1. What regular expressions do you use to extract all the dates in this blurb: http://bit.ly/regexexercise2 and to put them into the following format YYYY-MM-DD ?

First, I took the expressions from https://librarycarpentry.org/lc-data-intro/01-regular-expressions/index.html about Matching multiple date formats and took the solution from there:

\d{2}-\d{2}-\d{2,4}$

Then I tried to change it around until it matched all the numbers, and with the teachers help, I made an expression like this:

\d{1,2}.\d{1,2}..?\d{4}
https://regex101.com/r/hM8tMi/1

By changing the numbers inside { } it was able to highlight or find the date, months and year:

- {1,2} matches the previous token between 1 and 2 times, as many times as possible, giving back as needed
    - Which makes it able to count both date and month as they can be both 1 or 2 numbers
- {4} matches the previous token exactly 4 times
    - Which makes it count the year as it is always 4 numbers

The . between date and month makes it so there must be something between them, in the text dump it can be either . - / , and the ..? makes it so there can be multiple characters between the middle number and the 4 year number (I think).

Then to format the different dates to the format YYYY-MM-DD I went to change the FUNCTION in the left taskbar to Substitution instead of Match.

Afterwards I added a ( ) between every \d line so it looked like this:

(\d{1,2}).(\d{1,2})..?(\d{4})

By adding ( ) it would make the expressions inside different 'parts' or count them as separate from each other:

- First ( ) became (\d{1,2})
- Second ( ) (\d{1,2})
- And third was (\d{4})

Then in the Substitution menu I added $3-$2-$1 which meant it would count the third ( ), the 4 year numbers, as the first one, the second ( ) or the months into second in list and the date last.

Therefore making the data format into YYYY-MM-DD in the Substitution menu.

https://regex101.com/r/QmPDVq/1

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

From Voyant to R:

https://regex101.com/r/TKxaol/1

By adding \n in the regular expression it would highlight every case of making a new line or 'enter' on keyboard.

Then in the Substitution menu I added ", " which would change the format to match R as there no longer is any new line between every word.
It also adds " " between every word and adds a , and a space between every word.

From R to Voyant:

https://regex101.com/r/UYxBv7/1

By adding ", " in the regular expression it would highlight every case ", " so only the words were left.

Then in the Substitution menu I would add \n so that after every word it would create a new line so it matches the format in Voyant.

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

One of the most important principles for using spreadsheets for good data organization is being consistent in your work; By being consistent and using the same format every time you can drastically decrease the change of encountering errors and make it easier to analyze your data.

If you make use of multiple spreadsheets about different data or findings, it can help to keep the same format between them. It makes it easier for yourself, and others, to quickly understand your work and you remove the hassle of thinking about the format every time you make a new dataset.

In the article *Data Organization in Spreadsheets* by Broman K.M. & Kara H. Woo they list other basic principles such as write dates the same way, YYYY-MM-DD, to keep it understandable between countries. Another is to not leave a cell empty as it can be interpreted as missing data and not just no data or no result, and again keep it consistent the way you keep the data organized in the cells.

It is also important to save often and make use of multiple backups so you will not lose your work. If you must make calculations, you should also make it in a different spreadsheet and always have a copy of the raw data available – that way you will not lose it.

Additionally it is also very important to choose a good name for your file and again try to keep it consistent. That way your file will be easier to find and understood by others.