STAT GU4206/GR5206 Homework 5 [100 pts]
Due 11:59pm Friday, November 10th on Canvas

In this homework you'll explore transforming data with split/apply/combine and the `plyr` package.

Gross domestic product (GDP) is a measure of the total market value of all goods and services produced in a given country in a given year. The percentage growth rate of GDP in year t is

$$100 \times \left(\frac{GDP_{t+1} - GDP_t}{GDP_t} \right) - 100$$

An important claim in economics is that the rate of GDP growth is closely related to the level of government debt, specifically with the ratio of the government's debt to the GDP. The file `debt.csv` on the class website contains measurements of GDP growth and of the debt-to-GDP ratio for twenty countries around the world, from the 1940s to 2010. Note that not every country has data for the same years, and some years in the middle of the period are missing data for some countries but not others. **Throughout, use 3 significant digits for numerical answers!!** (That is, `signif(mydat,3)` is your friend).

```
debt <- read.csv("debt.csv", as.is = TRUE)
dim(debt)
head(debt)
```

1. Calculate the average GDP growth rate for each country (averaging over years). This is a classic split/apply/combine problem, and you will use `daply()` to solve it.

    a. Begin by writing a function, `mean.growth()`, that takes a data frame as its argument and returns the mean of the 'growth' column of that data frame.

    b. Use `daply()` to apply `mean.growth()` to each country in `debt`. Don't use something like `mean(debt$growth[debt$Country=="Australia"])`, except to check your work. You should not need to use a loop to do this. (The average growth rates for Australia and the Netherlands should be 3.72 and 3.03. Print these values.) Report the average GDP growth rates clearly.

2. Using the same instructions as problem 1, calculate the average GDP growth rate for each year (now averaging over countries). (The average growth rates for 1972 and 1989 should be 5.63 and 3.19, respectively. Print these values in your output.) Make a plot of the growth rates (y-axis) versus the year (x-axis). Make sure the axes are labeled appropriately.

3. The function `cor(x,y)` calculates the correlation coefficient between two vectors `x` and `y`.

    a. Calculate the correlation coefficient between GDP growth and the debt ratio over the whole data set (all countries, all years). Your answer should be -0.1995.

    b. Compute the correlation coefficient separately for each country, and plot a histogram of these coefficients (with 10 breaks). The mean of these correlations should be -0.1778. Do not use a loop. (Hint: consider writing a function and then making it an argument to `daply()`).

    c. Calculate the correlation coefficient separately for each year, and plot a histogram of these coefficients. The mean of these correlations should be -0.1906.

    d. Are there any countries or years where the correlation goes against the general trend?

4. Fit a linear model of overall growth on the debt ratio, using `lm()`. Report the intercept and slope. Make a scatter-plot of overall GDP growth (vertical) against the overall debt ratio (horizontal). Add a line to your scatterplot showing the fitted regression line.

5. There should be four countries with a correlation smaller than -0.5. Separately, plot GDP growth versus debt ratio from each of these four countries and put the country names in the titles. This should be four plots. Call `par(mfrow=c(2,2))` before plotting so all four plots will appear in the same figure.

    (Think about what this shows: individual relationships at the country level are sometimes concealed or "smudged out" when data is aggregated over all groups (countries). This conveys the importance of careful analysis at a more granular group level, when such groupings are available!)

6. Some economists claim that high levels of government debt cause slower growth. Other economists claim that low economic growth leads to higher levels of government debt. The data file, as given, lets us relate this year's debt to this year's growth rate; to check these claims, we need to relate current debt to future growth.

    a. Create a new data frame which just contains the rows of `debt` for France, but contains all those rows. It should have 54 rows and 4 columns (print the dimensions of your data frame). Note that some years are missing from the middle of this data set.

    b. Create a new column in your data frame for France, `next.growth`, which gives next year's growth **if** the next year is in the data frame, or `NA` if the next year

is missing. (`next.growth` for 1971 should be (rounded) 5.886, but for 1972 it should be `NA`. Print these two values.)

7. Add a `next.growth` column, as in the previous question, to the **whole** of the `debt` data frame. Make sure that you do not accidentally put the first growth value for one country as the `next.growth` value for another. (The `next.growth` for France in 2009 should be `NA`, not 9.167. Print this value.) **Hints**: Write a function to encapsulate what you did in the previous question, and apply it using `ddply()`.

8. Make a scatter-plot of next year's GDP growth against this year's debt ratio. Linearly regress next year's growth rate on the current year's debt ratio, and add the line to the plot. Report the intercept and slope to reasonable precision. How do they compare to the regression of the current year's growth on the current year's debt ratio?

9. Make a scatter-plot of next year's GDP growth against the current year's GDP growth. Linearly regress next year's growth on this year's growth, and add the line to the plot. Report the coefficients. Can you tell, from comparing these two simple regressions (from the current question, and the previous), whether current growth or current debt is a better predictor of future growth?