November 10th, 2021

A4: Exploratory Visual Analysis

DATA 511A

ROHIT LOKWANI

1. Objectives/Goals

This assignment is about using visualizations to explore data and pose questions. Effectively translating data into visualizations to pose new questions and answer them is extremely valuable for human-centered designers.

In this assignment, you will use Tableau, Python, or another tool of your choice to perform exploratory visual analysis on a real-world dataset we will provide you. The goal is to gain practice formulating and answering questions through the creation of graphs during the process of visual analysis.

Look at a dataset we provide → come up with a question → iteratively answer the question with visualizations → write up your process, describe your question and corresponding visualizations

Detailed Assignment Description: Link

2. Introduction

Good data visualizations have good data, the right choice for visualization, the color or information are simple and explicit, the data are accurately represented, and there is consistency in scales [1]. With these principles in mind, we work with the World Development Indicators dataset from the World Bank [2]. It has time-series data for a majority of the world's countries from 1960-2020. These development indicators together belong to either Health, Policies, Education, or other categories. The dataset in total has 383838 data points and 65 columns. The file size is 195 MB.

As stated in the assignment description, we start by iteratively framing and reframing our hypothesis. Here, we correlate education with socio-economic status, health, and family decision-making of women in Senegal throughout 2000-2019. We end up drawing a visualization dashboard with a set of multivariate visualization in a grid that justifies the hypothesis. The tools used for this assignment are Microsoft Excel, Python, and Tableau primarily.

3. Data Profile

Strategic search [7] is used to extract the data. The World Bank Web tool was used to create the required data profile from the raw data. As the tool had the functionality to filter out and look at the data manually, it was easy to get the required parameters and get a sense of the correlations, patterns, and trends. For creating the profile for our hypothesis, we filter the data for Senegal from 2000-2019. The subset of the population is fixed to women. The significant features include Health, Employment, mortality, fertility, and birth rates. The final dataset has 20 rows and 71 columns. The file size is reduced to 17 kB. Apart from the series name which is a nominal variable and years as the time-series data, the rest of the data comprises quantitative (ratio) variables.

The data quality is good. On manual checking and visual analysis, this particular subset of data has fewer missing values and almost no anomalies. Some of the important missing data include the average literacy rate and women in senior and middle management. This could have helped make the hypothesis stronger. Some of the curious things that were observed: Women were almost equally likely to accept domestic violence by their husbands in Senegal in the last 20 years. These trends changed quite a bit in other countries during that period.

3. Question Exploration

3.1 Question/Hypothesis

I predict that schooling had a positive correlation with Senegalese women's socio-economic lives (employment, family decision-making) and health during 2000-2019.

3.2 Process Discussion and Description

3.2.1 Data Extraction, Cleaning and Manipulation Process

The data filtering and visual analysis were done following Shneiderman's visual information seeking mantra [6]: Overview first, zoom and filter, then details-on-demand. The website tool gives the initial overview, allows to zoom and filter. It also provides metadata for each of the fields to know the domain, and understand the collection methodology, which is an important step in data preparation.[7]

I looked at the dataset on the whole initially and decided to start with an initial idea, which was the impact of education on women's employment and general health. To focus specifically, I decided to reduce the dimensions of the data and bring down the number of features. Skimming through the dataset helped develop a hypothesis about the increase in the percentage of women going to school would have correlated with their socio-economic, health, and decision-making skills and awareness in general. Since the data had a lot of missing data points for the entire range of years, we checked the data for completeness which was mostly in 2000-2019, hence decided to stick to that and make a better sense of recent trends. Then we decided to check the trends in the high-income, middle-income, and low-income countries. Countries like the US did not have a clear trend because of the high literacy rates already. The trends were quite visible in low and middle-income countries like Senegal and India respectively. Due to interesting insights, Senegal was finalized. The hypothesis started with predicting the impact of literacy rate in women on these factors. But due to the unavailability of data on literacy rate and advanced education, the hypothesis changed to schooling. On performing initial EDA, within schooling data, there is a correlation between preprimary, primary, and secondary schooling. We stick to primary due to the completeness of data and use certain factors like primary school completion and secondary school enrollment to support our hypothesis. In a nutshell, after filtering and extraction, an initial EDA was done to reframe the hypothesis, followed by fine-tuning, popularly known as progressive technique. We start from a pattern or exception to dimensionality reduction and investigation of cause in depth [8]. We start by looking for low-hanging fruit, similarities, and oddities by doing the EDA, where our initial histograms did not show any specific outliers

The tools used were the World Bank website tool for filtering and downloading the custom dataset. The second one was Microsoft Excel for cleaning the data. Pandas in python were used for data manipulation and pivoting the rows to columns and vice versa. Tableau was used for plotting visualizations.

3.2.2 Visualization Process

After developing the data profile, the hypothesis revolved around finding a correlation between women's education and self-esteem including employment, health, and awareness against harassment. To start with, it was important to understand how the education trend changed in recent years. The primary completion rate was the most complete field related to schooling and made the most sense as Senegal was a low-income country and advanced studies weren't a common thing. In this visualization process, we intentionally try to avoid any spurious correlations and any sort of predictive modeling or algorithmic transformations.

During the Exploratory Data Analysis, we use lines and points from Bertin's Graphical Vocabulary of "Retinal Variables", and his levels of organization to plot quantitative variables appropriately using position and slope [9]. The final visualization involved a lot more EDA and iterations, out of which we try to present the data in context, highlight only the significant features in the following sections to adhere to the word limits.

3.2.2.1 Initial Exploratory Data Analysis

In figure 1, we start by plotting the primary completion rate in females against years (Time-series) on the X-axis. It shows a trend of increase in the primary schooling completion rate in women during 2000-19

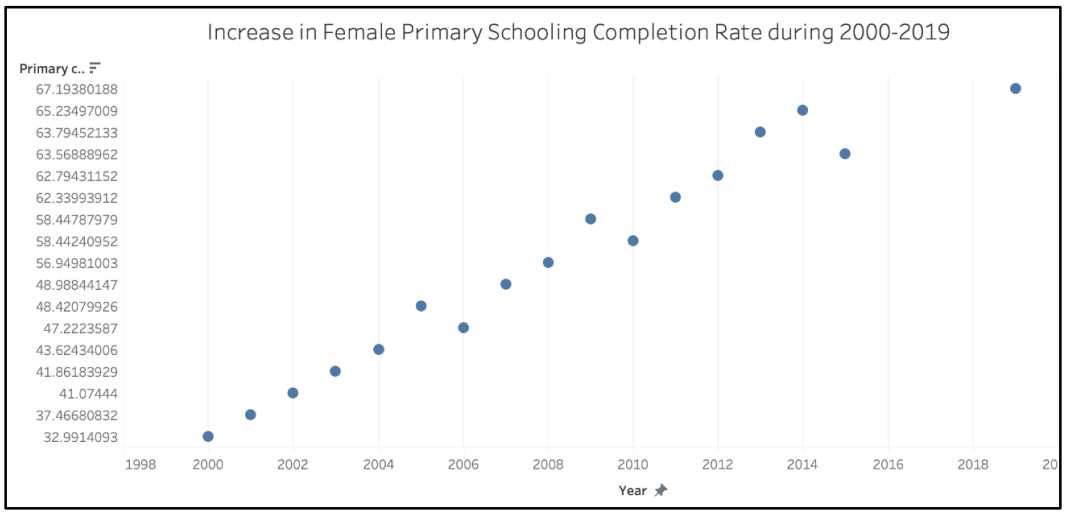

Figure 1: Scatterplot showing the change in the female primary completion rate for 2000-2019

We then check for the enrollment patterns in recent years for primary and secondary education to figure out if a similar trend existed for higher studies as well. The Y-axis in the following scatterplot, indicating quantitative (ratio) variables shows a similar trend in the rise of enrollments in recent years. Although figures 1 and 2, might not be perfectly correlated, they do show some positive correlation. With no

anomalous values, task zero in proving our hypothesis worked well for us. The missing values between 2015-2018 are left out for consistency and integrity.

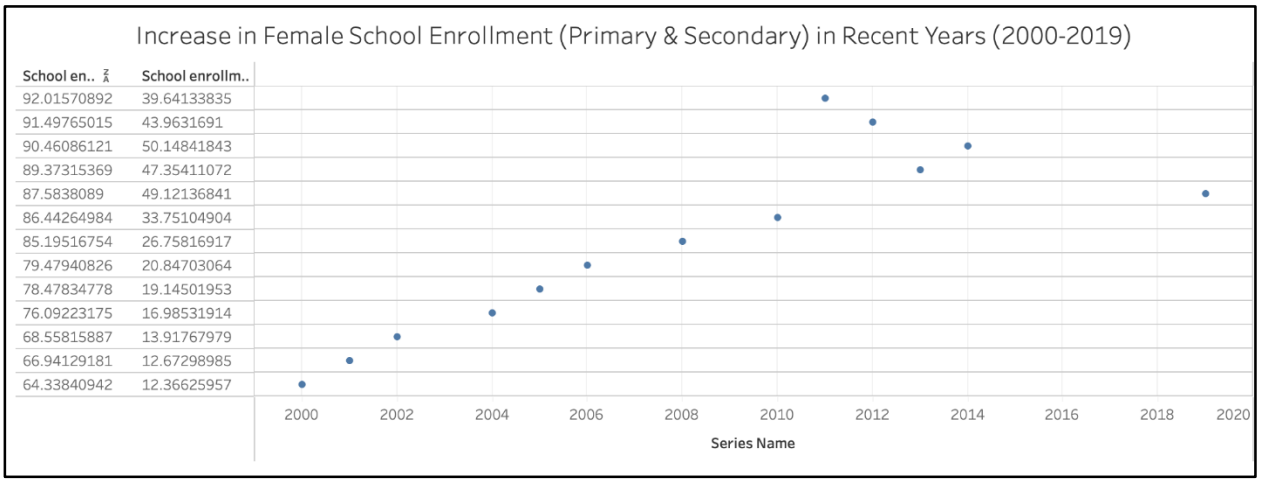

Figure 2: Scatterplot showing the change in female school enrollment rate for 2000-2019

3.2.2.2 Change in Hypothesis

For the actual proof of the hypothesis, we then decided to pick a field on the percentage of teenage mothers and tried correlating it with primary completion percentage, assuming that the trend would decrease. Although the scatterplot did not completely disprove the hypothesis, it did not show a clear trend as there were a lot of missing values. In figure 3, the lighter data points are missing values. We do not mislead the reader using occlusion [8] in the following graph. We just want the focus to be on detecting patterns in the available data.

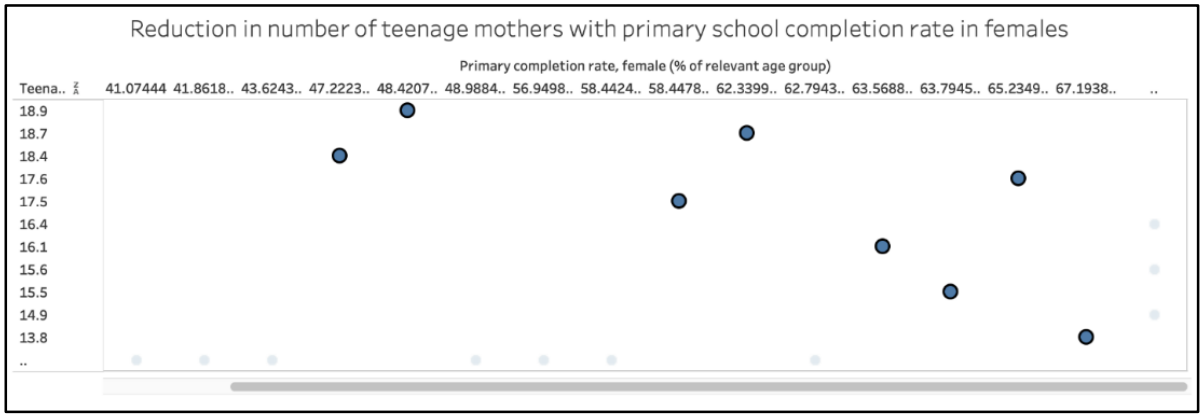

Figure 3: Scatterplot on finding a correlation between female primary completion rate and percentage of teenage mothers

Then the next step was to see with the increase in education rate, did the self-esteem of women grow in terms of tolerating domestic violence? With limited data from 2013-2019, that did not show any particular trends. If more data were available there are chances the trends or patterns would be appropriate. To avoid a perspective distortion due to the unavailability of data, we do not use this graph, instead reframe our

hypothesis and focus on data available at hand. The scope of the hypothesis was then reduced to socio-economic factors and health conditions.

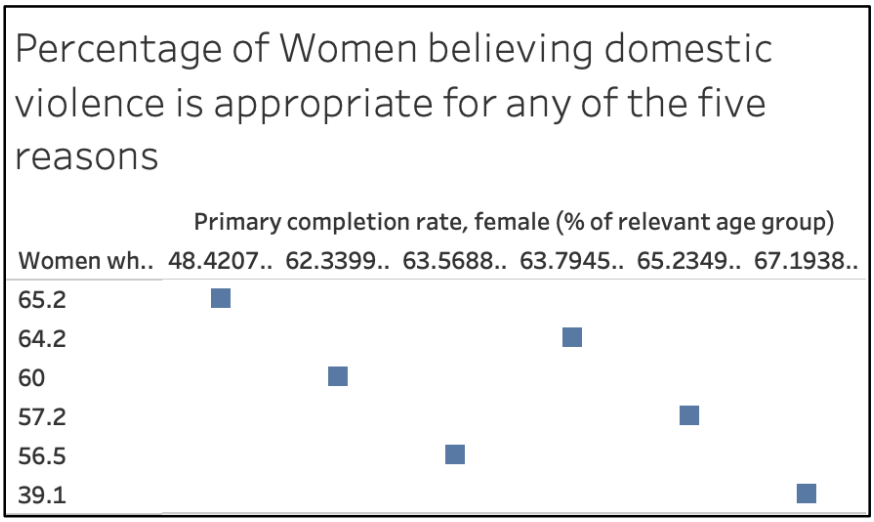

Figure 4: Scatterplot plotting female primary completion rate against the percentage of women thinking domestic violence is not incorrect

3.2.2.3 Contribution of EDA to Final Visualization

This section mentions EDAs that were improved later and finalized into a final data visualization.

As we stuck to finding the correlation between education and socio-economic factors, the employment features were majorly divided into three sectors: industry, services, and agriculture. The industry sector consists of mining and quarrying, manufacturing, construction, and public utilities (electricity, gas, and water). Industry and agriculture showed reducing trends with women's education [2] (Figure 5). The services sector consists of wholesale and retail trade and restaurants and hotels; transport, storage, and communications; financing, insurance, real estate, and business services; and community, social, and personal services [2]. Services showed an increasing trend. This could be quite intuitive as educated women could prefer working in Services over agriculture and industry labor work. Hence, one aspect of the hypothesis was supported.

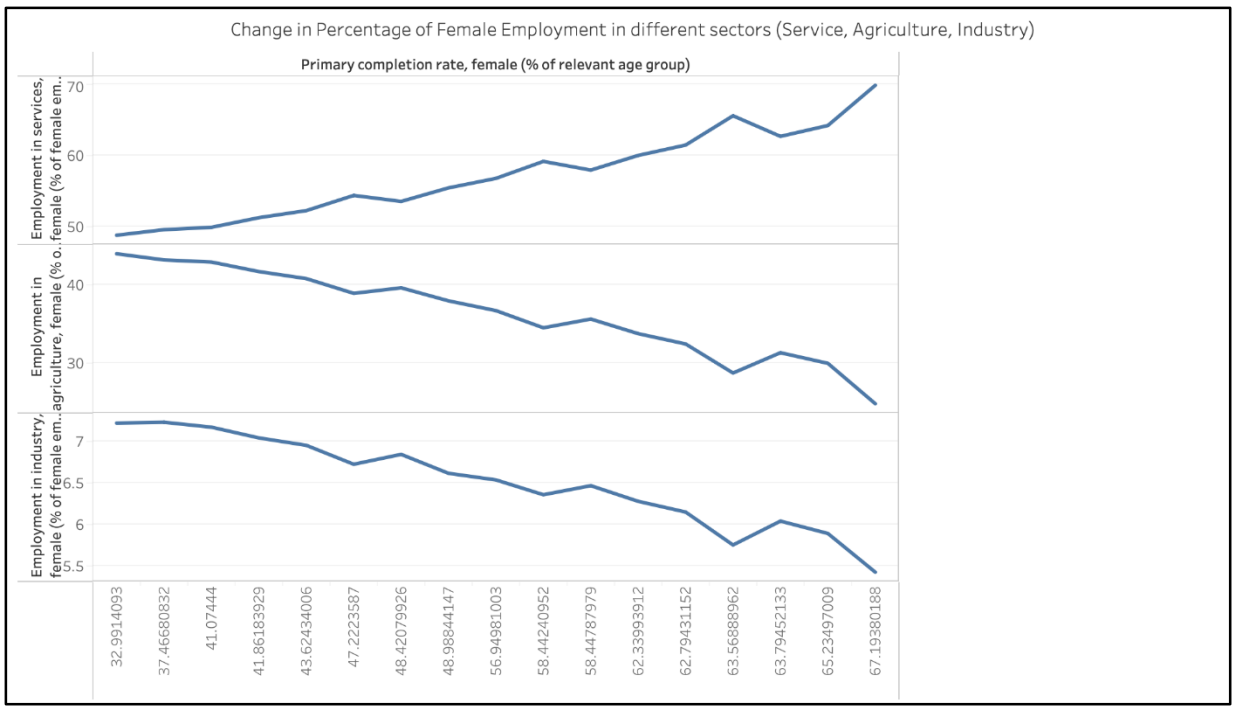

Figure 5: Line chart correlating female employment in different sectors with a primary completion rate

In other aspects of employment, we see that an increase in primary education rate correlated with decreased participation of women in the labor force and increased their number of seats in parliament. That could be related to the self-esteem of educated women (Figure 6). For figure 6, the goal in the final visualization is to reduce the data-ink ratio [10] of this diagram, we replace it with a line graph in the final visualization.

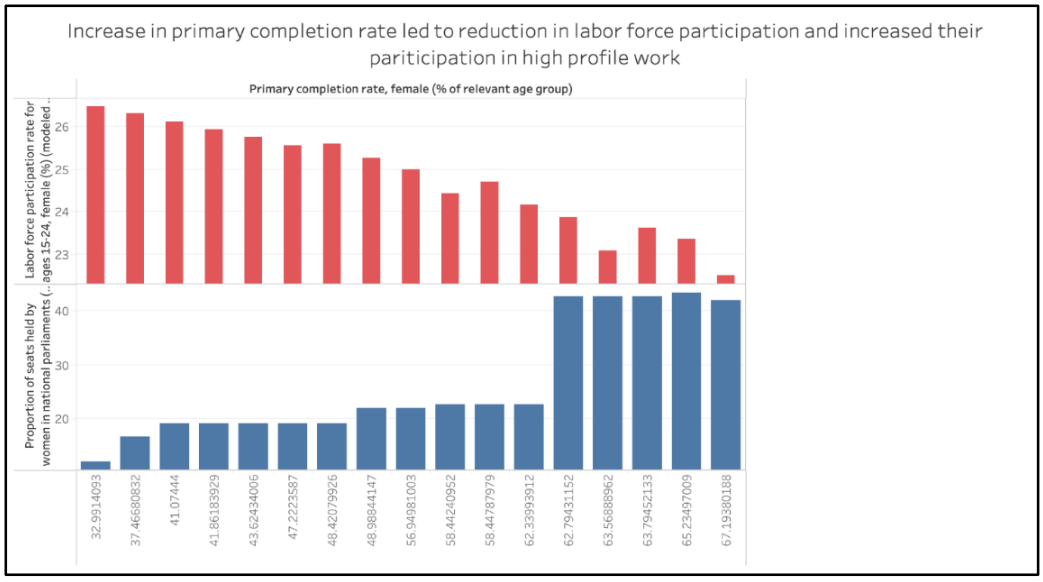

Figure 6: Bar chart correlating female participation in the labor force and Parliament with a primary completion rate

The next step was to determine how women handled the birth of children with the rise in education. This was divided into two main parts: 1. Fertility and Birth rates. 2. Mortality rates of infants and female children. This showed a decreasing trend as well, highlighting that women could be warier of the number of children and well-aware of pregnancy issues and made them open-minded/aware that might have helped them overcome the mortality rates of infants and females under the age of 5.

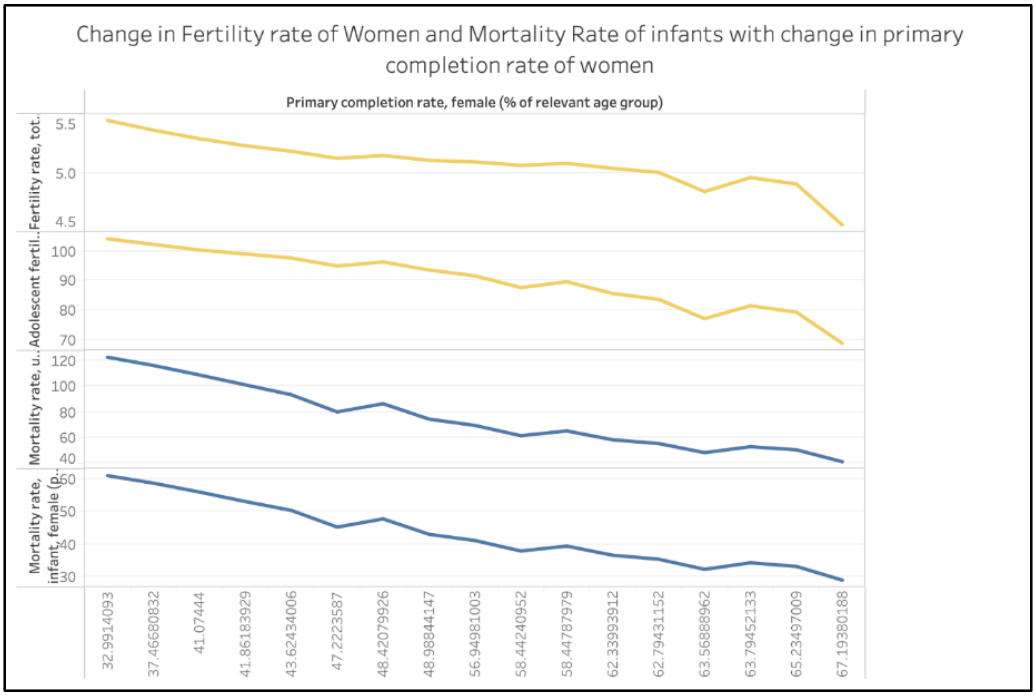

Figure 7: Line chart plotting change in fertility and mortality rates against primary completion rate

The last part of the hypothesis was to see that education was correlated with the physical and mental health of women. The indicators picked were mortality rates due to cancer and CVDs, suicide and unintentional poisoning, and on the other hand prevalence of anemia and HIV.

As we plot HIV, Anemia trends against primary education and, they did show some decrease in recent years although not massive. Similar was the case with mortality rates.

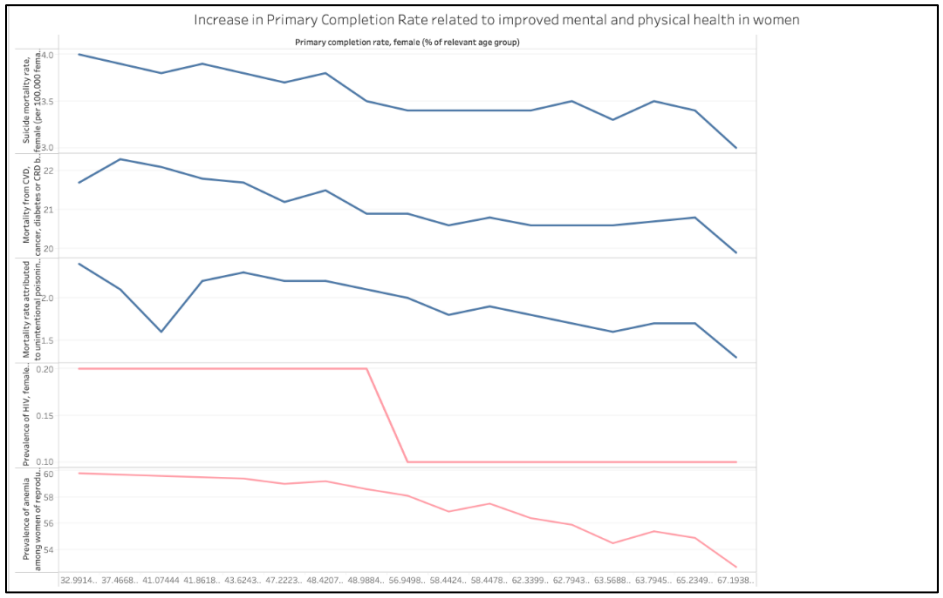

Figure 8: Multivariate line chart plotting correlation between improvement in the mental and physical health of women against primary completion rate

Some of the other trends that were observed while analyzing the data were reduction in the number of females dropping out of school and reduction in child marriages around age 15-18 showed decreasing trends.

We go by Tufte's definition of graphical excellence and combine visualizations based on their effectiveness of being perceived by the end user[10]. Hence, we shortlist the major factors and come up with a final visualization dashboard that supports our hypothesis. We stick to 2D plots and do not use any unjustified 3-D plots as it's hard for humans to interpret. We try to induce the viewer to think about the substance, rather than about methodology, graphic design, the technology of graphic productions, or something else. We avoid the area, volume, and tree-based encodings and prefer to keep the final visualization simple to make them more memorable.

We improve some of the non-uniform scales in the above EDAs in the final visualization and try to improve the overall graphical integrity of the final visualization.

3.3 Final Visualization

Following is the thumbnail view of the visualization with the respect to the document dimensions. We'll break these small multiples down to individual visualizations in a zoomed manner to understand them better. (From Top Left -> Top Right -> Bottom)

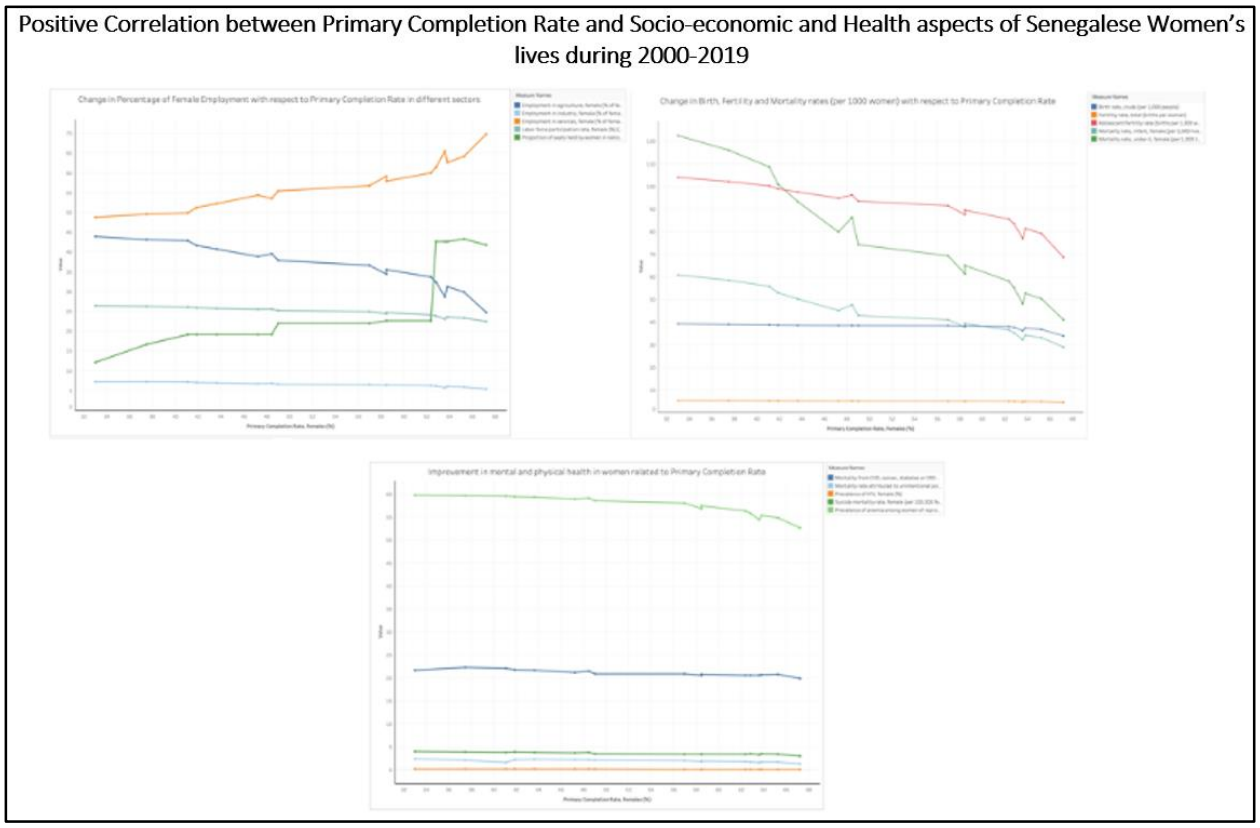

Figure 9: Final Visualization Dashboard Thumbnail

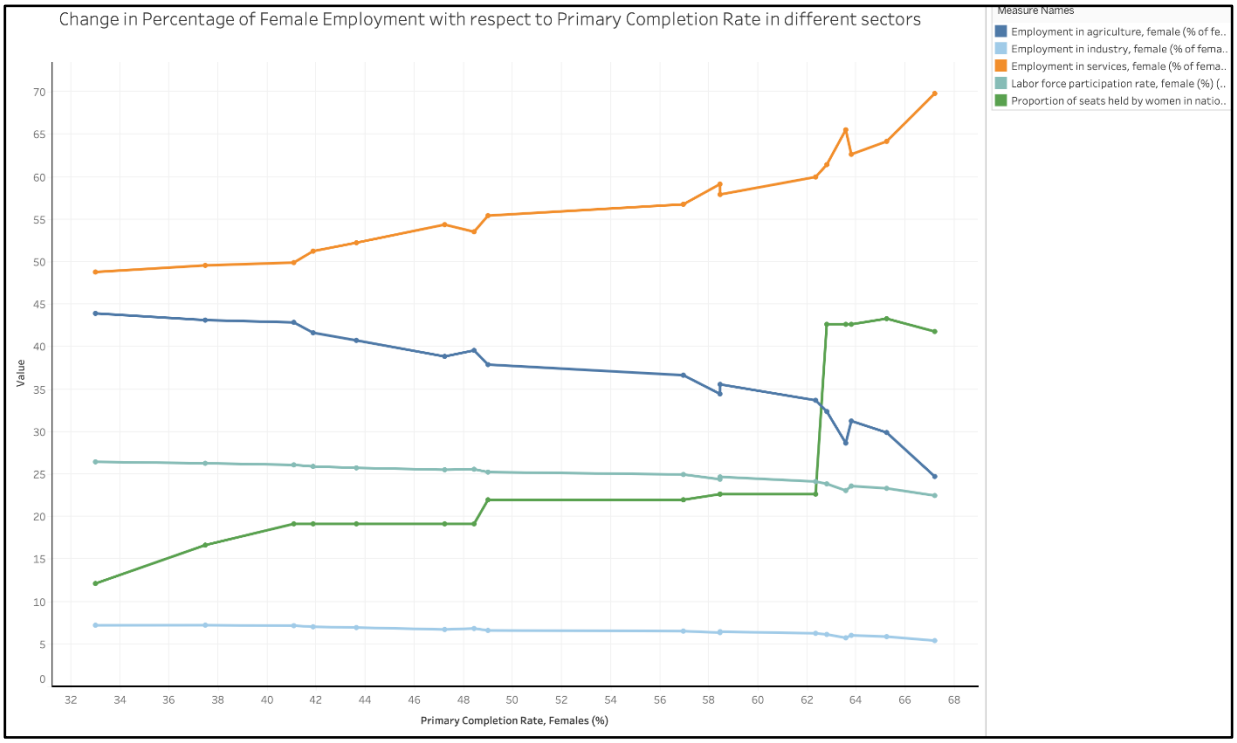

Figure 9.1: (Top-left) Multivariate line chart showing the change in different employment sectors for Senegalese women with the rise in primary completion rate

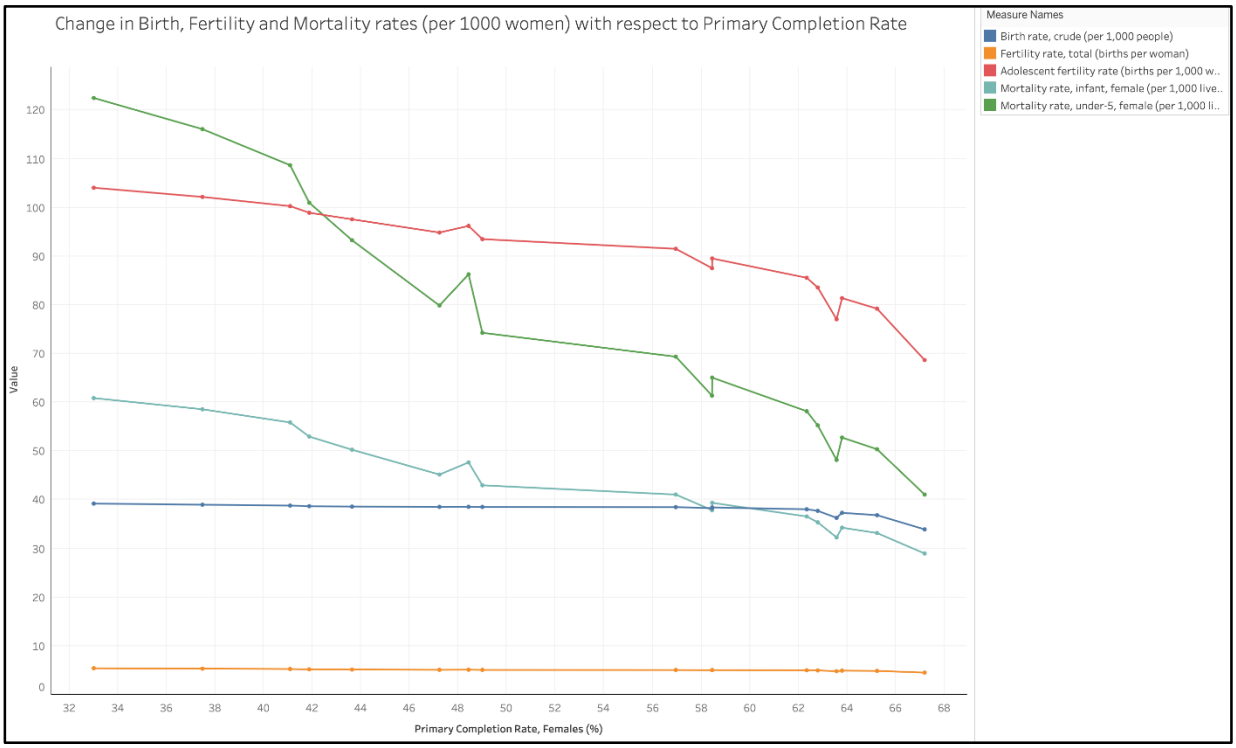

Figure 9.2: (Top-right) Multivariate line chart showing a decrease in birth, fertility, and mortality rates of Senegalese women with the rise in primary completion rate

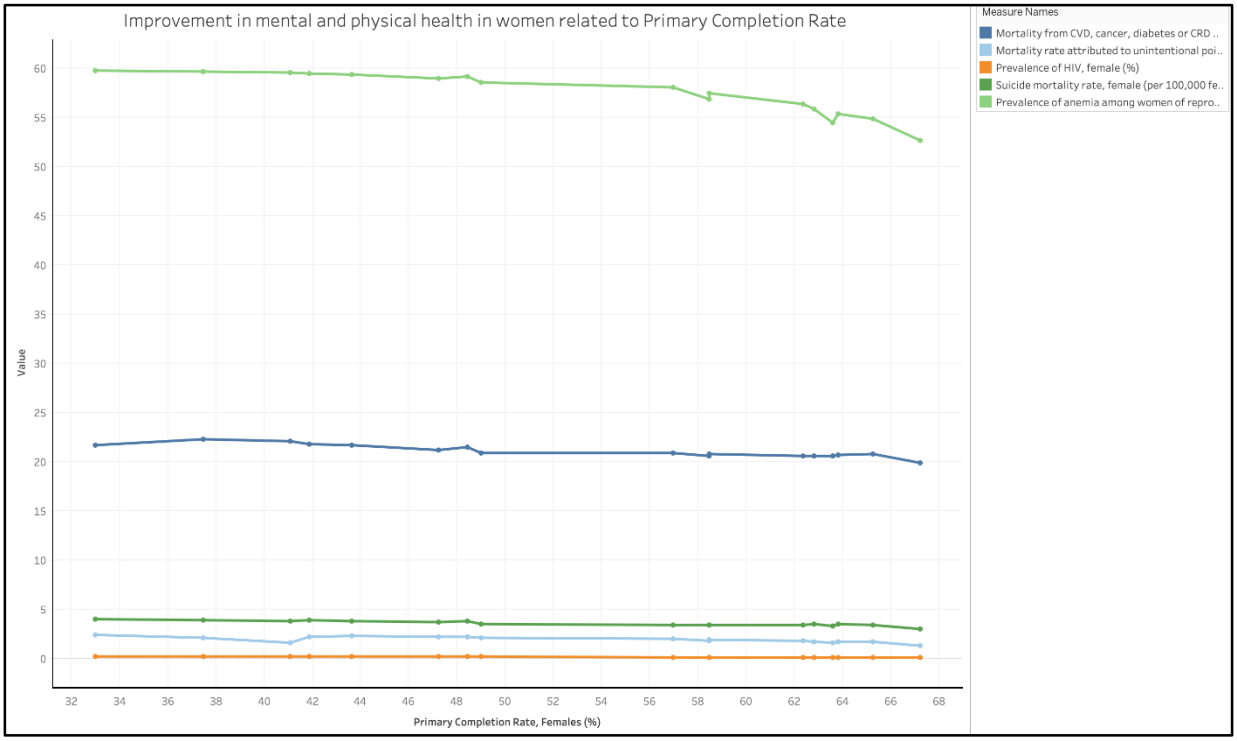

Figure 9.3: (Bottom) Multivariate line chart showing improvement in the physical and mental health of Senegalese women with the rise in primary completion rate

3.4 Visualization Description

The final visualization is a combination of the visualizations in section 3.2.2.3. In this section, we try to explain how the visualization was derived in brief and focus more on how it fits the visualization principles learned in class.

Multivariate visualizations are generally hard to read. Hence, motivated by Tufte's small multiples [4], Trellis Display, and Pair plots, we develop a grid of visualizations that support our hypothesis. With the least data-ink ratio, we follow Tufte's graphical excellence principles, where we showcase the greatest number of ideas, here 15 at a single glance with the least ink and minimal space. Since this document cannot be made interactive, the plots (figure 9.1, 9.2, 9.3) are vertically stacked for reading appropriately and panned and zoomed to understand the intricate details. In all three figures, we see that the visualization conveys the story and supports our hypothesis. Firstly, employment of women increases in high-profile jobs whereas reduced in agriculture, industry, and labor force. Secondly, the primary completion rate correlates negatively with all the birth and fertility rates, infant mortality rates. Thirdly, it shows a minor negative correlation with all the health issues faced by Senegalese women.

From Bertin's vocabulary for retinal variables, color/hue is used as a pre-attentive effect to differentiate between nominal variables. For example, different sectors in employment (services, industry amongst others) use different colors. The line charts are used as they are known to be rated high in Mackinlay's effectiveness ranking for plotting time-series/continuous quantitative variables. In these graphs, we do not

enter missing values mathematically for preserving integrity. The size and scales are consistent for the quantitative variables with minimum lie-factor. Titles are self-explainable. They also help detect patterns, fulfilling Tufte's graphical excellence and integrity principles. We keep the visualizations simple, aesthetic yet multivariate at the same time, thus enhancing easy pattern detection and inference operations and representing data in ways that support cognitive tasks. Hence, amplifying cognition [8]. Since the visualization is dealing with a lot of information in the form of multivariate data, it is worth noting that the patterns and trends are memorable but remembering yearly quantitative data in a tabular format would require immense cognitive effort. We further highlight principles of importance ordering when placing these visualizations in the final dashboard most effectively, thus gaining brownie points on Tufte's expressiveness and consistency definitions [10].

The overall positives from the visualization include it conveys a lot of data dimensions together effectively and can be easily seen at a single glance [5]. Although the data-ink ratio is minimized, we tried improving that further by removing the grids but it did not lead to effective comparisons. Overall, the visualization strongly supports our hypothesis.

4. Discussion/Conclusion

The activity started by looking at the whole data initially. The overall process looked quite unintuitive to proceed. Some of the initial EDA helped in narrowing down to a hypothesis. Multiple iterations of plotting, refining helped define a specific hypothesis with some limitations which worked out well on the dashboard. The final visualization was plotted keeping in mind all the visualization principles learned in class. In the end, with all the data and visualizations, we can conclude that the positive correlation between primary completion rate and socio-economic status and health of Senegalese women for 2000-2019 is strong. The visualization dashboard can further be improved by adding interactivity to it.

4.1 Critique of Tools:

Python, Microsoft Excel, and Tableau were used for this exercise. Python is convenient for cleaning and pivoting the data but hard to get a sense of the whole dataset at once. Microsoft Excel helps with the data inspection and first level of cleaning. Tableau with its UI is a quick way to perform the initial EDA as with Python the Lines of Code to get a visualization is generally high. Certain tasks like plotting multiple variables on a single chart are not that intuitive on Tableau and the initial graphs might be the best but not the most optimal in terms of visualization principles. With all the advantages and disadvantages, these tools when used together appropriately, help in a fluent end-to-end Exploratory Data Analysis.

5. Bibliography

1) The Differences Between Good Data Visualization and Bad Data Visualization Part 1 - Resagratia - Data Analytics
2) https://datacatalog.worldbank.org/search/dataset/0037712
3) https://databank.worldbank.org/reports.aspx?source=world-development-indicators
4) Better Know a Visualization: Small Multiple
5) Stephen Few's Now You See It, Analytics Press; 1st edition (April 1, 2009)
6) http://tephra.smith.edu/classwiki/images/c/cd/Informationvisualization2.pdf
7) DATA511 Week 1 Introduction
8) DATA 511Week 2 - Graphical Excellence and Integrity
9) Bertin Graphical Library
10) Edward Tufte's The Visual Display of Quantitative Information, 2nd ed., 2001.
11) Nathan Mannheimer's DATA-511 Weekly Slides