UECS3213 / UECS3453 Data Mining Assignment 2

Instructor: Dr. Simon Lau Boung Yew
Major: AM/SE
SESSION: Jan 2019

Title: Exploratory Data Analysis and Visualization

Learning Outcomes (Assessment = 10%)

At the end of this assignment, students are expected to achieve the following CLO:

- **CLO2**: Create programming solutions using data mining techniques for given problem (5%)
- **CLO3**: Evaluate performance of data mining solutions for a given problem (5%)

Introduction

Exploratory data analysis (EDA) is an approach in analyzing data sets to summarize their main characteristics, often with visual methods. In this assignment, we aim to put into practice about exploratory data analysis and some visualization techniques what we learned in Topic 2b in order to derive useful information / knowledge / pattern from raw data that is available on the public domain.

Instruction

This is an individual assignment.

Each student is required to perform the following:

1. Select and download THREE (3) different datasets (e.g. in raw delimited text of csv or text format) from an open database that is available for data such as (but not limited to) the following:
    - Weather data of a city
    - Temperature readings of a sensor
    - Heart beat rate readings of a patient
    - Traffic data of a city
    - Sales transaction data of a product of a company
    - Online user interaction data of a website
    - Economic data of a country
    - Intake data of UTAR students
    - Crime data of a city
    - etc. etc.

(Note: You may refer to database such as UC Irvine Machine Learning Repository: http://archive.ics.uci.edu/ml/index.php if you are not sure of what dataset to look for. Please consult the lecturer about the dataset you select if you have any doubt. Each of the dataset you select and download must at least contain more than 100 data points / records but also avoid datasets which is too large with size amounting to hundreds of MB or GB)

2. Carry out exploratory data analysis techniques you learned in Topic 2b and produce a report of NOT MORE than 12 pages which include the following:

   - **Section 1: Summary Statistics**
     - Summary of the descriptive statistics analysis of the data sample which may include (but not limited to) maximum/minimum values, mean, median, mode, standard deviation, truncated mean, AAD, MAD, IQR etc. whichever that is relevant to those particular datasets that you selected. Describe your analysis for the dataset based on the statistical parameters and values.
   - **Section 2: Visualization**
     - AT LEAST THREE (3) visual charts to display the data visually and meaningfully using visualization techniques that you have learned which may include (but not limited to) the following:
       - Histogram
       - Bar Graph
       - Line Graph
       - Box Plot
       - Scatter Plot
       - Contour Plot
       - Matrix Plot
       - Parallel Coordinates
       - Star / Radar Plot
       - etc.

3. For Section 2, each student MUST plot the charts programmatically using Python graphing library such as matplotlib, plotly, ggplot, seaborn, pygal etc. (Student is not allowed to use other tools such as Microsoft Excel, SPSS, R etc.). If necessary, you may perform some data pre-processing step such as data cleaning and data mining techniques such as classification, regression, clustering etc. to extract useful information and present them in a human-readable way. Describe briefly each of the charts that you have plotted by brief quantitative, logical and critical analysis.

Submission Procedure

1. Name your report (filename) as the following: **<Course>_<ID>_assign2.docx or .pdf** where <Course> and <ID> denote your course and university identity number, respectively.

2. Submit the softcopy of the report via wble.utar.edu.my on or before **22 March 2018 (Friday) at 5.00pm sharp**.

Marking Scheme

No.		Poor	Adequate	Proficient	Subtotal
Section 1: Summary Statistics					
1	Correctness	not reasonable, not precise, not understandable	not reasonable, precise, understandable	reasonable, precise, understandable	
		0 - 5	6 - 15	16 - 20	
2	Comprehensiveness / Completeness of analysis	incomplete, lack of information	partially complete, with some information	complete, informative	
		0 - 5	6 - 15	16 - 20	
Section 2: Visualization					
3	Visual	not tidily presented, visually not readable	moderately tidy, readable	visually appealing, rich with information	
		0 - 10	11 - 20	21 - 30	
4	Critical analysis	illogical, incorrect assumption	partially logical, correct assumption	logical, critical, correct assumption	
		0 - 10	11 - 20	21 - 30	
				Total	/100
				Assessment	/10%

UNIVERSITI TUNKU ABDUL RAHMAN

Assignment 2

Course Code: UECS3213 / UECS3453

Course Name: Data Mining

Lecturer: Dr. Simon Lau Boung Yew

Academic Session: 2019/01

Title: Exploratory Data Analysis and Visualization

Student ID	Student Name	Major
	Mark	/10