CS 171/CSCI E-64: Visualization
Homework 3, Problem 4: Analyze Data of your Choice with Tools of your choice

James Goodspeed – jgoodsp@fas.harvard.edu

1. Short Description of data set

The data I chose for this problem set comes from UNdata, the API from the United Nations. I chose to look at the correlation between per capita total expenditure on health by country as it relates to the life expectancy at birth for those countries for the year 2007.

http://data.un.org/Default.aspx

2. Process for obtaining the data

From the UNdata website I was able to obtain the Life Expectancy at birth data from the World Health Organization and export it as a csv file.
http://data.un.org/Data.aspx?d=WHO&f=MEASURE_CODE%3aWHOSIS_000001

I was also able to get the Per capita total expenditure on health for those same countries by also querying the World Health Organization data and exporting it as a csv file.
http://data.un.org/Data.aspx?d=WHO&f=MEASURE_CODE%3aWHS7_105

3. Refinements made to the data set

The first task was to remove some records from the Life Expectancy data. I wanted to separate out the data into Male and Female and the UNdata also had a category for 'Both sexes.' I was able to remove these lines with the following sed command:

```
sed '/Both sexes/d' life_expect.csv > life_expect-cleaned.csv
```

After that the data from the two files had to be combined so that the per capita expenditure was added to the end of the life expectancy data. To do this I wrote a simple shell script to combine the data:

```
#!/bin/bash

combined_file=combined.csv

# Loop through life expectancy file
while read line
do
    country_first=`echo $line | awk -F, '{ print $1 }'`

    # Loop through per capita file
    while read per_capita
    do
```

James Goodspeed – jgoodsp@fas.harvard.edu

```
        country_second=`echo $per_capita | awk -F, '{ print $1 }'`
        if [ "$country_first" == "$country_second" ]
        then
            per_cap=`echo $per_capita | awk -F, '{ print $3 }'`
        fi
    done < percapita.csv

    echo $line | awk -F, -v VAR=$per_cap '{ print $1 "," $3 "," $4 "," 
VAR }' >> $combined_file

done < life_expect-cleaned.csv
```

4. Explanation of the type of visualization chosen

Because I am correlating the life expectancy and per capita expenditure spent on health a scatter plot was the obvious choice to show this correlation. Using Tableau seemed to give more options in terms of customizing the visualization than was available using Many Eyes.

5. Relationships or trends discovered

Unsurprisingly a higher per capita expenditure on health care leads to a higher life expectancy in the countries measured. There are undoubtedly other factors that come into play but it seems obvious that the amount of money spent on health care plays a definitive role in the average life expectancy of a nation's citizens.

Another trend discovered is that women, on average, live longer than males in the countries measured.

6. Screenshots

The screenshot below shows that as per capita expenditure on health care goes up so does the life expectancy. Each circle represents the males (blue) or females (red) in the country being measured. Also shown is that women (the red circles) live, on average, longer than their male counterparts.

CS 171/CSCI E-64: Visualization

Homework 3, Problem 4: Analyze Data of your Choice with Tools of your choice

James Goodspeed – jgoodsp@fas.harvard.edu