The Fantastic 4 – Ben, Sharon, Renae, Mark

Analyzing the Life Expectancies for Females vs. Males

in the Known Countries of the World

and Causes Behind the Differences

COSC 481- Data Science

Final Project

4/24/19

**Proposal**

The question that our group has decided to tackle is, “What factors are correlated to life expectancy of females as a percentage of males?” in hope to discover ways the lower life expectancies could be increased. We will do this by graphing to see the other country outliers (both hi and low) and looking at those. We have decided that it would be helpful to examine outliers of “Survival rate to the last grade of primary schools: females as a % of males” as well. We will then determine why the higher ones may exist and why the lower ones may as well, based on the other variables. Its relevance to a societal need is that, if it turns out that the proportionately lower or higher life expectancy for women has a strong correlation with another variable, that could lead us to determining how the life expectancy could be increased by either increasing or decreasing other factors. For example, if there is a higher mortality rate for women who go through childbirth, that would most likely lead to a lower life expectancy of women as a percentage of men for the region. Perhaps that region needs to address its healthcare system. This is just an example, but these are the sorts of correlations that we are interested in exploring, leading to being able to predict female life expectancies based on those correlations. If we had the means, we could then use our information on what causes the lower life expectancies to help those in at-risk regions work their way up to having longer life expectancies.

The UNICEF data we are using has data in regard to the state of people around the world in the following topics (more detail can be found about these topics on the data set pdf under the section “Notes of specific tables” on p 151: Mortality Rates, Nutrition, Health, HIV/AIDS, Women, Child Protection, Early Childhood Development, Economic Indicators (UNICEF, 2017, Table 7).

We plan on comparing all of the UNICEF data in order to find correlations between different aspects of life and life expectancy between men and women.

We also have found a paper that is somewhat closely related to this topic and are interested in using its datasets in our research. These datasets show the difference in life expectancy between men and women, as well as the rates per cause of death of men and women (Sundberg et al, 2018).

UNICEF dataset available on our github or downloaded from the UNICEF website.

**Abstract**

Life expectancy has many variables at play, such as education level, where one lives, being female, among many others. Each country has different statistics for these variables to look into. We want to know; what factors are correlated to life expectancy of females as a percentage of males? Using the 2017 UNICEF dataset (UNICEF, 2017) and data from the United Nations Statistics Division (United Nations Statistics Division, 2019), we graphed the data by each variable by other variables looking for outliers both high and low, to determine why these outliers exist, and how they change based on other variables. Looking into “Survival rate to the last grade of primary schools: females as a % of males”, we focused on countries that lied below 99.1825 and above 103.206. This allows us to look into if increasing or decreasing factors will have an effect on life expectancy, hopefully to find ways to increase life expectancy. We could then use our information on what causes the lower life expectancies to help those in at-risk regions work their way up to having longer life expectancies. This project created multiple graphs of each variable with the other variables, showing many outliers throughout the graphs. From this, we focused on the more extreme outliers as factors most affecting life expectancy of females as a percentage of males, removing the other countries when discussing their respective variables. After identifying these outlying countries, we drew conclusions on how to help increase female life expectancy in the determined locations. Supporting these countries will hopefully help fix some of the disparities encountered in our analysis.

**How Our Approach Is Unique**

Many sources that we examined studied life expectancies of males vs. females in the United States, but we felt this was incomplete. The dataset that we found through the UNICEF Website, “The State of the World’s Children 2017,” (UNICEF, 2017) gave a lot of information. The dataset includes data on nutrition, health, HIV/AIDS, women, child protection, early childhood development, and economic indicators, all of which are studied for 202 different countries in world. However, this dataset does not include conclusions for this information; it just includes the data. When we came across this comprehensive dataset, we decided to look for a variable that we found interesting that also could be applied to the “Data Science for Social Good” requirement. Upon finding “Life Expectancies of Females as a Percent of Males,” we knew that we wanted to study it, but we did not quite have a question in mind yet. We ended up deciding to look at the correlations between "Life Expectancies of Females" and other variables on a country-by-country basis. We also expanded our focus to include studying “Survival Rate to the Last Grade of Primary: Females as a % of Males” from the same dataset as well.

The main goal of our project was to try to model life expectancies/mortality rates of females as a percent of males using the other variables. After doing so, our next goal was to address root causes in order to even out the mortality rates between females and males. We also looked at other data and studies about gender differences in mortality rates, as it is a generally accepted fact that women tend to live longer than men. We thought it was important to understand and address the reasons behind this discrepancy when studying gender differences in life expectancies. Our research and sources for this can be found in our “What Conclusions Are Already Available” section.

After beginning our study, we realized that the “Survival Rate to the Last Grade of Primary: Females as a % of Males” had more variation in its data than the Life Expectancy column. This was disturbing, as it should have been the one that was the most closely centered around 100%. We decided to take our original approach and shift our focus to the “Survival Rate to the Last Grade of Primary” column, as we felt those results were more significant.

We then discovered, as detailed in our “Successes and Failures during Data Modeling” section, that “Survival Rate” referred to continued enrollment in primary school – not mortality rate.

So our project took another turn. We decided to continue studying “Survival Rate to the Last Grade of Primary School: Females as a % of Males,” despite the new understanding that this referred to continued enrollment in school. Even though we were not as focused on causes behind child mortality rates, we thought it was still important that this column variable should be at 100%, and were troubled that it was not.

As the results from our heatmaps became more clear, however, we finally shifted our focus back to the “Life Expectancy of Females as a Percent of Males” variables, because the correlations relating to it were higher, and thus more significant. It was a rollercoaster of events, but we ended up right back where we started.

We also expanded our research to include other datasets from the United Nations that included information on GDP and other variables not included in our original dataset. These are listed in the “What Data Is Available” section below.

**What Data Is Available**

A general “Google” search on differences in longevity between men and women, or on reasons behind differences in gender for sustained enrolment rates to the last grade of primary school, will warrant many results. The main dataset that we used for our project was “The State of the World’s Children 2017,” (UNICEF, 2017) as listed above. This data came from the UNICEF main website. We also studied data from “Analyzing Whether Countries Are Equally Efﬁcient at Improving Longevity for Men and Women,” (Barthold et al, 2014), which included its own datasets. Finally, we included some statistics from the United Nations Statistics Division (United Nations Statistics Division, 2019).

**What Conclusions Are Already Available**

While doing research, we found a study on “Analyzing Whether Countries Are Equally Efﬁcient at Improving Longevity for Men and Women” (Barthold et al, 2014). The main purpose of this study was to find out if making the health care system more efficient in each country would increase the lifespans of each gender. Barthold also studied the current pattern of expenditures in the healthcare system, to see whether increased spending benefited men and women equally. They completed this study over 27 countries and found that life expectancy could be increased by improving the efficiency of the health care systems, without necessarily increasing health care expenditures. They also found that men benefitted much more from increases in health care expenditures than women did. The researchers stated that,

“For example, it is possible that men may receive a more accurate diagnosis, higher quality or more effective treatment, as has been indicated in literature on coronary heart disease and congestive heart failure” (Barthold et al, 2014, p. 5).

This conclusion falls into the biases in data science that we discussed in class. Perhaps the men receive a more accurate diagnosis because the data were tailored to them in the first place. Yet it is a generally accepted fact that females live longer than males, so we wanted to explore the reasons behind that first.

An interesting point is that this conclusion could also feed into “Survival Rate to the Last Grade of Primary School: Females as a Percent of Males” variable. Unfortunately, the UNICEF dataset (UNICEF, 2017) did not have any data for the United States in this column, but if given the chance, it would have been interesting to see whether or not the ratio was lower (meaning that males were more likely to make it all the way through primary school). If not primary school, it would be interesting to see what the gender difference rates for making it through medical school are, and how this would contribute to the results stated in the previous paragraph.

Why women tend to live longer than men may be able to be answered by a BBC article that we found (BBC, 2012). A group of scientists studied the mitochondria of fruit flies. They studied fruit flies because females tend to live longer than males regardless of the species, and fruit flies are easily studied (and have many generations in a relatively short period of time). During the study, there were many mutations of the DNA that supposedly affected how long the males would live and the length of the aging process.

One of the more interesting points of the study was that the same mutations did not affect the way that the females aged. Because mitochondria are passed down by females, the problems that affect males are never corrected. Natural selection will not eradicate the mutations because of how they are passed on. This means that, over time, men generally live a shorter time than women due to age-related illnesses caused by the mitochondria, speeding up the aging process in said males.

Now we return to our study of “Survival Rate to the Last Grade of Primary School: Females as a % of Males” (Huebler, 2008). Huebler is the head is the Head of the Education Standards and Methodology Section of the UNESCO Institute for Statistics. In his blog post, he discusses how this statistic reflects measurement of the UN Millennium Development Goal to realize worldwide primary education by 2015. Huebler’s blog states that in 2008, the retention rate for primary students were as low as 25% in some African countries for both boys and girls. In Northern Africa, Huebler states, the statistics are reversed from what one might expect: girls have a “survival” rate of 85%, compared with boys at 75%. This data was taken in 2008, so we compared it with the current tables. Nauru’s data is not available for 2017, and Uganda’s retention rate has fallen to 22% for boys and girls. Overall, enrollment retention rates have fallen for sub-Saharan Africa – from 64% overall in 2008 to 55% in 2017.

Huebler’s reference to the UN Millennium Development Goals caused us to pursue the results of those development goals (United Nations, 2015). In 2000 the United Nations created the UN Millennium Development Goals to provide direction to the assistance provided to developing countries. This was a list of eight global initiatives to end poverty and support human flourishing. According to the end state report, during this fifteen-year span, the number of people living in extreme poverty declined by half, with related improvements in maternal and infant mortality (United Nations Department of Public Information, 2015). Thus, even though we may have seen selected declines in primary school retention since Huebler’s summaries, overall the results of these initiatives were positive. After the deadline passed in 2015, the United Nations Sustainable Development Group revisited these goals with a new projected deadline of 2030. The seventeen new goals encompass healthcare, poverty, responsible land use, and education (United Nations, 2019).

**How Our Data Has Been Cleaned**

One problem that we discovered after downloading the datasets was that the formatting was quite a bit off for what python and pandas accepted. Many of the headings had to be reformatted, and we had to figure out how to load excel files with multiple sheets for data analysis. This proved to take more time than a lot of the actual data analysis.

Another part of our data cleaning was some steps we took after we did after a little bit of data analysis. We were mainly interested in the outliers, so we pulled just countries above and below the 75th and 25th percentiles, respectively, for our correlation analysis, meaning we cut the rest of the data before we did the remainder of the data analysis.

We collected the country names in two separate lists: countries that fell above the 75th percentile, and countries that fell below the 25th percentile for “Survival Rate to the Last Grade of Primary” column. Then, we joined all of our different sheets from our UNICEF dataset together. Once they were joined into a single dataset, we created one dataset with just the lower percentile country data, and another dataset with just the higher percentile country data.

After this, we looked at heatmaps of the correlations between the columns in the two separate datasets we created.

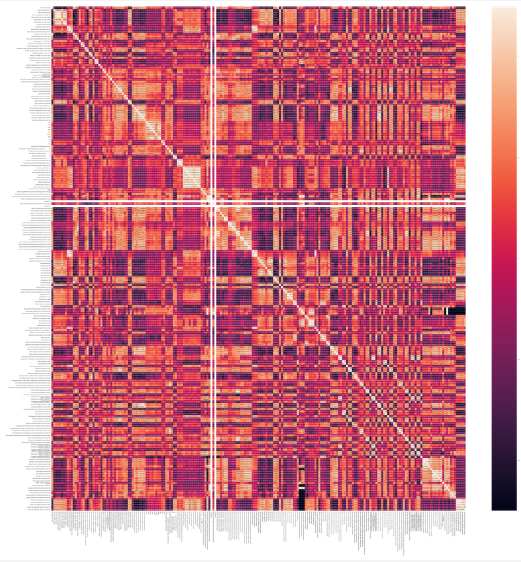
We ended up cleaning our data again to include all the countries, but broke up each Excel sheet into a different csv file in order to have heat maps that were more easily interpreted. We had some formatting issues there, but we managed to overcome them.

We also had to format our UN data (United Nations Statistics Division, 2019) in order to be able to properly work with it using Python and Pandas. By the time we got to that dataset, we had already done a lot of reorganization of our initial dataset, so the process was made easier by repetition. Through trial and error, we found that some datasets were most easily edited using Pandas (for instance, a merge which aligns country names), while some data cleaning is easier to edit in its native Excel (for instance, removing '--' from the dataset with a single "find/replace" keystroke).

**Inferences Made During/After Data Analysis**

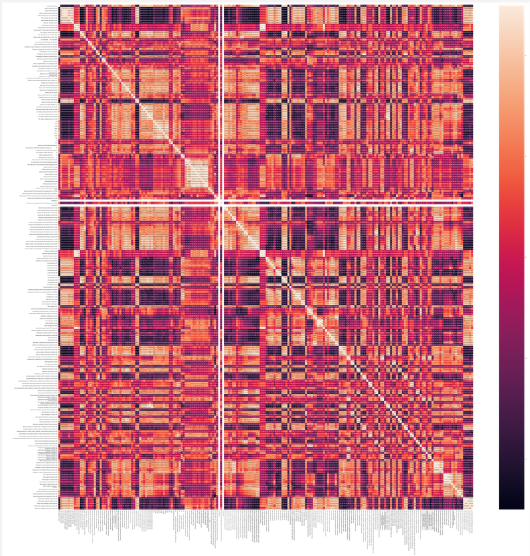
Our first bit of data analysis was to create boxplots of the data by country for “Life Expectancies of Females as a Percent of Males” and “Survival Rate to the Last Grade of Primary School: Females as a percent of Males.” Due to having so many outliers, our boxplots did not show up very well. We ended up printing out the 25th and 75th percentiles for both variables and doing our data cleaning based on that. The way we did our data cleaning is described under our “How Our Data Has Been Cleaned” section.

Once we had our cleaned data, we made heatmaps using python and pandas. For the upper 25% of countries in the primary school survival rates, this was our heatmap:



It has been pointed out that this does not look like anything that could be readily interpretable, (and instead, more like an aerial view of Mustafar), but through the isolation of specific values, the correlations become more clear.

We also completed a heatmap for the lower 25% of countries for the same variable. Similarly to the earlier heatmap, the isolation of specific variables made using this map possible:



It can be easily seen that both heatmaps look very similar. This was expected, as the same things that help some countries to thrive, would most likely be the same things that help the other countries to thrive.

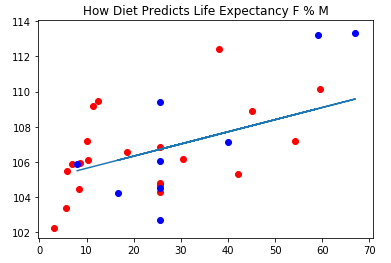
One of our group members was able to code to pull out the correlations above 0.8 and -0.8, but that gave us over 1000 correlations. That is when we realized that we were looking at correlations over the whole dataset, when in reality, we wanted to be looking for correlations in either the row or the column of “Survival Rate to the Last Grade of Primary School: Females as a Percent of Males” or “Life Expectancies of Females as a percent of Males” only, as these were the variables that we wanted to do our data modeling predictions off of. As was mentioned earlier, this method of isolating specific variables gave the hectic heatmap a more clear use. Appropriate changes to the code were made to allow for the access of specific variables to be checked for the desired correlation level.

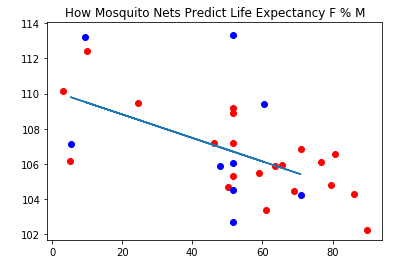
Due to this change, we ended up having fewer correlations to work with, so we merged the UN dataset to increase our number of higher correlations. Unfortunately, none of these correlations were over 0.5 or -0.5, which was suprising. We decided to use only the UNICEF dataset.

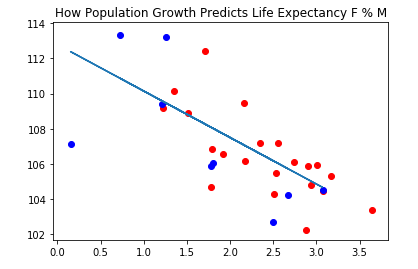
Despite the work we did to separate the countries between lows and highs, we realized that it was not giving us the results that we wanted. We then decided to focus on our entire country list of 202 countries, but still focused on trying to find correlations with the "Survival to the Last Grade of Primary." This did not yield any significant correlations, either.

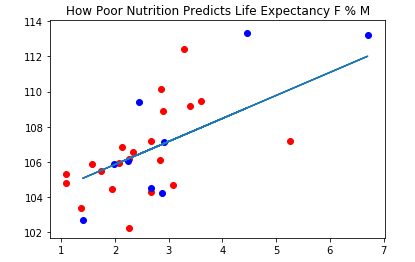
After finding no significant correlations, we shifted our focus back to “Life Expectancies of Females as a Percent of Males.” We had a lot of issues here, but we ended up with no significant correlations for the lower 25% countries. However we had a few correlations with the upper 25% of countries including: Minimum Acceptable Diet 6-23 Months (r = 0.7065), Households With At Least One ITN (insecticide-treated nets) (r = -0.7091), Population Annual Growth Rate (r = -0.6971), and Stunting Prevalence in Children Under 5 Ratio of Poorest to Richest (r = 0.7027). These are the variables we used in order to do the linear regression models to try to predict life expectancy information.

We ended up with four separate linear regression models using x and y train and test. The models we created are below:









The regression model with the strongest correlation was “How Mosquito Nets Predict Life Expectancy F % M,” with a correlation of almost -0.71.

**Successes and Failures During Data Modeling**

One of our first failures during our project was dealing with unwieldy formats of datasets from various sources. Our team had to work for a while trying to format the datasets in a way that could be interpreted in pandas dataframes (Fun fact: If you forget to type in “python” before “pandas,” all you get are pictures of cute pandas, which is not the worst problem to have. It is, perhaps, even a welcome relief when dealing with the stress that comes with an unruly dataset). This proved to take longer than the actual data analysis portion. Once we had the data initially cleaned, the data analysis commands went pretty quickly.

Another very annoying thing that we had to deal with was that we had to get rid of the “dash” marks in the Excel file for null data. Each time we tried to analyze the data using Pandas, Python would think that we had both floats and strings. After many attempts to edit this using Python, we found it was faster to "replace all” in Excel.

During the data analysis process, we noticed that one particular group member had some very interesting issues that occurred with the code. It seemed that, even when they copied and pasted the code of others, it still did different things on their computer than it did on everyone else’s machine. We would also like to point out that it did not matter which computer this group member used, as they all seemed to equally dislike him and halt his efforts. One of the most interesting turn of events was when Jupyter just stopped giving him access to his own files. It took the group some time to figure out a solution to this problem.

In addition to those, one very large problem that we encountered was that we were not actually studying what we originally intended. We were disturbed to see any sort of gap (either below or above 100) in “Survival Rate to the last grade of Primary School: Females as a % of Males.” Upon further exploration, we realized that “Survival Rate” was not talking about mortality, but instead, the children who made it in any sense to the last grade of primary school that started primary school. This result could definitely be dependent on children dying before they can finish school, but now we learned that was not the only thing we were researching. In our defense, “Survival Rate” did seem ominously similar to “Mortality Rate.” UNICEF could have used the term "retention" to alleviate misunderstanding.

Another problem that occurred was that for each step of our data analysis, we always had far more results than what we thought we would. This, however, was probably a good problem to have. For example, when we were looking for outliers, we ended up with more than just four or five, like we would if we were looking at a smaller dataset. When we did our heatmaps, we had to also analyze the actual correlations on their own, since we were analyzing so many variables at a time.

When we pulled the correlations out, we ended up with over 1000 for correlations of over 0.8 or -0.8. That was too much for us to really handle, so we had to keep upping the correlations. Eventually we made it to 1.0 and -1.0. We were not expecting correlations at this level (except for categories being compared to themselves). So our problem of so many correlations led us to find an interesting set of extremely strong correlations. Unfortunately, time did not allow us to change our focus and explore these correlations.

In another turn of events, we once again had a miscommunication of what it was we were studying, so our problem of “too many correlations” became not as much of a problem. We were looking at the entire heatmap when pulling correlations, when what we wanted to do was just pull from a single row or column. Realizing this narrowed down our correlations a bit.

**Conclusions from Data Analysis**

In 1996, Patricia Lone wrote for UNICEF, "Many forces combine to spell an early end to education for girls. Chief among them is poverty. ...When a poor family considers how much a daughter can help [at home], then the returns rarely seem to warrant the expenditure." (Lone, 1996) Based on this and our intuition, we expected that poverty would be a good indicator for girls dropping out of school earlier than boys.

Similarly, the European Commission is currently spending a million pounds on a project to help encourage young women in Malawi to stay in secondary school or to reenroll in school (International Cooperation and Development, nd.). This project cites a lack of knowledge about health and sexuality, as well as early marriages, as the reason for low education enrollment numbers among females.

Unfortunately, the data did not support either of these conclusions. Despite several different variables that should measure poverty (for instance: child nutrition, GDP per capita, and percentage of individuals living below the poverty line), and several different variables that should measure sexual health and child marriage rates ("Demand for family planning satisfied with modern methods," "Knowledge of HIV," "Child marriage"), NONE of these data points showed strong correlation for "Survival to the Last Grade of Primary."

One possible conclusion is that ever-so-typical answer, "It depends." Perhaps in Malawi, the major factor that makes females drop out of school is health related, while in other countries, the major factors have more correlation with markers of poverty. This would require further exploration to find pivot variables around which one might separate the countries in the dataset. With this more nuanced knowledge, UNICEF could focus certain interventions in areas where they are appropriate, rather than potentially "fixing" a problem that doesn't exist.

Another possible conclusion is that potentially relevant data are missing. As mentioned above, the United States has no data on "Survival to the Last Grade of Primary (Females as a percentage of Males)." Many developed countries don't collect data on things that are no longer an issue; so the data variation might all appear at the lower end, and may not give enough information for Python to accurately analyze.

The European Commission's project did bring up one factor we could not confirm: that is, that women tend to withdraw from school when they feel they do not have other female support (International Cooperation and Development, nd). The project seeks to ameliorate this effect by providing peer mentors to young women in school. Since this is not a data point measured by UNICEF or the UN, more research would be necessary to confirm their hypothesis.

After doing our correlation analysis for “Life Expectancies of Females as a Percentage of Males,” we ended up with a few moderately significant correlations, but our highest (or lowest in this case) was, surprisingly, Households With At Least One ITN (insecticide-treated net), with r = -0.7091. This would mean that, since these were for the upper 25% of countries (over 100% for life expectancy ratios), countries that had more households with mosquito nets tended to also have their men live longer, bringing the life expectancy ratio closer to 100%.

**Future Work/ Further Exploration**

Something that our group would love to be able to do is to actually provide support to these countries to fix some of the disparities that we found during analysis. In all of the “Data Science for Social Good” articles, after the conclusion was made, the groups doing the analysis used their findings to help people. We are sure that UNICEF probably has their own team of data analysts working on these things, whether they publish their findings or not, but using our data to help others definitely falls under the category of “things we would like to do in the future.”

Also, since one of the results that we found was that there was a moderately strong correlation between insecticide-treated nets and the lifespans of men, it would actually be very easy to help these countries. At least one of our group members has been a part of a fundraising effort to buy mosquito nets for at-risk countries. Seeing the correlations first-hand means that other group members will be more likely to contribute to these efforts in the future as well.

As a group, we also feel that more time would have allowed us to do more exploratory data analysis and predictive modeling. We would have loved to have studied correlations with more datasets, and we would have also liked to have done k means clustering modeling on the data as well. We ended up doing linear regression modeling, which was cool, but we would have liked to have done more with the predictive modeling. Perhaps those of us who take the Data Science II course will get that opportunity.

Bibliography

Barthold, D., Nandi, A., Mendoza Rodríguez, J. M., & Heymann, J. (2014). “Analyzing Whether Countries are Equally Efficient at Improving Longevity for Men and Women.” *American Journal of Public Health*, *104*(11), 2163–2169.

BBC. (2012). “Fruit Flies Offer DNA Clue to Why Women Live Longer.” *BBC News,* 2 Aug 2012. Retrieved from<https://www.bbc.com/news/health-19093442>.

International Cooperation and Development. (nd). "EU-ISEM Project Helping Girls to Complete Secondary Education." *European Commission.* Retrieved from https://ec.europa.eu/europeaid/case-studies/eu-isem-project-helping-girls-complete-secondary-education\_en.

Huebler, Friedrich. (2008) “Survival Rate to the Last Grade of Primary School.” *International Education Statistics.* Retrieved online from https://huebler.blogspot.com/2008/02/survival-rate-to-last-grade-of-primary.html

Lone, Patricia. (1996) "Commentary: Keeping Girls In School." *UNICEF.org.* Retrieved from <https://www.unicef.org/pon96/edgirls.htm> .

Sundberg, L., Agahi, N., Fritzell, J., & Fors, S. (2018). “Why is the gender gap in life expectancy decreasing? The impact of age- and cause-specific mortality in Sweden 1997-2014.” *International Journal of Public Health*, *63*(6), 673–681.

UNICEF. (2017). “State of the World’s Children.” New York, NY: United Nations Children’s Fund.

United Nations. (2019). “The Sustainable Development Agenda.” Retrieved from https://www.un.org/sustainabledevelopment/development-agenda/

United Nations Department of Public Information. (2015). “The Millennium Development Goals Report.” Retrieved from https://www.un.org/millenniumgoals/2015\_MDG\_Report/pdf/MDG%202015%20PR%20Regional%20LAC.pdf

United Nations Statistics Division. “UN Data: A World of Information.” Retrieved from http://data.un.org/