# Homework 5 - Regular Expression and Factor

Miao Yu

2023-11-20

### Requirements:

1. Submit your report in pdf knitted from R markdown.
2. Organize your report clearly by tasks, questions using different level of headers.
3. For each question, include the question itself, the code/result/graph to answer the question, and your answer in plain language.
4. You need to polish your graph details to reasonable visual comfort.

### Rubrics:

- Completeness: 20% - you lose up to 5% for each missing itemized tasks to answer.
- Correctness: 40% - you lose up to 5% for each incorrect graph/answer (that doesn't answer the proposed question reasonably well by data visualization).
- Appearance: 20% - you lose up to 2% for each unpolished details in your figure (missing title, labels, bad font size, typos in spelling, missing necessary units in labels etc.).
- Accuracy: 20% - you lose up to 4% for each inaccurate statements in your answer.

## 1. Regular expressions

#### a) Use the `words` data set, find all the words that match the following pattern:

- are exactly four letters long
- are either four or five letters long
- the second letter is "s" or "t"
- contains the pattern like "oxx" where "o" is one letter and "x" is another letter
- contains "a", "e" and "o" at the same time

#### b) Use the `sentences` data set, make the following plot

- a bar plot counting sentences with and without "the" (or "The").
- a scatterplot with $x$ being the average length of words in a sentence, and $y$ being the number of words starting with "a" or "e" or "i" or "o" or "u" in the sentence.

### c) Application

i) Download the Oxford English Dictionary as a ".txt" file from https://canvas.feitian.edu/files/9699/download?download_frd=1

ii) Read it into RStudio with `read_lines()` function (check how to use it by yourself)

iii) Turn the dictionary into a tibble and remove all blank lines

iv) Use regular expression to extract all words for each item in a separate column named "words"

## 2. Factors

### a) Use the `BankChurners.csv` to answer the following questions:

- Which features can be regarded as a factor?
- Which features can be regarded as an ordered factor (ordinal)?
- Read `BankChurners.csv` into RStudio, then change the columns that you answered above into factors or ordered factors.
- Visualize the effect of education level on averag utiliation ratio

### b) Use the `gss_cat` data set

- What are the levels of `marital` variable?
- Combine "Separated", "Divorced", "Widowed" into a new category "Once Married"
- Use the new levels, explore whether there is an effect of martial status on `tvhours`.