In this assignment, you will practice some basic R data preparation tasks on a data set that contains life expectancy, population, gdp, and location for several countries over many years. There are four data sets that you'll be using. Each is stored as a .csv and can be downloaded above and then read into R as a data frame. Here is a description of each data set.

ccA: life expectancy and population values for various countries in Europe and Asia over the course of several years
ccB: same data as ccA, but for Africa, the Americas and Oceania
gdp: gdp per capita for various countries and years
coords: lat/long coordinates for 252 countries

Perform the following tasks and summarize the results using the R Markdown template provided on the Canvas landing page. You should show your code and discuss your findings, but avoid nonsensical output, such as printing entire data frames. Upload a .pdf of your .Rmd file by the due date.

1. Use the `str()` command to examine the characteristics of each of the four databases. Then append `ccA` to `ccB` to make one large data frame and reexamine the output with `str()`.

2. Append the `gdp` data to the data frame you created in Number 1. Be careful here and make sure you examine the structure of the resulting data frame to ensure things are correct. Examine the output with `str()`.

3. How many times after 1980 did a country have gdp<20000?

4. Merge your data frame from 1 &2 above with the `coords` data via three different techniques: inner, right, and left merge. Compare the dimensions of each of the resulting data frame and explain any differences you see.

5. Suppose a researcher has a hypothesis that there is a relationship between gdp and distance from the equator, i.e. latitude. To explore this, the researcher would like to break the gdp into quartiles and then look at the mean latitude (in absolute value) for each quartile. To do this, perform the following tasks:

   a.) Using the data from the inner merge, use the `quantile()` function to determine the quartiles of gdp

   b.) Create a new variable in the data set called gdp.q based on the quartile that the a given observation's gdp value lies in. Print a frequency table of this new variable and explain how you can use it to demonstrate that your procedure is working properly. (The `findInterval()` function might be helpful).

   c.) Using dplyr's group_by and summarize functioins, find the mean absolute value of the latitude for each quartile. Based on your findings, draw a conclusion about the relationship between GDP and distance from the equator. Briefly speculate on the explanation for your findings. Why must you use absolute value?

6. (Bonus) Create a new data set that, for each year, lists the median gdp and also identifies the country whose gdp is nearest to the median. You should use the technique we reviewed in class to build a data frame within a loop. Below is some pseudo-code to get you started. I also believe the functions `unique()` and `which()` will be helpful.

```
create empty results data frame
extract the number of unique years
loop through all years
      -create a subset of the data of interest
      -calculate the median gdp
      -find the index of country nearest to the median
      -use the index to find the country name, continent, and actual gdp
      -store the results in a temp data frame
      -concatenate the results df with the temp df
```