ILLINOIS

STAT107 Data Science Discovery

Lab: Similarity

Man Fung (Heman) Leung

Fall, 2021

University of Illinois at Urbana-Champaign

- Please work in a group of 2–4 students
  - collaboration is important in data science!
  - meet new friends and discuss :)
  - let us know if you have any questions

Random fact of the day

Under standard (physics) assumptions, James Bond would have die in the plane jump scene in Goldeneye.

Practical experience of the day

Since Python is an interpreted language, it is naturally slower than compiled language. In practice, we can embed compiled language such as C/C++ in Python to enhance the speed.

Comment: Birthday

ILLINOIS

- Common/potential mistakes
  - did not follow instruction in 0.2
  - calculated wrong/did not calculate exact probability in Part 1/2
  - for 3.2, checking "==" is wrong (in test case)
  - forgot to do 4.4/did not guess in 4.5
  - for 4.5, checking "<=" or ">=" are both fine
  - for probabilities, some of you coded the number of success row directly
- Running the test cases successfully do not imply full score
  - example: 2.1a
  - some puzzles' output cannot be tested
  - but failing a test case usually imply point lost

Today's lab: Similarity

- Main page
- Retrieve the lab using git
- Complete the notebook
  - hints are available by double clicking the question cells
  - 2.2: change `exclude` to `include` in `select_dtypes`
  - 2.5: code is given but remember to do reflection below
  - 2.6: try `df[numcols].fillna(df[numcols].mean(axis=0))` for numeric columns
  - 4.2: use for-loop to iterate over all columns. Inside the loop, check if the current column is numeric or string. Then compute the score based on the notebook's description
- Submit your work. Feel free to:
  - ask us questions
  - leave whenever you finish the lab

Checking completion

ILLINOIS

Default total number of cells: 55

- 1.1 in cell 6
- 1.2 in cell 9
- 1.3 in cell 12 (reflection)
- 2.1 in cell 17
- 2.2 in cell 20
- 2.3 in cell 23
- 2.4 in cell 25–26
- 2.5 in cell 28, 30 (reflection)
- 2.6 in cell 32–33 (reflection)
- 3.1 in cell 35
- 3.2 in cell 38
- 4.1 in cell 42 (textual)
- 4.2 in cell 44
- 4.3 in cell 46
- 5.1 in cell 48
- 5.2 in cell 50
- 5.3 in cell 53