Data Analysis and Visualisation

2810ICT— Software Technologies
School of ICT
Griffith University
Trimester 2, 2017

Due at the end of Week 12, Friday 22nd September, 5pm

Assignment Description

purpose of this assignment

This assignment will test your ability to use python to interact with databases, excel workbooks, perform data analysis using NumPy, and use matplotlib to generate graphs. Raw data will be provided and will need to be stored in a database. You will then need to write scripts to query the database and pull out subsets of the data. These subsets will then need to be further analysed and have graphs procedurally generated to display relevant information. All of the investigations should yield results that you can present in a professional report.

It is important to note that submission of this assignment is a requirement for passing the course. Late submissions will be marked according to Griffith University's assessment policy. 10% of the overall mark will be deducted for each business day late. After 5 days, no submissions will be accepted.

Submission Requirements

This assignment must be submitted online via L@G under the assessment page. Only 1 submission per group is needed. Your submission should include;

- A word document (preferred), PDF Is also acceptable, other formats are not allowed. This should contain your results presented as a scientific report.
- .py files containing your code. You should submit a separate .py file for each problem.
- A readme.txt file for your scripts

Problem Statement

As more and more industries are becoming data driven, being able to process a large volume of raw data and produce a concise and insightful summary is becoming more and more important. As a data consultant for a political candidate, you are tasked with processing some global temperature data and producing a report summarising some of the information.

The raw data is provided in 3 excel spreadsheets: (A) temperature by state, (B) temperature by country, (c) temperature by major city. You will need to complete the below tasks and present your results in a report.

For each python script, you should handle the case where the script has already been run and therefore the data already exists. This could mean checking to see if the table already existed in the database, or a specific workbook/worksheet already exists. In each case, you should decide what to do (display error ? Create a book/sheet with a different name ? Delete the existing version and run the script ?).

You should write a readme.txt file to accompany your scripts. Prepare a brief usage guide, any requirements and assumptions, and document what each script does and any other important info (for example; how does each script deal with database tables/excel sheets already existing).

Task 1 – Access the workbooks and create a database

Create a Python script (db_create.py) to perform the following tasks:

- Open the excel files.
- Create a SQLite database with three tables, one for each excel file. Each column in the excel files should correspond to a column in the tables. Make sensible decisions for attribute types.
- Import the data from the excel files to the corresponding tables in the database.

import data from excel to database table;

Task 2 – Query the database

找出南半球所有不同的city：
select distinct cityName, country, geolocation
from tableName where latitude %N ordered by country

Create a Python script (sql_temp.py) to perform the following tasks:

- List the distinctive major cities located in southern hemisphere ordered by country to the console and then write their name, country, and geolocation into a new database table called "Southern cities".
- Find the maximum, minimum and average temperature of Queensland for year 2000 and print this information to the console.

select max(avgTemp), min(avgTemp), avg(avgTemp) from
tableName where state = 'Queensland' and date = '2000%'
or date = '%2000'

Task 3 – Excel via Python

Create a Python script (excel_temp.py) to perform the following tasks:

- Create a new workbook named “World Temperature.xlsx”.
- Create a sheet named “Temperature by city”.
- Query the database and calculate the yearly mean temperature of each city in China. Note some data may be missing.
- Write the relevant data into the worksheet you created. World Temperature.xlsx
- Generate a line chart for the above data.

```
select AverageTemperature, city, strftime('%Y', date) as year
from GlobalTemperatureByMajorCity,
          where country = 'China' order by year
```

Task 4 – Numpy in Python

Create a Python script (numpy_temp.py) to perform the following tasks:

- Open the World Temperature workbook.
- Create another sheet called “Comparison”.
- Calculate the mean temperature of Australian states for each year (using the temperature by state table in your database).
- Calculate the mean temperature of Australia for each year (using the temperature by country table in your database).
- Calculate their differences between each state and the national data for each year.
- Use MatPlotLib to plot the difference across years.
- Write the data into the sheet.

Task 5 – Report

After developing all of your scripts, you need to take the information and present it in a brief executive-style scientific report. You may use the provided template as is, modify it, or come up with your own. Do not simply paste data from the console/excel into each section in the template – present the data in a professional manner.

Marking

This assignment is worth 40% of your final grade. The assignment will be marked out of 100 and marks will be allocated as follows:

- Task 1 - Creating the database and importing the data (20 marks)
- Task 2 - Querying the database (15 marks)
- Task 3 - Excel in Python (20 marks)
- Task 4 – NumPy and Matplotlib (25 marks)
- Task 5 - Report - (20 marks)

Additionally, marks will be deducted for incorrect:

- Poor English/grammar
- Lack of comments in the code
- Poor presentation