlates to the teenage pregnancy.

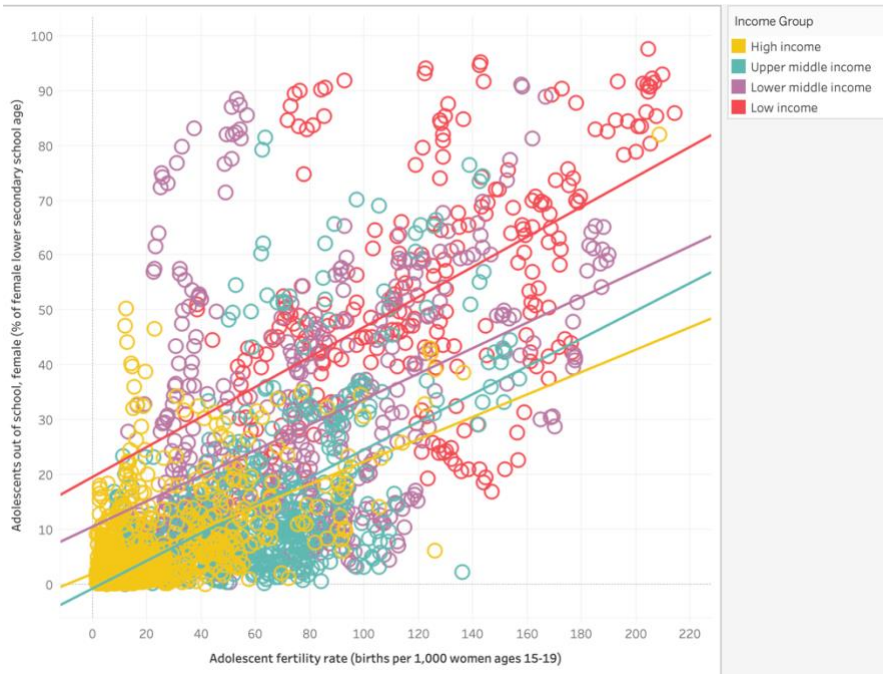


Figure 3.6 - Showing the trend in clusters of income group.

As I revisited indicator and attributes while looking for ways to prove my hypothesis, I came across the indicator "Mortality rate, neonatal (per 1,000 live births)". Although teenage pregnancies are very risky putting both the mother and the infant at risk. I was not fully convinced about the correlation, so I decided to explore it further. I brought in the fourth variable into the graph in the most possible way, as size factor. Surprisingly they are interrelated as seen in the figure 3.7. I removed trend lines to improve data-to-ink ratio.

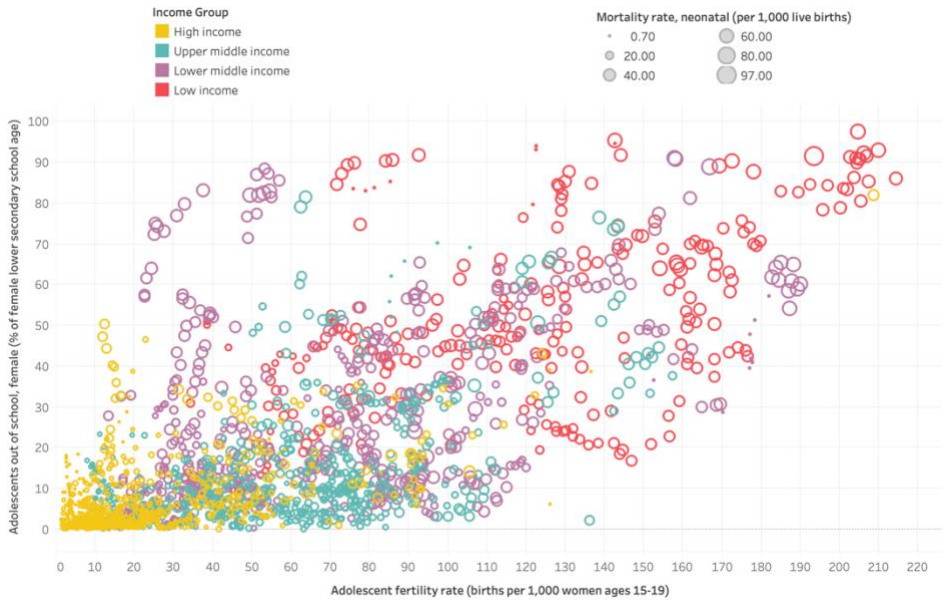


Figure 3.7: Exhibiting the relationship with Neonatal Mortality rate.

4. Conclusion

For the final visualization, I chose to create a visual that exclusively featured income groups, omitting the more granular year-level details as the goal was not see the trend over time. I sought to determine whether the observed correlation remained consistent. The resulting visualization is graphically excellent and upholds the graphical integrity (figure 4.1). With few outliers, low-income countries experienced a higher incidence of adolescent pregnancies, which, in turn, led to an increase in neonatal mortality rates. Of particular significance is the direct relationship between fertility rates and the phenomenon of adolescents being out of school, as they can mutually influence one another.

As I did the exploratory analysis, I used various options available in Tableau desktop and Tableau prep to construct enlightening visual representations in the support of my hypothesis. Leveraging the options in the tool, I was able to validate every part of my hypothesis. I would say Tableau is an immensely user-friendly tool for data visualization that comes with an extensive help document and community that answers most technical questions. With multiple iteration and a little help from Tableau document I was able to present the information and relationships in a clear and easily comprehensible manner.

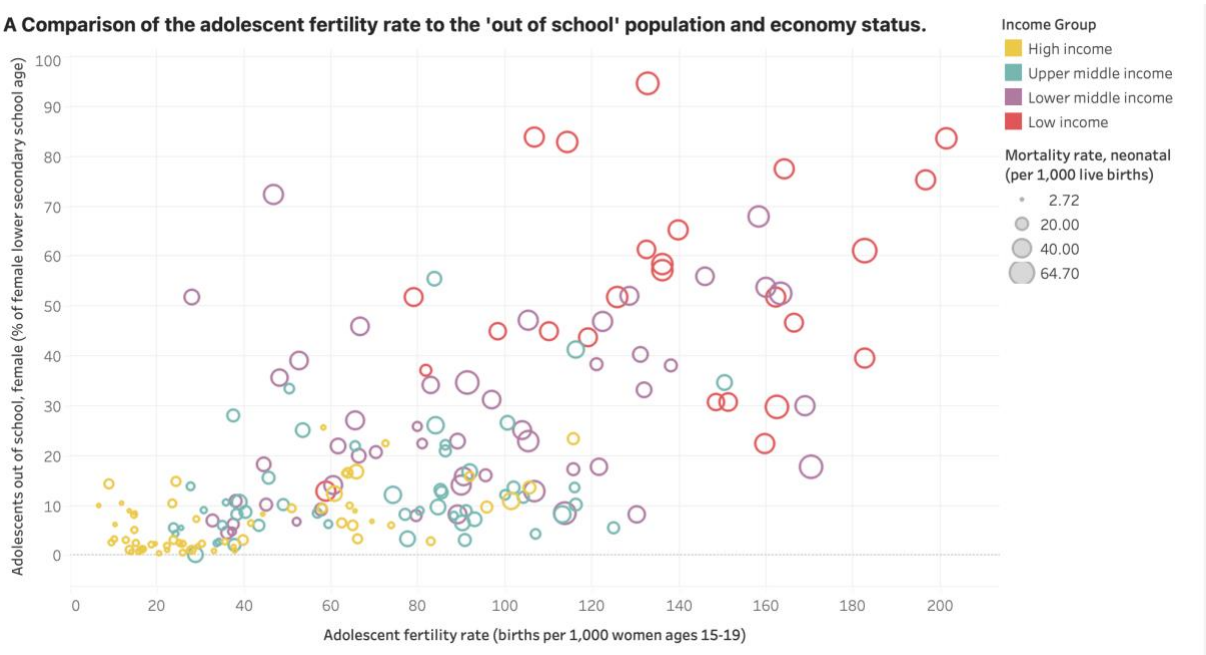


Figure 4.1: A visualization depicting relationship between adolescent fertility rate, access to primary education, economy status and neonatal mortality rate.

In addition to acquiring the skills for data cleaning and encoding, I also developed the knack of focusing only on the required subset of data that is relevant to the problem and the purpose. This skill comes in handy while working in the real-world application where we are often presented with large datasets. To conclude, I would assert that both the quality and quantity of data, in conjunction with the choice of tools, exert a significant influence on the narration of any visual story.

5. References

1. "Data Catalog." World Development Indicators. Accessed November 7, 2023. https://datacatalog.worldbank.org/search/dataset/0037712.

2. "Adolescent Pregnancy." World Health Organization. Accessed November 7, 2023. https://www.who.int/news-room/fact-sheets/detail/adolescent-pregnancy.

3. Heer, Jeffrey, and Ben Shneiderman. "Interactive Dynamics for Visual Analysis." *Communications of the ACM* 55, no. 4 (2012): 45–54. https://doi.org/10.1145/2133806.2133821.

4. Lecture Slides for Data 511 Au 23: Data Visualization for Data Scientists

5. Few, Stephen. *Now you see it: Simple visualization techniques for quantitative analysis*. Oakland, CA: Analytics Press, 2015.

6. Fry, Ben. *Visualizing data*. Sebastopol (Calif.): O'Reilly, 2008.

7. "Build a Scatter Plot." Tableau. Accessed November 7, 2023. https://help.tableau.com/current/pro/desktop/en-us/buildexamples_scatter.htm.