Exercises: Introduction to R

Exercise 1

What are the values after each statement in the following?

```
mass <- 50  # mass?
age <- 30  # age?
mass <- mass * 2  # mass?
age <- age - 10  # age?
mass_index <- mass/age  # massIndex?
```

Exercise 2

See `?abs` and calculate the square root of the log-base-10 of the absolute value of `-4*(2550-50)`. Answer should be `2`.

Exercise 3

- Use the `c()` function to create/assign a new object that combines the `weights` and `animals` vectors into a single vector called `combined`.
- What happened to the numeric values? *Hint*: What's the `class()` of `combined`?
- Why do you think this happens?

Exercise 4

Sum the integers 1 through 100 and 501 through 600 (e.g. 1+2+...+99+100+501+502+...+599+600)

Exercise 5

1. What country and what years had a low GDP (<500) but high life expectancy (>50)?
2. What's the average GDP for Asian countries in 2002? How does that compare to European countries in the same year? To the Americas?

Exercise 6

Using the `with()`, do the following:

1. Compute the average GDP in billions for all Asian countries in 2007.
2. Do the same for Europe in 2007.

Hint: GDP per capita is the GDP divided by the population size. So to get GDP, you'd multiple `gdpPercap*pop`. To get that in billions, divide by 1,000,000, or more easily expressed in R using scientific notation: `1e9`.

Exercise 7

Plot GDP in trillions (`gdpPercap*pop/1e9`) on the y-axis versus population size in millions on the x-axis for all countries in the Americas. Use solid (`pch=16`) "blue" points, and give the plot a title and legends.

Exercises: Advanced Data Manipulation

Exercise 1

Load the `malebmi.csv` data. This is male BMI data from 1980 through 2008 from the gapminder project. If you downloaded the zip file from the repository, it's located in `workshops/lessons/r/data`. Alternatively you can donload it directly from http://bioconnector.org/data/.

1. Load this into a data frame object called `bmi`.
2. Turn it into a `tbl_df`.
3. How many rows and columns does it have?

Don't remove it – we're going to use it later on.

Exercise 2

Here's a warm-up round. Try the following.

What was the population of Peru in 1992? Show only the population variable. *Hint: 2 pipes; use filter and select.*

```
## Source: local data frame [1 x 1]
##
##        pop
## 1 22430449
```

Which countries and which years had the worst five GDP per capita measurements? *Hint: 2 pipes; use arrange and head.*

```
## Source: local data frame [5 x 6]
##
##            country continent year lifeExp      pop gdpPercap
## 1 Congo, Dem. Rep.    Africa 2002    45.0 55379852       241
## 2 Congo, Dem. Rep.    Africa 2007    46.5 64606759       278
## 3          Lesotho    Africa 1952    42.1   748747       299
## 4    Guinea-Bissau    Africa 1952    32.5   580653       300
## 5 Congo, Dem. Rep.    Africa 1997    42.6 47798986       312
```

What was the average life expectancy across all contries for each year in the dataset? *Hint: 2 pipes; group by and summarize.*

```
## Source: local data frame [12 x 2]
##
##    year mean(lifeExp)
## 1  1952          49.1
## 2  1957          51.5
## 3  1962          53.6
## 4  1967          55.7
## 5  1972          57.6
## 6  1977          59.6
## 7  1982          61.5
## 8  1987          63.2
## 9  1992          64.2
## 10 1997          65.0
## 11 2002          65.7
## 12 2007          67.0
```

Exercise 3

Which five Asian countries had the highest life expectancy in 2007? *Hint: 3 pipes.*

```
## Source: local data frame [5 x 6]
##
##             country continent year lifeExp       pop gdpPercap
## 1             Japan      Asia 2007    82.6 127467972     31656
## 2 Hong Kong, China      Asia 2007    82.2   6980412     39725
## 3            Israel      Asia 2007    80.7   6426679     25523
## 4         Singapore      Asia 2007    80.0   4553009     47143
## 5       Korea, Rep.      Asia 2007    78.6  49044790     23348
```

How many countries are on each continent? *Hint: 2 pipes.*

```
## Source: local data frame [5 x 2]
##
##   continent n_distinct(country)
## 1    Africa                 52
## 2  Americas                 25
## 3      Asia                 33
## 4    Europe                 30
## 5   Oceania                  2
```

Separately for each year, compute the correlation coefficients (e.g., `cor(x,y)`) for life expectancy (y) against both log10 of the population size and log10 of the per capita GDP. What do these trends mean? *Hint: 2 pipes.*

```
## Source: local data frame [12 x 3]
##
##    year cor(log10(pop), lifeExp) cor(log10(gdpPercap), lifeExp)
## 1  1952                 0.1543                           0.748
## 2  1957                 0.1584                           0.759
## 3  1962                 0.1376                           0.771
## 4  1967                 0.1482                           0.773
## 5  1972                 0.1322                           0.789
## 6  1977                 0.1142                           0.814
## 7  1982                 0.0944                           0.846
## 8  1987                 0.0732                           0.874
## 9  1992                 0.0593                           0.856
## 10 1997                 0.0636                           0.864
## 11 2002                 0.0746                           0.825
## 12 2007                 0.0653                           0.809
```

Compute the average GDP (not per-capita) in billions averaged across all contries separately for each continent separately for each year. What continents/years had the top 5 overall GDP? *Hint: 6 pipes. If you want to arrange a dataset by a value computed on grouped data, you first have to pass that resulting dataset to a funcion called* ***`ungroup()`*** *before continuing to operate.*

```
## Source: local data frame [5 x 3]
##
##   continent year meangdp
## 1  Americas 2007     777
## 2  Americas 2002     661
## 3      Asia 2007     628
## 4  Americas 1997     583
## 5    Europe 2007     493
```

Exercises: Advanced Manipulation

Exercise 1

Re-create this same plot from scratch without saving anything to a variable. That is, start from the `ggplot` call.

- Start with the `ggplot()` function.
- Use the gm data.
- Map `gdpPercap` to the x-axis and `lifeExp` to the y-axis.
- Add points to the plot
- Make the points size 4
- Map continent onto the aesthetics of the point
- Use a log10 scale for the x-axis.

Exercise 2

1. Make a scatter plot of `lifeExp` on the y-axis against `year` on the x.
2. Make a series of small multiples faceting on continent.
3. Add a fitted curve, smooth or lm, with and without facets.
4. **Bonus**: using `geom_line()` and and aesthetic mapping `country` to `group=`, make a "spaghetti plot", showing *semitransparent* lines connected for each country, faceted by continent. Add a smoothed loess curve with a thick (`lwd=3`) line with no standard error stripe. Reduce the opacity (`alpha=`) of the individual black lines.

Exercise 3

1. Make a jittered strip plot of GDP per capita against continent.
2. Make a box plot of GDP per capita against continent.
3. Using a log10 y-axis scale, overlay semitransparent jittered points on top of box plots, where outlying points are colored.
4. **BONUS**: Try to reorder the continents on the x-axis by GDP per capita. Why isn't this working as expected? See `?reorder` for clues.

Exercise 4

1. Plot a histogram of GDP Per Capita.
2. Do the same but use a log10 x-axis.
3. Still on the log10 x-axis scale, try a density plot mapping continent to the fill of each density distribution, and reduce the opacity.
4. Still on the log10 x-axis scale, make a histogram faceted by continent *and* filled by continent. Facet with a single column (see `?facet_wrap` for help). Save this to a 6x10 PDF file.