CPS188 : Term Project : Winter 2023

Introduction:
This project is conceived as a team project. Groups of 4 people are recommended but groups of 2 and 3 can be allowed (a group cannot be more than 4 - don't ask!). Group members can be from different lab sections but must be from the same professor (Dr. Hamelin sections 1-10; Dr. Mustafiz sections 11-15; Mr. Ufkes sections 16-18). Each group will appoint a group leader that will submit the project on behalf of the group. Only one submission per group will be permitted so make sure the cover page clearly indicates the members of the group.

Group formation is managed by D2L under Communications > Groups. You can form your own groups until Friday March 17 at 5:00pm. After that deadline, students without a group will be randomly assigned. So if you want to choose your project partners, do it before the deadline!

Description:
In this project you will make calculations and conclusions based on real data collected by Statistics Canada about the prevalence of diabetes in Canada in its four most populated provinces (Ontario, Quebec, British Columbia and Alberta).

The data file contains the percentages of the population that have diabetes in each of the four provinces and also the country as a whole (excluding territories) for age groups 35 years of age and above. Data were collected between 2015 and 2021.

You can find the actual data file by following this link: statscan_diabetes.csv

The file format is CSV (comma spaced values). For this project, the relevant columns will be *REF_DATE* (year), *GEO* (country/province), *Age group* (35-49 years, 50-64 years, 65 years and over), *Sex* (Females/Males), and *VALUE* (the % of population that has diabetes). In your C program, you will need to open the file "as is", read the data and do the required computations. You might have to open and close the file more than once to keep the programming manageable.

It is up to you to read the data from the file and put it into your C program. This operation must be done with C! Do not copy and enter data by hand! The data in the file is in string format so some numerical values (like the percentages) will have to be converted into double values using the **atof** function from the stdlib library.

Some data may be missing because they were not collected by Statistics Canada. You must ignore those missing data in your calculations (do not substitute 0 for the values - that would change the results!).

You are to make a report showing tables, graphs and conclusions based on data using C programming and GNUPlot functionalities.

Required elements:
The entire project must be presented as one single program (excluding the graph scripts) . Divide your code into sections, one for each question, and add text comments to identify which question is answered in that section.

All computations and determinations are to be done in C using the imported data file. For the graphs, all the labels, legends and titles must be generated by the GNUPlot script. Nothing can be done by hand.

Calculations:

1. Determine the following averages of the percentage of the population that are diagnosed with diabetes. Present your outputs clearly within your program with labels (explanatory text), not just the numbers by themselves.
    a. Provincial averages (Ontario, Quebec, British Columbia, Alberta). One average per province (for all years and age groups).
    b. One national (*Canada excluding territories*) average for all years and age groups.
    c. Yearly averages (2015, 2016, 2017, 2018, 2019, 2020, 2021). One average per year (all age groups together) for each province and the whole country (*Canada excluding territories*) for a total of 35 averages.
    d. The average percentage of diabetes among age groups (35-49, 60-64, 65+). One average per age group (all years) for each province and the whole country (*Canada excluding territories*) .
2. Determine which province has the highest percentage of diabetes (all years and age groups together as calculated in question **1a**) and which province has the lowest percentage.

3. Indicate the provinces that have diabetes percentages above the national average (*Canada excluding territories*) and the provinces that are below the national average.

4. Indicate which year and province has the highest percentage of diabetes. Do the same for the lowest percentage. In case of a tie, you can mention multiple years and provinces.

Graphs:

5. Make a graph (simple line plot) of the diabetes percentages for the years 2015 to 2021 (all age groups together). Make a single graph with the four provinces and the national average (indicated as *Canada excluding territorie*s) (5 lines). Use different line styles and/or colours for each line plot making sure the national line stands out from the other four. Label the axes clearly and add a title and a legend to your graph.

6. Make a graph (bar chart) that shows the average percentages of diabetes among the three age groups for the entire country. Label the axes clearly and add a title and a legend to your graph.

Important notes about the report and its submission:

I. All computations and plots are to be done with C and GNUPlot only.

II. You are to write a report. Your report must have an introduction about the purpose of the report and its presentation.

III. The report must be detailed, well presented and attractive. Don't be afraid to use colours to emphasize parts of the report. Be creative in the use of tables, graphs and images. Points will be awarded to the exactness of the computations, appearance, ease of reading (use font sizes that are easy to read and use adequate line spacing and margins), and the quality of the English language.

IV. The report consists of the answers provided by <u>each</u> of the program requirements (the actual outputs from the program as cut/paste or screenshots - make sure they are easily readable in high resolution). For the graphs, generate a PNG file and paste it to your report.

V. For <u>each</u> question, have one short explanation paragraph to describe the C or GNUPlot operations you used to answer the question.

VI. Be original! Plagiarism will be dealt with severely to the full extent of TMU's academic integrity regulations. The Turnitin system will be used to help the markers in their assessment of originality.

VII. Your report must have a conclusion. You must report in the conclusion about your experience doing this project and how you would do things differently if you had to do this again.

VIII. Your report must have a cover page that clearly shows your names and section numbers.

IX. At the end of your report you will paste the complete code of your C program and the GNUPlot codes for each graph.

X. The report must be in PDF format only. You can use GoogleDoc to write the report but download it as PDF before submitting (penalties will apply if submitted in any other format - do not attach a .gdoc file). Also change the extension of the C program from .c to .txt as .c files are not supported by D2L. Submit your report (.pdf) and the code (.txt) file on D2L.

XI. Projects must be submitted on or before the date specified in the D2L dropbox. Late assignments will not be accepted for marking. If you are concerned about getting the assignment in on time, submit it early. Technical excuses will not be accepted.

Have fun!