SCHOOL OF COMPUTER SCIENCE & INFORMATICS
COURSEWORK ASSESSMENT PROFORMA

MODULE & LECTURER: CM2105 Data Processing & Visualisation, H. Liu

DATE SET: 30th October 2017

SUBMISSION DATE: 8th December 2017 (9:30am)

SUBMISSION ARRANGEMENTS:
Your coursework programme – your code and results should be contained within an Jupyter Notebook named [CW-your student number.ipynb] – should be submitted via Learning Central by 9:30am on the submission date.

TITLE: Data Processing & Visualisation Coursework

This coursework is worth 50% of the total marks available for this module. The penalty for late or non-submission is an award of zero marks. You are reminded of the need to comply with Cardiff University's Student Guide to Academic Integrity. Your work should be submitted using the official Coursework Submission Cover sheet.

INSTRUCTIONS

Q1. The US annual "County Health Rankings" provide information of how health is influenced by where people live, learn, work and play. They provide a starting point for change in communities. A data analytics team attempts to estimate the premature death (i.e., "**Years of Potential Life Lost Rate (YPLLR)**" based on number of deaths under age 75) in Florida. Other variables that they believe offer some insight on the premature death include:
– Teen births, i.e., "**Teen Birth Rate (TBR)**" (Teen births/females ages 15-19 * 1,000);
– Violent crime, i.e., "**Violent Crime Rate (VCR)**" (violent crimes/population * 100,000) ;
– Adult smoking, i.e., "**Percentage Smokers (PS)**" (Percentage of adults that reported currently smoking).
The text file named "**2017Health.txt**" (available on Learning Central) contains the data. Shown below is the form of the data.

State	County	Years of Potential Life Lost Rate	Teen Birth Rate	Violent Crime Rate	Percentage Smokers
Florida	Alachua	6633	19	579	16
Florida	Baker	8270	58	360	19
Florida	Bay	9168	50	508	18
Florida	Bradford	10346	61	461	18
Florida	Brevard	7722	25	518	16
Florida	Broward	5737	23	441	15
Florida	Calhoun	6415	59	130	19
Florida	Charlotte	7353	30	219	14
…	…	…	…	…	…

1) [**cell1 – 1 mark**] Download the file "CW-your student number.ipynb" from Learning Central, and upload it to your IPython Notebook. Change the title of the file using your student number. Write code to read the given data in text format (i.e., "2017Health.txt") into required tabular data structure: make the county (i.e., Alachua, Baker…) be the index of the returned data structure; the first column of the returned data structure represents the "Years of Potential Life Lost Rate"; the second column represents the "Teen Birth Rate"; the third column represents the "Violent Crime Rate"; and the last column represents the "Percentage Smokers".
   – **Display** the returned tabular data structure in your programme.

2) [**cell2 – 5 marks**] Write code to analyse the data contained in the variable called "Percentage Smokers".
   – **Print** the "mean of Percentage Smokers".
   – **Print** the "minimum of Percentage Smokers".
   – **Print** the "maximum of Percentage Smokers".
   – **Print** the "standard deviation of Percentage Smokers".
   – **Print** the "95% confidence interval of Percentage Smokers".

3) [**cell3 – 9 marks**] Write code to plot a bar graph that uses bars to compare the "Percentage Smokers" of North Florida, Central Florida and South Florida. North Florida: use the measures of the following counties: Duval, Alachua, Leon, Flagler, Marion; Central Florida: use the measures of the following counties: Orange, Polk, Hillsborough, Pinellas, Brevard; South Florida: use the measures of the following counties: Miami-Dade, Broward, Lee, Palm Beach, Sarasota.
   – **Visualise** a single plot: the horizonal axis shows the data categories being compared (i.e., North Florida, Central Florida and South Florida); and the vertical axis represents the mean measure of Percentage Smokers.
   – Add **error bars** to the bar graph, showing the 95% confidence interval.
   – Add appropriate **title**, **horizontal axis label** and **vertical axis label** to the bar graph.

4) [**cell4 – 4 marks**] Based on the following two predictor variables: "Teen Birth Rate (TBR)" and "Percentage Smokers (PS)", write code to build a linear regression model to estimate the "Years of Potential Life Lost Rate (YPLLR)".
   – **Print** the resulting linear equation in the programme.

5) [**cell5 – 6 marks**] Based on the error of prediction (i.e., the **absolute error/difference** between the measured "Years of Potential Life Lost Rate" and the predicted "Years of Potential Life Lost Rate"), compare the following two linear models

   Model A: YPLLR = 60.6 × TBR + 5297.06
   Model B: YPLLR = 1.36 × VCR + 7254.3

   and advise the data analytics team which model should be used. Write code to perform appropriate statistical data analysis.
   – **Print** the mean absolute error (MAE) of the model A.
   – **Print** the mean absolute error (MAE) of the model B.
   – **Print** the main results of the data analysis processes, including normality test and statistical significance test.
   – **Print** ONE sentence, stating your conclusion and justification on the difference in performance between two models, in terms of predicting the "Years of Potential Life Lost Rate".

SUBMISSION INSTRUCTIONS

All submission should be via Learning Central unless agreed in advance with the Director of Teaching. The current electronic coursework submission policy can be found at:
http://www.cs.cf.ac.uk/currentstudents/ElectronicCourseworkSubmissionPolicy.pdf

Description		*Type*	*Name*
Cover sheet	*Compulsory*	*One PDF (.pdf) file*	*[student number].pdf*
Q1	*Compulsory*	*One IPython Notebook file (.ipynb)*	*CW-student number.ipynb*

CRITERIA FOR ASSESSMENT

Credit will be awarded against the following criteria.

Your CODE and RESULTS should be contained within an IPython Notebook that analyses and visualises a given dataset (should be obtained via Learning Central: 17/18-CM2105 Data Processing and Visualisation). This coursework assesses the intended learning outcomes of 1, 2, 3, 4.

The breakdown of marks (total=25 marks) will be for correctly computing and visualising:

1) [cell1 – 1 mark]: Manipulate data and display the restructured tabular data (see "Sample output 1" below).
2) [cell2 – 5 marks]: Produce the required results of descriptive statistics.
3) [cell3 – 9 marks]: Create and visualise the required graph.
4) [cell4 – 4 marks]: Conduct and report the required results of regression analysis.
5) [cell5 – 6 marks]: Conduct and report procedures and results of statistical data analysis.

Sample output 1

	Years of Potential Life Lost Rate	Teen Birth Rate	Violent Crime Rate	Percentage Smokers
Alachua	6633	19	579	16
Baker	8270	58	360	19
Bay	9168	50	508	18
Bradford	10346	61	461	18
Brevard	7722	25	518	16
Broward	5737	23	441	15
Calhoun	6415	59	130	19
Charlotte	7353	30	219	14

Feedback on your performance will address each of these criteria.

FURTHER DETAILS

Feedback on your coursework will address the above criteria and will be returned in approximately four weeks. This will be supplemented with oral feedback in lectures and labs. If you have any questions relating to your individual solutions talk to the lecturer.