W44: Regular expressions and spreadsheets

**DESCRIPTION**

Upload your answers/solutions to the problems below. Beware of making the submission legible and understandable to another reader:

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 ?**

Juan Ponce de León sights Florida for the first time, on 3.27, 1513

Giovanni da Verrazzano explored the Atlantic coast of North America under French employ, on 4.17.1524

The Roanoke Colony was found deserted, on 8/15/1590

John Smith founded the Jamestown settlement, on 5/14, 1607

The Dutch laid claim to the territories of New Netherland, on 11.11.1614

The Massachusetts Bay Colony founded, on 3-4-1629

* \d = First, I wanted to match a digit from 0-9
* {1,2} = Then I needed to match 1 or 2 digits, because it varies between 1 and 2. I use curly brackets because the numbers I put “1,2” inside them because I need to match 1 or 2 digits because it variates.
* . = All the numbers have a different character on the 2. or 3. place.
* \d = Again, I need to match a digit from 0-9
* {1,2} = And again, I need to match either 1 or 2 digits
* . = Again, it varies witch digit that comes next
* ? = The question mark matches the previous token 0 or 1 times as many times as possible. Because in some of the dates there are a space, but not in all of them
* . = Then again, there is a different digit on this spot
* \d = We need to mark the year, which is written the same in all the dates
* {4} = Lastly we have to match the different 4 digits, but the value is still 4 in all of them.

Et billede, der indeholder tekst

Automatisk genereret beskrivelse

Here is a link to my R solution on the dates exercise: <https://regex101.com/r/9iLFnt/1>

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)**

* I copied the stop word list from Voyant into R
* \n = matches a new line. Because we want to mark the line so we can replace it. We want the words in one line only.
* In the substitution line I put in “, “ so I replace the new line with a comma and quotation signs.

**Et billede, der indeholder tekst

Automatisk genereret beskrivelse**

Here is a link to my R solution on the stop word list exercise: [**https://regex101.com/r/qJ6yWh/1**](https://regex101.com/r/qJ6yWh/1)

* “, “ = Then to transfer the stop list back to Voyant in have inserted the stoplist from R into my test string, I need to match the commas and the quotation marks
* \n = Then I add a substitution and add the newline again
* I have to delete the first and the last quotation mark manually, because the regular expression can’t catch it.

**Et billede, der indeholder tekst

Automatisk genereret beskrivelse**

Here is a link to my R solution on the stop word list exercise: <https://regex101.com/r/h0Q1Sn/1>

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

[**https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1375989**](https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1375989)

In the example above with the dates written in completely different ways, it makes it clear how important consistently data organization is for the reader. It is therefore important to select one simple way of writing the dates, like YYYY-MM-DD to avoid complicated data.

The basically principles of making good data are firstly to be consistent, like mentioned above. Secondly you should give your data good names, so it’s easy to recognize. Don’t use spaces but use other tokens like underscores. Don’t use special digits because they often have a meaning in the “data language”.

Fill in all the cells. It makes it more manageable. The article says the same thing about putting only one thing in the cell, it’s more manageable. The cells should also be rectangle like so:

**Et billede, der indeholder bord

Automatisk genereret beskrivelse**

It’s also a good idea to create a data-dictionary, that contains a list over the exact data names and an explanation, the measured units and expected maximum and minimum values. This is meta data, and it makes it easier for the reader to understand how the data was made. The data should be in the data cells only, and the graphs and calculations in other cells. This provides junk in the data. Of course, you need to save your data safe place and back it up. Microsoft Excel, Google Sheets and LibreOffice Calc are good tools to create a spreadsheet with calculations, analyses, and visualizations. It’s best to type in your data manually, because copy paste can cause mistakes.

W45: Open Refine

**DESCRIPTION**

Upload your answers to these questions:

1. **Create a \*tidy\* spreadsheet/table listing the names of Danish monarchs with their birth- and death-date and start and end of their reign. They should be sortable by year of birth. Suitable source website is for example**[**here**](https://kongehuset.dk/monarkiet-i-danmark/kongerakken)**, but you can also use another source, provided you reference it. (Collaboration is welcome. Remember to attach this spreadsheet to Brightspace submission)**
   * **Et billede, der indeholder bord

     Automatisk genereret beskrivelse**
2. **Does OpenRefine alter the raw data during sorting and filtering?**
   * No you have to edit it yourself.
3. **Fix the**[**interviews dataset**](https://ndownloader.figshare.com/files/11502815)**in OpenRefine enough to answer this question: "Which two months are reported as the most water-deprived/driest by the interviewed farmer households?"**
   * **Et billede, der indeholder bord

     Automatisk genereret beskrivelse**
   * It’s this column I’ve been asked to fix
   * The problem with this column is that several months are in the same cell in some of the cells, and the digits like ‘’, ; []. I need to remove them by doing this regular expression: value.replace(“[“,””).replace(“]”,””).replace(“’”,””).replace(“ “,””)

**Et billede, der indeholder bord

Automatisk genereret beskrivelse**

* + So I edit the cells by clicking on “Edit Cells” and “Transform”
  + Et billede, der indeholder bord

    Automatisk genereret beskrivelse
  + Now it looks like this, but we still have a problem with the multiple months in the same cells
  + **Et billede, der indeholder bord

    Automatisk genereret beskrivelse**
  + So now I need to custom text facet by adding this regular expression value.split(";")

****

* + Now I can see that the two driest months reported are October and September.

Et billede, der indeholder tekst

Automatisk genereret beskrivelse

* + This is the script of my solutions.

W46: Start with R

**DESCRIPTION**

**Instructions:** For this assignment, you need to answer a couple questions with code and then take a particular screenshot of your working environment.

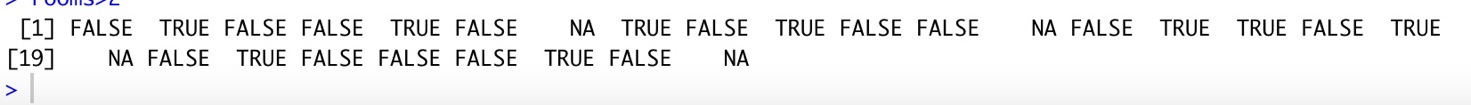
You can submit the solutions including the URL to the screenshot typed up in a doc/pdf to Brightspace OR upload the document with solutions and the screenshot to your repository on Github and submit here (to Brightspace) only your Github URL (make sure your HW files are immediately findable there).

1**) Use R to figure out how many elements in the vector below are greater than 2.**

**rooms <- c(1, 5, 2, 1, 3, 1, NA, 3, 1, 3, 2, 1, NA, 1, 8, 3, 1, 4, NA, 1, 3, 1, 2, 1, 7, 1, NA)**

**Et billede, der indeholder tekst

Automatisk genereret beskrivelse**

****

**-**To find the vectors greater than 2 I count the “TRUE”. And I count 9 vectors bigger than 2

**2) Which function tells you the type of data the 'rooms' vector above contains?**

**-** I use the function “class(rooms) to define the type of data Et billede, der indeholder tekst

Automatisk genereret beskrivelse and then it tells me that tells me that the type of data is “numeric”

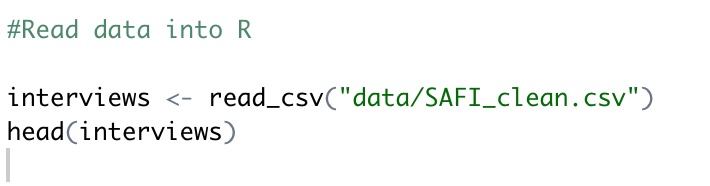
**3) What is the result of running the median() function on the above 'rooms' vector?**

**-** Et billede, der indeholder tekst

Automatisk genereret beskrivelseI get the result “NA” because the median of the data is NA

**4) Submit the following image to Github: Inside your R Project (.Rproj), install the 'tidyverse' package and use the download.file() and read\_csv() function to read the SAFI\_clean.csv dataset into your R project as 'interviews' digital object (see instructions in https://datacarpentry.org/r-socialsci/setup.html and 'Starting with Data' section). Take a screenshot of your RStudio interface showing**

**a) the line of code you used to create the object,**

****

**b) the 'interviews' object in the Environment, and**

**Et billede, der indeholder bord

Automatisk genereret beskrivelse**

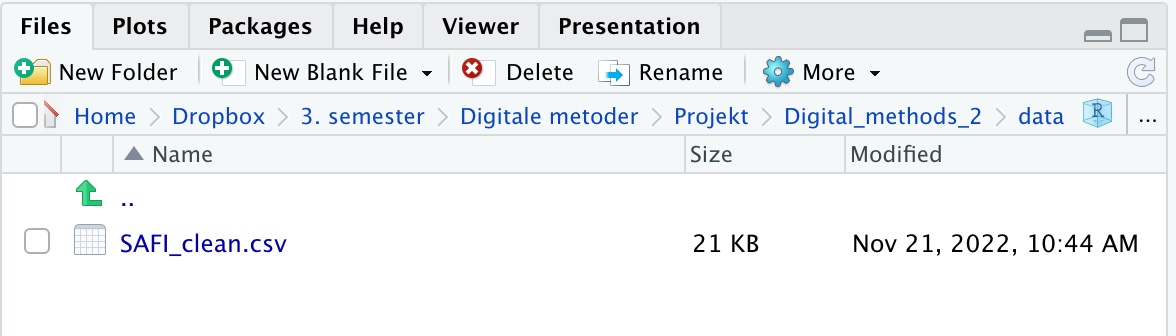
**c) the file structure of your R project in the bottom right "Files" pane.**

****

**Save the screenshot as an image and put it in your AUID\_lastname\_firstname repository inside our Github organisation (github.com/Digital-Methods-HASS) or equivalent. Place here the URL leading to the screenshot in your repository.**

**Et billede, der indeholder tekst

Automatisk genereret beskrivelse**

****

**Et billede, der indeholder bord

Automatisk genereret beskrivelse**

[**https://raw.githubusercontent.com/Digital-Methods-HASS/AU691180\_Tarp\_Anna/main/interview%20.jpg**](https://raw.githubusercontent.com/Digital-Methods-HASS/AU691180_Tarp_Anna/main/interview%20.jpg)

W48: Web scraping, text mining and mapping with R

**DESCRIPTION**

For this assignment, **choose ONE** of the following:

2.2) use the 'rvest' library to scrape data of your interest (football statistics in Wikipedia? trends of self-inflicted death in musicians? global population by country in https://www.worldometers.info/world-population/population-by-country/ )

<file:///Users/annatarp/Dropbox/3.%20semester/Digitale%20metoder/Projekt/Police_killings_assignment/Assignment-5-.html>

Final project digital methods

Nuclear weapons

Why are we afraid of Putin’s nuclear weapons when the stockpile is much lower today than in the 1980’s?

# Introduction

We didn’t thought Putin would invade the Crimean Peninsula. But he did in. We didn’t thought Putin would invade Ukraine. But he did. We don’t think Putin will be using his nuclear weapons. But we never know what Putin’s next step will be. This gave me the motivation to investigate Russia’s, and thus Putin’s nuclear arsenal today. My hypothesis is that the Russian nuclear stockpile today is significantly smaller than in the 1980’s when the Cold War was at its peak, and that the nine disarmament agreements made between Russia and USA helped decrease both stockpiles. But even though Russia’s nuclear possession is much lower in 2022, why are should we be afraid of a nuclear war?

The war in Ukraine is has been raging since the Russian army invaded Ukraine 24’th of February 2022 after President Putin announced “a special military operation in Ukraine” in a TV-speech.[[1]](#footnote-1) The war is a tragedy. Both for the Ukrainian population and for the world peace in general. We know that Russia possesses nuclear weapons, and in a decree from 2020 is it emphasized that Russia reserves the right to use their nuclear warheads, but also that these primarily are a political, strategic weapons with the purpose of creating fear.[[2]](#footnote-2) Nevertheless, we see a tendency from Russia that reveals Putin’s psychological game. When Ukraine is fighting back on the battlefield and it affects the Russian army in a negative way, Putin uses his nuclear weapons as a threat.[[3]](#footnote-3) In the decree from 2020 Russia has four main scenarios for the using nuclear weapons risk the world peace: The first scenario is if Russia has been victim of nuclear weapons or other weapons of mass destruction. The second is if another country destroys Russia’s ability to protect themselves from a nuclear attack. That could as an example be the Russian radar systems, airfields, missile silos and other Russian nuclear related properties. The third scenario for Russia to use their nuclear weapons is if Russia is being attacked by a nuclear weapon. And lastly, Russia will be using nuclear weapons if a conventional war is threatening Russia’s existence as a country.[[4]](#footnote-4) The last scenario is worth noticing concerning the current war in Ukraine. A nuclear bomb is not the only catalysator of a nuclear war. A false alarm, a close call or an error is enough to trigger a new war that can put the whole world’s destiny at risk.

# Historical context and goal

Putin’s frightening statements and hints about using Russia’s nuclear weapons in the war against Ukraine brings terrible associations back to the Cold War and the nuclear arms race between the Soviet Union and USA. There are currently nine countries who is in possession of nuclear weapons- These are China, France, India, Israel, North Korea, Pakistan, Russia, Great Britain, and USA. Historians talk about the “balance of terror” which means that the countries in possession of nuclear weapons should stay afraid of a nuclear war to never use their own warheads.[[5]](#footnote-5) Considering the raging war in Ukraine forced by Russia, I find it interesting to find some data about Russia’s possession of nuclear weapons compared to the eight other countries’ stockpiles. I found some useful data on <https://ourworldindata.org/nuclear-weapons> about nuclear stockpiles in different countries over the years 1945-2022. All their data is free to use for anybody. Since I’m interested in looking at the Russian nuclear stockpile compared to the American stockpile, I created two tidy datasets of the two countries stockpiles from 1945-2022. I was curious if the nine nuclear disarmament agreements between Russia and USA had any effect on the stockpiles. The first agreement was the Nuclear Test Ban Treaty from 1963 that banned nuclear tests in the atmosphere. The Non-Proliferation Treaty in 1968 was to prevent the nuclear technology to spread. Then the first Strategic Arms Limitation Talk SALT agreement in 1972 added a limit to the legal possession of the countries’ stockpiles. In 1979 the SALT II treaty contained a wider spectrum of nuclear weapons than the SALT I treaty. In 1987, the two superpowers began to realize that the Cold War was varnishing to an end. The Intermediate-range Nuclear Forces (the INF-treaty) was introduced and abolished the medium-range missiles in Europe. In 1991 the Conventional Forces Europe (The CFE-treaty) involved a significant disarmament within the conventional field. In 1991 and 1993 where the START I and START II introduced. It stands for Strategic Arms Reduction Talks and were to reduce the nuclear weapons to 2/3. The last two nuclear disarmament agreements were in 2002 and 2010. In 2002 Russia and USA made an agreement containing that both countries should disarm their stockpiles within ten years to maximum 2200 weapons each country. In 2010 President Barack Obama sign a treaty with disarming further 30% of the stockpiles.[[6]](#footnote-6)

# Software framework

I created this project on my 1,5-year-old MacBook Air, 8 Gb RAM, which runs macOS Monterey version 12.0.1 operation system. I wrote my code in the desktop version of R version 4.2.2 (2022-10-31) and RStudio version 2022.07.2+576.

# Data Acquisition and Processing

As I mentioned before, I used the data from the online database Our World in Data. And more specifically I used the datasets from the article “Nuclear Weapons” made by Max Roser, Bastian Herre and Joe Hassel. <https://ourworldindata.org/nuclear-weapons>. I was able to download the data I needed as csv files and edit them and make them tidy in Microsoft Excel.

# Empirical Results

Husk at slette der hvor jeg har importeret datasættene

# Critical evaluation

# Conclusions

# Bibliography

Andersen, Jacob Svendsen and Sebastian Skov. "Atomspøgelset Er Genopstået Med Ukrainekrigen – Hvorfor Og Hvor Galt Kan Det Gå?" *Politiken*, 14. november 2022 2022, Politiken Fortæller.

Roser, Max. "Nuclear Weapons: Why Reducing the Risk of Nuclear War Should Be a Key Concern of Our Generation." 3. marts. Our World in Data, 8. december 2022 2022.

"Faktalink." Krigen i Ukraine, iBureauet, Updated Januar 2022, 2022, accessed 8. december, 2022.

Søndergård, Lars. "Nedrustningsaftaler." Accessed 4. Januar 2023. <https://samfundsfag.dk/begreber/international-politik/generel-ip/nedrustningsaftaler/>.

# References

# ***Table 1 – Software metadata***

|  |  |  |
| --- | --- | --- |
| **Nr** | **Software metadata description** |  |
| S1 | Current software version | R version 4.2.2 (2022-10-31)  RStudio version 2022.07.2+576 |
| S2 | Permanent link to Github repository where you put your script or R project | <https://github.com/AnnaTarp/Nuclear_possessions_project.git> |
| S3 | Legal Software License | “Licenses: All visualizations, data, and articles produced by Our World in Data are open access under the [Creative Commons BY license](https://creativecommons.org/licenses/by/4.0/). You have permission to use, distribute, and reproduce these in any medium, provided the source and authors are credited. All the software and code that we write is open source and made available via GitHub under the permissive [MIT license](https://github.com/owid/owid-grapher/blob/master/LICENSE.md). All other material, including data produced by third parties and made available by Our World in Data, is subject to the license terms from the original third-party authors.” <https://ourworldindata.org/about#legal> |
| S4 | Computing platform / Operating System | MacBook Air, 8 Gb RAM, which runs macOS Monterey version 12.0.1 operation system. |
| S5 | Installation requirements & dependencies for software not used in class | *I used this package in RStudio: library(stringr)* |
| S6 | If available Link to software documentation for special software |  |
| S6 | Support email for questions | **202107538@post.au.dk** |

|  |  |  |
| --- | --- | --- |
| **Nr** | **Metadata description** | ***Please fill in this column*** |
| D1 | USA.csv | *The American stockpile of nuclear weapons from 1945-2022* |
| D2 | Russia.csv | *The Russian stockpile of nuclear weapons from 1945-2022* |
| D3 | 9\_lande.csv | *The stockpiles of the nine countries possessing nuclear weapons* |

# Appendix

Translation of the variables in the dataset

1. Country = Country
2. Year = Year
3. Nuclear\_weapons = Stockpile of nuclear weapons
4. Entity = Country
5. nuclear\_weapons\_depl\_strat = Deployed strategic nuclear weapons
6. nuclear\_weapons\_depl\_nonstrat = Deployed nonstrategic nuclear weapons
7. nuclear\_weapons\_reserve\_nondepl = Nondeployed reserved nuclear weapons
8. nuclear\_weapons\_retired = Retired nuclear weapons

<https://ourworldindata.org/nuclear-weapons-risk>

Der skal ikke kun en atombombe sprængning til for at udløse en atomkrig, da en falsk alarm også kan udløse det.

”Balance of terror” betyder at de lande der er i besiddelse af atomvåbens ledere er så bange for en atomkrig at de aldrig vil bruge deres atomvåben. Vi kan ikke vide om Putin bryder denne balance. The world still possesses weapons that can kill humanity several of times. This is why false alarms, errors, and close calls are so crucial to monitor: they are the incidents that can push the ‘balance of terror’ out of balance and into war.

<https://politiken.dk/udland/art8987577/Dette-kan-Putin-opn%C3%A5-med-sine-atomv%C3%A5ben>

»Jeg vil gerne minde dem, der giver den slags udtalelser om Rusland, at vort land også ligger inde med forskellige ødelæggelsesvåben, og i nogle tilfælde er de mere moderne end dem, Nato-landene har«, sagde han og fortsatte:

»De, der prøver at afpresse os med atomvåben, skal vide, at vinden kan vende«.

<https://nyheder.tv2.dk/live/2022-12-07-seneste-nyt-om-krigen-i-ukraine>

Putin siger også, at truslen for en atomkrig er stigende. Han understreger i samme ombæring, at Rusland betragter atomvåben som et afskrækkelsesvåben. (7. december 2022)

<https://www.diis.dk/publikationer/atomaftalerne-smuldrer-nedrustning-stadig-noedvendig>

Atomprøvestopaftalen (1963)

Ikke-spredningsaftalen (1968)

SALT I (Strategic Arms Limitation Talks) fra 1972

 SALT II fra 1979

 INF-aftale (Intermediate-range Nuclear Forces) fra 1987

CFE-aftale (Conventional Forces Europe) fra 1991

START I og II (Strategic Arms Reduction Talks) fra 1991 og 1993

Nedrustningsaftale maj 2002: Rusland og USA indgik aftale, der gik ud på, at begge lande forpligtes til inden for ti år at reducere antallet af strategiske atomvåben med omkring to tredjedele til et maksimum på 2.200 for hvert land

I 2002 trak USA sig ud

I marts 2010 underskrev USA´s præsident Obama og Ruslands præsident Medvedev en aftale om yderligere nedskæringer af deres atomvåbenarsenaler med 30% i forhold til [Moskvatraktaten](https://samfundsfag.dk/begreber/international-politik/oevrige/traktat/) fra 2002.

If we look at the nuclear weapons on a historical perspective again, i would like to see if the seven disarmament agreements has any effect on the American and Russian nuclear stockpiles during the Cold War. After the Cold Wars's ending, the two super powers made three more disarmament agreements.

1. "Faktalink," Krigen i Ukraine, iBureauet, updated Januar 2022, 2022, accessed 8. december, 2022. [↑](#footnote-ref-1)
2. Jacob Svendsen and Sebastian Skov Andersen, "Atomspøgelset er genopstået med Ukrainekrigen – hvorfor og hvor galt kan det gå?," *Politiken*, 14. november 2022 2022, Politiken Fortæller. [↑](#footnote-ref-2)
3. Andersen, "Atomspøgelset er genopstået med Ukrainekrigen – hvorfor og hvor galt kan det gå?." [↑](#footnote-ref-3)
4. Andersen, "Atomspøgelset er genopstået med Ukrainekrigen – hvorfor og hvor galt kan det gå?." [↑](#footnote-ref-4)
5. Max Roser, "Nuclear weapons: Why reducing the risk of nuclear war should be a key concern of our generation," (3. marts, Our World in Data, 8. december 2022 2022). [↑](#footnote-ref-5)
6. Lars Søndergård, "Nedrustningsaftaler." https://samfundsfag.dk/begreber/international-politik/generel-ip/nedrustningsaftaler/. [↑](#footnote-ref-6)