

STAT107 Data Science Discovery

LAB: EXPERIMENTAL DESIGN

Man Fung (Heman) Leung

Spring, 2022

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
- Attendance form
 - submit before you leave the lab

Random fact of the day

The **lipstick effect** states that consumers are more willing to buy less costly luxury goods during an economic crisis. Is it possible to apply randomization if we have a lipstick sales dataset?

- Main page
- Hints:
 - Array index in Python begins at 0
 - A dataframe `df`'s index can be retrieved via `df.index`
 - The part of a dataframe `df` not included in another dataframe `sub` can be selected via `df.drop(subset.index)`
 - Two dataframes `sub1` and `sub2` can be combined via `pd.concat([sub1, sub2])`
 - Other functions that you may find useful: `.head(n)`, `.tail(n)`, `.merge(sub1, sub2, how="outer")`
- Submit your work. Feel free to:
 - ask us questions
 - leave whenever you finish the lab

- Common mistakes last semester:
 - 2.1: using `control_random = sample(frac=0.5)` and `treatment_random = sample(frac=0.5)`. This is wrong because the same row may be included in both variables. We have updated the test case this semester to detect this mistake.
 - 3.3: combining `block1` and `block2` before random sampling. This is wrong because you may sample more from `block1` than from `block2`. Correct stratified random sampling should sample from `block1` and `block2` separately