Electronic Content Usage & the COVID-19 Pandemic

Contents

This exploratory data analysis is divided into the following five parts.

1. Introduction: Basic background of this research
2. Sanity Check and Data Cleaning: Detailed decisions made during data cleaning
3. Preliminary insights
4. Conclusion
5. Next steps for further investigations.

Introduction

Library assessment is a set of processes undertaken by library staff to measure whether resources and services have met library users' expectations and identify how a library might need to improve. Assessment librarians will use their approaches to ensure the library's collections, services and spaces reach their goals, meet user needs, and continue to improve. One of the methods is conducting data analysis by using Counting Online Usage of Networked Electronic Resources (COUNTER) usage report. The reports follow the industry standard for recording and reporting electronic resources use and governs formatting. These ensures that usage from different vendors/platforms can be compared and shows what is being accessed and how much. Currently, many industries are experiencing the COVID-19 pandemic and some of them have to be closed due to safety issues, for example, libraries. In recent decades, scholarly materials have increasingly moved to an electronic format, and are accessible online to subscribers. So, our purpose is to investigate whether the closure of the libraries due to the COVID-19 pandemic result in changes to usage of electronic resources in specific subject area. We will use the COUNTER usage reports that show use of licensed content by UofT affiliated users and one subject classification dataset. In order to figure out the impact of COVID-19, we will use two periods of reports to compare whether there are changes, which includes Jan-Apr,2019 and Jan-Apr,2020.

Sanity Check and Data Cleaning

For combined dataset:

Overall, the combined dataset have 587742 observations and 33 variables, including vendor, collection, file, title, publisher, platform, isbn, print_issn, online_issn, yop, metric_type, reporting_period_total, jan.2019, feb.2019, mar.2019, apr.2019, jan.2020, feb.2020, mar.2020, apr.2020, publisher_id, uri, data_type, access_type, is_archive, journal_doi, proprietary_identifier, reporting_period_html, reporting_period_pdf, book_doi, and issn,.

For the variable metric_type, there are two levels, which are Total_item_Requests and Unique_item_Requests. Unique_item_Requests means investigations made during unique user sessions. However, it may be affected by users' devices, locations, browser and so on. Unique_item_Requests may not be accurate. As a result, we decide to only consider Total_item_Requests.

After that, we begin to deal with missing value. First at all, we use sapply function to get the ratio of the number of missing values to total number of the total data. Since the variables publisher_id, uri,

data_type, access_type, is_archive, journal_doi, proprietary_identifier, reporting_period_html, reporting_period_pdf, book_doi, issn miss more than 80% data, so we decide to remove these variables. It is because we will not use these variables for analysis, as there are too many missing values may resulting in inaccurate outcomes. The rest of variables are stored in total dataset, with 251822 observations and 19 variables.

Then, due to the fact that there are two periods of time, 2019 and 2020, so we use str_detect function to select the data separately by 2019 and 2020. As a result, there are 123191 observation and 16 variables in the data of 2019, in addition, 128631 observation and 16 variables in the data of 2020. We also imply unique function to see whether the separation is complete and accurate. We realize that missing value is implicit, so we use fct_explicit_na to give missing values an explicit factor level. Furthermore, using group_by function, there are 5 major vendors for both 2019 and 2020, but there are 232 different publishers with one missing in 2019 and 238 publishers in 2020. However, fewer publishers listed in the report don't mean few publishers in the library system, which indicates that users haven't used these materials in 2020.

	x
vendor	0.00
collection	0.00
file	0.00
title	0.00
publisher	0.00
publisher_id	0.83
platform	0.00
doi	0.12
proprietary_id	0.43
print_issn	0.63
online_issn	0.69
uri	0.89
metric_type	0.00
reporting_period_total	0.00
jan.2019	0.52
feb.2019	0.52
mar.2019	0.52
apr.2019	0.52
jan.2020	0.48
feb.2020	0.48
mar.2020	0.48
apr.2020	0.48
isbn	0.15
yop	0.18
data_type	0.96
access_type	0.96
is_archive	0.96
journal_doi	1.00
proprietary_identifier	1.00
reporting_period_html	1.00
reporting_period_pdf	1.00
book_doi	1.00
issn	1.00

For elsevier_jnlsubject dataset:

Our research question is investigating the changes in specific subject areas. However, for the time being, we only have the subject data for the vendor Elsevier. There are 10612 observations and 9 variables in total.

Top Level, Primary Level and Secondary Level are regard as main subject level, primary subject level, and secondary subject level. So, we determine to investigate the data we have for the subject classification.

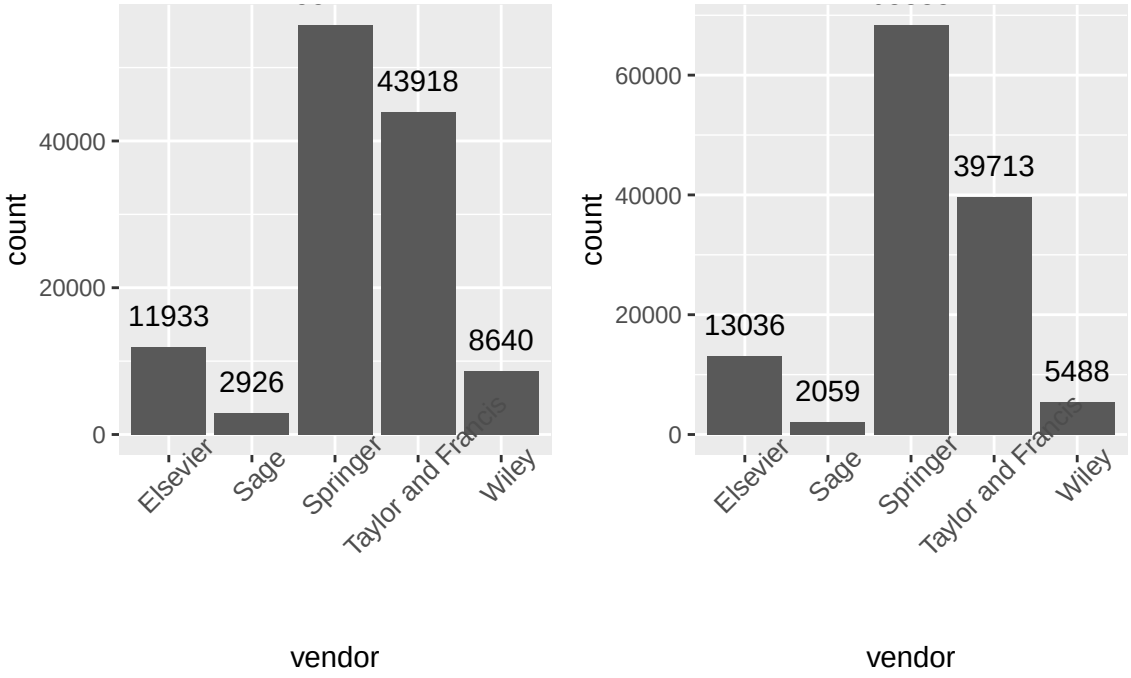
First thing I do for the data is to convert character to factor, which is necessary for further analysis. Then, to deal with missing values, we use sapply function to get the ratio of the number of missing values to total number of the total data. Since the variables Top Level, Primary Level, Secondary Level, Full Title, Unformatted ISSN, Product ID, Status don't exist the missing values, so we will remain all the variables for this time.

	x
Top Level	0.00
Primary Level	0.00
Secondary Level	0.00
Full Title	0.00
ISSN	0.00
Unformatted ISSN	0.00
Product ID	0.00
Status	0.00
Change History	0.56

Preliminary insights

Firstly, we would like to see whether the changes occurs within the vendors. So, we used group_by function to summarize the number of each vendors occurs in data2019 and data2020. In order to make it visualized, we also plot a bar plot, which can clearly see the changes. Two of them are increasing, and three of them are decreasing, based on that, we cannot make conclusion.

vendor	n19	n20
Elsevier	11933	13036
Sage	2926	2059
Springer	55774	68335
Taylor and Francis	43918	39713
Wiley	8640	5488

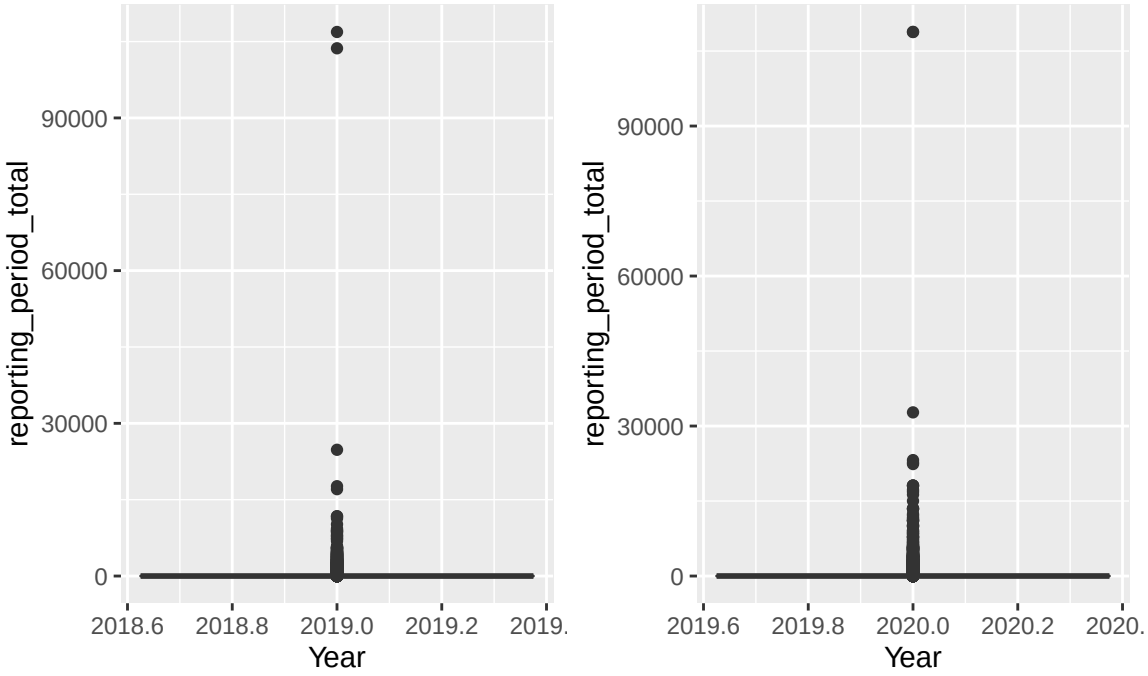


Then, we try to discover the variables publishers, and apply group_by function to summarize the number of each vendors occurs. However, there are 232 publishers, some of them are increasing, and some of them are decreasing. We cannot make conclusion as well.

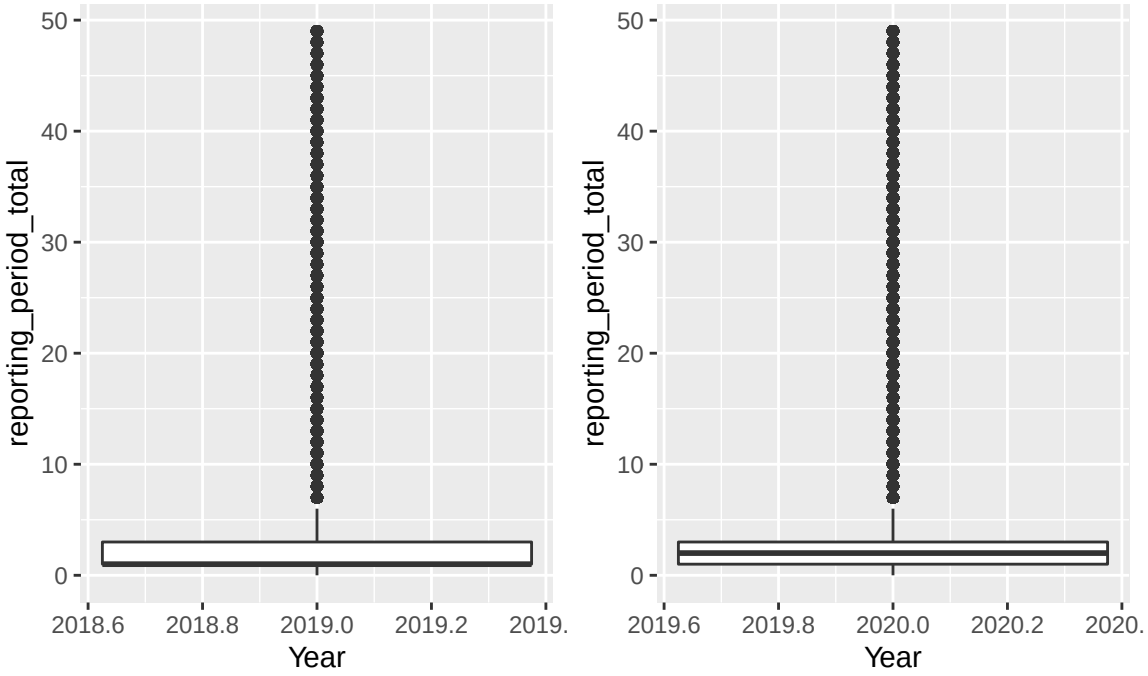
```
## # A tibble: 232 x 3
##    publisher                           n19   n20
##    <chr>                             <int> <int>
##  1 Routledge                         38136 35689
##  2 Springer International Publishing 17727 25733
##  3 Elsevier                          11234 12166
##  4 Springer Berlin Heidelberg         8670  9651
##  5 Wiley                              8638  5486
##  6 Springer New York                  5695  6027
##  7 Springer Netherlands               4202  4188
##  8 Palgrave Macmillan UK              2990  3537
##  9 Springer US                        2863  3027
## 10 SAGE Publications                  2842  2001
## # ... with 222 more rows
```

Next, we want to see the pattern of reporting_period_total both Jan-Apr in 2019 and 2020.

Boxplots of total reporting_period_total by Year



Boxplots of total reporting_period_total <50 by Year

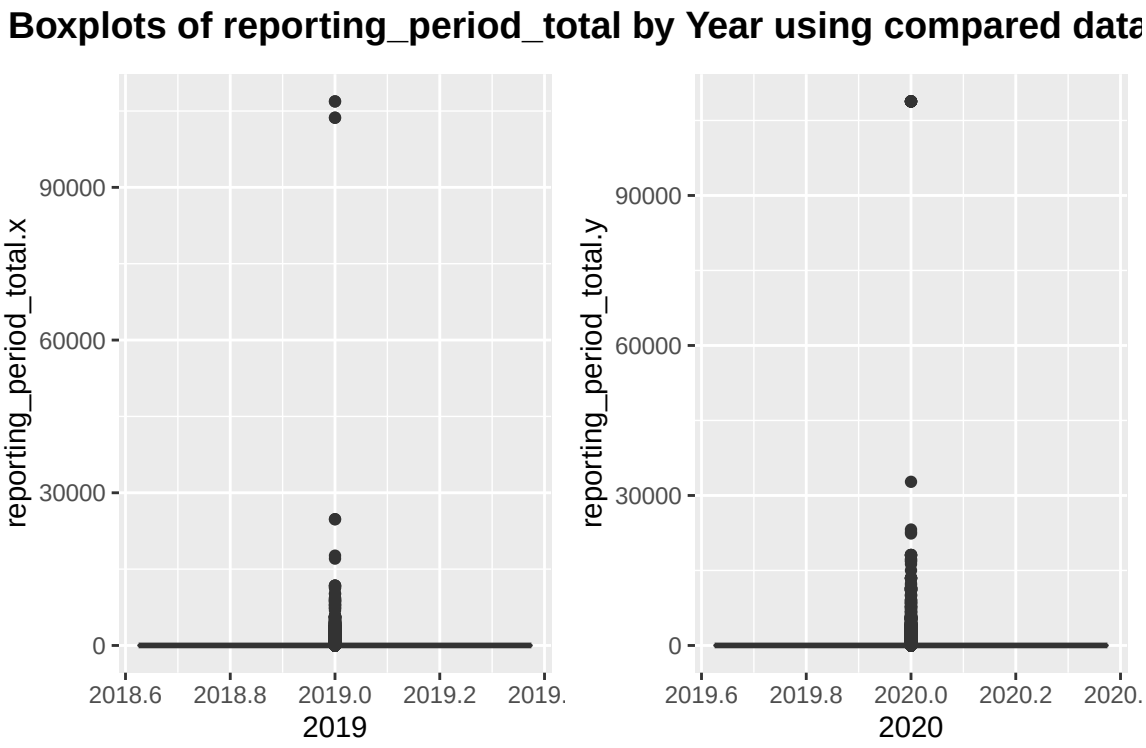


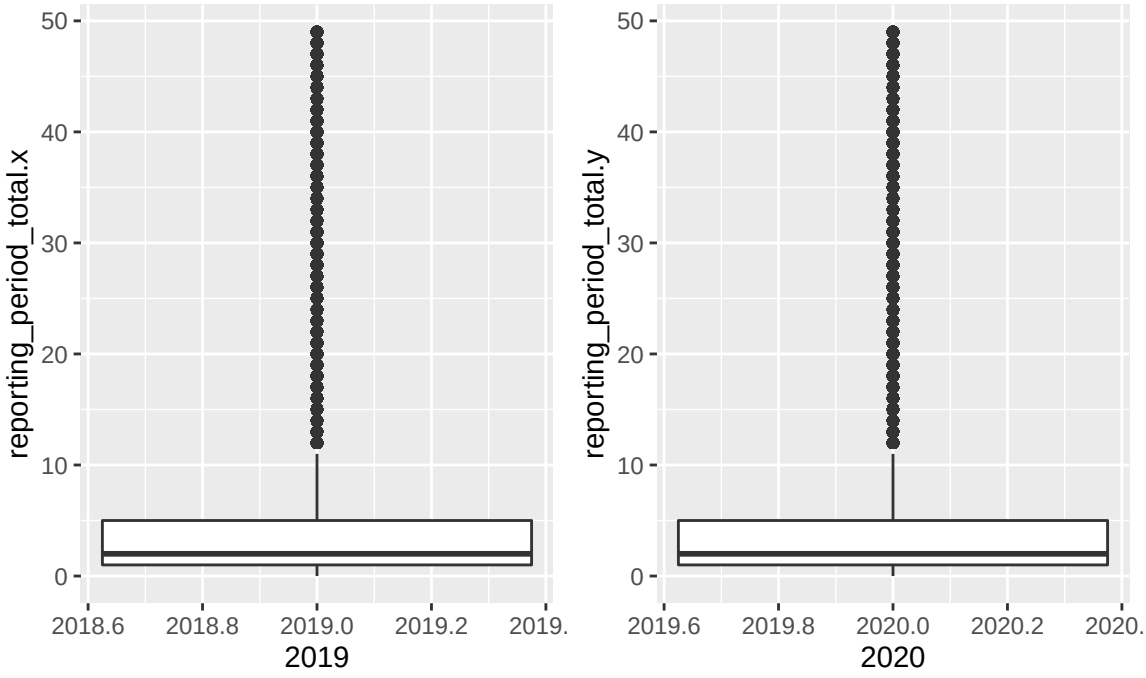
min	mean	median	max
0	29.19030	1	106891
0	34.61605	2	108815

According to the box plots of reporting_period_total, there are many extremely large outliers affecting

the pattern of box plot so that we are not able to discover. Then, we try filtering reporting_period_total less than 50 to avoid larger outliers, but still there are many outliers as well for both years. From the box plots, we cannot discover any big differences, but a extremely right-skewed data. After conducting description statistics table, mean for 2020 29.19030 is larger than 34.61605 in 2019. The minimum, median and maximum values are quite similar. This appears that there may be an increase of online usages of journals and eBooks due to COVID-19.

Furthermore, as we mention in Sanity Check and Data Cleaning part, we also want to compare the data that being requested for both two periods of time. So, we repeat the same box plots.





min	mean	median	max
0	36.64443	2	106891
0	49.67055	2	108815

We still are able to conclude any differences by box plots, but for description statistics table, for 2020, mean 49.67055 is clearly larger than mean 36.64443 in 2019. The minimum, median and maximum values are also quite similar.

Then, investigation of vendors are also needed to compare with the previous part.

vendor	Year.x	min	mean	median	max	Year.y
Elsevier	2019	0	138.675227	15	24801	NA
Elsevier	NA	0	169.375023	11	32733	2020
Sage	2019	0	95.088889	22	7716	NA
Sage	NA	1	146.317037	19	7927	2020
Springer	2019	1	28.070439	2	106891	NA
Springer	NA	0	38.235599	2	108815	2020
Taylor and Francis	2019	0	8.312672	1	2212	NA
Taylor and Francis	NA	1	22.683770	1	108815	2020
Wiley	2019	0	147.087885	22	17066	NA
Wiley	NA	1	153.595313	32	18080	2020

By vendor, means of 2020 reporting_period_total for each 5 major vendors are all higher than 2019. This means that 5 major vendors may experience increases in usage during the same period in two different years. In order to proof this, we need further investigation.

```
## # A tibble: 380 x 7
```

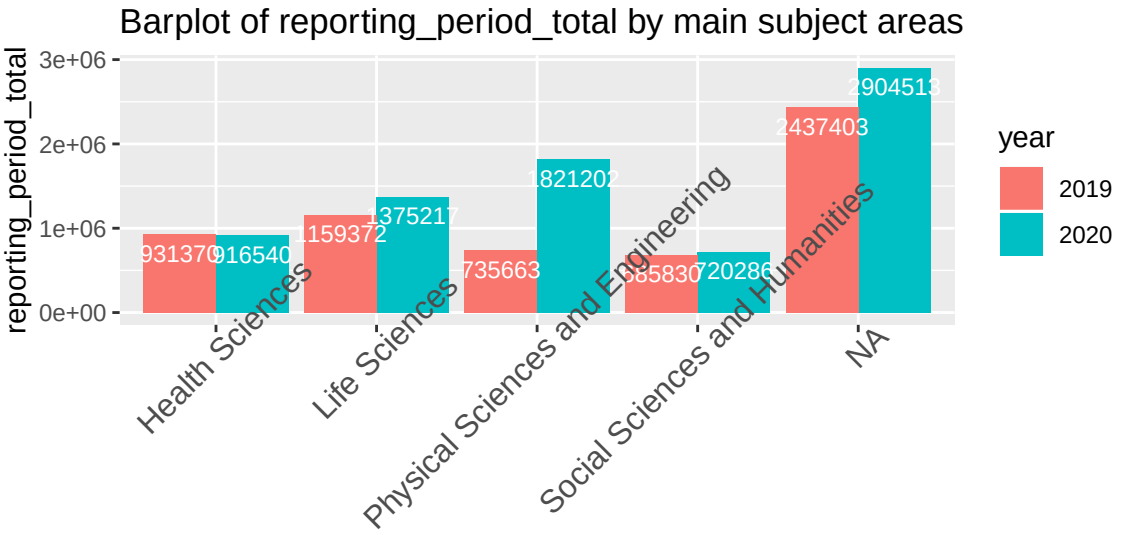
```
## # Groups:   publisher [190]
##    publisher                     Year.x   min   mean median   max Year.y
##    <fct>                          <dbl> <int>  <dbl>  <dbl> <int>  <dbl>
##  1 3D Research Center              2019     4  8.5      8.5    13     NA
##  2 3D Research Center                NA     6  6        6       6   2020
##  3 A K Peters/CRC Press            2019     0  0.211    0       1     NA
##  4 A K Peters/CRC Press              NA     1  4.89     1      35   2020
##  5 Academi                         2019     1  1.67     1       3     NA
##  6 Academi                           NA     1  1.5      1.5     2   2020
##  7 Academic Press                  2019     1  4.58     3      38     NA
##  8 Academic Press                    NA     1 14.8      2     132   2020
##  9 Academic Publication Council    2019     6  9.5      9.5    13     NA
## 10 Academic Publication Council      NA    10 10       10      10   2020
## # ... with 370 more rows
```

```
## [1] 109
```

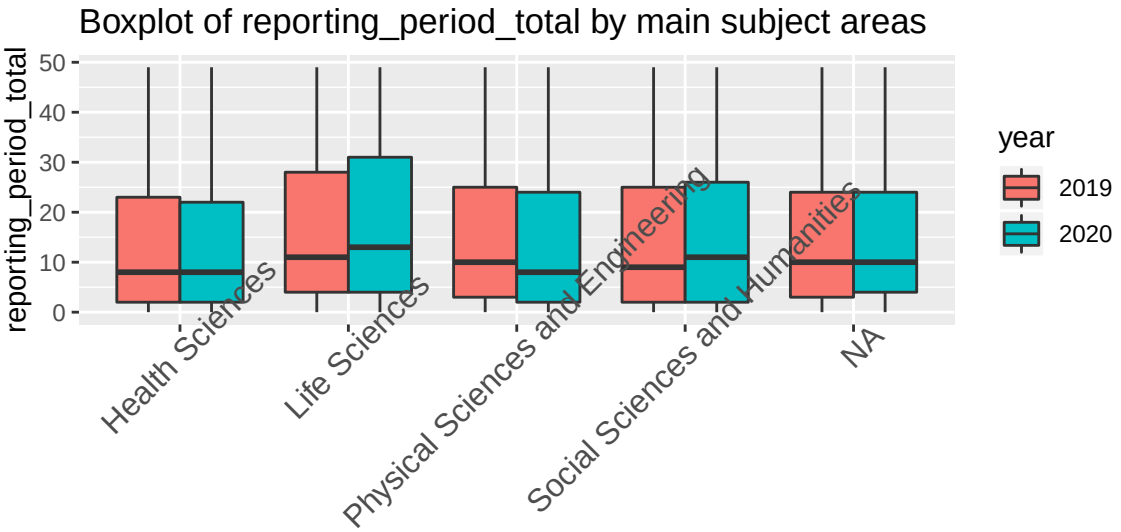
After all the discoveries, we are going to investigate more deeply into subject areas part.

```
##                            Top Level min     mean median    max
## 1                    Health Sciences   0 254.3337     70  17649
## 2                      Life Sciences   0 397.9993    108  24801
## 3 Physical Sciences and Engineering   0 259.3102     83   5702
## 4    Social Sciences and Humanities   0 435.1713    183   5702
## 5                               <NA>   0 209.2730     30 420931
##                            Top Level min     mean median    max
## 1                    Health Sciences   0 251.5203     57  22687
## 2                      Life Sciences   0 475.5246     91  23244
## 3 Physical Sciences and Engineering   0 639.4670     88  32860
## 4    Social Sciences and Humanities   0 457.9059    173  13728
## 5                               <NA>   0 252.4785     40 455093
```

Based on the summary table of reporting_period_total by Top_Level, despite slightly decreasing in Health Science areas, the other 3 subjects and missing values all experienced increases on mean of reporting_period_total. Especially, Physical Science and Engineering has increased more than twice in 2019. This can also be told by the following bar plot: reporting_period_total by main subject areas.



Main Subject



Top Level

Then, we also plot a box plot of reporting_period_total which is less than by main subject areas. There is a slight difference: Physical Science and Engineering has a higher mean when removing some big outliers.

Because of short and tight time, investigations have not been done. More will be provided.

Conclusion

Based on the current data we have and investigations we have done, we may conclude that the closure of the libraries due to the COVID-19 pandemic results in changes to usage of electronic resources in specific subject area. However, if removing some large outliers, the conclusion may not be worked. Also, because of only one subject classification from Elsevier, it is not strong enough to state the conclusion. In order to verify my conclusion, more investigation will be needed.

Next steps

1. Investigations of the relationship between reporting_period_total and other variables. e.g. more reporting_period_total may be due to the variable vendor. Bigger vendor will result in higher views.
2. Dependency in the data Participants who work at health science area are more likely access the journals and eBooks from health science area. e.g. GLMM model
3. Investigation on primary subject, secondary level.

Potential Challenges:

1. Any titles that show up in 2019 but not in 2020 are those that were available for access but 0 views in 2020. Similarly, assume the same for those that show up in 2020 but not in 2019, except those published in 2019 and 2020. More cleanings are needed.
2. Due to a large number of missing values, some important messages may be missed.
3. Due to the fact that we only have the vendor Elsevier's subject classification, the conclusions are not persuasive.

Next, we will begin to investigate the relationship between reporting_period_total and other variables. First, doing more data cleaning to solve the first potential challenge listed above. Then, trying to plot more figures to see the relationship with more variables. Also, investigating more deeply on subject, such as primary level and secondary level. If there are no more new data provided, manually subject assigned is necessary for future analysis.