UW STATR 510A - HW03

Melissa Gaughan

2022-10-15

Problem #1

Write a function called Median() which calculates the median of a vector of quantitative measurements and is robust to NA's in the vector. Test that your function gives the same answer as median() for any of the columns in the Country data set (which is full of NA's). Note that you will have to test to see if the length of the vector is even or odd. You may want to use a combination of modular division `%%` (see `?"%%"`) and the `if(TEST)` command, which will run a line (or lines) of code if the logical test is correct. An example is below:

```
a <- 1

if(a == 1) {
  print("It's true!")
}
```

```
## [1] "It's true!"
```

```
if(a!=1) {
  print("It's false!")
}
# No output here, because the test has failed
# The "if(a!=1)" could have been simply replaced with "else"

rm(a)
```

```
# Student Answer
test <- c(2, 4, 6, 8, 10, NA)

test_even <- c(0, 2, 4, 6, 8, 10, NA)


Median <- function(x) {
if(!is.numeric(x)){
  stop("Function requires numeric input")
}else{
 x_no_na <- na.omit(x) %>%
   sort()
 length_x <- length(x_no_na)
 if(length_x %% 2 == 1 ){
   x_median <- x_no_na[(length_x+1)/2]

 } else {
   x_median <- (x_no_na[length_x/2] +x_no_na[(length_x/2) + 1] )/2

 }
 x_median
}
}

Median(test)
```

```
## [1] 6
```

```
median(test, na.rm = T)
```

```
## [1] 6
```

```
Median(test_even)
```

```
## [1] 5
```

```
median(test_even, na.rm = T)
```

```
## [1] 5
```

```
#Median(letters)

country_test <- country_data %>%
  summarise(across(where(is.numeric), .fns = Median))

country_test_median <- country_data %>%
  summarise(across(where(is.numeric), .fns = median, na.rm = T))

country_test == country_test_median
```

```
##      Area Literacy Population  GDP Water Birthrate
## [1,] TRUE     TRUE       TRUE TRUE  TRUE      TRUE
```

Problem #2

We talked in lecture about "skewness" (sometimes denoted γ) being a characteristic of a distribution of numbers related to its asymmetry.

Part A

Much as the mean and variance of a population can be calculated using the definition of population mean $\overline{X}$ and population variance s_x^2, the skewness g of a population is given by the formula below:

$$g = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2)^{\frac{3}{2}}}$$

Write a function called Skew which computes the skewness of a vector of measurements (and make it robust to NA's).

Part B

Qualitatively, positive skewness indicate that there are more extremely high observations than low observations, and vice versa for negatively skewed distributions. Use your function to calculate the skewness of the following variables in the CountryData.csv dataset: global GDP, literacy and birth rates.

Which of these is most and least skewed, and in which direction? Does the skewness measure give you what you would expect, given what you know (and histograms you can plot) of these variables?

Answer: GDP is the most skewed. It has a positive skew of 1.8704649, indicating that some countries have much higher GDPs than others. Birth rate is the least skewed, with a γ value of 0.75 this does not surprise me, as wealth inequality is a clear problem in the global economic system. Birth rates being less skewed than literacy or GDP is sensible because there is a natural limit on how many children a child-bearing person can have.

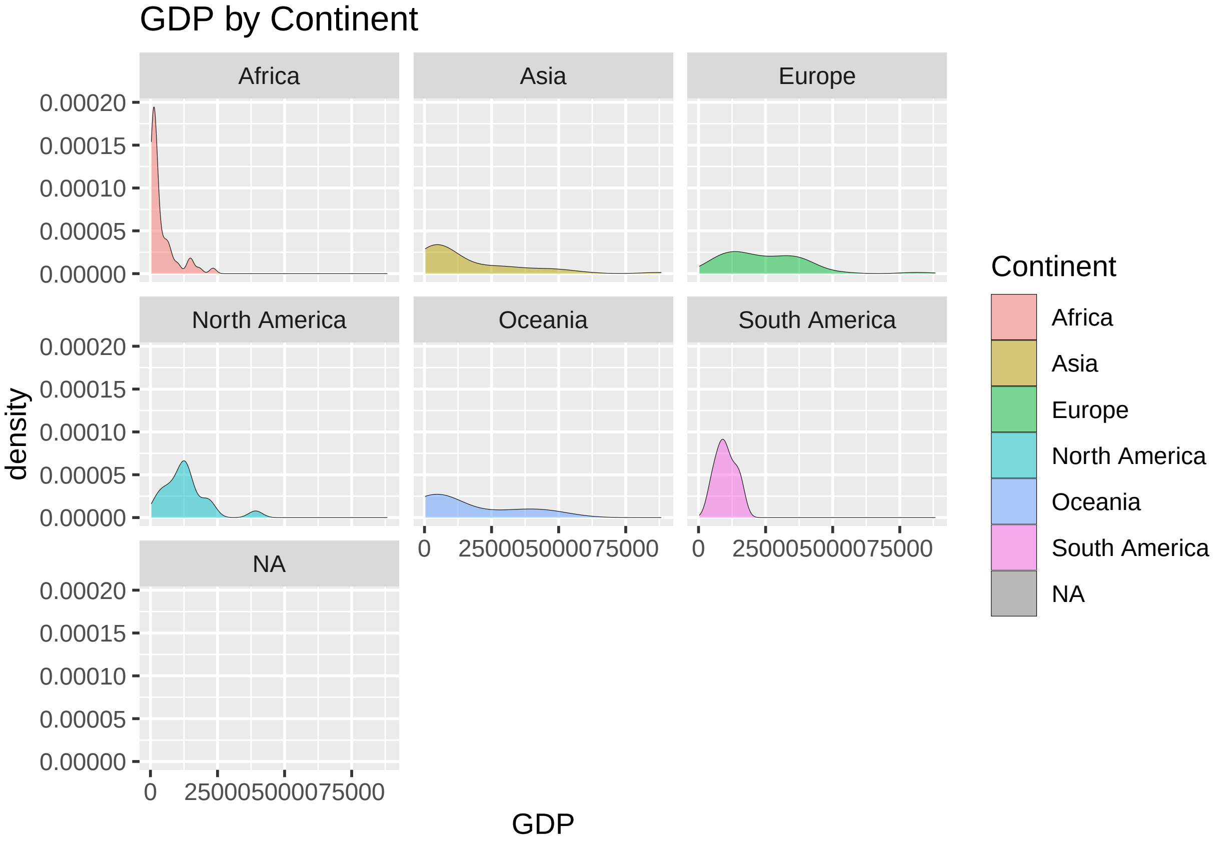
Part C

Generate a single `data.frame` that reports the skewness for GDP, literacy, and birth rates for each of the six continental regions (such that rows correspond to continents and there is a separate column corresponding to skewness values for each variable). (In Lecture03 Part A we discussed functions that call a function separately on different subsets of your data. Use one of these functions.)

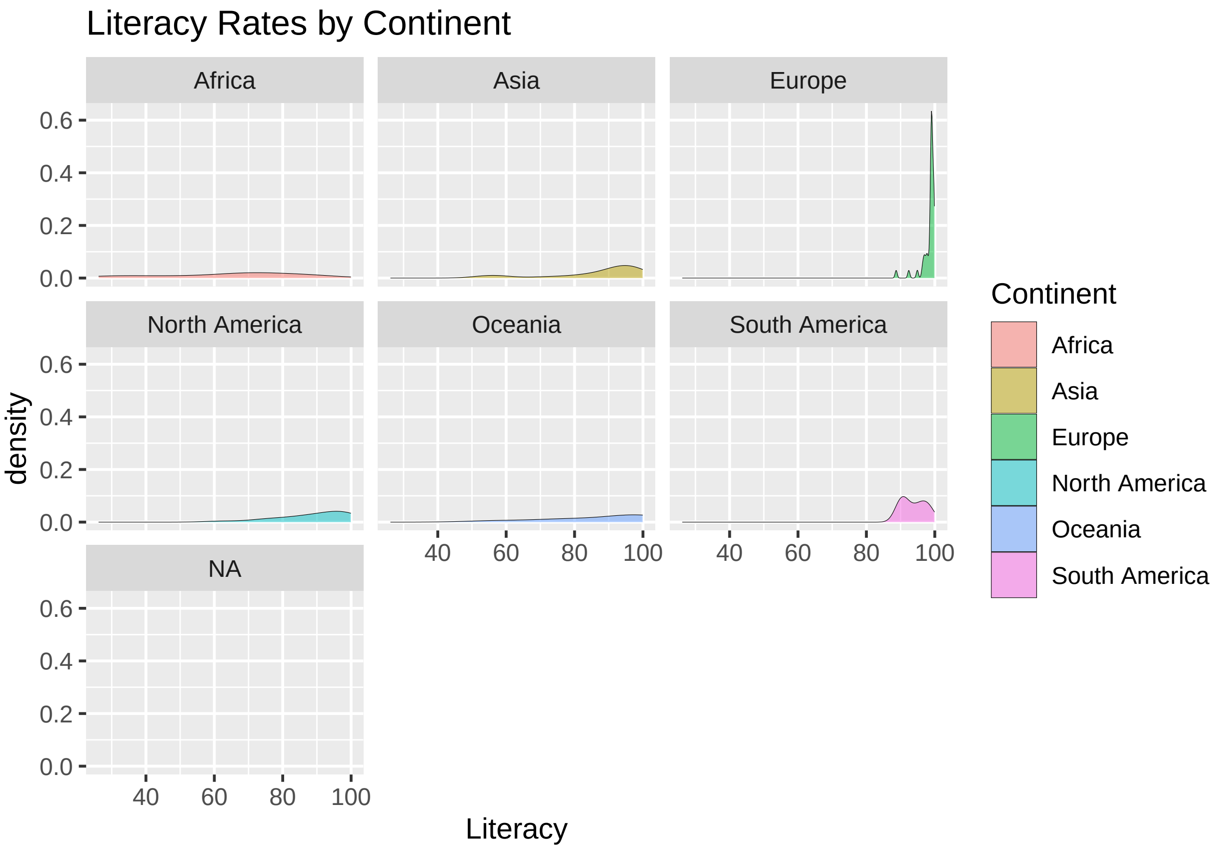
```
## Warning: Removed 48 rows containing non-finite values (stat_density).
```

```
## Warning: Groups with fewer than two data points have been dropped.
```

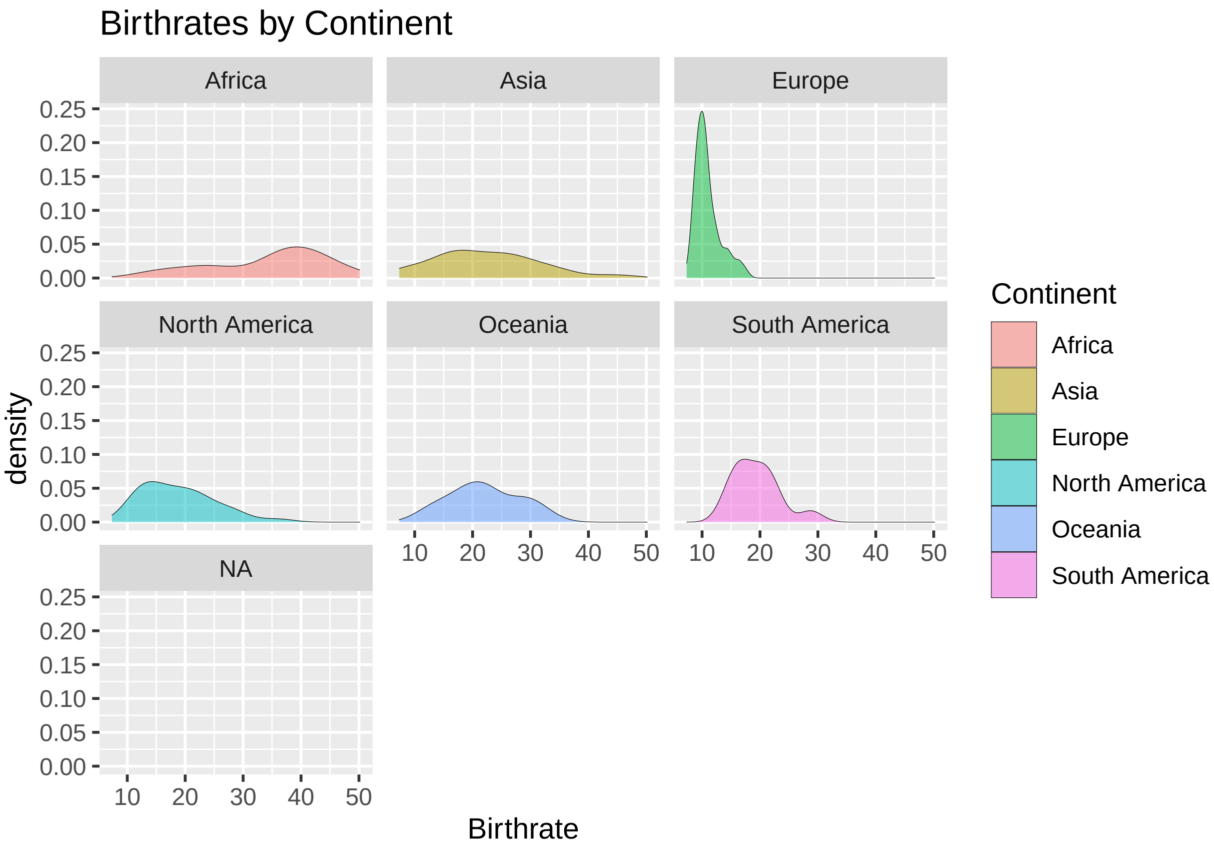
```
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf
```



```
## Warning: Removed 53 rows containing non-finite values (stat_density).
```



```
## Warning: Removed 16 rows containing non-finite values (stat_density).
```



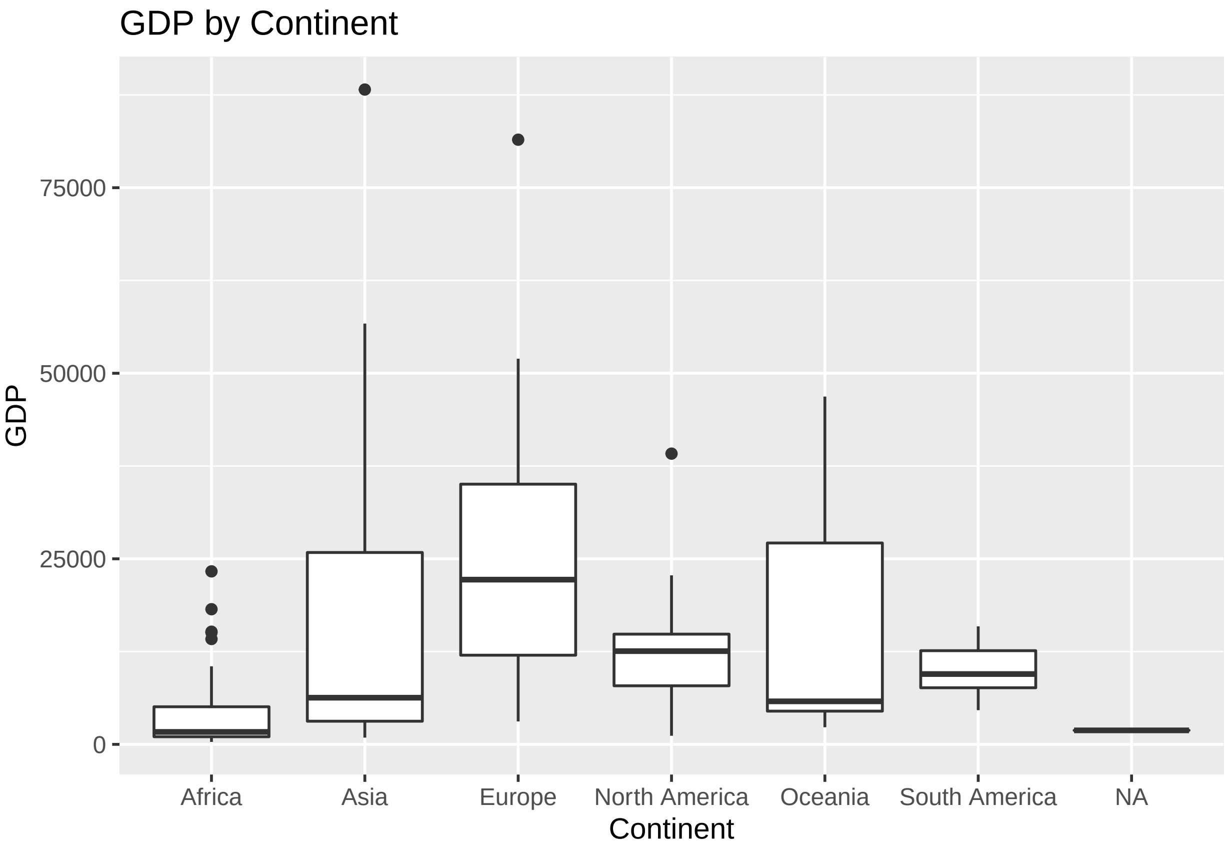
Report any differences in the signs of the skewness between the global and regional results. How would interpret the wide difference in, e.g., the skewness of GDP?

Skew for literacy, birthrate and GDP are clearly different for each continent when compared to the global skew. GDP has stark differences by continent, and each continent has different patterns of modality. For example, Africa's GDP is multimodal and North America's is bimodal. The clear differences in the skew and shape of the data at a continent level makes me question if the global GDP calculations are useful in showing the nuances of the data. I would say that the wide differences in skewness indicate differing levels of wealth disparity by continent. It is clear that wealth and economic power are not distributed equally in most parts of the world. Also, it is important to note that missing observations do matter in this analysis; the shape of a continent's distribution might look very different if missing data were added back into the dataset.

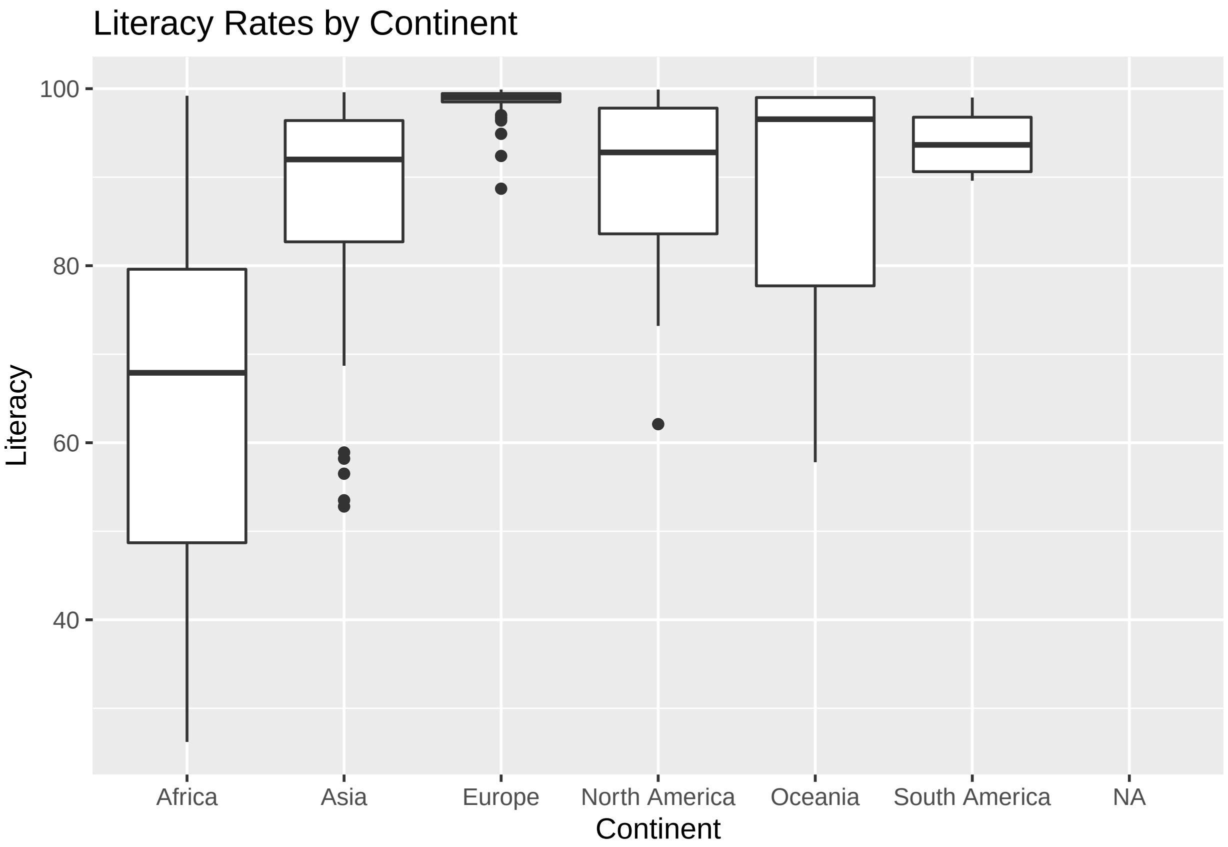
Part D

Now produce three separate boxplot plots (one for GDP, one for literacy, and one for birthrate) where the x-axis is continent and the y-axis is the countrydata variable. I.E. a boxplot for GDP would be produced with the code `boxplot(countrydata$GDP ~ countrydata$Continent)`. Make sure you label the y-axis with the `ylab` argument and add a title with the `title()` function (this is a function that can add to an existing plot).

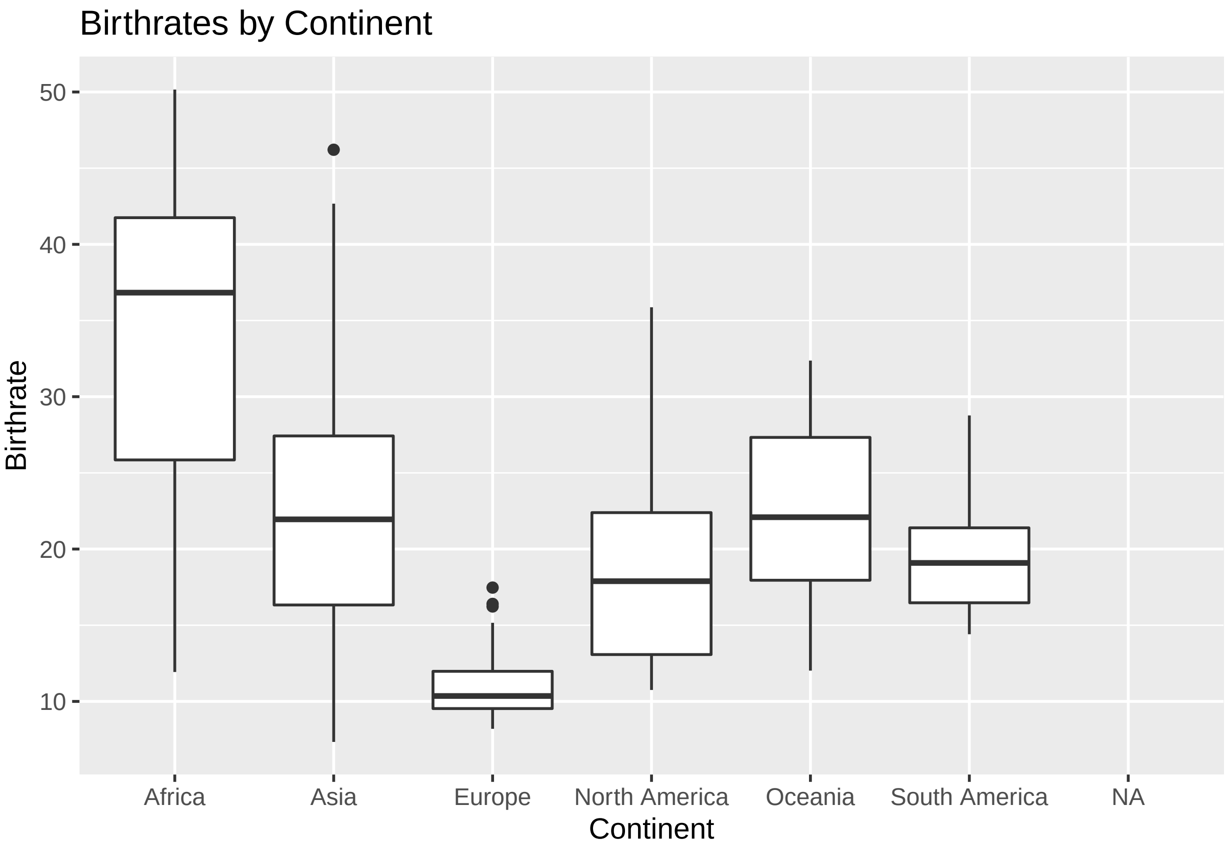
```
## Warning: Removed 48 rows containing non-finite values (stat_boxplot).
```



```
## Warning: Removed 53 rows containing non-finite values (stat_boxplot).
```



```
## Warning: Removed 16 rows containing non-finite values (stat_boxplot).
```



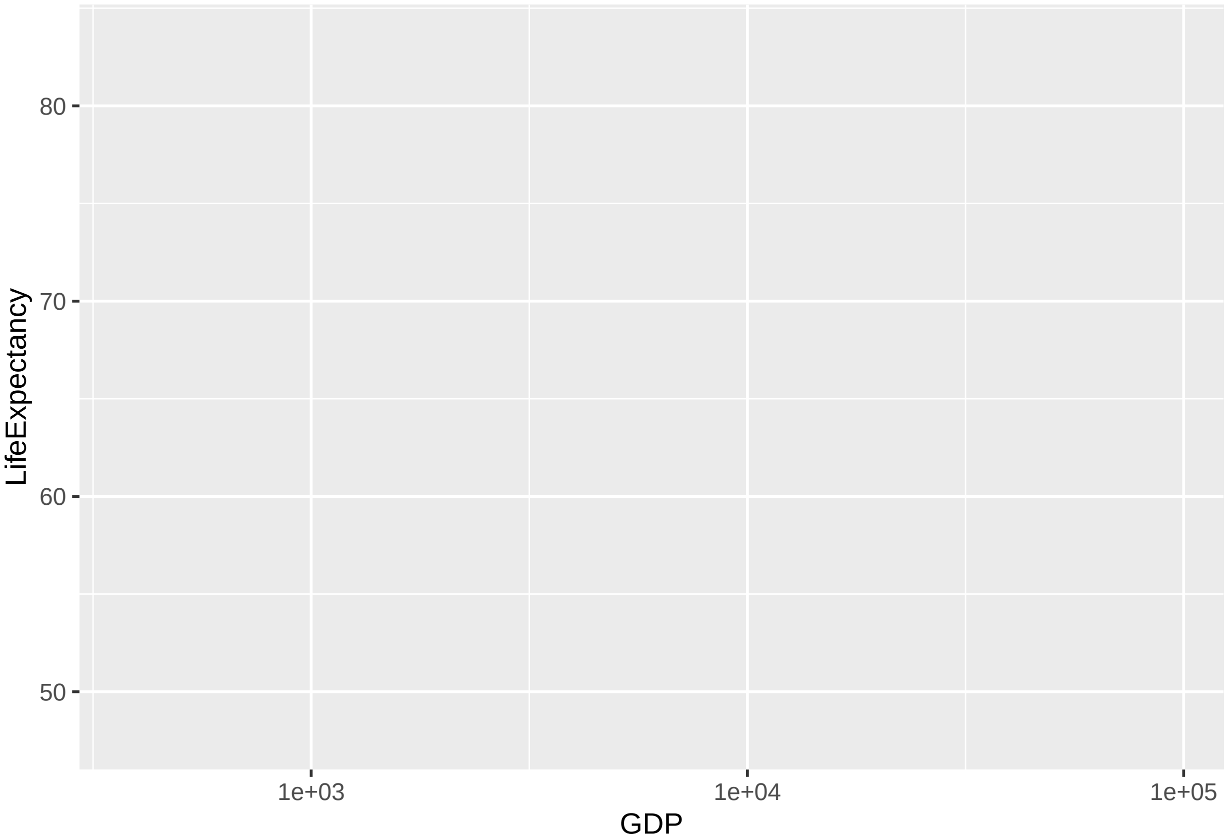
What visual feature of the boxplots seems most closely related to high or low values of skewness? **Answer:** The outliers (the dots) are closely related to measures of skewness.

Problem #3

Figure 1 shows the "Gapminder World Map 2010", which displays the condition of countries along an axis of wealth (GDP per capita) and health (life expectancy at birth). The corresponding data for 2011 has been downloaded and organized into a csv file. The goal of this question is to recreate in R as closely as is feasible the Gapminder World Map using the GDPLifeExpectancyRegion2011.csv (note that the more or less matching region data is in the Region2 column). Please read the lab3_part1.html file before attempting this question. It goes into great detail on how to create a similar plot. Try to include the following features:

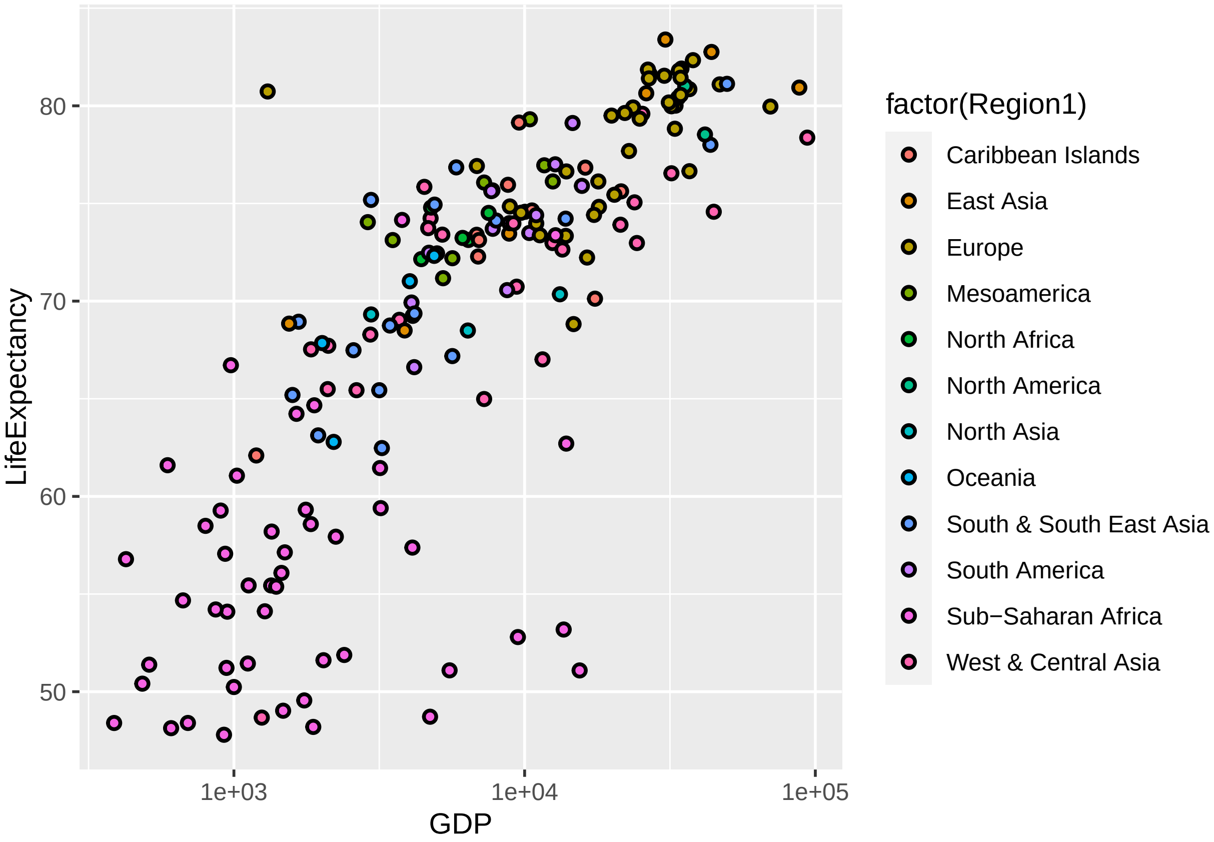
Part A

Start by creating a completely blank plotting window via `plot(Data$GDP, Data$LifeExpectancy, type="n", xaxt="n", yaxt="n")`. Argument `type="n"` will create a plot according to the range of the data without actually plotting it. Arguments `xaxt="n"` and `yaxt="n"` will create blank x and y axes. Try specifying argument `log="x"` in this plot call to get a logged x-axis (this won't be immediately apparent until you add axis ticks).



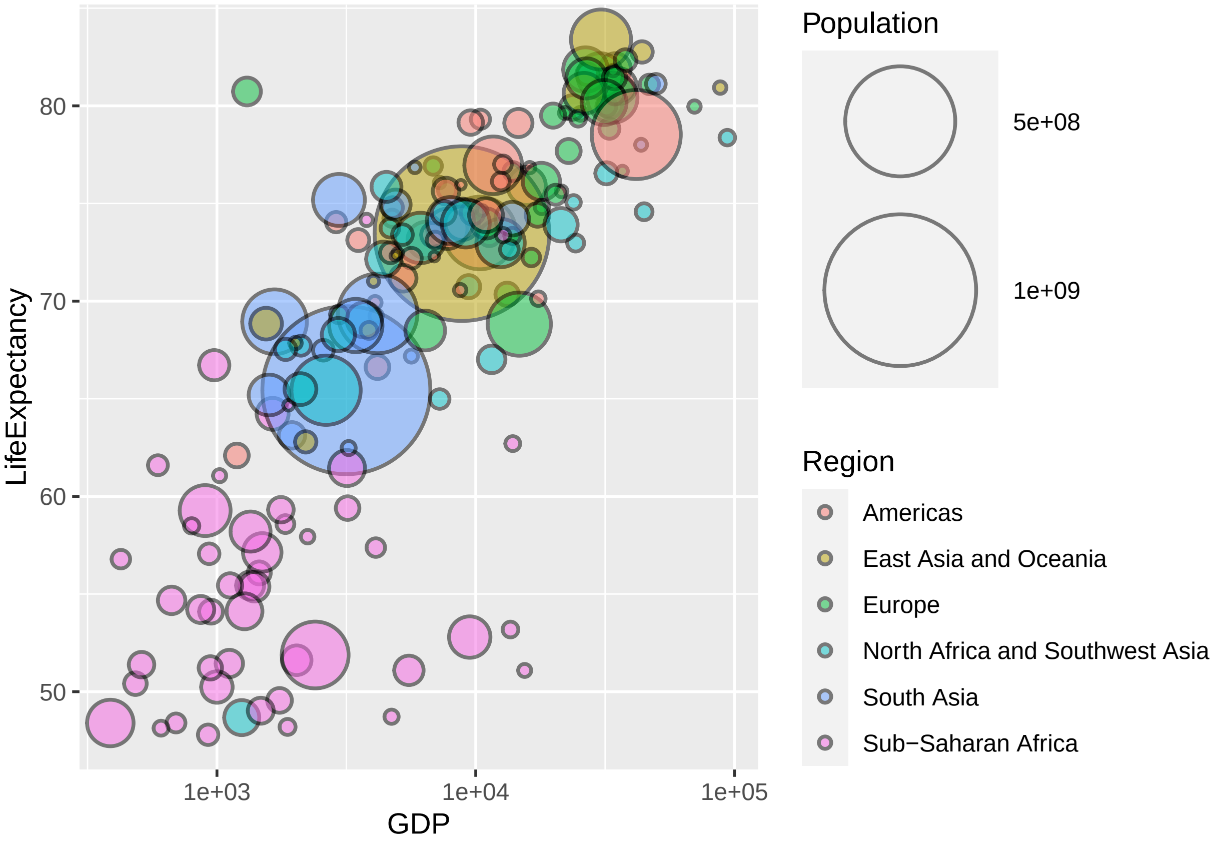
Part B

Make similar colors for the regional data, and similar types of circles (filled circles with black outlines). Argument `pch=21` works well to make this kind of point using the points() function. Colors can be specified by name. For example: `colors <- c("yellow", "red", "orange", "turquoise", "lightblue", "blue")`



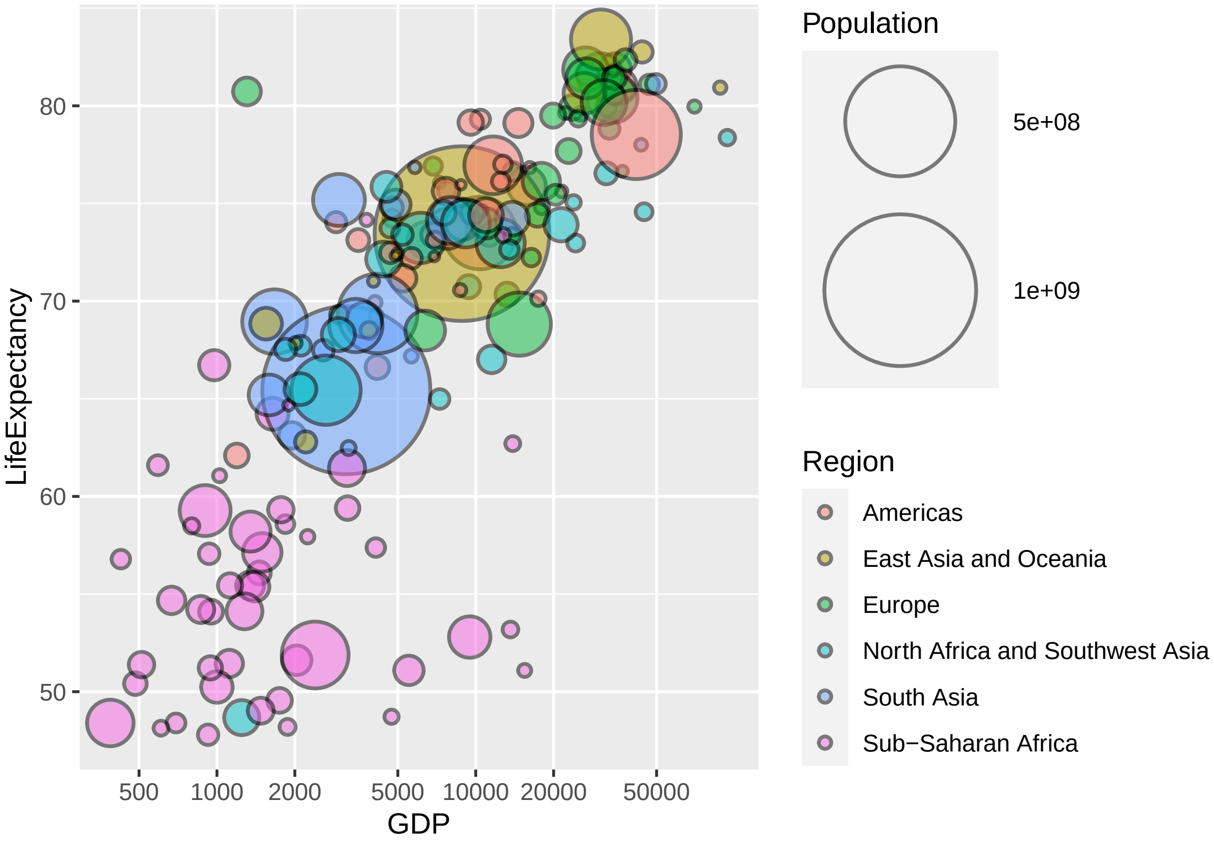
Part C

Similar scaling for the sizes of the circles. I recommend trying. Note that argument `bg` specifies the fill color of these points and that it needs to be the same length as the number of observations (I.E. `nrow(Data)`). Try to figure out what `colors[Data$Region2]`does and why its what you want here.



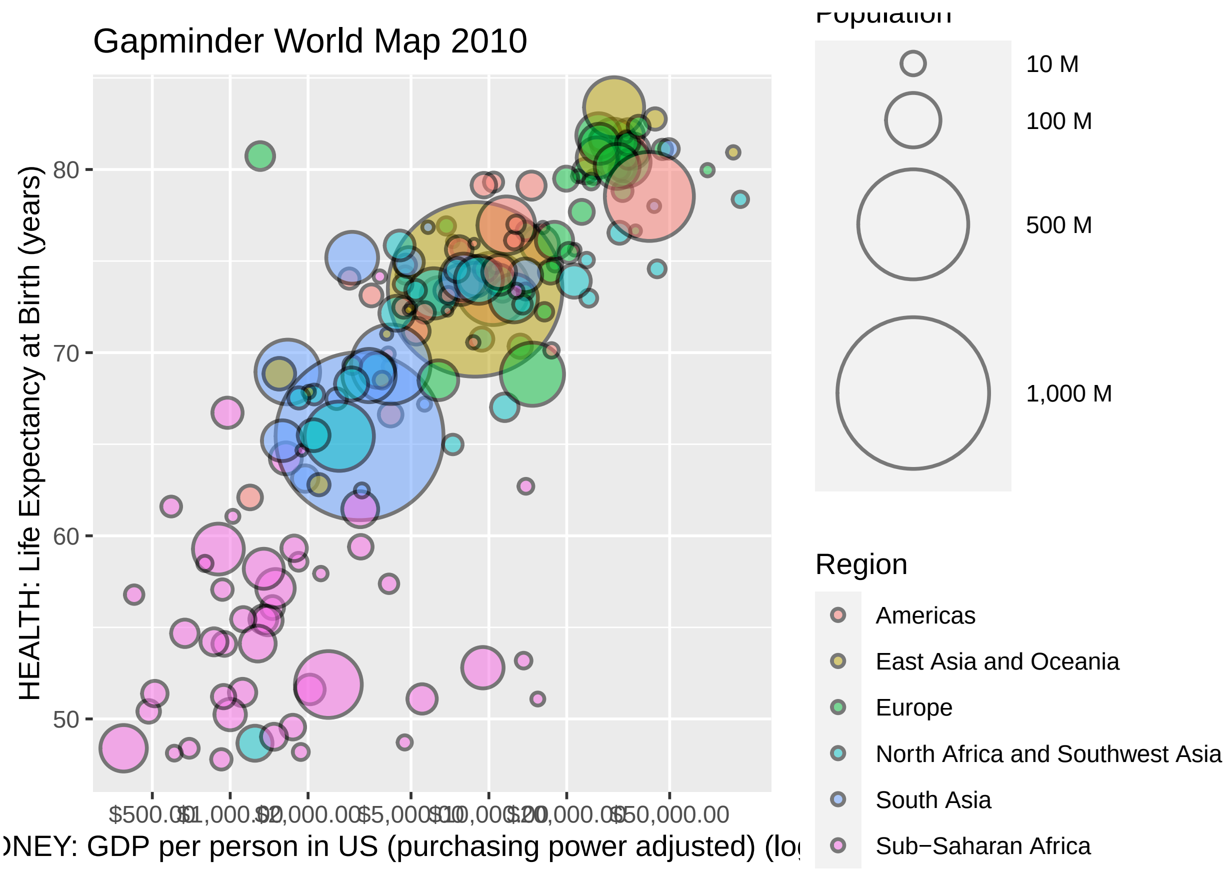
Part D

Similar grid lines (use the `abline(v = c(1,2,3))` and `abline(h = c(1,2,3))` functions).



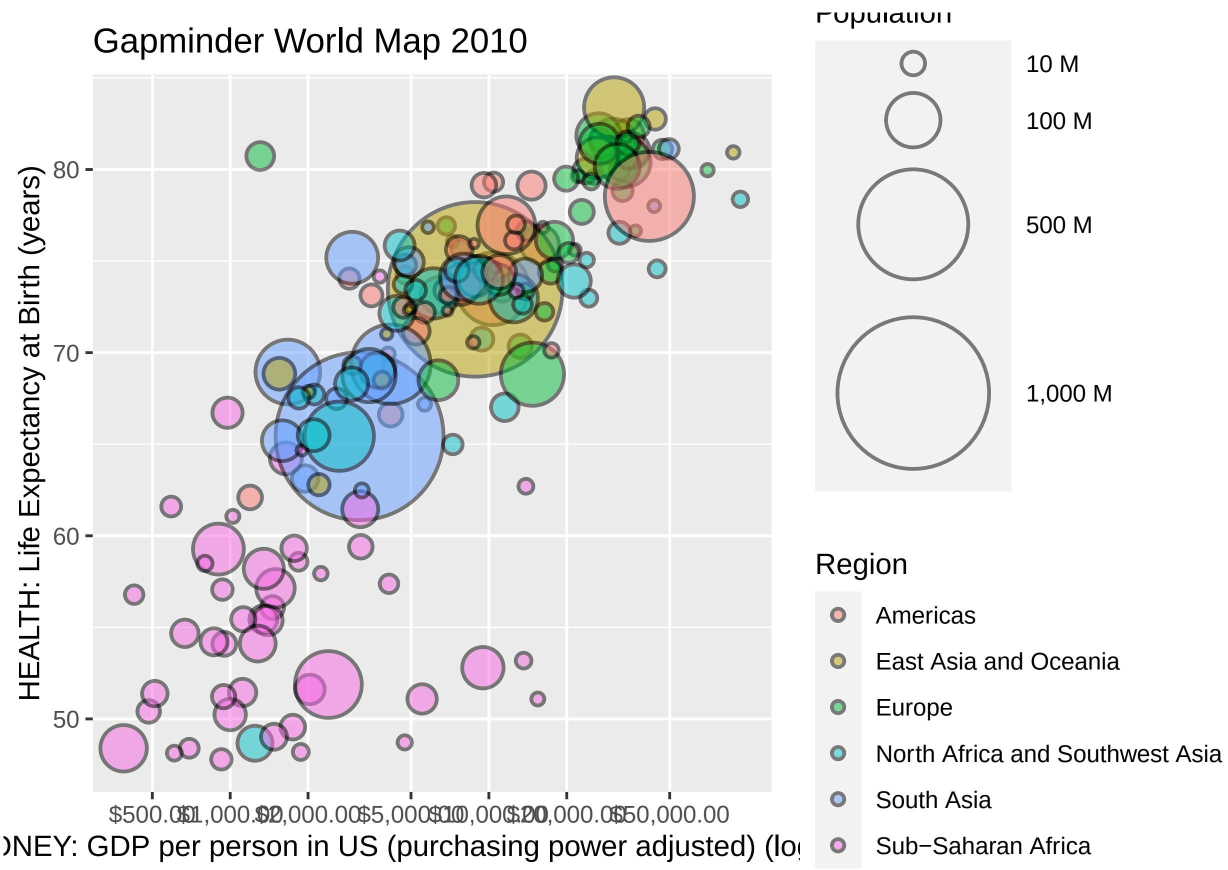
Part E

Similar labeling text on the margins (I.E. "Money GDP per person in US (purchasing power adjusted) (log scale)") You can produce these axes labels using the `xlab` and `ylab` arguments or using the `mtext()` function.



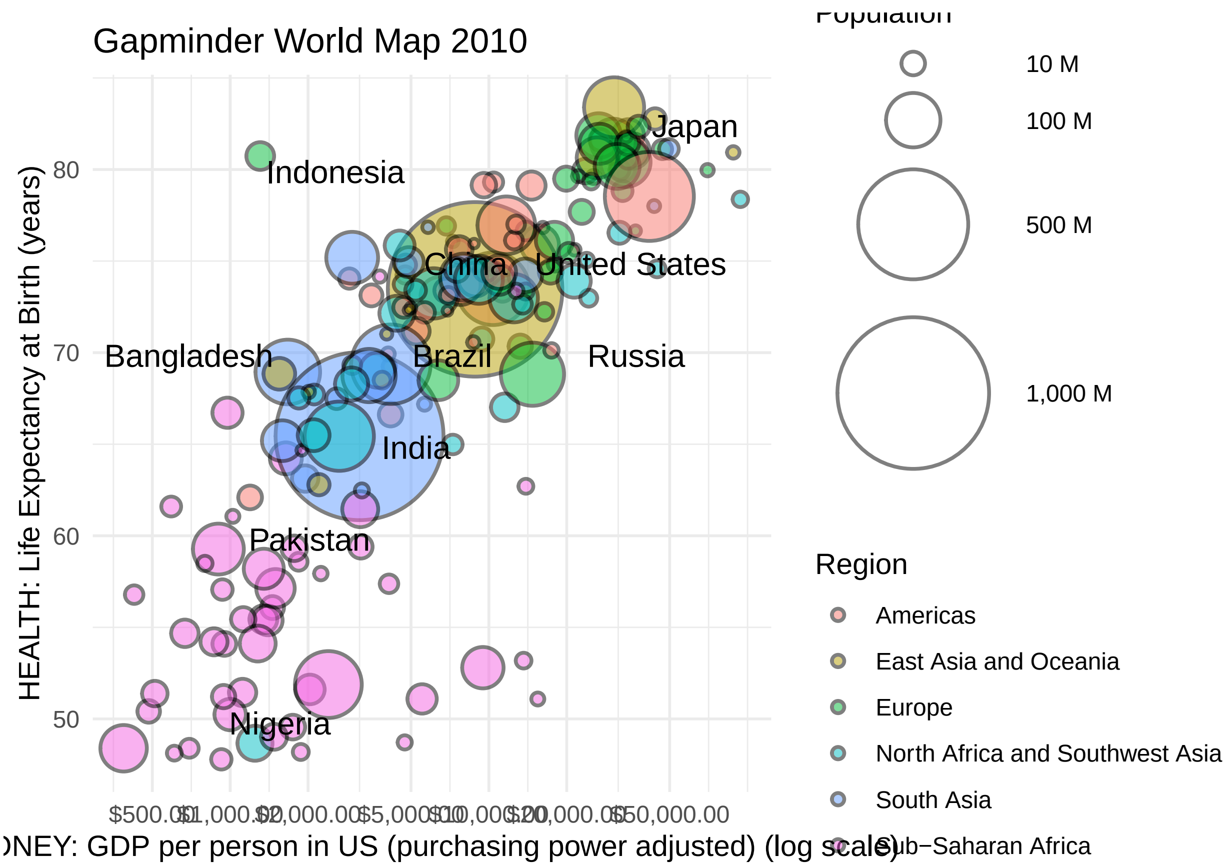
Part F

Two legends: one identifying the colors with the regions, and one matching circles to particular population sizes. Making this legend is tricky. We can discuss in office hours.



Part G

Label the 10 largest countries.



Reproducing the following features is difficult in R and not recommended for this assignment:

- The salmon colored exterior border color
- The alternating black and white bars in the axes/plot borders.
- The map legend (a simple mapping between point color and text is fine–see the lab3_part1.html file)
- The compass rose and background art fish/dragon.
- Matching the the exact font of “Gapminder World Map 2010” (try to include this text though)
- Additional text and copyright images in the bottom right of the legend.
- Individually placed text labels for each country. It is sufficient to label the 10 largest countries using `text()`

Note also that tweaking figures is a bottomless enterprise! Try to get a rough approximation, and learn some things along the way, but don’t go for perfection.

Problem #4

Reading: * *

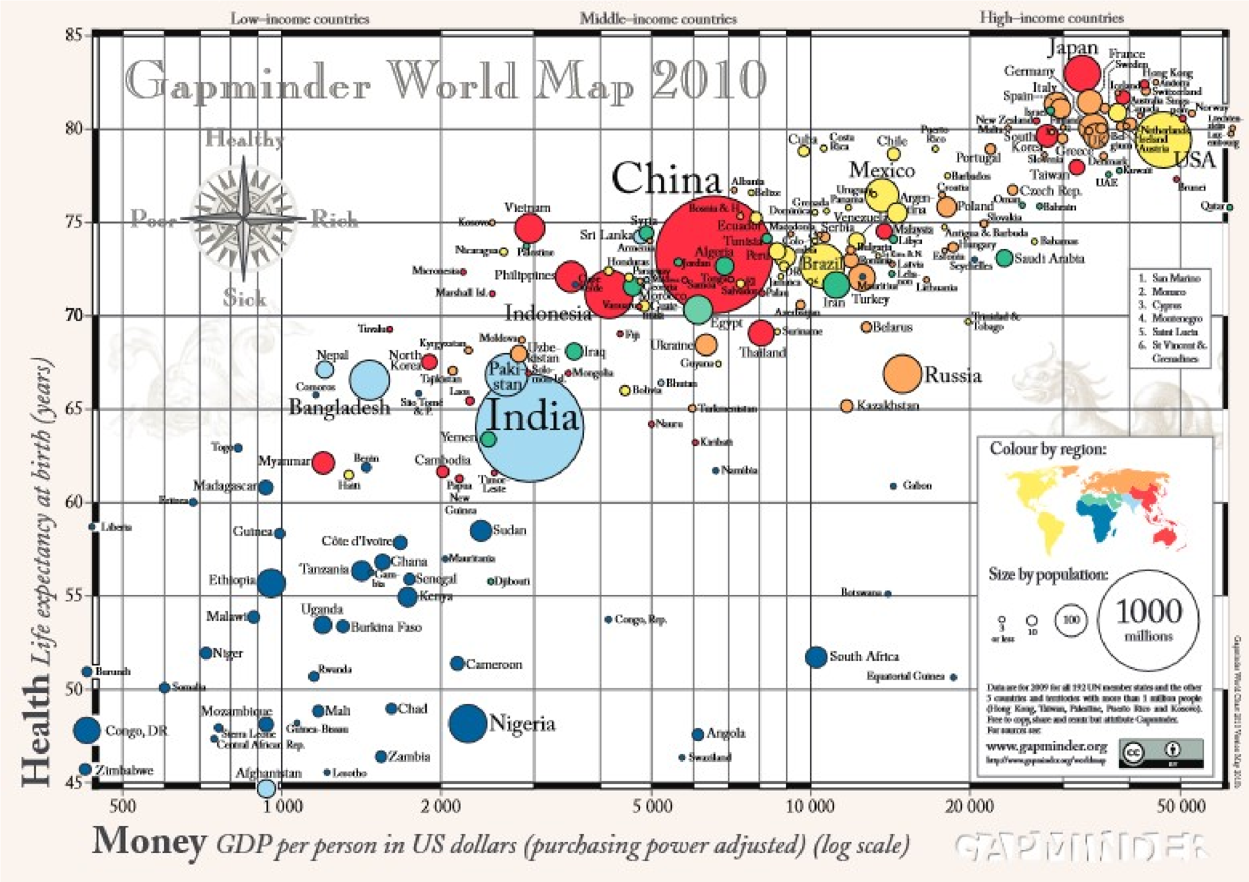


Figure 1: Gapminder 2010