A4: Exploratory Visual Analysis

Logan O'Brien

Introduction:

In this paper I outline the work and findings I performed in my exploratory visual analysis. In the project, I explored the World Development Indicators (WDI) dataset to learn what attributes of a country are highly correlated with fertility rate around the world. During this process, I cleaned the data and made initial observations about it, consulted several sources for guidance, calculated the correlation between various indicators in the dataset, and visualized my findings.

Data Profile:

The WDI dataset comes from the World Bank Group. From my brief investigation, I believe World Bank Group is a worldwide nonprofit supporting the development of client countries by helping them improve their data measurement and recording capabilities. The dataset, available here: https://datacatalog.worldbank.org/search/dataset/0037712 contains a csv file called *WDIData.csv*. The file is approximately 204MB and contains 67 fields/columns and 383,572 rows. The fields consist of a country name, country code, indicator name, indicator code, a column for each year from 1960 to 2021, and an empty column. Examining the data variables and dimensions, I observed that *Country Name*, *Country Code*, *Indicator Name*, and *Indicator Code* contain categorical data represented as strings and the year columns from 1960-2021 contained ratio data in the form of decimal values. *WDIData.csv* is organized such that for each row, a given country and indicator name are specified and the indicator value for all years fills the remainder of the row. If no indicator value was recorded for a certain country and year, the value is null. Upon inspection I determined that there are 266 countries represented and 1,442 unique indicator names in the dataset.

After opening the .csv, in Tableau Prep, one thing I noticed right away was that a lot of the indicator values were missing. Over time, though, more information was collected resulting in fewer missing values. For instance, in 1960 only about 9% of the rows had values, whereas in 2018 53% of all rows had values. The sheer volume of missing values presented challenges when working with the dataset. Additionally, I also noticed that it made more sense to arrange the data so that for a given year and country, you could see all corresponding indicator values. As such, I planned to pivot the year columns into rows (to form a single year column) and the indicator values rows to many columns. Although the *WDIData.csv* dataset was not perfect, the wide array of diverse indicators contained within presented an exciting treasure trove to explore.

Question Exploration:

After digging into the dataset, the question I chose to answer was: **In 2020, what interesting indicators are strongly correlated with the fertility rate, and what does that correlation look like?**

While the crux of my exploratory analysis centered on this question, my work began before the question was framed. I started my analysis by cleaning the *WDIData.csv* dataset in Tableau Prep, where I eliminated unnecessary columns and pivoted the data as described above. Initially I had some difficulty performing the pivot transformations since the dataset was quite large, but with some ingenuity, I succeeded. Additionally, I learned through a data cleaning demonstration video that Professor Nathan recorded using the same dataset, that the WDI dataset contained several country groupings that were not actual countries, so I removed.

I also received some insight from Zach Price, one of my classmates. He shared that for part of his exploratory analysis he used a correlation matrix to identify which indicators for a given year were most strongly correlated with a specific indicator – I think life expectancy. He then visualized a subset of the indicators and their correlation via a parallel coordinate plot (PCP). I liked this approach and decided to incorporate it into my own project. However, I selected a different indicator to focus on, namely the *Fertility rate, total (births per woman)* indicator. Additionally, I wished to go beyond what Zach shared with me by comparing multiple years and observing how that impacted the correlation between the same set of indicators. Note: for the sake of brevity, when I refer to the fertility rate for the remainder of this paper, I am specifically referring to the *Fertility rate, total (births per woman)* indicator.

After determining the question I wanted to answer and the procedure to answer it, the first thing I needed to do was decide which years I would focus on for my analysis. After some basic data exploration in Tableau, guided by Professor Nathan's recording, I determined that the fertility rate data was captured between 1960 and 2020. I decided to focus on the endpoints of that range to see what interesting comparisons I could make. Now, before I could construct my parallel coordinate plots, I needed to have a means of narrowing down the list of 1,442 distinct indicators to several interesting ones that were strongly correlated with fertility rate. To facilitate this, I built 2 correlation matrices using a mixture of Excel and Python. The first, built on the WDI dataset filtered for 1960 and the other built from data filtered for 2020. After building the tables, I examined the 1960 table to identify some of the indicators with the strongest correlation (positive or negative) to the fertility rate and found several interesting ones to plot in my PCPs.

At this point, I knew which indicators I wanted to examine in parallel coordinate plots. I began by constructing a PCP for the 1960 data in Python using the *plotly.express* package (see Figure 1 below). Unsurprisingly, the diagram displayed strong correlation between the selected indicators. However, there were several issues with the diagram. First, the column headings were unreadable due to overlapping labels. To address this, after researching the problem, I decided to try another PCP function. In the *plotly.graph_objects* package, I used the *Parcoords()* function to rename the column axes headings to make them more readable and generated a new plot (Figure 2). Comparing the two results, the second method mostly fixed the heading overlap, and I later learned how to angle them to completely resolve the issue.

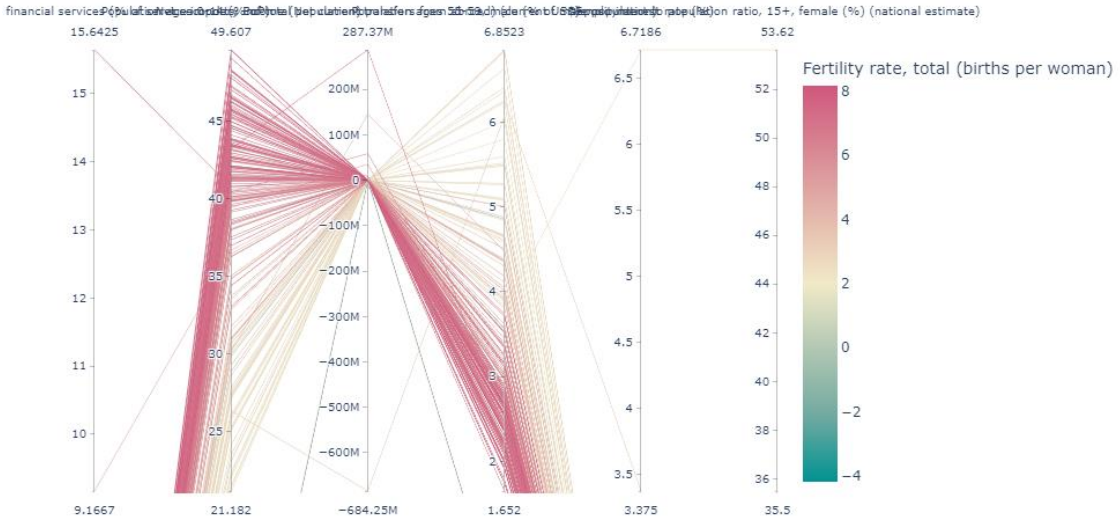

Figure 1: First PCP for the 1960 Data

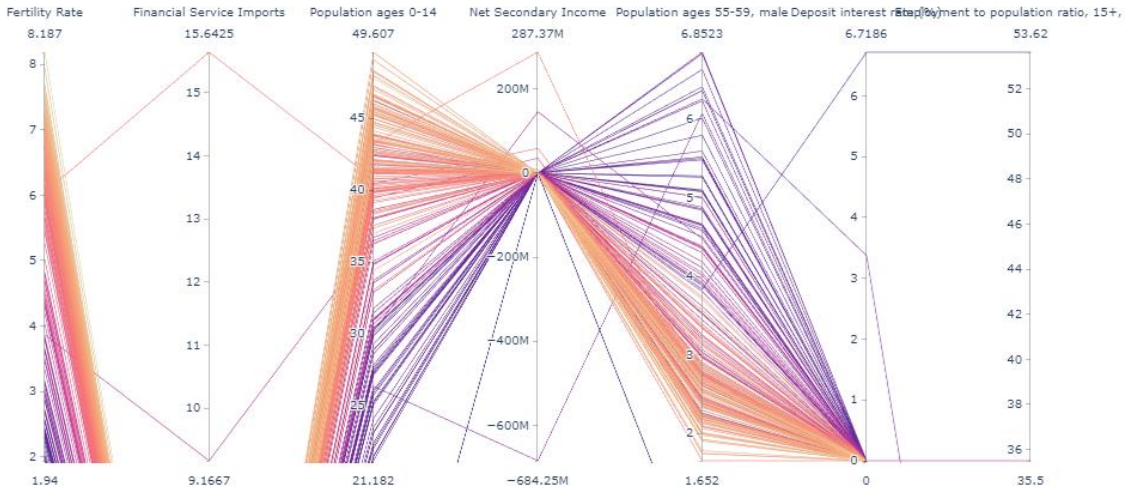

Figure 2: PCP Built via *Parcoords()* Function

Another significant issue I identified in both PCPs was that a lot of the lines appeared to go below the bottom of the diagrams. Through some trial and error, I realized that this was because there were null values present in the data. Both PCP plotting functions I used treat null values as zero. However, for many of the columns, the minimum value in the dataset corresponding to that indicator was greater than zero. As a result, many of the lines dropped below the diagram. By looking at Figure 2, it is clear that the indicators labelled *Financial Service Imports*, *Deposit interest rate (%)*, and *Employment to population ratio, 15+, female* are composed almost entirely of missing (null) values. Upon reopening the .csv for the 1960 data, I confirmed that this was indeed the case. To address this, I decided to look again at the

correlation table I built in Excel for the 1960 data and select some new indicators that contained few null values. The new PCP is shown in Figure 3 below.

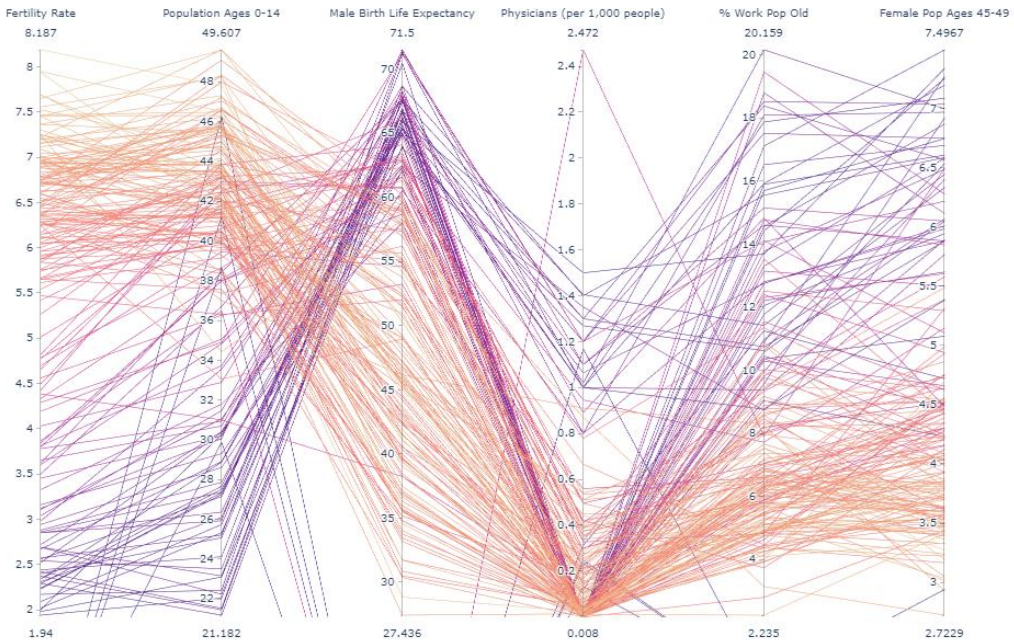

Figure 3: PCP With New Indicators for 1960 Data

After visualizing these correlation relationships between indicators in the 1960 data, I wanted to visualize the 2020 data in the same way. Initially, I started with the same set of indicators as the previous visualization (Figure 3) to compare how the correlation between the indicators changed from 1960 to 2020. The result of this plot is in Figure 4 below. One significant difference I noticed was that interestingly, the indicators seemed to be more strongly correlated in the 2020 parallel coordinate plot (this can be seen by the fact that there are less lines crossing each other). Another change I picked up on was the dramatic "pinching" of the lines on the *Physicians (per 1,000 people)* axis of the 2020 plot. I quickly realized that this was because the indicator was almost completely empty. I found this quite interesting as prior to this point, I had assumed that the completeness of the data recorded for the various indicators would improve significantly from 1960 to 2020, but this is clearly not the case for this indicator. To address this issue, I visited my 2020 correlation table and selected several new, highly correlated indicators that were not available in the 1960 dataset and plotted a new PCP (see Figure 5 below).

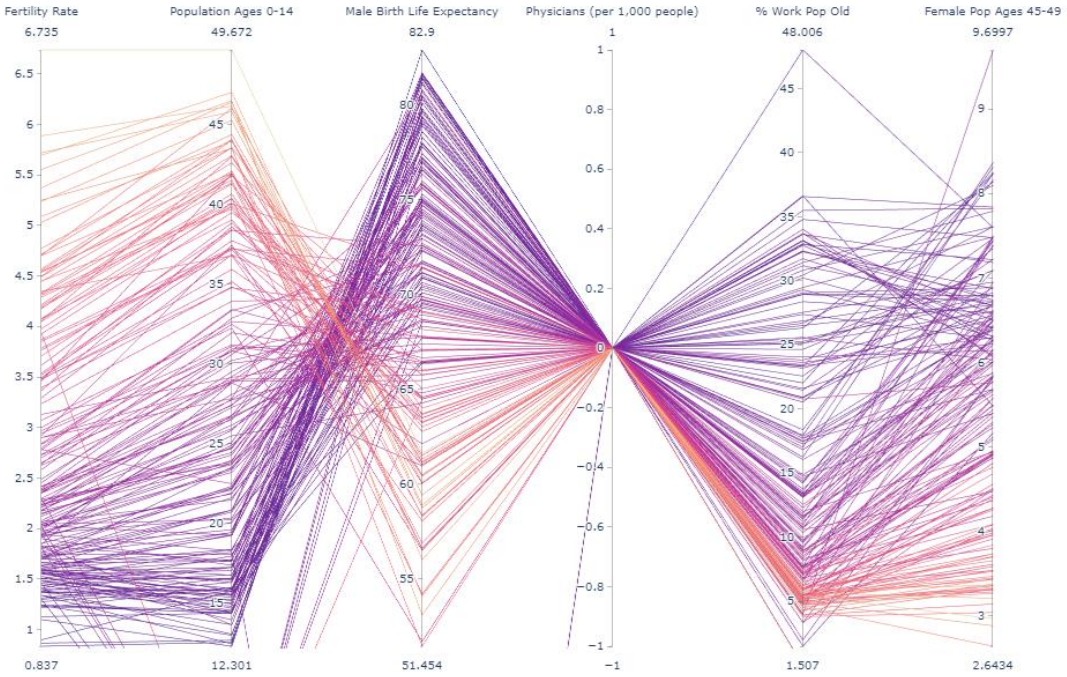

Figure 4: 2020 Data PCP With Same Indicators as the Updated 1960 PCP

Correlation Between Fertility Rate and Other Interesting Indicators in 2020

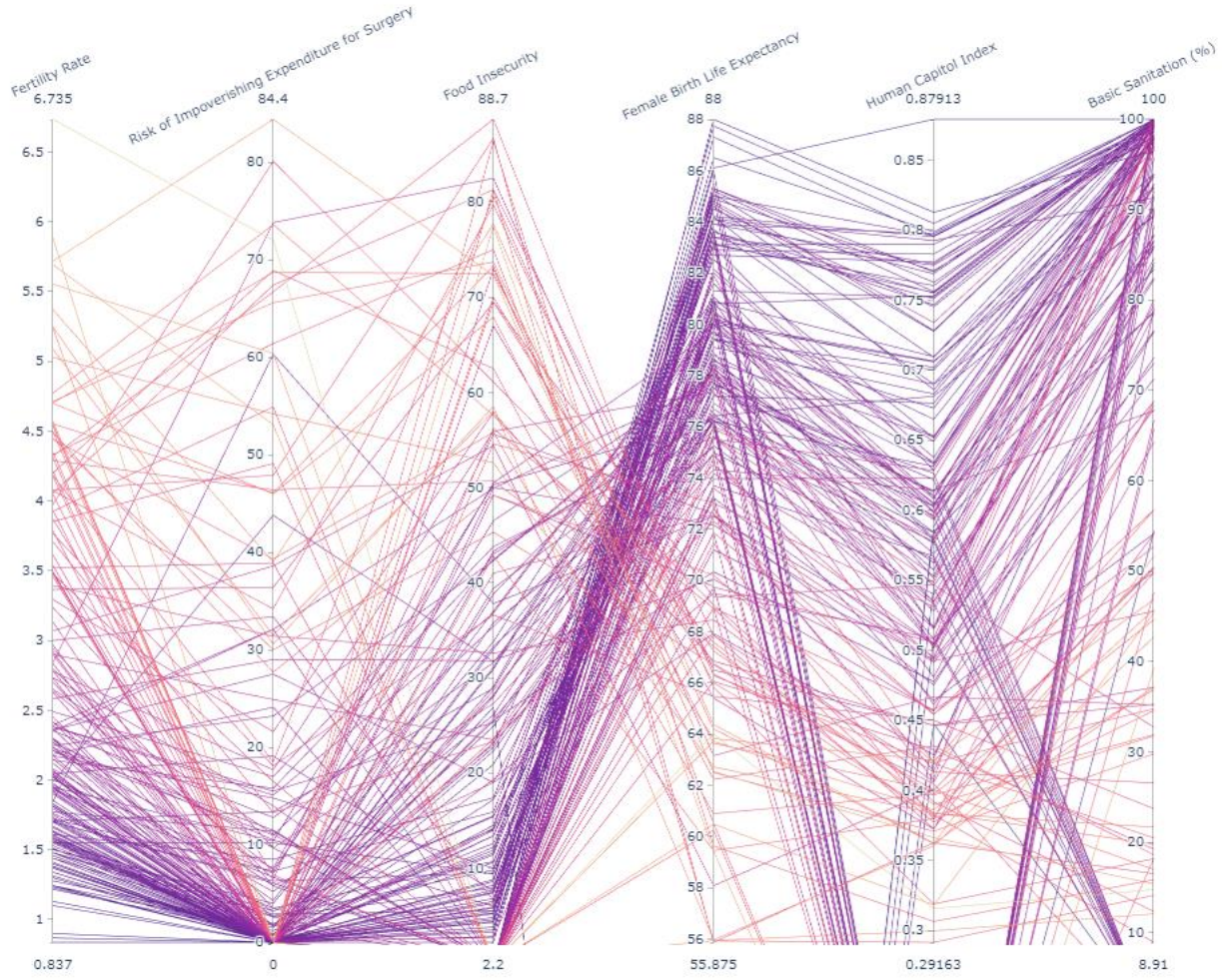

Figure 5: Final Parallel Coordinate Plot

The parallel coordinate plot presented above in Figure 5 represents the culmination of my efforts in this exploratory analysis project. You may recall that the question I sought to answer was "In 2020, what interesting indicators are strongly correlated with the fertility rate, and what does that correlation look like?" I believe this diagram nicely addresses this question by showing the correlation between *Fertility Rate* and five other interesting indicators. The visualization achieves this through a series of clever operations. First, each line drawn through the vertical axes represents a row of the dataset. Additionally, the hue of the lines is encoded on the relative range of possible values of *Fertility Rate.* In this case, the hue scale goes from purple to red to yellow as the fertility rate increases. Through this feature, viewers can discern how the other indicators' values are correlated with fertility. For instance, we can see that *Risk of Improving Expenditure for Surgery*, and *Food Insecurity* are positively correlated with fertility, because the upper ranges of those columns are filled with orange lines representing medium-high fertility rate. By similar reasoning, the remaining three indicators are negatively correlated with fertility rate.

Discussion/Conclusion:

In this exploratory analysis project, I used the World Development Indicators dataset to examine what datapoints about a country are strongly correlated with fertility rate. Along the way, I cleaned and explored the data, created correlation matrices for the dataset filtered by 1960 and 2020, and then generated a series of parallel coordinate plots, culminating in a plot highlighting the correlation between fertility rate and other interesting indicators for 2020. I learned that the following were some of the indicators highly correlated with fertility:

- Risk of impoverishing expenditure for surgical care (% of people at risk)
- Prevalence of moderate or severe food insecurity in the population (%)
- Life expectancy at birth, female (years)
- Human capital index (HCI) (scale 0-1)
- People using at least basic sanitation services (% of population)

Additionally, I gained a greater appreciation for the strengths and weaknesses of the various tools I utilized during this project. I used Tableau for most of my data cleaning and initial data exploration and was mostly content with its performance. My only complaint was that it seemed to struggle when working with large (>2GB) data sources while performing multiple transformations in the same flow. In addition to Tableau, I used Excel for general .csv exploration and for some correlation matrix work. I quickly realized however, that Python can perform intense computing tasks much more efficiently than Excel. Python was also very useful in creating the parallel coordinate plots. When I started the project, I expected to do most of my analysis in Tableau, however, I transitioned a large portion to Python amid the project and I now have a much greater appreciation for Python. I thoroughly enjoyed completing this exploratory visual analysis and learned a lot along the way.

References/Acknowledgements:

- I found the following webpage very helpful in creating my PCPs: https://www.analyticsvidhya.com/blog/2021/11/visualize-data-using-parallel-coordinates-plot/