THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY
Department of Computer Science and Engineering
COMP4331: Introduction to Data Mining
Fall 2021 Assignment 1
Due time and date: 11:59pm, October 20 (Wed), 2021.

IMPORTANT NOTES

- **Your grade will be based on the correctness, efficiency and clarity.**
- **Late submission: 25 marks will be deducted for every 24 hours after the deadline.**
- **ZERO-Tolerance on Plagiarism: All involved parties will get zero mark.**

You are given a student performance dataset, you can download it from the UCI Machine Learning Repository:

`https://archive.ics.uci.edu/ml/datasets/student+performance`.

The downloaded zip file contains three files, please **DO** use the data "**student-por.csv**" for the following tasks. A notebook template is provided for you on Google Colab:

`https://colab.research.google.com/drive/1Qbz_HnJh5bKbcX5KV-pLDZN4hF2kRsBT?usp=sharing`.

Tasks:

1. Report the mean, standard deviation, mode, median, and the five-number summary of the attributes *age* and *G1*.
2. For the attribute *G2*, use seaborn or matplotlib to show boxplots for
   (a) *G2*;
   (b) *G2* for various values of *sex* (show *sex* in x-axis);
   (c) *G2* for various values of *Dalc* (show *Dalc* in x-axis).
3. For attributes *age* and *absences*, use seaborn or matplotlib to show their histograms (use 8 equal-sized bins).
4. Use seaborn or matplotlib to show the scatter plot for attributes *G2* and *G3* (show *G2* in x-axis). Report the correlation coefficient between *G2* and *G3*.
5. Consider the 10 attributes {*studytime, traveltime, age, absences, health, Walc, Dalc, famrel, goout, G2*}, report the top 5 attributes that are most correlated (either positively or negatively) with *G1*.
6. χ^2-test (using the χ^2 table shown on the last page). Show your steps clearly (including contingency table, χ^2-value, and p-value) in the report. You can use the built-in function from scipy.
   (a) By performing the χ^2-test at a significance level of 0.01, are the attributes *internet* and *romantic* independent of each other?

(b) By performing the χ^2-test at a significance level of 0.01, are the attributes *sex* and *romantic* independent of each other?

7. Normalization:

   (a) Normalize attribute *studytime* to the range $[0, 1]$ using min-max normalization. You can use the built-in function from scikit-learn.

   (b) Normalize attributes {*G1, G2, G3, Dalc, Walc*} to mean zero and standard deviation one using z-score normalization. You can use built-in function from scikit-learn.

   (c) Output your results in parts (a) and (b) above to the csv file data_normalized.csv (use "," as field delimiter and include column names in the header).

8. PCA: In this question, use only the attributes {*G1, G2, G3, Dalc, Walc*} after the normalization in Task 7b.

   (a) Plot the cumulative explained variance with the number of principal components.

   (b) Transform the data by PCA, by using the smallest number of PCA components such that the proportion of explained variance is at least 0.9.

   (c) For the transformed data, output the transformed dimensions to the csv file data_reduced.csv without the header (use "," as field delimiter), and report the five-number summary for each obtained dimension. You can use the built-in function from scikit-learn.

9. Consider the original data. At first glance, this dataset has no missing value. However, for *G2* and *G3*, some students got zero scores. We assume that this is because they missed the exams, and these zero scores can be viewed as missing values. Please fill in the missing values with the corresponding attribute mode, and show the scatter plot for attributes *G2* and *G3* again (show *G2* in x-axis). You can use the "DataFrame.replace" function from pandas. After filling in the missing values, report the correlation coefficient between *G2* and *G3* again, and compare it with the value obtained in Task 4.

Submission Guidelines

Please submit (i) a report report.pdf which includes the answers for Q1 to Q6, Q8 ((a) and 5-number summary in (c)), and Q9 (the scatter plot and correlation coefficient analysis). Your steps in Q6 should be clear, (ii) a python notebook assignment1.ipynb for your code, and (iii) the output data files data_normalized.csv and data_reduced.csv. Zip all the files to YourStudentID_assignment1.zip (e.g., 12345678_assignment1.zip). Please submit the assignment by uploading the compressed file to Canvas. Note that the assignment should be clearly legible, otherwise you may lose some points if the assignment is difficult to read. **Plagiarism will lead to zero point on this assignment.**

Degrees of Freedom	Probability										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
	Nonsignificant								Significant		