

## Module 5 Assignment Questions

Note that the answers to each of these questions should be the direct result of running appropriate commands and not involve any further processing, including manual work. Answers without the commands used to achieve them will not get any grade. Make sure you have MySQL installed on your system and have installed the required databases listed below.

Databases required:

- The first is the “world” database
  - The second is the “auto” database
1. Answer the following questions using the “world” database (See Blackboard assignment prompt). Present your SQL queries and/or processes that you used to derive the answers.
    - 1.1 How many countries became independent in the twentieth century? **(10 points)**
    - 1.2 How many people in the world are expected to live for 75 years or more? **(10 points)**
    - 1.3 List the 10 most populated countries in the world with their populations as a percentage of the world population. [Hint: You can first find the population for the world and then use it for percentage for countries, so something like: select Population/5000000000 from Country ....] **(10 points)**
    - 1.4 List the top 10 countries with the highest population density. [Hint: For population density, you can try something like: select Population/SurfaceArea from Country where....] **(10 points)**
    - 1.5 How many countries are there in each “Region” ? Write a SQL query that produces a list of regions with a column for country counts for each region and order the count descending. **(10 points)**
    - 1.6 What countries have more than 10 languages represented? Write a SQL query, using the "HAVING" clause, that produces the list of countries that have greater than 10 languages. Group by "CountryCode" and order by language count descending. **(10 points)**

2. Answer the following questions using the “auto” database.

2.1 Use Python to explore the relationship of different variables to models per gallon (mpg). Find out which of the variables have high correlation with mpg. Report those values. Build a regression model using one of those variables to predict mpg. Do the same using two of those variables. Report your models along with the regression line equations. **(10 points)**

2.2 Use R to understand how horsepower and weights are related to each other. Plot them using a scatter plot and color the data points using mpg. Do you see anything interesting/useful here? Report your observations with this plot. Now let us cluster the data on this plane in a “reasonable” number of groups. Show your plot where the data points are now colored with the cluster information and provide your interpretations. **(10 points)**