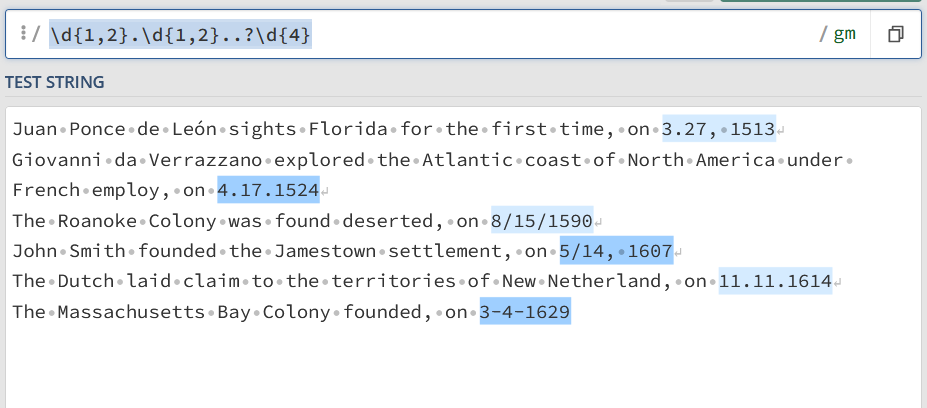
# What regular expressions do you use to extract all the dates in this blurb: [http://bit.ly/regexexercise2](http://bit.ly/regexexercise2" \t "_blank) and to put them into the following format YYYY-MM-DD ?

I use the expression: \d{1,2}.\d{1,2}..?\d{4} to find the dates already in the text

I use \d as it matches digits and curly brackets around the numbers as the first number can be both 1 digit and to digits. I use “.” as “.” can mean any character as what between the first and second number can be both an actual period and / and – and “.” catches all of them.



Then I group them together with parentheses like this: (\d{1,2}).(\d{1,2})..?(\d{4})

The parantheses means that I group specific number together shown in the different colours in the picture beneath

Et billede, der indeholder tekst

Automatisk genereret beskrivelse

Then I get to substituting the different formats into YYYY-MM-DD: by choosing substitution and then using this expression: $3-$1-$2

“$” shows which group I wish to use, and since I wish the year to be first which is group 3 that is the first one and then group 1 as that is the month and then finally group 2 as the day

Et billede, der indeholder tekst

Automatisk genereret beskrivelse

Source: <https://regex101.com/r/eerxHc/1>

# Write a regular expression to convert the stopwordlist (list of most frequent Danish words) from Voyant in [http://bit.ly/regexexercise3](http://bit.ly/regexexercise3" \t "_blank) into a neat stopword list for R (which comprises "words" separated by commas, such as [http://bit.ly/regexexercise4](http://bit.ly/regexexercise4" \t "_blank)). Then take the stopwordlist from R [http://bit.ly/regexexercise4](http://bit.ly/regexexercise4" \t "_blank) and convert it into a Voyant list (words on separate line without interpunction)

## R to Voyant

In R the stopword list is separated by “, “ so first I search for that and as it can be seen you miss the first one, which I haven’t figured out how to remove other than manually.

Et billede, der indeholder bord

Automatisk genereret beskrivelse

After that I choose substitution (red arrow) and use the expression \n which means new line.

The blue arrow shows where the first “ has to be removed manually

Et billede, der indeholder tekst

Automatisk genereret beskrivelse

Sources: <https://regex101.com/r/eerxHc/1>

## Voyant to R

In this I use \n to find all the line changes, when I did this in class I used \s because it matches spaces and new lines, but that might not be reproducible in other cases than this specific one.

Et billede, der indeholder tekst

Automatisk genereret beskrivelse

Then I use substitution again and replaces the new lines with “, “

Et billede, der indeholder tekst

Automatisk genereret beskrivelse

Again I run into the issue at the beginning where I have to manually add “ at the first one for it to be done.

Source: <https://regex101.com/r/aDN0cS/1>

# In 250 words, answer the following question: "What are the basic principles for using spreadsheets for good data organisation?"

The one that imprinted on me the most is that if I want my spreadsheet to make good data later, I need to use the same words or terms when I write about the same thing[[1]](#footnote-1), no random capitalizations or weird shortenings both of which I am prone to do. Generally, the characters used should be consistent whether it is the terms or whether hyphen or underscores are used or how dates are written, although with dates it is also a good idea to use the format YYYY-MM-DD and also to remember that spreadsheets can do weird thing with dates or numbers that look like dates[[2]](#footnote-2).

It is also a good idea to use terms that are self-explanatory, so that other people seeing the data can use it without too much difficulty or double check the work previously done[[3]](#footnote-3).

Another thing is to leave no empty cells and write NA or something similar in cells with no information although this is something that is discussed within the data-community. I think I agree with not leaving cells empty as that makes it easier to see an empty cell as a mistake[[4]](#footnote-4).

It is also a good idea to leave the raw data alone and do the calculations or changes in a different document, as the calculations often require you to open and type in the file often and therefor have more opportunities to do something that could mess up the data[[5]](#footnote-5).

Challenge (OPTIONAL)!Can you find all the instances of 'Dis Manibus' invocation in the EDH inscriptions in <https://bit.ly/regexexercise5>? Beware of the six possible canonical versions of the Dis Manibus formula!

I didn’t do this one

1. Karl W. Broman & Kara H. Woo (2018) Data Organization in Spreadsheets, The American Statistician, 72:1, 2-10, DOI: [10.1080/00031305.2017.1375989](https://doi.org/10.1080/00031305.2017.1375989) s. 3 [↑](#footnote-ref-1)
2. Karl W. Broman & Kara H. Woo (2018) Data Organization in Spreadsheets, The American Statistician, 72:1, 2-10, DOI: [10.1080/00031305.2017.1375989](https://doi.org/10.1080/00031305.2017.1375989) s. 3 [↑](#footnote-ref-2)
3. Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, et al. (2017) Good enough practices in scientific computing. PLOS Computational Biology 13(6): e1005510. <https://doi.org/10.1371/journal.pcbi.1005510> S. 3 [↑](#footnote-ref-3)
4. Karl W. Broman & Kara H. Woo (2018) Data Organization in Spreadsheets, The American Statistician, 72:1, 2-10, DOI: [10.1080/00031305.2017.1375989](https://doi.org/10.1080/00031305.2017.1375989) s. 5 [↑](#footnote-ref-4)
5. Karl W. Broman & Kara H. Woo (2018) Data Organization in Spreadsheets, The American Statistician, 72:1, 2-10, DOI: [10.1080/00031305.2017.1375989](https://doi.org/10.1080/00031305.2017.1375989) s. 7 [↑](#footnote-ref-5)