STAT107 Data Science Discovery

Lab: Simpson's Paradox

Man Fung (Heman) Leung

Spring, 2022

University of Illinois at Urbana-Champaign

Collaboration

ILLINOIS

- 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
- Attendance form
  - you can come up if you do not want to use this form
  - submit before you leave the lab

Random fact of the day

Do you know why grocery stores produce and sell their own brands, e.g., Good & Gather (Target) or Great Value (Walmart)? This is related to the anchoring effect.

Practical experience of the day

What is the difference between ‘df["colname"]‘ and ‘df.colname‘? I would recommend you always use ‘df["colname"]‘; see this post.

Midterm 1 information

ILLINOIS

- Register on CBTF
  - Feb 15–17
  - 50 minutes
  - Python available via a zero-point question
  - Same questions as in practice midterm
- Feb 18 (Fri) lecture cancelled
  - Wed/Thur lab attendance still required

Comment on lab_intro

- Attendance is assumed in the first week (10 points)
- 2.2: `(4*3)**2` is different from `4*3**2`. You should also use `print((4*3)**2)` instead of `print("(4*3)**2")` as the puzzle asks for the result. As a side note, `^` does not work as exponentiation in Python
- 3.1: an error is expected as stated in the puzzle
- 3.4: leap year can be ignored as stated in my slides. However, your code has to be logical. For example, you cannot put the number 315360000 directly without steps

Comment on lab_pandas

ILLINOIS

- Score $\leq 10 \iff$ lack of attendance
- 1.1b/1.2b: Name of variable $\neq$ value of variable. String/number is acceptable but you can read here if you want to know the name of specific data type in Python
- 1.2a: `current_year` should be a number so `"2022"` is not ok
- 1.4: an error is expected as stated in the puzzle
- 2.1: use the variable `current_year` as stated in the puzzle
- 2.2: result should not be `"1744.219"`. Must use `int()` or `float()`
- 3.3: `"Stat"` is different from `"STAT"` as string is case sensitive. Result should be `True`/`False` only
- 4.2: use `mydf[mydf["Number"] >= 300]` instead of `df[df["Number"] >= 300]` because the upper-level courses in **your major** are required. `mydf[df["Number"] >= 300]` is logically wrong but I did not take off point if you use it

Today's lab: Simpson's Paradox

ILLINOIS

- Main page
- Hints:
  - Read the questions carefully
  - 2.1 approach 1: `sum(df_discovery['Recommend'] == "Yes")/len(df_discovery)`
  - 2.1 approach 2: `len(df_discovery[df_discovery['Recommend'] == "Yes"])/len(df_discovery)`
  - 3.1 numerator: `sum((df_discovery["Recommend"] == "Yes") & df_discovery["Gender"].isin(male))`
  - 3.1 denominator: `sum(df_discovery["Gender"].isin(male))`
- Submit your work. Feel free to:
  - ask us questions
  - leave whenever you finish the lab