DA607 Week 03 Assignment

Dan Fanelli

February 13, 2016

Pre Reqs:

- The table: **tb** was created using the script: https://raw.githubusercontent.com/fandang/DA607/master/Wk03/tb.sql
- The **tb** table was loaded with data from the csv: https://raw.githubusercontent.com/fandang/DA607/master/Wk03/tb.csv
- The **population** db table was loaded using **populations.csv** (https://raw.githubusercontent.com/fandang/DA607/master/Wk03/population.csv)
- The 2 tables were tested with SQL to be sure that the countries (the join column) were in sync:

```
select count(country) from population where country not in (select country from tb)

select count(country) from tb where country not in (select country from population)
```

- The results for both came back **0**, so the tables join columns seem to be in sync.
- The following query was run against the DB:

```
select population.country, population.year,
(IFNULL(tb.child, 0)+IFNULL(tb.adult, 0)+IFNULL(tb.elderly, 0)) as fatalities,
population.population,
((IFNULL(tb.child, 0)+IFNULL(tb.adult, 0)+IFNULL(tb.elderly, 0))/population.population) as rate
from population, tb
where population.country = tb.country and population.year = tb.year
group by population.country, population.year, population.population
order by rate desc
```

- The results of that query were exported to a csv and uploaded to: https://raw.githubusercontent.com/fandang/DA607/master/Wk03/joined_results.csv
- After that, the summary data was used to create a plot:

```
library(ggplot2)

dataUrl <- "https://raw.githubusercontent.com/fandang/DA607/master/Wk03/joined_results.csv"
df <- read.csv(dataUrl, header = TRUE)
head(df)
```

```
##       country year fatalities population   rate
## 1    Swaziland 2010       4817    1193148 0.0040
## 2 South Africa 2009     187046   50889543 0.0037
## 3 South Africa 2010     179291   51452352 0.0035
## 4    Swaziland 2006       3862    1118253 0.0035
## 5 South Africa 2008     154539   50267488 0.0031
## 6 South Africa 2007     136944   49602778 0.0028
```

Now, subset the data (pick a random few countries) to get an overall idea of the data, and plot it with GGPlot.

```
subsetDF <- subset(df, country=='Swaziland'|country=='Bangladesh'|country=='South Africa'|country=='Gua

ggplot(subsetDF,aes(x=year,y=rate,colour=country,group=country)) + geom_line()
```