Introduction to programmatic business analytics, 12/2020, exam instructions:

0. **General rules/info:**
    - **You can use Python or R, or both, in the exam. All support material is allowed, but note that I will check programmatically that all students have unique scripts, and if not, <u>all students with similar scripts may fail the course</u>**. That is, not just those who copied, but also those who let their scripts to be copied will fail the course, unless you have a very good explanation for why the scripts are similar. You've been warned.
    - **You will get zero points for doing anything else than what is requested in these instructions.** You can do more than is needed, but it won't help you.
    - **Bonus steps are optional, and the bonus points count towards the final course grade.** So, if you for example got 30 points from DataCamp assignments, and maximum points from the exam (50 + 10) -> total points: 90 -> grade: 5.
    - **<u>Deadline</u> for sending the .py- and/or .R-files and the written document to the course instructor via personal email is <u>20.12., at 23:59</u>. If you miss the deadline, you'll get -10 points per each additional day it takes for you to send the files.**
1. Download the exam data files from: https://lut-my.sharepoint.com/:u:/g/personal/pontus_huotari_lut_fi/EYIihD2y6PxOjKSzeHxjKj0BEFz7u7y7Bv4sj27nTR1oiA?e=vBmQMI. The data are split into two comma-separated csv-files, containing 60 000 StackOverflow posts with quality ratings and other info.
2. Read the two .csv-files into separate data frames. (max. 5 points)
3. Combine the two data frames into one. (5 points) If you cannot combine the two data frames, you can proceed with either one.
4. Rename the column "Y" as "Post category". (5 points)*** If you cannot rename the column, just continue with the original name.
5. Replace values on the "Post category" column as follows: "HQ" -> "High-quality post", and both "LQ_EDIT" and "LQ_CLOSE" -> "Low-quality post". (max. 5 points) If you cannot replace the values, just continue with the original values.
6. Create a new column "Python in tags", which equals True if <python> is found on the "Tags" of a post, and otherwise False. (5 points)
7. BONUS: Remove html, or in other words extract pure text on the "Body" column (5 points).
8. BONUS: Remove excess whitespace (e.g., more than one space between words) from texts on the "Body" column (5 points).
9. Count the number of words in the posts on the "Body" column. (5 points) If you cannot do this, count the number of characters in the posts. (2.5 points)
10. Create a box plot with the following properties (HINT: you can do this whole thing with one command from a suitable package in both Python and R):
    - the number of words (or characters, if you could not count the words) in the posts on the y-axis, separated by post category (i.e., high- or low-quality) on the x-axis (max. 4 points)
    - for each post category, split the results further by whether the posts are Python-related, marking this info in the legend (2 points)
    - outliers removed from the visualization (2 points)
    - confidence intervals ("notches") shown for the median values (2 points)
11. Write max. one page (if longer with the default document settings in Word -> -5 points) document to which you attach the box plot, and explain verbally what the differences in the median values mean and how reliable the plotted results are alone. (max. 10 points)

*** If you complete the exam with R, you do not need to use spaces in column names