The Fantastic 4 – Ben, Renae, Sharon, Mark

Analyzing the Survival Rates to the Last Grade of Primary School

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 the page marked 151 (6th page in the pdf)): Mortality Rates, Nutrition, Health, HIV/AIDS, Women, Child Protection, Early Childhood Development, Economic Indicators

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. The link for that paper can be found here: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6015620/

UNICEF dataset available on our github: https://github.com/bigbbv/COSC\_481-Major\_Project/blob/master/SOWC-2017-statistical-tables.pdf

It can also be downloaded from the UNICEF website:

https://data.unicef.org/resources/state-worlds-children-2017-statistical-tables/

**Abstract**

**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,” need a source here gave a lot of information. The dataset includes data on nutrition, health, HIV/AID, 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 on looking at the correlations that variable shares with 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 what some countries needed to improve 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 why that is 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%, so our focus became more tied to that data. We decided to take our original approach and just shift that to be being focused on “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” just means, “Did they reach the last grade of primary school if they enrolled in school to begin with.” We had interpreted it as meaning, to put it bluntly, “did they die?”

So, our project took another turn. We decided to study both the “Life Expectancies of Females as a Percent of Males” and the “Survival Rate to the Last Grade of Primary School: Females as a % of Males” categories, which is what we originally wanted to study. So, we went in a big circle, but ended up back to where we started, which we still consider a unique approach.

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

**What Data Is Available**

Just doing a general “Google” search on differences in longevity between men and women, or on why some countries have differences in gender for the survival rate 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,” need a source here, 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,” need a source here, which included its own datasets. Input Sharon’s data here.

**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” This can be found at \_\_\_\_\_\_\_\_\_. The main purpose of this was to find out if making the health care system more efficient in each country would increase the lifespans of each gender. They also studied if the current pattern of expenditures in the health care system benefitted men or women more, or if increasing spending in health care helped the two groups equally. They did this study over 27 countries and found that life expectancy could be increased by improving the efficiency of the health care systems without necessarily just increasing health care expenditures. One interesting bit of information that they found was 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.” (source) pg. 5

This conclusion falls into the biases in data science that we discussed in class. Perhaps the men receive a more accurate diagnoses because the data was 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.

Why women live longer than men may be able to be answered by a BBC <https://www.bbc.com/news/health-19093442> article that we found. A group of scientists studied fruit flies, specifically, their mitochondria. They studied fruit flies because females tend to live longer than males regardless of the species. During the study, there were many mutations of the DNA that supposedly affected how long the males would live and how quickly the aging process took.

One of the more interesting points of the study was that the same mutations don’t seem to affect the way that the females aged. An important aspect is that mitochondria are passed down by females, so the problems that affect males are never collected. Natural selection will not eradicate the mutations, due to the process of them being passed down. 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.

We also studied “Survival Rate to the Last Grade of Primary School: Females as a % of Males.”

<https://huebler.blogspot.com/2008/02/survival-rate-to-last-grade-of-primary.html>

**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.

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 rest of the data analysis.

We collected the country names in two separate lists, ones that fell above the 75th percentile and 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, and the additional datasets that we found that are listed in our “What Data is Available,” 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 our datasets (multiple datasets due to having the higher value countries and lower value countries in their own datasets).

**Inferences Made During/After Data Analysis**

**Successes and Failures During Data Modeling**

One of our first failures during our project was dealing with weird formats of datasets from various sources. Our team had to work for a while trying to format the datasets in a way that they could be understood in pandas data frames. 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 “- “for null data. Each time we tried to analyze the data using pandas, python would think that we had both floats and strings, so we had to deal with that, and ended up just fixing it on the actual dataset instead of using coding techniques.

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’. So that was an interesting obstacle we had to work with throughout the project.

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. Now, the result from that could definitely be dependent on children dying before they can finish school, but now we know that is not the only thing we are researching.

**Conclusions from Data Analysis**

**Future Work/ Further Exploration**

Something that our group would love to be able to do is actually help to support 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 actually used their findings to help people. Now, 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 would fall under the category of “things we would like to do in the future.”

Bibliography