STAT 6430
SAS PROJECT 1

You must submit your work online via Collab before 10am on Friday, August 7.

Each group must submit one pdf file and one SAS program.

All the work submitted for assessment must be entirely the work of the individuals within your group - do not seek help from others or copy any code. If you have questions about the project you may contact Dr Tomas.

Make sure to include a cover page on your text document which includes a signed honor pledge (e-signatures are acceptable). This should read:

> *"On my honor, I have neither given nor received help on this assignment from anybody outside of the group to which I was assigned."*

Assessment

Each project will be graded out of a total of 100 points, based on the assessment criteria below.

Criterion	Description	Points
Code Clarity	How easy your code is to understand (sufficient comments etc)	5
Programming Style	Use of comments, titles, indentation, spacing, etc. Your code might also contain tests to make sure that data steps and procedures have run as expected.	15
Results	Does your code generate the correct answers and output?	45
Efficiency	Are you results obtained in an efficient/straight-forward way, without much more code than is necessary?	5
Report	Should be clear, comprehensive, concise, easy to follow, and aligned with the code.	30

Format of the Project Submission

One member of each group should submit the following two electronic documents via the Collab site to represent the work of the entire group:

1. A SAS program in the form of a `.sas` file that gives the code used to complete the steps below. Please use comments to clearly demarcate the code you used to answer each question.
   You do not need to include all the code you used to make mistakes, test the output, etc, but only include what is needed to perform your final analysis. Ideally, your final program files should run through without errors.
   Please place any library references and file references near the beginning of your program so that they can be easily modified by the grader.

   Files should be named in the following format: `STAT6430_GroupNumber`, where `GroupNumber` is equal to the number assigned to your group (see the Excel spreadsheet).

2. A PDF document that describes the process used to create the final dataset `FinalData.csv`. This should be a reference document that could be used by your boss, colleagues, or a data analyst to understand the steps taken to get from the raw data files to the final "analysis" dataset. It should not include any SAS code, but should clearly indicate the data sources used and how they were modified and combined to produce the final dataset. You may include pseudo-code or graphics if it helps to get your message across clearly. As a guide, this document should be 1–3 pages long.

A good question to ask when reviewing your report is:

> Could a good R programmer (say) with no SAS knowledge produce a dataset identical to the final dataset we produced, just using the information in our report?

If the answer is "no" then you should probably revise your report to add more detail or clarity.

THE ANALYSIS TASK

The aim of this project is to create a dataset containing information on rates of municipal, agricultural, and industrial water use by country.

Available Data

Raw data relevant to this task can be found at the following locations:

- `aquastat.csv` contains the most recent available data on total volumes of municipal, industrial, and agricultural water use by country. This data was generated by a query to the Food and Agriculture Organization's AQUASTAT database at `www.fao.org/nr/water/aquastat/data`.
- `Data_Extract_From_World_Development_Indicators_Data.txt` contains 2014 data on population and GDP, as well as a number of development indices. This data was generated by a query to the World Bank's World DataBank at `databank.worldbank.org/data/`.

Take a look at each file in text editor to see how they are formatted. Make sure to look at both the top *and* bottom of the files.

STEP-BY-STEP

1. Import and prepare water data.
   (a) Import the `aquastat.csv` data into SAS. If you use the Import Wizard make sure to include the generated PROC IMPORT code in your program.
   (b) Only keep variables corresponding to country name, global region, agricultural withdrawl, industrial withdrawl, municipal withdrawl, and total withdrawl. After this step your dataset should have six variables.
   (c) Make sure each variable has an appropriate type, format, name, and label.
2. Import and prepare population data.
   (a) Import the population data into SAS.
   (b) Only keep observations corresponding to GDP, GDP per capita, Agriculture value added, and total population.
   (c) Only keep the variables indicating the name of the statistic, the country name, the country code, and the value of the statistic.
   (d) Use the TRANSPOSE procedure to reshape this dataset into one row per country.
   (e) Make sure each variable has an appropriate type, format, name, and label.
3. Combine the water and population datasets to create a single SAS dataset that has one row per country. There should be 11 variables: country name, country code, global region, agricultural withdrawl, industrial withdrawl, municipal withdrawl, total withdrawl, GDP, GDP per capita, agriculture value added, and total population.
   (a) Print out the names of countries in the water dataset that did not match any country in the population dataset.

(b) Print out the names of countries in the population dataset that did not match any country in the water dataset.

(c) Don't worry about the other countries, but make sure that America has the same name in both datasets then run the merge again.

4. Create three new variables that give the agricultural, industrial, and municipal water withdrawal per capita. That is, add the variables we are interested in!

5. Export the final dataset as a csv file called `FinalData.csv`.

6. Summary

   (a) Run PROC CONTENTS to display the details of your final SAS dataset.

   (b) Print out (to the results viewer) an ordered list of the top-10 water users per capita.

   (c) Print out an ordered list of the top-10 municipal water users per capita.

   (d) Print out an ordered list of the top-10 agricultural water users per capita.

   (e) Print out an ordered list of the top-10 industrial water users per capita.

   Each printed dataset should only contain the country name, the value of the statistic, and the row number indicating the ranking. Make sure that everything you output to the Results Viewer has an appropriate title.