Data Mining Lab, *The Second*

Note

- **Preferably use Python and external libraries like Pandas.**
- **Avoid excessive imperative-style programming.**

The (morphed) data supplied in the CSV represents the mark sheet of <beep> course taken at <beep> university. The roll numbers have been anonymized, just in case they contain data of some current industry leaders. Load the data and write programs to answer the following:

1. Find the top 10 pairs of columns with the largest correlation.
2. Find all pairs of columns that are statistically similar. Find all pairs of columns that are statistically different. Define the threshold to a suitable value.
3. Print all details of all students between roll number <prefix>100 and <prefix>150, where <prefix> is img_2020 or imt_2020.
4. Print details of 1st five columns of all IMT students.
5. Print the list of roll numbers with at least one negative mark in the evaluation.
6. Print the list of roll numbers with positive marks in all evaluations.
7. Replace all negative marks with zeros.
8. Use a max-min normalization to normalize all marks between 0 and 1, and add the marks to add a total column. Print the roll numbers of all students in the top 10% percentile and the bottom 10% percentile.
9. Plot a histogram of marks of all columns.
10. Consider a new table (or amend the old one) with only those columns with a wide distribution of marks as per the histogram. In the new table, find students at the top *k* percentile of the first column and simultaneously the bottom *k* percentile of either of the remaining columns. Also, find students at the top *k* percentile of any column and simultaneously the bottom *k* percentile of either of the remaining columns.