Part 2: Analysing the UN City Population Dataset

You are given the "UN city population" dataset. Perform the following analysis using Pig:

Question 1. Find the number of countries in the dataset.
Question 2. List the countries together with the number of cities in each country.
Question 3. List countries in ascending order of female-to-male ratio, throughout the Years (*1).
Question 4. List the top 10 most populated cities according to the most recent data in the dataset (*2).
Question 5. List the top 10 cities which have the highest population change per year in percentage since the start of the survey (*3)

Notes:

• You must use Pig.
• Annotate your program code properly so that the marker can understand how it works. The annotation also contributes to the grade.
• State any assumption you made.
• If you cannot complete a task, an incomplete solution may still give you partial credit.

Part 3: Analysing Datasets of Your Choice

In this part of the coursework, you need to:
• Find a dataset, or multiple datasets in the following domains: UK living cost, world travel, or the NHS.
    o You cannot use any dataset outside these domains.
    o You can combine multiple datasets from these domains.
    o Dataset(s) must be public domain and of a considerable size.
        § A dataset cannot be too small. (e.g. Just a few lines.)
        § There is no need to go for a GB or TB-sized dataset unless the dataset is very interesting.
    o DO NOT choose a dataset similar to the one in Part 2.
• Propose 3 analysis tasks to perform on the dataset(s).
    o Your proposed analyses should be insightful. e.g. give useful information for decision making.
    o DO NOT propose tasks similar to those in part 2.

**1. You should have one entry for each country (not city) in each year, as the ratio may change throughout the years.*
**2. Do not fix the year but let your script find most recent population figure for each city. As we cannot guarantee that data are available for all cities in each year, you may end up comparing city A's population with city B's in different years. This is fine as far as the figures are the most recent ones for both A and B.*
**3. By "population change per year in percentage" since "the start of the survey", we mean: If the earliest figure of city X is P0 and the latest figure is Pn after N years, the total change in percentage C=(PnP0)/P0*100% over N years. And the change per year is C/N. We need to take the number of years into account as we cannot guarantee that all cities' data span the*

same period of time. e.g. It makes no sense to compare city A's change in 1 year with city B's change in 10 years.

Tasks Format Part 2 & 3

Your submission should include the followings:

- All Pig scripts for the tasks in both Part 2 and 3.
    - DO NOT combine the Pig scripts into one file. The scripts should be provided as separate text files which are ready to be executed.
- For part 3, a report in PDF format, with a maximum of 2500 words (5 pages of A4).
    - Describe your chosen dataset.
        - § Explain your motivation in selecting this dataset.
        - § State clearly the source of the dataset. e.g. Its URL.
        - § Describe its format. e.g. The meanings of the fields.
    - For the 3 proposed analyses.
        - § Describe and state the motivation of each analysis.
- Pay attention to the motivation, value, and technical challenges in each analysis. Unless it is an interesting question, a simple/generic task (e.g. counting total, calculating average) may not give you a good grade.
        - § Implement each analysis in Pig.
        - § State the result of each analysis.
        - § Interpret and discuss the analysis result.
- DO NOT just repeat the numbers from the output or give generic comments. You should: