Running the Sketch:

The sketch for this assignment is located in the pyramidPlots code directory.

The following are a list of instructions and commands to use the interface implemented. I tried to make the interface as intuitive as possible, though there are a few features implemented that may need further explanation.

1. Upon running the sketch, the user is presented a window with a title, a heading with "+" and "-" controls, an empty plot with axis labels and the text "Choose Country", and a timeline with a vertical black bar indicating the current year. This is rather uninteresting because there is no data being displayed yet, apart from the world population text. To assign data to the plots, click the "Choose Country" link near the top, which should be underlined on mousehover.
    a. Doing so should cause a side panel to slide in from the right, which gives a list of available country data to choose from.
    b. Use the letters A-Z to filter the country selections by first letter. By default, the "*" option is highlighted, which is used to indicate "all countries".
    c. Click through the listed country options, and then click the "OK!" button at the bottom to confirm your selection and close the side panel. The name of the selected country will now appear instead of the default "Choose Country" text in the plot window, with its data filled in to the corresponding pyramid plot and timeline graph.
2. Use the "+" and "-" controls at the top of the window to increase/decrease the number of plots shown. Note that the number of pyramid plots cannot exceed 4. Additional plots will also appear empty initially, and require the user to select the corresponding country for that plot using the instructions specified in Step 1. Note that the user can change the country at any time by clicking the country name, and that removing plots and then adding them back will not require reselecting that country.
3. As the user assigns country data to the pyramid plots, the timeline window will also update accordingly. Each pyramid plot number is represented by a unique color indicated by the circle to the left of the selected country name. The purpose of the bottom plot is both to serve as a timeline (explained in Step 4) and a line graph (an example of a scented widget). The population number (in billions) is plotted as a line graph for each selected country. The y axis scaling will also adjust according to the max and min values of the selected country populations over 1950-2015.
4. The vertical black bar on the timeline indicates the current selected year (also shown in the top left corner of the window). The user can click and drag the black bar, or click above or on one of the indicated years. Doing so will update the pyramid plots, the world population in the upper right, and the population shown in the top right of each pyramid plot window. Note that this population value is identical to the numerical y-value of the selected year on the line graph.
5. Click the "l" key to toggle displaying the pyramid axis information. Toggling this option off is not necessarily a good data visualization tool, though it can be useful when just admiring the shape

of some of the pyramid plots or reducing some of the clutter when viewing 4 pyramid plots at once.

Overview

The goal of this assignment was to design and develop a visualization tool showing how the population distribution of different countries by age has changed over time. Specifically, the visualization technique I chose to focus on and improve is the population pyramid. This ultimately involved researching and evaluating existing population pyramid tools, gathering/interpreting the necessary data, designing my own population pyramid framework and interface, implementing this framework, and finally exploring interesting data characteristics.

Understanding Pyramid Plots

A population pyramid, also known as an age pyramid, is a graphical diagram explaining the distribution of age groups of a population. This usually involves drawing two back to back histograms or line graphs, with the population on the x-axis and the age partitions on the y-axis. As a convention, the male population breakdown is shown as the left-hand plot, and the females are shown on the right. Although the designs usually vary, the x-axis population may also be expressed as a percentage of the total region's population rather than a raw number. Overall, the population pyramid is a traditional way of visualizing and explaining the age structure of a population group. This is because using a pyramid can effectively break down a population by age and sex, while explaining a great deal of information about a population. In fact, data analysts often use the general shape of a population pyramid to draw conclusions about a country's fertility and mortality rates. In Figure 1, four general shapes of population pyramids are displayed, which each represent a different generalization or trend about a particular population group. In Figure 1(a-b), examples of expansive pyramids with a wide base and narrow top are shown, indicating that the population has high birth and death rates. Figure 1c demonstrates a more stationary pyramid with lower fertility and death rates. Figure 1d is similar to 1c except that there is a much lower birth rate. A more realistic depiction of some of these shapes in effect are shown in Figure 2.

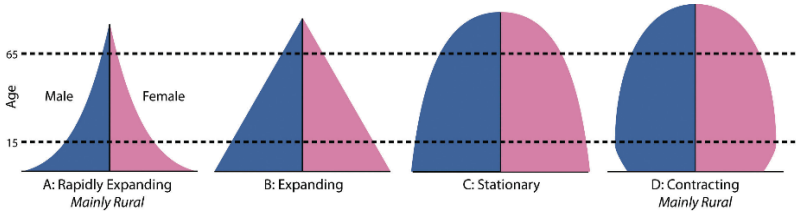

Figure 1

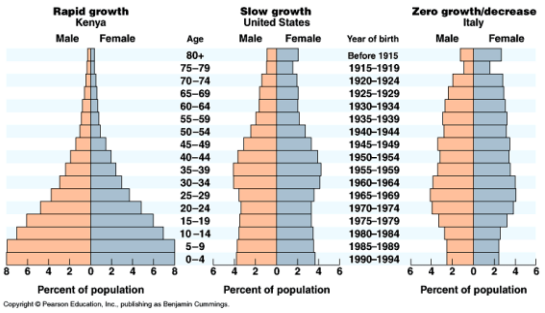

Figure 2

Examples of Population Pyramid Tools

To begin the assignment, I looked at some existing population pyramid plotting tools using various population data. However most of the examples used a similar data source, which is explained in more detail in the next section. Ultimately, my project's design is based on some of the pros and cons of the design decisions portrayed in the following population pyramid visualizations.

- USA Population Animation

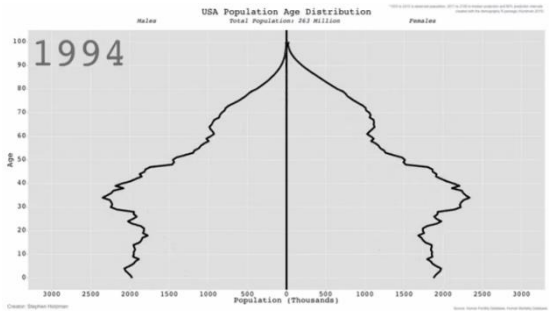

Figure 3

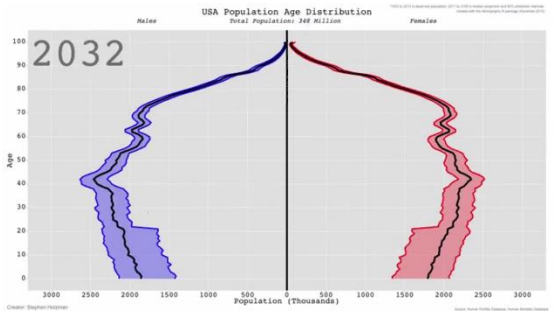

Figure 4

The first visualization I looked at is an animated gif of the population age distribution of the U.S. over time. The first problem with this visualization is the fact that it relies heavily on animation to demonstrate time varying data, which violates a general design principle of preferring eyes over memory. Not only is it difficult to compare each year's data with the previous year, but it is nearly impossible to compare the plots of any two non-consecutive years. In addition, it is difficult to determine the bin size for the population age groups on the y axis. It looks like the

age groups are broken down by individual years, though it is unclear if this data is interpolated or given directly by the data source (most sources seem to break down age groups into 5 year intervals). Finally, they are doing some interesting data prediction calculations by coloring the confidence intervals with blue and green, as shown in Figure 4. In this way, I think the graph does a good job at distinguishing those years that are based on true data and those that are predicted. However, the graph fails to mention exactly what type of predictive method is being used. Overall, I really like the simplicity of the design and clear axis labeling/titles.

- IndexMundi

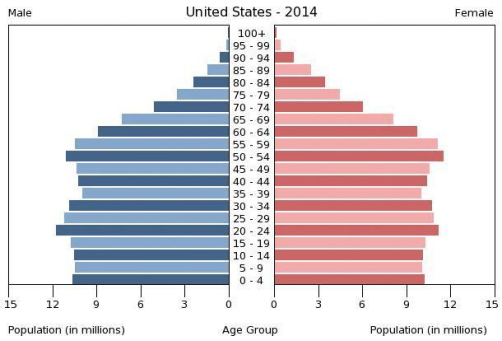

Figure 5

The second visualization I looked at is from a website called IndexMundi, which contains country statistics, charts, etc. from multiple data sources. Overall, I think that this visualization (Figure 5) follows design principles more closely than the previous example, though I believe that some of this is at a cost of data effectiveness. For example, this visualization solves the problem of using animation by only showing a single year's data. However, this isn't nearly as interesting. In addition, the most noticeable difference is the use of histograms instead of line graphs. Perhaps a more rational approach for visualizing these types of data is to use a histogram to depict distributions. However, I believe that in general the line graph approach is actually more effective than a histogram for population pyramids, which can highlight the overall shape of the pyramid much better. The population pyramid is not often used to compare sex distribution within age groups, but rather show a glance of the past, current, and future distribution of an entire population, often in contrast with other population groups. Therefore since this example only shows only a single year, the histogram is more effective. However in the case use of developing a generalized tool for comparing population pyramids, I argue that a line graph is actually more effective. Finally, another notable design choice here is the placement of the age ranges in the middle of the graph. While this eliminates the need to compare y-axis values across the male and female plots, it also isolates the two side-by-side gender plots in a way that makes the general shape of the graph even more difficult to interpret.

- Populationpyramid.net

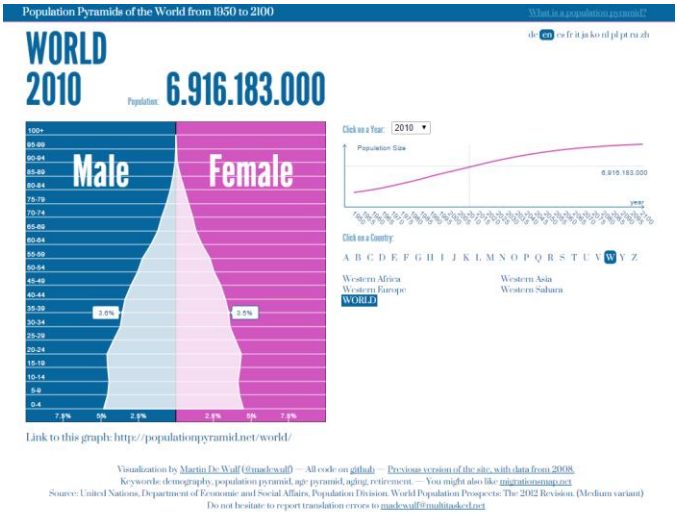

Figure 6

Unlike the previous examples, the above website (Figure 6) attempts to create a population pyramid framework similar to my original project proposal. Not only is the user able to compare data across multiple years, but they are also able to select data from multiple countries. Also, the graph uses a lot of interaction and animated transitions to make the interface appealing. The coloring of the male and female plots also help distinguish these two categorical attributes. I really like the scented widget on the right hand side that serves as both a timeline selection and line graph. They also implement filtering of the countries to allow for a more intuitive search. Finally, the x-axis labels on the pyramid plot use percentages of the population instead of raw numbers, unlike the previous two examples. This makes for more easy comparisons between countries and time periods, avoiding having to scale each graph individually. The main disadvantage of this website is that it becomes very hard to compare multiple countries quickly. Again, it is violating the idea of eyes over memory. Next, the graph does not make it clear which dates are predicted, and which come from the data source. It is somewhat misleading to include dates up to 2100 without even indicating what type of prediction method they are using. Finally, it is very misleading how they handle the data from years 1950-1985, due to inconsistencies with how the U.N. collects age data from ages 80 and over. In other words, before 1990, data for ages 80 and over are provided ONLY as an aggregate for age 80+. This is discussed further in the next section.

Data Analysis

The data I chose to visualize for this project comes from the United Nations Population Division (http://esa.un.org/unpd/wpp/DVD/). The website lists many spreadsheets and csv formatted data files for population fertility, mortality, age, gender, and other population attributes since 1950. The data is collected every 5 years, however some data files interpolate the same data by single calendar years. It is therefore easy to be misled by the different data listings and thus it was important to first analyze the data being presented before starting to design my population pyramid interface. The data I chose to download gives the quinquennial population by five-year age groups, sex, major area, region, or country

from 1950-2100. The raw data file is formatted in the following way, with the most important attributes used in my visualization highlighted in yellow:

Field name	Description
LocID	Numerical Location Code
Location	Name of country, region, major area or other aggregate
VarID	Projection variant code 2=Medium; 3=High; 4=Low; 5=Constant fertility; 6=Instant-replacement; 7=Zero-migration; 8=Constant-mortality; 9=No change
Variant	Projection variant name
Time	Calendar year (1 July)
MidPeriod	Mid-Period
SexID	Sex code (1=Male, 2=Female, 3=Both)
Sex	Male, Female, Both
Pop_0_4	Population in age group 0-4 (thousands)
Pop_5_9	Population in age group 5-9 (thousands)
Pop_10_14	Population in age group 10-14 (thousands)
...	
Pop_75_79	Population in age group 75-79 (thousands)
Pop_80_100	Population in age group 80+ (thousands) ***
Pop_80_84	Population in age group 80-84 (thousands)
...	
Pop_95_99	Population in age group 95-99 (thousands)
Pop_100	Population at age 100+ (thousands)

****Note that before 1990, data for age 80 and over are provided ONLY as aggregate for age 80+.*

The location ID is an ordinal attribute. At first glance it may appear to be categorical, however each country is assigned an ID in alphabetical order. This attribute's use is pretty self-explanatory, giving the location name and corresponding number to be used in my visualization tool. The original data, however, includes locations by region and major area in addition to just countries, and therefore it was important to develop a script to separate the important country information from the rest. The VarID is a very important categorical attribute, indicating the type of projection used to predict data values from 2015 and on. My initial project proposal included this type of projection, however after critiquing the example visualizations in Figures 3-6, I found that including such data in general is very misleading if not giving it careful thought. I did quite a bit of research into distinguishing between the different projection types included in the data, but I found no way of justifying one particular method over another. These data assumptions can be found on the U.N. website (http://data.un.org/Resources/Methodology/PopDiv.htm#_I._Assumptions_underlying_1). Each of these projection methods rely on a ton of assumptions about the future behavior of population groups, something that I am not qualified to make judgements on for every country. Therefore in an effort to avoid misleading the user, I chose to omit this feature from my original proposed work. SexID (categorical) is obviously important when designing a population pyramid, though it is important to note

that I stored the “both” field values when parsing the data in order to more quickly calculate the population totals for every country. The majority of the data comes from Pop_0_4 – Pop_100, giving quantitative ratio attributes (floats) for the population in each age range (in 1000s). Thus this needed to be converted to a percentage in my design for the population pyramid tool.

In addition to my design sketch, I wrote a Python script to parse through the U.N. data and extract the necessary attributes and countries of particular interest. This script is included in the directory as “parseData.py”.

Interface Design & Implementation

As indicated in my project proposal, some goals for my project included 1) the ability to select from population data across multiple countries using small multiples (allowing the user to make country-to-country comparisons). This was a major disadvantage of the populationpyramid.net visualization. 2) Overview+detail line graph indicating the selected year and population trend. This is an example of linked highlighting. There is also a bar to indicate the current year which would give you the data points on mouse hover. Ideally, you will be able to click on each of the individual years, or drag the vertical bar to more easily view a larger timespan. 3) Ability to add and subtract country plots to compare. Doing this will also resize the plot accordingly to fill up the space optimally.

These three visualization tasks were implemented in my population pyramid viewer, and are reflected in the first section of the paper. Details about how these tasks were actually implemented in the interface are also included there. A final feature that I suggested in my proposal was to calculate the future prediction values from 2020 and on. However, I found this to be actually a limitation of my design and as discussed in the previous section, I chose to not implement this feature due to problems with misleading the user.

My initial sketch for the assignment was included in the project proposal as the below figure (Figure 7). As shown by actually running my final Processing sketch, the basic ideas for this mockup were ultimately implemented in the final project with some minor differences in the interface. For example, in the mockup there is no side menu bar. However, I did mention in my proposal that I wanted to add additional sorting and filtering to the data, which can be shown in my final Processing sketch in the side menu’s alphabetical indexing and sorted alphabetical ordering.

One of the major design concerns was how to add and remove additional population pyramid plots using the space available. In my proposal, I wanted to resize the plots to fill up the space optimally. Overall, I think my project does a good job at using up all of the space available. However, there may be some issues with too much clutter when comparing the plots for 4 countries at once. In my final project, adding 4 plots will cause the axis labelling to be much smaller and thus hard to read. Some ways to fix this are either to view the plot in the larger single view, or by typing the ‘l’ button to stop drawing the grid lines. For the single pyramid plot view, I decided to draw the axis labels on either side of the male and female plots, since the increased size of the plot area can make it difficult to scan across the screen to determine the associated age range for a particular data point. However, doing the same thing for the smaller plots creates too much clutter, so this feature was omitted from those.

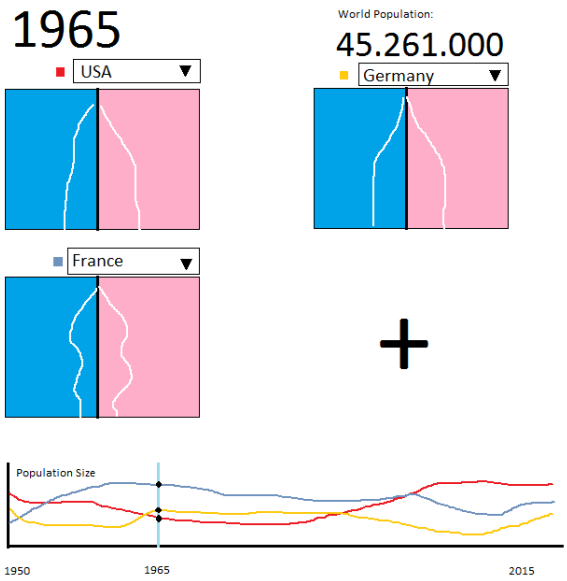

Figure 7

Limitations

In addition to some of the limitations of my design explained in earlier sections, one of the main limitations addressed in my design was the fact that in the U.N. data, data for ages 80 and over are provided only as an aggregate for ages 80+. This means that the age ranges for the years 1950-2015 are not consistent. By not addressing this issue (as in populationpyramid.net), it can appear that prior to 1990 there did not exist any people over the age of 85. This is obviously not true, and so I wanted to design my interface to handle this issue. The obvious way to handle this in my design is to make all of the bin sizes consistent, where years 1990-2015 would only consider ages 80 or higher in the "80+" bin. This is probably the "safe" way to handle this issue, though I tried to preserve data expressiveness of the original U.N. data by not losing all of this information. For example, for some users it may be of particular interest to just look at the trend of 100+ year olds from 1990 and on. Therefore I chose to address this issue in my design by keeping the y axis scaling constant, but drawing/hiding the appropriate y axis labels accordingly. This may not be the most effective design choice, and therefore I'm open to any feedback.

Along the same problem, I had trouble deciding how to display the population data for those countries prior to 1990. In a way, the way I am currently drawing the population pyramid on this domain is to draw the lines even where there is no data. The intention of this is to explicitly show that the values for that domain are undefined. If I were to cut off the graph at the 80+ tick mark, then the fact that the pyramid would then have a hole in it might be confusing for some users. On the other hand, leaving it this way may appear misleading.

Some other minor design limitations in my visualization include not being able to compare more than 4 countries at once, which I think is okay since comparing more than 4 at a time seems impractical. Also, I only chose to include the country data for a subset of the listed U.N. countries/regions. It is definitely true that some users may find other countries of more interest, but the point of my project is to demonstrate how this would be done at a larger scale. When adding the entire dataset, the Processing

program ran very slow and reduced the response time for the interaction implemented. Therefore, I omitted importing all country data for practical reasons, but tried to design my interface to show that including all of them at once would be no different (using the side panel sorting and filter functions). Another consideration is adding the WORLD dataset, which is of course not a country. Therefore adding this particular dataset to one of the pyramid plots will drastically change the scaling of the timeline y axis at the bottom of the sketch, making it difficult to determine individual values on the line graph. Note that in my original project proposal, I intended to allow the vertical bar to indicate the current population data point on mouse hover for each of the colored lines. The reason I chose not to implement it this way (and instead added y axis labels) was due to this distortion.

Data Exploration

The remaining figures and text show my visualization tool in use, using the 10 country data included in the project submission. Some interesting conclusions are made in the following figures about the trends of population distributions over time and when compared to other countries, which was the objective goal of my project.

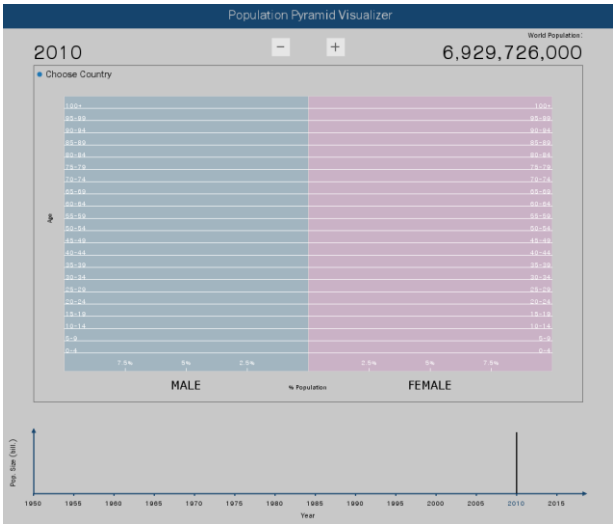

Figure 8

Figure 9

Figure 8 shows the Processing sketch when initially opened. Therefore there is no country data chosen at this point. Figure 9 shows an example of opening the side panel and selecting among the alphabet filters. The current pyramid shows the population age distribution of the U.S. in 1990. Note how the axis labels continue to 100+. The lower-most bulge in the graph is interesting to follow through the timeline, as this represents the baby boom generation in the United States (an increased birth rate in post-WWII era).

Figure 10

Figure 11

Figures 10-11 show the ability of the tool to expand to at most 4 pyramid plots. Figure 10 shows a number of Asian countries when plotted against the world population distribution. You can see the interesting spikes in the distribution in North Korea and China in 2010, and note the younger population is quite low for both genders. In contrast, the distribution in India has an expanding pyramid shape with a wide base, showing a large younger population. In the plot for North Korea, you can see the low numbers of older males, with a significantly greater number of older females. Finally, Figure 11 compares the population distributions for Sierra Leone vs Mexico. Despite their large differences in population size, comparing the distributions show that Sierra Leone has a rapidly expanding pyramidal shape, whereas Mexico's is far more stable in 2015.