ROBERT GORDON UNIVERSITY ABERDEEN

School of Computing
Coursework Assignment

Surname	ELOM
First name	CHIZOBA FAVOUR
Matriculation Number	1911632
Contact phone number	+44 7887 887510
Course + Year	Advanced Data Management, 2020
Module Co-ordinator	Kit Ying Hui
Module Number + Name	CMM524 Advanced Data Management
Coursework Title	CMM 524, Advanced Data Management
Coursework Part	Part 2 of 3
Due Date	15 December 2020 (16:00)
Feedback Due Date	By 27 January 2021

The University operates a Fit to Sit Policy which means that if you undertake an assessment then you are declaring yourself well enough to do so. Further details are available at: www.rgu.ac.uk/academicregulationsstudentforms

Declaration ** *This **must** be affirmed by adding your name below with the date of submission*

I acknowledge that by submitting the work, accompanied by this front cover, I take responsibility for the ownership of the submitted work.
I confirm:

- **that the work undertaken for this assignment is entirely my own and that I have not made use of any unauthorised assistance**
- **that the sources of all reference material has been properly acknowledged.**

Student Signature	Elom Chizoba Favour
Date Submitted	10th December, 2020

Marker's Comments	
Marker	**Grade**

****see over for regulations on plagiarism**

** An extract from the University Regulations

6. Academic Misconduct

Refer also to Schedule 3.3 of this Regulation for guidance on this procedure.

6.1 **Academic Misconduct** is defined as any attempt by students to gain an unfair advantage in assessments and examinations. Examples of academic misconduct include plagiarism, cheating, falsifying data, collusion, bribery or attempted bribery, personation or any other activity intended to provide an unfair advantage.

(i) **Plagiarism** is the practice of presenting the thoughts or writings of another or others as original, without acknowledgement of their source(s). All material used to support a piece of work should be carefully referenced and should not normally be copied directly unless as an acknowledged quote. Text translated into the words of the individual student should in all cases acknowledge the source.

(ii) **Cheating** includes:

- the taking of any unauthorised material into an examination;
- obtaining copy of "unseen" papers in advance of an examination;
- communicating or attempting to communicate in any way with another student during an examination;
- copying or attempting to copy from another student during an examination or in the production of coursework;
- wilful deception in any element of an examination or assessment.

(iii) **Falsification of data** consists of the misrepresentation of the results of experimental work or the presentation of results from fictitious work.

(iv) **Collusion** is the representation of unauthorised group work as that of an individual student.

(v) **Bribery** is the paying, offering or attempted exchange of an inducement for information or material intended to advantage the recipient in an examination or assessment.

(vi) **Personation** consists of a substitute taking the place of a student in an examination.

A student who aids and abets a fellow student to commit academic misconduct shall be deemed to have committed academic misconduct and will be dealt with accordingly.

Part 2 Report

This report is written to provide information about the approaches taken in using pig language to analyse the *UN city population* dataset. The dataset is a record of population census for various countries and a series of questions will be answered from the computed queries.

The attributes used for the analysis are the ones considered and assumed to be most relevant in answering the questions. Having stated that, it is also important to note that not all attributes of the dataset were used.

Question 1:

Result: **163**

From the *part2_q1.pig* file, there are 163 countries from which the census data was obtained. This result was obtained by generating a subset of the dataset that contains only the country attribute. The subset was then **GROUP**ed and subsequently **COUNTED**ed.

Question 2:

Result: *part2_q2_result.txt* file

The output for question 2 is contained in the file indicated. It contains the countries and the respective cities in each country. For a clearer understanding, the output was further ordered in descending order, to easily identify the country with the greatest number of cities, which happens to be China. The query and annotations can be found in the *part2_q2.pig* file.

Question 3:

Result: *part2_q3_result.txt* file

The ascending order of female to male ratio is obtained from the analysis in this section. The result is in the file indicated. From the output, Pitcairn in 1991 has the highest female to male ratio of 1.36 which was lower in 1985, at 1.06. Whereas Kuwait as of 2005 has a low female to male ration of 0.21 as opposed to 0.45 in 1995. The query and annotations can be found in the *part2_q3.pig* file.

Question 4:

Result:

(Mexico,2010,MEXICO - CIUDAD DE,28967922)

(India,2001,Mumbai (Bombay),28412836)

(India,2001,Delhi,22756642)

(Mexico,2010,Tlalnepantla,20770252)

(India,2001,Kolkata (Calcutta),17778573)

(China,2000,Shanghai,14348535)

(Turkey,2012,Istanbul,13596781)

(Colombia,2005,BOGOTA - D.C.,13542016)

(France,2010,PARIS,12703949)

(Brazil,2010,Rio de Janeiro,12640892)

The analysis for this section is contained in the *part2_q4.pig* file. The result above shows the list of top 10 most populated cities, considering only the most recent data in the dataset. From the result, Mexico in 2010, had MEXICO - CIUDAD DE city, with a population of 28,967,922. Making it the most topmost in the list. On the other hand, Brazil in 2010, had Rio de Janeiro city, with a population of 12,640,892. Making it the 10^{th}.

Question 5:

Result:

(Mexico,MEXICO - CIUDAD DE,867.16)

(Mexico,León (de los Aldama),281.14)

(Pakistan,Hyderabad,234.6)

(Mexico,Guadalajara,175.94)

(Republic of Korea,Icheon,152.43)

(Mexico,Matamoros,149.61)

(Mexico,Juárez,62.68)

(Malaysia,Shah Alam,33.17)

(Malaysia,Petaling Jaya,32.28)

(Mexico,Reynosa,31.84)

The analysis for this section is contained in the *part2_q5.pig* file. It is geared towards obtaining the top 10 cities with the highest population change in the year, since the start of the survey. MEXICO - CIUDAD DE city also came to the top of the list, with a population change of 867.16.