Weekly homework

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

To extract all the specific dates, I am using this regular expression:

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

Et billede, der indeholder tekst

Automatisk genereret beskrivelse

I have now extracted the dates. Now I want to put them into the following format: YYYY-MM-DD.

First, i will be using this regular expression: (\d{1,2}).(\d{1,2})..?(\d{4})

This will now allow me to use the substitution function in the regex101 program, where i can switch the dates so they will fit in YYYY-MM-DD.

Et billede, der indeholder tekst

Automatisk genereret beskrivelse

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)***into a neat stopword list for R (which comprises "words" separated by commas, such as***[***http://bit.ly/regexexercise4***](http://bit.ly/regexexercise4)***). Then take the stopwordlist from R***[***http://bit.ly/regexexercise4***](http://bit.ly/regexexercise4)***and convert it into a Voyant list (words on separate line without interpunction)***

In the first part of this task, we want to convert the vertical stopwordlist (Voyant), into a horizontal list (R). To do so, I am going to insert the Voyant list in the text string, and use the regular expression: *\n.* This makes a new list in the substitution. But this is without “ and ,.

Therefor I will write: *“,”* in the substitution, and then it will divide the words in the list as it is listed in the graphic below.

Et billede, der indeholder tekst

Automatisk genereret beskrivelse

In the second part of the task, I am now asked to convert a horizontal (R) stopwordlist into a vertical (Voyant) list.

Therefor am I going to write the regular expression: *“, “.* That is with a space between , and “.

In this way I can list all these in the text string and substitute them with *\n (newline).* This will give me the results in the graphic below.

Et billede, der indeholder tekst

Automatisk genereret beskrivelse

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

To have good data organisation requires a great amount of working with your spreadsheets. And there are some basic principles required in relation to having a great spreadsheet.

First - it is very important to be consistent working in a spreadsheet. No matter what you do, and no matter the purpose, it is important that you do it consistently. Consistency is demanded to organize your spreadsheet in a way to prevent you, or any other person, from making mistakes. For instance: If you are writing “male” in your spreadsheet, then continue to write it in the same way. Do not write “Male” or “MALE”.

Secondly - choose good names for anything. It is important to choose a good name. Because it can lead to misunderstandings in data, just as I mentioned above, but also for other thing e.g., name of the spreadsheet. As a basic ground rule, do not use spaces - either in the document or in file name 🡪 It is harder for the programmer!

Also, it is very important to write the date as YYYY-MM-DD, because it is standard in most places and will therefore not lead to any misunderstandings. It will also be strongly recommended because it can cause trouble in the spreadsheets.

Next basic rule is, do not leave any cell empty. It is important to use a common code for missing cells, either because there is nothing to write, or due to missing data. A common code, such as NA, will not lead to any misunderstandings.

Another thing, related to having good names, is that you should only put one thing in a cell. No spaces or whatsoever, just one thing. Alternatively, it is possible to assemble a word in a cell with “-“. E.g., Plate-Well

Another basic, which is very important, is to make the spreadsheet rectangle. In this way the columns are responding to one another. It will be readable and clear. For example:

Et billede, der indeholder hvid

Automatisk genereret beskrivelse

Next thing is that it will be very nice to create a data dictionary. In that way it is very helpful and nice for understanding the data variables. The most important things must be included, helping to understand the data and give an overview.

And now a couple of things that you really cannot do! Never do calculations in the raw data. It is very likely to mess up the whole dataset. Do it somewhere else. Another thing is, that you also cannot make/use color or highlighting as data. The numbers/text must be cleansed for such.

Now for some of the most important: Make backups! Do not lose all your work because you did not save or make a backup. It is crucial.

Last things, it is important to use data validation to avoid errors. Also - save the data in plain text files.