

WorkingWithDataAssignment

João Mota

Table of contents

1	Now we do the same for the other years before we merge them	2
2	Quarto	18
3	Running Code	18

```
/*  
A few hours of trial and errors can save you a few minutes of reading the proper documentation :)  
https://quarto.org/docs/output-formats/pdf-basics.html  
Go to terminal tab down there and type quarto install tool tinytex  
NOTE TO SELF!!!! using quarto is the same as playing restart Rstudio simulator 2022 because nothing is  
properly recached and they have a worse garbage collector than assembly so if you still get the same error  
after changing the just restart rstudio and remember to never ever ever change the initial format or add  
anything close to it because it will break the pdf and start generating html also please be smart and read  
https://quarto.org/docs/reference/formats/pdf.html for the formatting  
*/
```

```
library(haven)  
library(tidyverse)
```

```
-- Attaching packages ----- tidyverse 1.3.2 --  
v ggplot2 3.3.6      v purrr  0.3.4  
v tibble  3.1.8      v dplyr  1.0.10  
v tidyr   1.2.1      v stringr 1.4.1  
v readr   2.1.3      v forcats 0.5.2  
-- Conflicts ----- tidyverse_conflicts() --  
x dplyr::filter() masks stats::filter()  
x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)  
library(geometry)  
#install.packages('fontspec')  
library(formatR)
```

```

## Very important documentation for the 2018 data set //it is a
## surprise toll that will help us later
technicalAnnex2018 = "https://doc.ukdataservice.ac.uk/doc/8406/mrdoc/pdf/8406_cyber_security_b

## this is the loading the first year of this level of survey data set
## after burning my entire brain, replacing it with the backup one and
## also burning that one I discovered that it is just these lines that
## aren't being formatted in pdf because they are absolutely huge but
## at least it works for the other ones #FicaADica I assume it was
## thanks to formatR ?? I won't bother to redo every single bloody step
## again, enough pain and stack for the day
dataCyberSecuritySurvey2018 = read_spss("C:/AppliedDataScienceAndStatistics/Applied-Data-Scien

## adding the variable year because none of the data sets have any
## proper way to distinguish between the years of each survey
dataCyberSecuritySurvey2018$year = "2018"

```

1 Now we do the same for the other years before we merge them

```

## loading the second year of this level of survey data set
dataCyberSecuritySurvey2019 = read_spss("C:/AppliedDataScienceAndStatistics/Applied-Data-Scien

## adding the variable year because none of the data sets have any
## proper way to distinguish between the years of each survey
dataCyberSecuritySurvey2019$year = "2019"

## loading the third year of this level of survey data set
dataCyberSecuritySurvey2020 = read_spss("C:/AppliedDataScienceAndStatistics/Applied-Data-Scien

## adding the variable year because none of the data sets have any
## proper way to distinguish between the years of each survey
dataCyberSecuritySurvey2020$year = "2020"

## loading the forth year of this level of survey data set
dataCyberSecuritySurvey2021 = read_spss("C:/AppliedDataScienceAndStatistics/Applied-Data-Scien

## adding the variable year because none of the data sets have any
## proper way to distinguish between the years of each survey
dataCyberSecuritySurvey2021$year = "2021"

## loading the fifth and final year of this level of survey data set
dataCyberSecuritySurvey2022 = read_spss("C:/AppliedDataScienceAndStatistics/Applied-Data-Scien

## adding the variable year because none of the data sets have any

```

```

## proper way to distinguish between the years of each survey
dataCyberSecuritySurvey2022$year = "2022"

## Now that we have all data loaded lets start by tidying up data set
## by data set start from 2018

## for some sweet sweet documentation about the questions starting from
## page 26 TODO comment in case of fire or debugging
browseURL(technicalAnnex2018)

## This entire code snippet is tidying up the type of organisation for
## the 2018 survey renaming the bloody variables to a more java like
## name

dataCyberSecuritySurvey2018TidyName = rename(dataCyberSecuritySurvey2018,
  institutionTypes = "samptype")

## if institutionTypes is 1 it is a business if it is 2 it is a
## charity and in the future 3 is for schools and education

## daily reminder that there is a boolean type but it is called logical
## Numeric -\tSet of all real numbers Integer -\tSet of all integers, Z
## Logical - \tTRUE and FALSE Complex -\tSet of complex numbers
## Character -\t"a", "b", "c", ..., "ç", "#", "~", ..., "1", "2", ...etc

## it is a string so lets make it a proper numeric code

dataCyberSecuritySurvey2018TidyName$institutionTypes = as.integer(dataCyberSecuritySurvey2018

## typex is 1-2 for businesses and 3 for charities so redundant and can
## be removed

dataCyberSecuritySurvey2018TidyName = dataCyberSecuritySurvey2018TidyName %>%
  select(-typex)

## dataCyberSecuritySurvey2018TidyName never forget if R can't show all
## displayed text from a computation it breaks both the rendering and
## ##the refreshing of the rendered code for some reason ~/_(_)/~
## future edit anything and everything breaks for no reason at all,
## just kill it and reopen refer to the first NOTE TO SELF for more
## information

```

```
## see questioner documentation start from page 27
```

```
technicalAnnex2019 = "https://assets.publishing.service.gov.uk/government/uploads/system/uploa  
## TODO comment in case of fire or debugging  
browseURL(technicalAnnex2019)
```

```
## see questioner documentation start from page 31
```

```
technicalAnnex2020 = "https://assets.publishing.service.gov.uk/government/uploads/system/uploa  
## TODO comment in case of fire or debugging  
browseURL(technicalAnnex2020)
```

```
## see questioner documentation start from page 28
```

```
technicalAnnex2021 = "https://assets.publishing.service.gov.uk/government/uploads/system/uploa  
## TODO comment in case of fire or debugging  
browseURL(technicalAnnex2021)
```

```
## see questioner documentation start from page 36
```

```
technicalAnnex2022 = "https://assets.publishing.service.gov.uk/government/uploads/system/uploa  
## TODO comment in case of fire or debugging  
browseURL(technicalAnnex2022)
```

```
## trying not to get arrested for DDoSing the uk goverment by making a  
## request to all the pdfs after rendering the page for the nth because  
## I can't code nor debug (challenge impossible) bonus points if I get  
## an exeter ip banned because of it
```

```
## time to recycle the code for the 2018 survey that gets a 'neat' code  
## of the institution types
```

```
## This entire code snippet is tidying up the type of organisation for  
## the 2019 survey renaming the bloody variables to a more java like  
## name
```

```
dataCyberSecuritySurvey2019TidyName = rename(dataCyberSecuritySurvey2019,  
      institutionTypes = "samptype")
```

```
dataCyberSecuritySurvey2019TidyName$institutionTypes = as.integer(dataCyberSecuritySurvey2019  
str(dataCyberSecuritySurvey2019TidyName$institutionTypes)
```

```
int [1:2080] 1 1 1 1 1 1 1 1 1 1 ...
```

```
## typex is redundant be we already have an indentifies for each type  
## of institution and can be removed same for questtype since this  
## questioner has more redundancy than amazon and google data centers  
## combined
```

```
dataCyberSecuritySurvey2019TidyName = dataCyberSecuritySurvey2019TidyName %>%  
  select(-one_of("typex", "questtype"))
```

```
## I continue to save the planet by recycling as much as I can, mostly  
## recycled code from the previous snippet today though this time we do  
## have the concept of education institutions as our code just annoy me  
## after I thought they should be converted to boolean like a getter in  
## java
```

```
dataCyberSecuritySurvey2020TidyName = rename(dataCyberSecuritySurvey2020,  
  institutionTypes = "samptype")
```

```
dataCyberSecuritySurvey2020TidyName$institutionTypes = as.integer(dataCyberSecuritySurvey2020  
str(dataCyberSecuritySurvey2020TidyName$institutionTypes)
```

```
int [1:1900] 1 1 1 1 1 1 1 1 1 1 ...
```

```
## typex is redundant be we already have an indentifies for each type  
## of institution and can be removed same for questtype since this  
## questioner has more redundancy than amazon and google data centers  
## combined
```

```
dataCyberSecuritySurvey2020TidyName = dataCyberSecuritySurvey2020TidyName %>%  
  select(-one_of("typex", "questtype"))
```

```
## saving the planet one recycled snippet of code at a time
```

```
dataCyberSecuritySurvey2021TidyName = rename(dataCyberSecuritySurvey2021,  
  institutionTypes = "samptype")
```

```
dataCyberSecuritySurvey2021TidyName$institutionTypes = as.integer(dataCyberSecuritySurvey2021  
str(dataCyberSecuritySurvey2021TidyName$institutionTypes)
```

```
int [1:1900] 1 1 1 1 1 1 1 1 1 1 ...
```

```

## typex is redundant be we already have an indentifies for each type
## of institution and can be removed same for questtype since this
## questioner has more redundancy than amazon and google data centers
## combined

dataCyberSecuritySurvey2021TidyName = dataCyberSecuritySurvey2021TidyName %>%
  select(-one_of("typex", "questtype"))

## this comment was already dealt by the garbage collector unlike the
## previous ones

dataCyberSecuritySurvey2022TidyName = rename(dataCyberSecuritySurvey2022,
  institutionTypes = "samptype")

dataCyberSecuritySurvey2022TidyName$institutionTypes = as.integer(dataCyberSecuritySurvey2022

str(dataCyberSecuritySurvey2022TidyName$institutionTypes)

int [1:2157] 1 1 1 1 1 1 1 1 1 1 1 ...

## questtype is redundant be we already have an indentifies for each
## type of institution and can be removed

dataCyberSecuritySurvey2022TidyName = dataCyberSecuritySurvey2022TidyName %>%
  select(-questtype)

## now that we have started the data wrangling we will categorize all
## institutions by size remember that for some wicked reason they use
## -97 for missing values for anything without a proper missing value
## code for each question I will start by simply nulling every single
## -97 so we can see how much is missing and then possibly make a table
## with custom missing values for each like I did in C (remember to
## start from -1000 to -1999 like standard ACLs)

##### 2018
#####
numberOfCycles = length(dataCyberSecuritySurvey2018TidyName$sizea)

dataCyberSecuritySurvey2018TidyNameSize = dataCyberSecuritySurvey2018TidyName

## apparently we have to be careful because an already inserted NA on
## the variable breaks the
for (i in 1:numberOfCycles) {
  if (dataCyberSecuritySurvey2018TidyNameSize$sizea[i] == -97) {
    dataCyberSecuritySurvey2018TidyNameSize$sizea[i] = NA
  }
}

```

```

    }
    if (dataCyberSecuritySurvey2018TidyNameSize$sizeb[i] == -97) {
        dataCyberSecuritySurvey2018TidyNameSize$sizeb[i] = NA
    }
}

##### 2019
##### 

numberOfCycles = length(dataCyberSecuritySurvey2019TidyName$sizea)

dataCyberSecuritySurvey2019TidyNameSize = dataCyberSecuritySurvey2019TidyName

## apparently we have to be careful because an already inserted NA on
## the variable breaks the
for (i in 1:numberOfCycles) {
    if (dataCyberSecuritySurvey2019TidyNameSize$sizea[i] == -97) {
        dataCyberSecuritySurvey2019TidyNameSize$sizea[i] = NA
    }
    if (dataCyberSecuritySurvey2019TidyNameSize$sizeb[i] == -97) {
        dataCyberSecuritySurvey2019TidyNameSize$sizeb[i] = NA
    }
}

##### 2020
##### 

numberOfCycles = length(dataCyberSecuritySurvey2020TidyName$sizea)

dataCyberSecuritySurvey2020TidyNameSize = dataCyberSecuritySurvey2020TidyName

## apparently we have to be careful because an already inserted NA on
## the variable breaks the
for (i in 1:numberOfCycles) {
    if (dataCyberSecuritySurvey2020TidyNameSize$sizea[i] == -97) {
        dataCyberSecuritySurvey2020TidyNameSize$sizea[i] = NA
    }
    if (dataCyberSecuritySurvey2020TidyNameSize$sizeb[i] == -97) {
        dataCyberSecuritySurvey2020TidyNameSize$sizeb[i] = NA
    }
}

##### 2021
##### 

```

```

numberOfCycles = length(dataCyberSecuritySurvey2021TidyName$sizea)

dataCyberSecuritySurvey2021TidyNameSize = dataCyberSecuritySurvey2021TidyName

## apparently we have to be careful because an already inserted NA on
## the variable breaks the
for (i in 1:numberOfCycles) {
  if (dataCyberSecuritySurvey2021TidyNameSize$sizea[i] == -97) {
    dataCyberSecuritySurvey2021TidyNameSize$sizea[i] = NA
  }
  if (dataCyberSecuritySurvey2021TidyNameSize$sizeb[i] == -97) {
    dataCyberSecuritySurvey2021TidyNameSize$sizeb[i] = NA
  }
}

##### 2022
##### 

numberOfCycles = length(dataCyberSecuritySurvey2022TidyName$sizea)

dataCyberSecuritySurvey2022TidyNameSize = dataCyberSecuritySurvey2022TidyName

## apparently we have to be careful because an already inserted NA on
## the variable breaks the
for (i in 1:numberOfCycles) {
  if (dataCyberSecuritySurvey2022TidyNameSize$sizea[i] == -97) {
    dataCyberSecuritySurvey2022TidyNameSize$sizea[i] = NA
  }
  if (dataCyberSecuritySurvey2022TidyNameSize$sizeb[i] == -97) {
    dataCyberSecuritySurvey2022TidyNameSize$sizeb[i] = NA
  }
}

## we don't need neither the combined regions (since those are for
## business analyzes and we don't do those) same for sector_comb1 and
## 2.

## region_comb? throw it in the trash. sector_comb1? throw it in the
## trash. sector_comb2? throw it in the trash.

##### 2018
##### 

dataCyberSecuritySurvey2018TidyNameSize = dataCyberSecuritySurvey2018TidyNameSize %>%
  select(-one_of("region_comb", "sector_comb1", "sector_comb2"))

##### 2019

```



```
#####
```

```
dataCyberSecuritySurvey2019TidyNameSize = dataCyberSecuritySurvey2019TidyNameSize %>%  
  select(-one_of("region_comb", "sector_comb2"))
```

```
##### 2020
```

```
#####
```

```
dataCyberSecuritySurvey2020TidyNameSize = dataCyberSecuritySurvey2020TidyNameSize %>%  
  select(-one_of("region_comb", "sector_comb2"))
```

```
##### 2021
```

```
#####
```

```
dataCyberSecuritySurvey2021TidyNameSize = dataCyberSecuritySurvey2021TidyNameSize %>%  
  select(-one_of("region_comb", "sector_comb2"))
```

```
##### 2022
```

```
#####
```

```
dataCyberSecuritySurvey2022TidyNameSize = dataCyberSecuritySurvey2022TidyNameSize %>%  
  select(-one_of("region_comb", "sector_comb2"))
```

```
## removing social media questions that are irrelevant because they are  
## absolutely terrible metrics to understand the digitalization of an  
## institution Note to self: if I have time get all of these type of  
## functions in try catch because them breaking up with the select  
## error is no good and it makes me cry every time I have to manually  
## run a part of the snippet and see which is one the bad one  
## https://r-lang.com/r-trycatch-function/ ## #FicaADica
```

```
##### 2018
```

```
#####
```

```
dataCyberSecuritySurvey2018TidyNameSize = dataCyberSecuritySurvey2018TidyNameSize %>%  
  select(-(online1:online11))
```

```
##### 2019
```

```
#####
```

```
dataCyberSecuritySurvey2019TidyNameSize = dataCyberSecuritySurvey2019TidyNameSize %>%  
  select(-(online1:online11))
```

```
##### 2020
```

```
#####
```

```
dataCyberSecuritySurvey2020TidyNameSize = dataCyberSecuritySurvey2020TidyNameSize %>%
```

```

select(-(online1:online11))

##### 2021
##### 

dataCyberSecuritySurvey2021TidyNameSize = dataCyberSecuritySurvey2021TidyNameSize %>%
  select(-(online1:online11))

##### 2022
##### 

dataCyberSecuritySurvey2022TidyNameSize = dataCyberSecuritySurvey2022TidyNameSize %>%
  select(-(online1:online14))

## removing the question about the mobile usage because it also is a
## terrible indicator of a company digitalization

##### 2018
##### 

dataCyberSecuritySurvey2018TidyNameSize = dataCyberSecuritySurvey2018TidyNameSize %>%
  select(-mobile)

##### 2019
##### 

dataCyberSecuritySurvey2019TidyNameSize = dataCyberSecuritySurvey2019TidyNameSize %>%
  select(-mobile)

##### 2020
##### 

dataCyberSecuritySurvey2020TidyNameSize = dataCyberSecuritySurvey2020TidyNameSize %>%
  select(-mobile)

##### 2021
##### 

dataCyberSecuritySurvey2021TidyNameSize = dataCyberSecuritySurvey2021TidyNameSize %>%
  select(-mobile)

##### 2022
##### 

dataCyberSecuritySurvey2022TidyNameSize = dataCyberSecuritySurvey2022TidyNameSize %>%
  select(-mobile)

```

```
## question about the attitude and outsourcing of cyber security have
## been removed the the surveys starting from 2020 so it doesn't make
## sense to keep them in the 2018 and 2019 data set
```

```
## I will start doing some proper garbage collection and this time I am
## not just taking myself out I will only ever have the original data
## and the most recent modified one
```

```
##### 2018
```

```
#####
```

```
dataCyberSecuritySurvey2018TidyNameSize = dataCyberSecuritySurvey2018TidyNameSize %>%
  select(-(outsource:attitude4))
```

```
##### 2019
```

```
#####
```

```
dataCyberSecuritySurvey2019TidyNameSize = dataCyberSecuritySurvey2019TidyNameSize %>%
  select(-(outsource:attitude4))
```

```
## since we want to have access to some proper data we will tidy the
## questions about how big of a priority is cyber security and how
## often are the higher ups updated about it this could really use some
## try catches because for the some weird reason -97 == NA does not
## return true or false, this is like javascript levels of bad
```

```
## also excepting this very first one the removals will be at the end
## so they are computed as if they were a transaction because try and
## catch is a lie to sell more lines of codes
```

```
## thanks to a blessing for our lord not finding the object only gives
## a warning and we ignore those as long as it still lets run the rest
## of the code
```

```
rm(dataCyberSecuritySurvey2018TidyName)
rm(dataCyberSecuritySurvey2019TidyName)
rm(dataCyberSecuritySurvey2020TidyName)
rm(dataCyberSecuritySurvey2021TidyName)
rm(dataCyberSecuritySurvey2022TidyName)
```

```
##### 2018
```

```
#####
```

```
numberOfCycles = length(dataCyberSecuritySurvey2018TidyNameSize$priority)
```

```
dataCyberSecuritySurvey2018TidyNameSizeCyber = dataCyberSecuritySurvey2018TidyNameSize
```

```

## apparently we have to be careful because an already inserted NA on
## the variable breaks the
for (i in 1:numberOfCycles) {
  if (dataCyberSecuritySurvey2018TidyNameSizeCyber$priority[i] == -97) {
    dataCyberSecuritySurvey2018TidyNameSizeCyber$priority[i] = NA
  }
  if (dataCyberSecuritySurvey2018TidyNameSizeCyber$update[i] == -97) {
    dataCyberSecuritySurvey2018TidyNameSizeCyber$update[i] = NA
  }
}

##### 2019
#####

numberOfCycles = length(dataCyberSecuritySurvey2019TidyNameSize$priority)

dataCyberSecuritySurvey2019TidyNameSizeCyber = dataCyberSecuritySurvey2019TidyNameSize

## apparently we have to be careful because an already inserted NA on
## the variable breaks the
for (i in 1:numberOfCycles) {
  if (dataCyberSecuritySurvey2019TidyNameSizeCyber$priority[i] == -97) {
    dataCyberSecuritySurvey2019TidyNameSizeCyber$priority[i] = NA
  }
  if (dataCyberSecuritySurvey2019TidyNameSizeCyber$update[i] == -97) {
    dataCyberSecuritySurvey2019TidyNameSizeCyber$update[i] = NA
  }
}

##### 2020
#####

numberOfCycles = length(dataCyberSecuritySurvey2020TidyNameSize$priority)

dataCyberSecuritySurvey2020TidyNameSizeCyber = dataCyberSecuritySurvey2020TidyNameSize

## apparently we have to be careful because an already inserted NA on
## the variable breaks the
for (i in 1:numberOfCycles) {
  if (dataCyberSecuritySurvey2020TidyNameSizeCyber$priority[i] == -97) {
    dataCyberSecuritySurvey2020TidyNameSizeCyber$priority[i] = NA
  }
  if (dataCyberSecuritySurvey2020TidyNameSizeCyber$update[i] == -97) {

```

```

        dataCyberSecuritySurvey2020TidyNameSizeCyber$update[i] = NA
    }
}

##### 2021
##### 

numberOfCycles = length(dataCyberSecuritySurvey2021TidyNameSize$priority)

dataCyberSecuritySurvey2021TidyNameSizeCyber = dataCyberSecuritySurvey2021TidyNameSize

## apparently we have to be careful because an already inserted NA on
## the variable breaks the
for (i in 1:numberOfCycles) {
    if (dataCyberSecuritySurvey2021TidyNameSizeCyber$priority[i] == -97) {
        dataCyberSecuritySurvey2021TidyNameSizeCyber$priority[i] = NA
    }
    if (dataCyberSecuritySurvey2021TidyNameSizeCyber$update[i] == -97) {
        dataCyberSecuritySurvey2021TidyNameSizeCyber$update[i] = NA
    }
}

##### 2022
##### 

numberOfCycles = length(dataCyberSecuritySurvey2022TidyNameSize$priority)

dataCyberSecuritySurvey2022TidyNameSizeCyber = dataCyberSecuritySurvey2022TidyNameSize

## apparently we have to be careful because an already inserted NA on
## the variable breaks the
for (i in 1:numberOfCycles) {
    if (dataCyberSecuritySurvey2022TidyNameSizeCyber$priority[i] == -97) {
        dataCyberSecuritySurvey2022TidyNameSizeCyber$priority[i] = NA
    }
    if (dataCyberSecuritySurvey2022TidyNameSizeCyber$update[i] == -97) {
        dataCyberSecuritySurvey2022TidyNameSizeCyber$update[i] = NA
    }
}

## garbage man? Well, of course I know him. He is me.

rm(dataCyberSecuritySurvey2018TidyNameSize)
rm(dataCyberSecuritySurvey2019TidyNameSize)
rm(dataCyberSecuritySurvey2020TidyNameSize)

```

```
rm(dataCyberSecuritySurvey2021TidyNameSize)
rm(dataCyberSecuritySurvey2022TidyNameSize)

## questions about reason of investment in cybersecuirty were removed
## form the pre-pilot survey in 2020

##### 2018
##### 

dataCyberSecuritySurvey2018TidyNameSizeCyber = dataCyberSecuritySurvey2018TidyNameSizeCyber %>
  select(-(reason1:reason27))

##### 2019
##### 

dataCyberSecuritySurvey2019TidyNameSizeCyber = dataCyberSecuritySurvey2019TidyNameSizeCyber %>
  select(-(reason1:reason28))

## the rest were already deleted
```

```
1 + 1
```

```
[1] 2
```

```
1 + 1
```

```
[1] 2
```

```
1 + 1
```

```
[1] 2
```

```
1 + 1
```

```
[1] 2
```

```
1 + 1
```

```
[1] 2
```

```
1 + 1
```

```
[1] 2
```

$1 + 1$

[1] 2

$1 + 1$

[1] 2

$1 + 1$

[1] 2

$1 + 1$

[1] 2

$1 + 1$

[1] 2

$1 + 1$

[1] 2

$1 + 1$

[1] 2

$1 + 1$

[1] 2

$1 + 1$

[1] 2

$1 + 1$

[1] 2

$1 + 1$

[1] 2

$1 + 1$

[1] 2

$1 + 1$

[1] 2

$1 + 1$

[1] 2

$1 + 1$

[1] 2

$1 + 1$

[1] 2

$1 + 1$

[1] 2

$1 + 1$

[1] 2

$1 + 1$

[1] 2

$1 + 1$

[1] 2

$1 + 1$

[1] 2

$1 + 1$

[1] 2

$1 + 1$

[1] 2

$1 + 1$

[1] 2

$1 + 1$

[1] 2

$1 + 1$

[1] 2

$1 + 1$

[1] 2

$1 + 1$

[1] 2

$1 + 1$

[1] 2

2 Quarto

Quarto enables you to weave together content and executable code into a finished document. To learn more about Quarto see <https://quarto.org>.

3 Running Code

When you click the **Render** button a document will be generated that includes both content and the output of embedded code. You can embed code like this:

```
1 + 1
```

```
[1] 2
```

You can add options to executable code like this

```
[1] 4
```

The `echo: false` option disables the printing of code (only output is displayed).