CS 470 Homework 1

Due by Thursday, February 6, 2020 at 8:00 PM

Submission instructions

Submit your assignment through the QTest system, using course ID: **CS470** and exam ID: **hw1**. Upload a single ZIP archive file named **hw1.zip**, containing at least three files: a PDF file with your solutions, the LaTeX source file used to generate it, and a CSV file containing the modified/extended dataset. This assignment must be typeset using LaTeX. No email submissions are accepted. No late submissions are accepted.

At the top of your solution, include a section named "Collaboration statement" in which you acknowledge any collaboration, help, or resource you used or consulted to complete this assignment.

1 Data exploration and pre-processing

The file `grades.csv` contains a dataset of anonymized scores from 5 semesters of the course "CS 170: Introduction to Programming I." This homework archive also includes the syllabus of that course, so you can better understand the context and interpret the data. Perform all the data exploration and pre-processing tasks described below.

Task 1 (12 points): Attribute description. Describe all the attributes included in the dataset. For each attribute, explain its meaning, and determine the type of attribute (categorical, ordinal, numeric interval, numeric ratio) – be careful that some of these attributes may be deceiving! If there is a "grey area" explain all the possibilities, and always motivate your choices.

Task 2 (12 points): Missing values. There are several missing values in the dataset. Describe the different possible meanings of missing values for each attribute (or set of attributes), and describe multiple ways to deal with those missing values. Compare the pros and cons of each solution you describe.

Task 3 (12 points): Re-encoding. The attributes *Semester* and *Section* are not encoded (represented) in a very smart way. Explain why. Then describe a better way to re-encode those attributes, and perform such re-encoding in the dataset.

Task 4 (12 points): Scaling and z-scoring. For each attribute representing a score, calculate three new attributes:

1. An attribute with the scores re-scaled to the interval [0, 100].
2. An attribute with the scores normalized using the z-scoring method, using the mean and standard deviation from all semesters combined.
3. An attribute with the scores normalized using the z-scoring method, using for each student the mean and standard deviation from only the students in their same semester.

Discuss and compare results (1), (2), and (3), reflecting on their meaning, and the cases in which it could be appropriate to perform such transformations.

Task 5 (12 points): Summary statistics. For each attribute where it makes sense to do so, calculate the most common summary statistics: mean, standard deviation, and 5-number summary.

Task 6 (15 points): Charts. Generate 10 or more plots from the data: at least 3 box plots, at least 3 histograms, and at least 3 scatter plots. At least one plot should include a derived attribute (i.e., and attribute that you don't have already in the dataset but needs to be calculated from some other attributes). For each plot write a detailed description and a discussion of what can be observed and noticed about the data.

Task 7 (10 points): Tools and languages. Describe, compare, and discuss the pros and cons of all the various tools and languages that you used (or tried out) to complete the tasks in this homework assignment.

Task 8 (15 points): LaTeX document. Include all your results, charts, discussions, and the code/formulas you wrote to perform your calculations in a well organized PDF document generated using LaTeX. Make sure you submit both the PDF file and the LaTeX source.

Grading criteria

- 10, 12 or 15 points (depending on the task) for each task that is complete, correct, and well justified; 6 points if the task is partially incomplete, or with minor mistakes, or not well justified; zero points if the task has major omissions, major errors, or it has wrong or totally missing justification.
- For task 8: 15 points if the document is properly typeset using LaTeX, and demonstrates good usage of document structure, sections, mathematical notation, and figures; 7 points if the document is typeset using LaTeX but it is not very well organized; zero points if the document is not typeset using LaTeX, or if the LaTeX source code is not provided.
- -10 points for missing collaboration statement.