

STAT107 Data Science Discovery

LAB: BIRTHDAY

Man Fung (Heman) Leung

Spring, 2022

University of Illinois 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**
 - you can come up if you do not want to use this form
 - submit before you leave the lab

Practical experience of the day

1. Debugging via printing variables
2. Searching documentation via library + what you want to do
3. Searching general solution via language + what you want to do

- Check email for score decomposition
- Why `count()` sometimes does not give the correct number of rows? The short answer is due to NA in some cells; see the [documentation](#)
- Why plot in Python instead of Excel? Excel stores limited data only and provides less flexibility
- Do not reuse variable names unless you are sure you will not use the values stored inside later
- Do not put your individual reflection under group discussion! I may overlook if you do that
- One of you found that the best payed employee is a guy in a mystery (???) department. He is actually the Chancellor of our school (if you do not recognize his name haha)

- If you see weird looking long text without space in your reflection, the reason is usually the use of two "\$"’s, which generates latex output in between
- 1.3/1.4/2.2: plot salary only. I can accept box plot with multiple columns but histogram with multiple columns is not readable
- 2.1: do not reset index more than once here
- 2.2: the direct way to find max is `max()`. `nlargest()` is indirect (as the whole row is displayed) but still acceptable
- 3.1: you need to display two box plots separately
- 4.1: you need to **use Python** to **answer at least one new question** to get full score. Data independent questions are not preferred but fine as long as you have some correct code

- **Main page**
- Hints:
 - 1.1: read the maths in Puzzle 1.2 if you have no idea
 - 2.1: read **birthday problem** on Wikipedia if you have no idea
 - 3.2: instead of checking `== 2000`, check `<= 2000` instead
 - 4.4: any reasonable estimate is fine. You can check it in Puzzle 4.5
- Submit your work. Feel free to:
 - ask us questions
 - leave whenever you finish the lab