Homework 0

Name: Put your name here

I worked with:

Click the "Knit" button in RStudio to knit this file to a pdf.

```
# load necessary packages
library(ggplot2)
library(dplyr)
library(gapminder)
library(babynames)
```

Problem 1: Gapminder

The `gapminder` package includes a data frame called `gapminder`, containing information about different countries from 1952 to 2007. We are interested in the year 2007 and we are going to save this in a new data frame `gapminder_2007` below. We use data wrangling using code from the `dplyr` package. We will learn about in detail later this later.

```
gapminder_2007 <- gapminder %>%
  filter(year == 2007)
```

Run `View(gapminder_2007)` in your console to explore this data. An alternative method for exploring a data frame is by using the `glipmse()` function:

```
glimpse(gapminder_2007)
Rows: 142
Columns: 6
$ country   <fct> "Afghanistan", "Albania", "Algeria", "Angola", "Argentina", ~
$ continent <fct> Asia, Europe, Africa, Africa, Americas, Oceania, Europe, Asi~
$ year      <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, ~
$ lifeExp   <dbl> 43.828, 76.423, 72.301, 42.731, 75.320, 81.235, 79.829, 75.6~
$ pop       <int> 31889923, 3600523, 33333216, 12420476, 40301927, 20434176, 8~
$ gdpPercap <dbl> 974.5803, 5937.0295, 6223.3675, 4797.2313, 12779.3796, 34435~
```

The function below plots the life expectancy Vs GDP per capita (in USD) of all the countries in the `gapminder` dataset. You can see these individual points color-coded by continent and the size of the points correspond to the population of the country they correspond to. We will dissect this function in detail in the coming weeks.

```
ggplot(data = gapminder_2007,
       mapping = aes(x = gdpPercap, y = lifeExp, size = pop, color = continent)) +
  geom_point() +
  # Note: this was not required for this problem set:
  labs(x = "GDP per capita (in USD)",
       y = "Life expectancy",
       size = "Population",
       color = "Continent",
       title = "Gapminder data",
       subtitle = "2007 values")
```

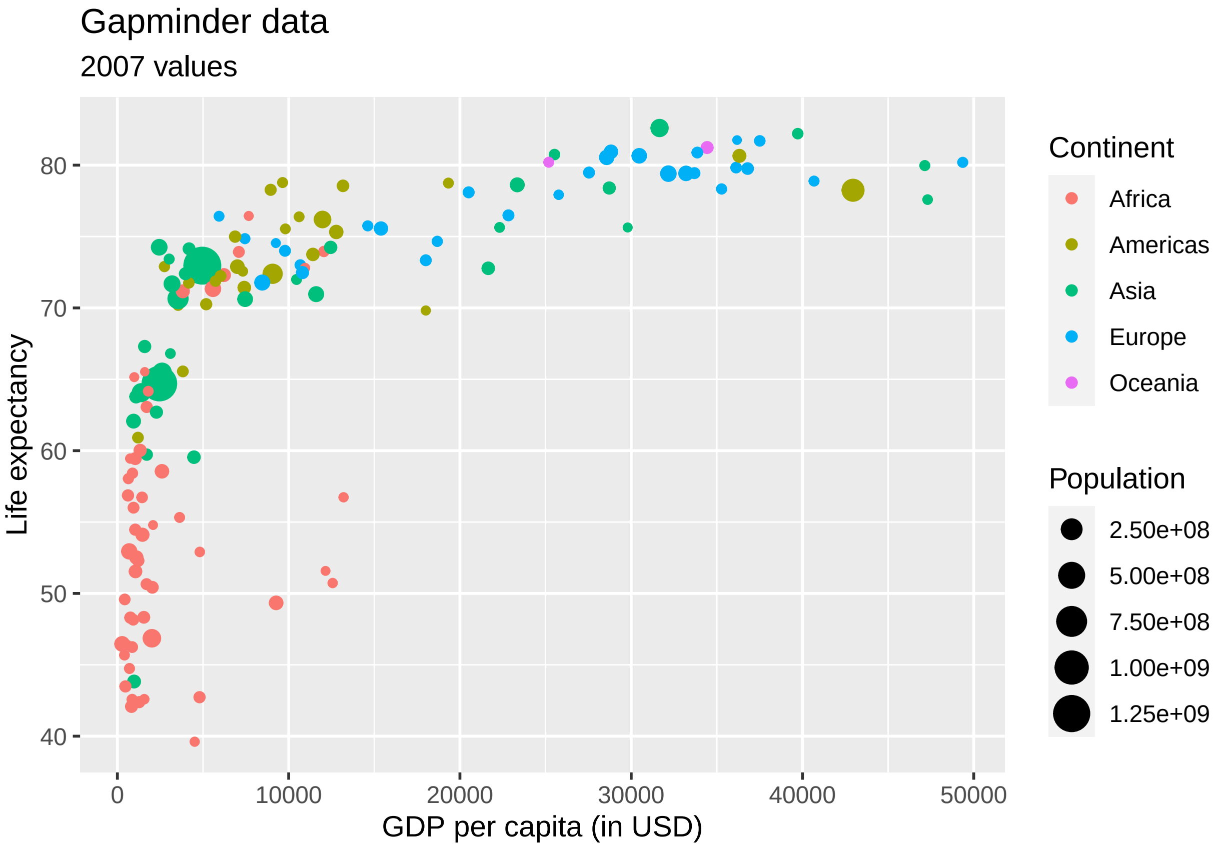

Figure 1: A nice visualization

a. Comment on the two biggest green circles that you see in the Figure 1 above.

Answer:

Problem 2: Babynames

The `babynames` package provides data about the popularity of individual baby names from the US Social Security Administration. Data includes all names used at least 5 times in a year beginning in 1880.

In the code chunk below, we extract the data corresponding to `Aimee` or `Sammy` from `babynames` dataset and store it in a new dataset `babynames_Aimee_Sammy`.

```
babynames_Aimee_Sammy <- babynames %>%
  filter(name == "Aimee" | name == "Sammy")
```

a. What do you see in Figure 2 below? We will learn the functions in detail later.

```
ggplot(babynames_Aimee_Sammy, aes(x=year, y=prop, col=sex)) +
  geom_line() +
  facet_wrap(~name) +
  labs(x = "Year", y = "Proportion", color = "Sex", title = "Comparison of Aimee and Sammy")
```

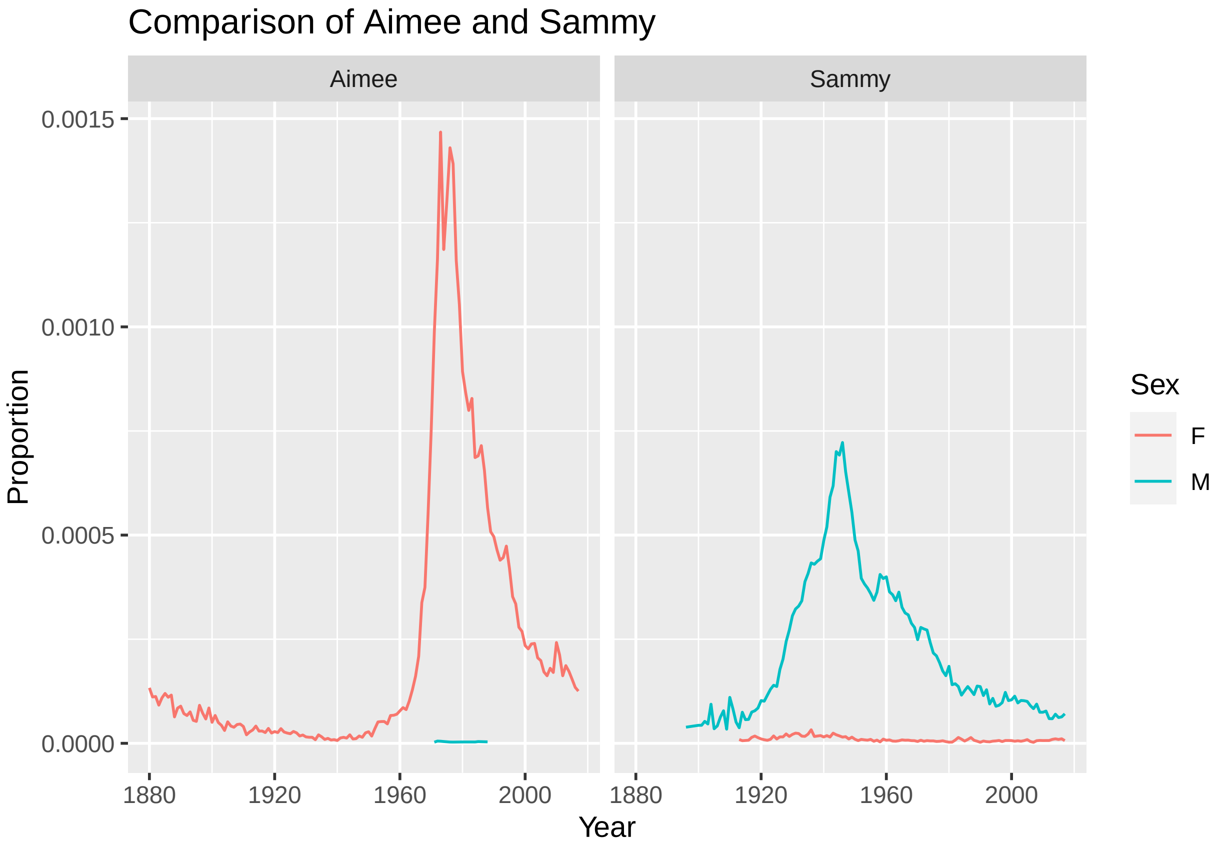

Figure 2: A nice comparison

Answer: