Wellcome Trust Data Cleaning

EA

April 18, 2018

Contents

The purpose of this script is to clean the Wellcome Trust dataset and answer the associated Challenge questions from Lesson 4. I have added a section looking at the range of the data, removing an outlier, and looking at missing data. This is outside the scope of Lesson 4 but illustrates the data cleaning process.

Inspect your dataset. Are the data types what you expect?

Read in the Wellcome Trust APC (article processing charge) dataset.

Take a look at the entire dataset before removing 'Article title'. The backticks around Article title allow for selecting a column name with a space present - we will get rid of these spaces shortly). I am removing the Article title column because a) we don't need it to answer our questions and b) there are some Greek letters make it so we would have to install another package to make our document render to pdf.

```
#dat <- read_xlsx("data/University returns_for_figshare_FINAL.xlsx")
dat <- read_xlsx("data/University returns_for_figshare_FINAL.xlsx") %>%
  select(-`Article title`)

glimpse(dat)
```

```
## Observations: 2,127
## Variables: 4
## $ `PMID/PMCID`                                          <chr> "PMC3378...
## $ Publisher                                             <chr> "Elsevie...
## $ `Journal title`                                       <chr> "Academy...
## $ `COST (£) charged to Wellcome (inc VAT when charged)` <dbl> 2379.54,...
```

I expect the publisher, journal and article titles to be characters, and cost to be numeric. The PMID/PMCID might be numeric or character if the different ids were split into separate columns and the prefix id types were removed.

Identify any immediate problems.

Question 1: List 3 problems with this dataset that require data cleaning.

```
head(dat)
```

```
## # A tibble: 6 x 4
##   `PMID/PMCID`                Publisher   `Journal title`  `COST (£) charge~
##   <chr>                       <chr>       <chr>                       <dbl>
## 1 "PMC3378987\r\n"            Elsevier    Academy of Nutr~             2380.
## 2 PMCID: PMC3780468           ACS (Amerc~ ACS Chemical Bi~             1295.
## 3 PMCID: PMC3621575           ACS (Amerc~ ACS Chemical Bi~             1295.
## 4 <NA>                        American C~ ACS Chemical Bi~              947.
## 5 PMID: 24015914 PMC3833349 American C~ ACS Chemical Bi~             1268.
## 6 : PMC3805332                American C~ ACS Chemical Bi~             2287.
```

1. There is no structured vocabulary and as a consequence many publisher and Journal titles have multiple entries (abbreviations vs full names, spelling errors) instead of one journal or publisher.
2. There are multiple variables in one column for different ids (PMID, PMCID) where each should have a separate column.
3. Horrible column names ie.'COST (£) charged to Wellcome (inc VAT when charged).
4. There are newline characters in our cells.

Clean up column names.

First of all let's change the column names to not have spaces and weird characters in them. I am also going to convert them to lowercase. I am not too concerned about including the units for cost as it is referenced in the README for the dataset.

```
colnames(dat) <- gsub("/", "_", colnames(dat))
colnames(dat) <- gsub(" .*", "", colnames(dat))
colnames(dat) <-tolower(colnames(dat))
```

Investigate the range of your data.

Let's find the least expensive publication.

```
dat %>% arrange(cost) %>% head()
```

```
## # A tibble: 6 x 4
##   pmid_pmcid  publisher                     journal                   cost
##   <chr>       <chr>                         <chr>                    <dbl>
## 1 <NA>        American Society for Nutrition American Society for N~  45.9
## 2 3543450     Public Library of Science     PLoS One                 122.
## 3 <NA>        Sciedu Press                  Journal of Biomedical ~  135.
## 4 PMC 3536734 Landes Bioscience             Channels                 160.
## 5 <NA>        JSciMed Central               Journal of Neurology &~  160.
## 6 <NA>        Sciedu Press                  "International Journal~  187.
```

There is a very low value (45.9) and it is possible other funders contributed to this publication (as mentioned in the README). Let's see if there are any other publications by this publisher to tell us if this is reasonable for the publisher.

```
dat %>% arrange(cost) %>% filter(publisher == "American Society for Nutrition")
```

```
## # A tibble: 1 x 4
##   pmid_pmcid publisher                       journal                   cost
##   <chr>      <chr>                           <chr>                    <dbl>
## 1 <NA>       American Society for Nutrition American Society for Nu~  45.9
```

Since there is only one entry from this publisher, let's check its value in relation to the mean and standard deviation of the dataset as a whole.

```
dat %>% summarize(sd = sd(cost), mean = mean(cost))
```

```
## # A tibble: 1 x 2
##      sd  mean
##   <dbl> <dbl>
## 1  807. 1826.
```

The mean less the lowest cost divided by the standard deviation tells us that the point is 2.2 standard deviations from the mean.

```
(1826-45.9)/807
```

```
## [1] 2.205824
```

Since there is only one entry from this publisher, and its cost is less than 3 standard deviations from the mean, we will keep it in our dataset.

Let's find the most expensive publication.

```
dat %>% arrange(desc(cost)) %>% head()
```

```
## # A tibble: 6 x 4
##   pmid_pmcid                          publisher            journal      cost
##   <chr>                               <chr>                <chr>       <dbl>
## 1 <NA>                                MacMillan            NA         13200.
## 2 543219                              public.service.co.uk Public Se~  6000.
## 3 PMID: 23041239 /PMCID: PMC3490334 Elsevier             The Lance~  5760.
## 4 23541370 PMC3744751                 Elsevier             Elsevier    4800.
## 5 <NA>                                Elsevier             Lancet      4800.
## 6 "PMCID:\r\n    PMC3627205\r\n"     Elsevier             The Lancet  4800.
```

It looks like the most expensive publication is not a journal, but a book. It also appears to be more than twice as expensive as our next most expensive cost. The book is obviously an outlier as a value (greater than 3 standard deviations from the mean) and for a fair comparison of cost (books vs journal articles). We will remove it from the dataset.

Do you have missing data?

While we are removing the book from our dataset, we should check if there are any other cases where journal is *NA*? We can look for 'incomplete' entries in the journal column. It is probably a good idea to look for incomplete entries in the entire dataset as well.

```
dat[!complete.cases(dat$journal),]
```

```
## # A tibble: 0 x 4
## # ... with 4 variables: pmid_pmcid <chr>, publisher <chr>, journal <chr>,
## #   cost <dbl>
```

```
dat[!complete.cases(dat),]
```

```
## # A tibble: 199 x 4
##    pmid_pmcid publisher                          journal                cost
##    <chr>      <chr>                              <chr>                 <dbl>
##  1 <NA>       American Chemical Society          ACS Chemical Biology   947.
##  2 <NA>       Springer                           Acta Neuropathologica 1884.
##  3 <NA>       Springer                           Advances in Experime~ 1928.
##  4 <NA>       Springer                           Advances in Experime~ 1928.
##  5 <NA>       Springer                           Advances in Experime~ 1928.
##  6 <NA>       Springer                           Advances in Experime~ 1928.
##  7 <NA>       Cambridge University Press         Ageing & Society      1695.
##  8 <NA>       Wiley                              American Ethnologist  1870.
##  9 <NA>       Wiley                              American Ethnologist  1871.
## 10 <NA>       American Psychiatric Association American Journal of ~ 2352.
## # ... with 189 more rows
```

What is going on? We know that there is a case where journal is *NA* and there are values in the id column. It appears that in data collection "*NA*" was entered as a character string instead of simply being an empty entry like with the identifiers. This was likely due to a rule that 'something' had to be entered into the journal field. We can then remove it from our dataset using the `filter` function.

```
dat <- dat %>% filter(journal != "NA")
```

We are not concerned with PMIDs as for today as we do not need this column to answer any of our questions.

Data-Cleaning

Okay! Let's do some data cleaning! For this dataset we are ONLY going to clean the **publisher** column. We will answer questions about the journal column using `grep` to grab the information we want without cleaning the entire column.

This will illustrate how easy our questions will be to answer once we have the data in the appropriate format, and how the majority of our time is spent data cleaning.

First let's look at the publishers we have. We can see how many different publishers there are by looking at the number of factor levels. You could instead look to see which values of publisher are unique. Grouping by factor levels is convenient in this case to see the names alphabetically.

```
levels(as.factor(dat$publisher))
```

```
##   [1] "ACS"
##   [2] "ACS (Amercian Chemical Society) Publications"
##   [3] "ACS Publications"
##   [4] "AGA Institute"
##   [5] "AMBSB"
##   [6] "American Association of Immunologists"
##   [7] "American Chemical Society"
##   [8] "AMERICAN CHEMICAL SOCIETY"
##   [9] "American Chemical Society Publications"
##  [10] "American College of Chest Physicians"
##  [11] "American Physiological Society"
##  [12] "American Psychiatric Association"
##  [13] "American Psychiatric Publishing"
##  [14] "American Psychological Association"
##  [15] "American Public Health Association"
```

```
##  [16] "American Soc for Biochemistry and Molecular Biology"
##  [17] "American Society for Biochemistry and Molecular Biolgy"
##  [18] "American Society for Biochemistry and Molecular Biology"
##  [19] "American Society for Investigative Pathology"
##  [20] "American Society for Microbiology"
##  [21] "American Society for Microbiology \r\n"
##  [22] "American Society for Nutrition"
##  [23] "American Society of Haematology"
##  [24] "American Society of Hamatology"
##  [25] "American Society of Hematology"
##  [26] "American Society of Human Genetics (Elsevier)"
##  [27] "American Society of Microbiology"
##  [28] "American Speech-Language-Hearing Association"
##  [29] "ASBMB"
##  [30] "ASBMB Cadmus"
##  [31] "ASBMB/Cadmus"
##  [32] "ASBMB/Cenveo Publisher Services"
##  [33] "ASBMC /CENVEO"
##  [34] "ASM"
##  [35] "ASM (American Society for Microbiology)"
##  [36] "Association for Research in Vision & Ophthalmology"
##  [37] "Bentham Science Publishers"
##  [38] "Benthan Science Publishers"
##  [39] "Berhahn Books"
##  [40] "Biochem Journal"
##  [41] "Biomed Central"
##  [42] "BioMed central"
##  [43] "BioMed Central"
##  [44] "BioMed Central Limited"
##  [45] "BioMed Central Ltd"
##  [46] "Biophysical Society"
##  [47] "Bioscientifica"
##  [48] "BioScientifica"
##  [49] "Blackwell Publishing Ltd/Wiley"
##  [50] "BMC"
##  [51] "BMJ"
##  [52] "BMJ group"
##  [53] "BMJ Group"
##  [54] "BMJ Journals"
##  [55] "BMJ Publishing Group"
##  [56] "BMJ PUBLISHING GROUP"
##  [57] "BMJ Publishing Group Ltd"
##  [58] "BMJ Publishing Group Ltd & British Thoracic Society"
##  [59] "Brill"
##  [60] "British Medical Journal"
##  [61] "Byophysical Society"
##  [62] "Cadmus"
##  [63] "CADMUS JOURNAL SERVICE"
##  [64] "Cadmus Journal Services"
##  [65] "CADMUS JOURNAL SERVICES"
##  [66] "Cambridge Journals"
##  [67] "Cambridge Uni Press"
##  [68] "Cambridge Univ Press"
##  [69] "Cambridge University Press"
```

```
##  [70] "Camdus Journal Services"
##  [71] "Cell Press"
##  [72] "Cenveo Publisher services"
##  [73] "Cenveo Publisher Services/ASM JV1"
##  [74] "COACTION"
##  [75] "Cold Spring Habour Press"
##  [76] "Cold Spring Harbor"
##  [77] "Cold Spring Harbor Laboratory Press"
##  [78] "Cold Spring Harbor Press"
##  [79] "Cold Spring Harbor Publications"
##  [80] "Company of Biologist"
##  [81] "Company of Biologists"
##  [82] "Company of Biologists Ltd"
##  [83] "Copyright Clearace Center"
##  [84] "CSHLP"
##  [85] "CUP"
##  [86] "Darmouth Journal Services"
##  [87] "Dartmouth Journal Services"
##  [88] "Dartmouth Journals"
##  [89] "Elseveier Science"
##  [90] "Elsevier"
##  [91] "ELSEVIER"
##  [92] "Elsevier (Cell Press)"
##  [93] "Elsevier / Cell Science"
##  [94] "Elsevier B.V."
##  [95] "Elsevier Ltd"
##  [96] "Elsevier/Cell Press"
##  [97] "Endocrine Society"
##  [98] "European Respiratory Society"
##  [99] "European Society of Endocrinolog"
## [100] "FASEB"
## [101] "Federation of American Societies for Experimental Biology"
## [102] "Federation of American Societies for Experimental Biology (FASEB)"
## [103] "Federation of the American Society of Experimental Biology"
## [104] "Ferrata Storti Foundation"
## [105] "Frontiers"
## [106] "Frontiers Media"
## [107] "Frontiers Media SA"
## [108] "Frontiers Research Foundation"
## [109] "Future Medicine"
## [110] "Future Medicine Ltd"
## [111] "Future Science"
## [112] "Hindawi"
## [113] "Hindawi Publishing Corporation"
## [114] "Humana Press (Springer Imprint)"
## [115] "Impact Journals"
## [116] "Impact Journals LLC"
## [117] "Informa Healthcare"
## [118] "Informa Healthcare communications"
## [119] "Institute of Physics"
## [120] "International AIDS Society"
## [121] "International Union Against tuberculosis and Lung Disease"
## [122] "International Union Against Tuberculosis and Lung Disease"
## [123] "International Union of Crystallography"
```

```
## [124] "International Union of Crystallography (iucr)"
## [125] "IOP Publishing"
## [126] "IOS Press"
## [127] "Ivyspring International Publisher"
## [128] "J Med Internet Research"
## [129] "John Wiley"
## [130] "John Wiley & Sons"
## [131] "JOHN WILEY & SONS"
## [132] "John Wiley & Sons Inc"
## [133] "John Wiley & Sons Ltd"
## [134] "John Wiley & Sons, Inc."
## [135] "John Wiley and Sons"
## [136] "John Wiley and Sons Ltd"
## [137] "Johns Hopkins University Press"
## [138] "Journal of the American Physiological Proceedings of National Academy of Sciences"
## [139] "Journal of Visualized Experiments"
## [140] "JoVE"
## [141] "JSciMed Central"
## [142] "Karger"
## [143] "KARGER"
## [144] "Landes Bioscience"
## [145] "Landes Biosciences"
## [146] "LWW"
## [147] "Mary Ann Liebert"
## [148] "MARY ANN LIEBERT INC"
## [149] "Mary Ann Liebert, Inc. Publishers"
## [150] "MDPI"
## [151] "MIT Press"
## [152] "MIT PRESS OPEN ACCESS"
## [153] "MY JOVE CORP"
## [154] "My JOVE corporation"
## [155] "MYJoVE Corporation"
## [156] "National Academy of Sciences"
## [157] "National Academy of Sciences of the United States of America"
## [158] "National Academy of Sciences USA"
## [159] "National Academy of Sciences, USA"
## [160] "Nature"
## [161] "Nature PG"
## [162] "Nature Publishing"
## [163] "Nature publishing group"
## [164] "Nature Publishing Group"
## [165] "NATURE PUBLISHING GROUP LTD"
## [166] "NPG"
## [167] "Open Access Reg Ltd"
## [168] "Optical Society of America"
## [169] "OUP"
## [170] "Oxford Journals"
## [171] "Oxford Journals (OUP)"
## [172] "Oxford Univ Press"
## [173] "Oxford University Press"
## [174] "OXFORD UNIVERSITY PRESS"
## [175] "Oxford University Press\r\n"
## [176] "Oxford University Press (OUP)"
## [177] "Oxford Univesity Press"
```

```
## [178] "Palgrave MacMillan"
## [179] "Pion"
## [180] "Plos"
## [181] "PLoS"
## [182] "PLOS"
## [183] "PLoS (Public Library of Science)"
## [184] "PLoS Public Library of Science"
## [185] "PNAS"
## [186] "PNAS Author Publication"
## [187] "Policy Press"
## [188] "Portland press"
## [189] "Portland Press"
## [190] "Portland Press Ltd"
## [191] "PORTLAND PRESS LTD"
## [192] "Proceedings of the National Academy of Sciences (PNAS)"
## [193] "Public Library of Science"
## [194] "public.service.co.uk"
## [195] "Publisher Society for Endocrinology"
## [196] "PubMed"
## [197] "PubMed Central"
## [198] "Research Media Ltd"
## [199] "Royal College of Psychiatrists"
## [200] "Royal Society"
## [201] "Royal Society for Chemistry"
## [202] "Royal Society of Chemistry"
## [203] "RSC"
## [204] "RSC Publishing"
## [205] "Sage"
## [206] "SAGE"
## [207] "Sage Publications"
## [208] "SAGE Publications"
## [209] "Sage Publications Inc"
## [210] "Sage Publications Ltd"
## [211] "Sage Publishers"
## [212] "Sage Publishing"
## [213] "Sciedu Press"
## [214] "Scientific Research Publishing"
## [215] "Sheridan Press"
## [216] "Society for Endocrinology"
## [217] "Society for General Microbiology"
## [218] "Society for Genermal Microbiology"
## [219] "Society for Leukocyte Biology"
## [220] "Society for Neuroscience"
## [221] "Society for Neurosciences"
## [222] "Society for Publication of Acta Dermato-Venereologica"
## [223] "Society of General Microbiology"
## [224] "Society of Leukocyte Biology"
## [225] "SOCIETY OF NEURO SCIENCES"
## [226] "Society of Neuroscience"
## [227] "Springer"
## [228] "Springer - Verlag GMBH"
## [229] "Springer Science + Business Media"
## [230] "Springer Verlag"
## [231] "Springer-Veriag GmbH"
```

```
## [232] "Springer-Verlag GmbH"
## [233] "SPRINGER-VERLAG GMBH"
## [234] "Springer-Verlag GMBH & Ci"
## [235] "Springer-Verlag GmbH, Heidelberger Platz 3, D-14197 Berlin"
## [236] "T&F"
## [237] "Taylor & Francis"
## [238] "Taylor & Francis Journals"
## [239] "Taylor and Francis"
## [240] "The American Chemical Society Petroleum Research Trust"
## [241] "The American Physiological Society"
## [242] "The American Society for Biochemistry and Molecular Biology"
## [243] "The American Society for Biochemistry and Molecular Biology, Inc"
## [244] "The American Society of Pediatrics"
## [245] "The Boulevard"
## [246] "The company of Biolgists"
## [247] "The company of Biologists"
## [248] "The Company of Biologists"
## [249] "THE COMPANY OF BIOLOGISTS"
## [250] "The Company of Biologists Ltd"
## [251] "The Endocrine Socety"
## [252] "The Endocrine Society"
## [253] "THE ENDOCRINE SOCIETY"
## [254] "The Endrocrine Society"
## [255] "The Journal of Visualized Experiments"
## [256] "The Royal College of Psychiatrists"
## [257] "The royal Society"
## [258] "The Royal Society"
## [259] "The Sheridan Press"
## [260] "Transcript Verlag"
## [261] "University of the Basque Country Press"
## [262] "Wiley"
## [263] "Wiley & Son"
## [264] "Wiley Blackwell"
## [265] "Wiley Online Library"
## [266] "Wiley Subscription Services"
## [267] "Wiley Subscription Services  Inc."
## [268] "Wiley Subscription Services Inc"
## [269] "Wiley Subscription Serviices Inc"
## [270] "Wiley VCH"
## [271] "Wiley-Blackwell"
## [272] "Wiley-Blackwell, John Wiley & Sons"
## [273] "Wiley-VCH"
## [274] "Wiley/Blackwell"
## [275] "Wliey-Blackwell"
## [276] "Wolters Kluwer"
## [277] "Wolters Kluwer Health"
## [278] "Wolters Kluwer N.V./Lippinott"
## [279] "Wolters Kluwers"
```

```
length(levels(as.factor(dat$publisher)))
```

```
## [1] 279
```

There appears to be 280 publishers, however we can see that many are duplicated through inconsistencies in data collection. For example I would guess that 'Sage', 'SAGE', 'Sage Publications', 'SAGE Publications',

'Sage Publications Inc', 'Sage Publications Ltd', 'Sage publishers' and 'Sage Publishing' are all the same publisher.

The first thing I am going to do is change everything to lowercase and look at how many matches were case-dependant.

```
dat$publisher <- tolower(dat$publisher)

length(levels(as.factor(dat$publisher)))
```

```
## [1] 255
```

There are now 255 publishers instead of 279.

We can also see some typos in here that can be corrected.

```
dat$publisher <- str_replace_all(dat$publisher, "darmouth.*", "dartmouth")
dat$publisher <- str_replace_all(dat$publisher, "wliey.*", "wiley")
dat$publisher <- str_replace_all(dat$publisher, "endrocrine", "endocrine")
dat$publisher <- str_replace_all(dat$publisher, "biolgists", "biologists")
dat$publisher <- str_replace_all(dat$publisher, "socety", "society")
dat$publisher <- str_replace_all(dat$publisher, "genermal", "general")
dat$publisher <- str_replace_all(dat$publisher, "neuro\\s?science[s]?", "neuroscience")
dat$publisher <- str_replace_all(dat$publisher, "elseveier.*", "elsevier")
dat$publisher <- str_replace_all(dat$publisher, "of neuroscience", "for neuroscience")
dat$publisher <- str_replace_all(dat$publisher, "habour", "harbor")
dat$publisher <- str_replace_all(dat$publisher, "landes biosciences", "landes bioscience")
dat$publisher <- str_replace_all(dat$publisher, "biolgy", "biology")
dat$publisher <- str_replace_all(dat$publisher, "soc\\b", "society")
dat$publisher <- str_replace_all(dat$publisher, "hamatology|haematology", "hematology")
dat$publisher <- str_replace_all(dat$publisher, "cadmus", "camdus")
dat$publisher <- str_replace_all(dat$publisher, "benthan", "bentham")
dat$publisher <- str_replace_all(dat$publisher, "berhahn books", "berghahn")
dat$publisher <- str_replace_all(dat$publisher, "byophysical", "biophysical")
dat$publisher <- str_replace_all(dat$publisher, "asbmc.*", "asbmb")
dat$publisher <- str_replace_all(dat$publisher, "endocrinolog", "endocrinology")
dat$publisher <- str_replace_all(dat$publisher, "clearace", "clearance")
```

Next, I am going to shorten several publishers to their 'base' name. For example, everything above to do with Sage will be under 'sage'. Everything to do with Wiley will be 'wiley'.

```
#anything before or after wiley will be removed
dat$publisher <- str_replace_all(dat$publisher, ".*wiley.*", "wiley")
dat$publisher <- str_replace_all(dat$publisher, ".*elsevier.*", "elsevier")
dat$publisher <- str_replace_all(dat$publisher, ".*pnas.*", "pnas")
dat$publisher <- str_replace_all(dat$publisher, ".*royal college of psychiatrists.*",
                                 "royal college of psychiatrists")
dat$publisher <- str_replace_all(dat$publisher, ".*jove.*", "jove")
dat$publisher <- str_replace_all(dat$publisher, ".*faseb.*", "faseb")
dat$publisher <- str_replace_all(dat$publisher, ".*?company of biologists.*",
                                 "the company of biologists")
#anything after the name will be removed
dat$publisher <- str_replace_all(dat$publisher, "sage.*", "sage")
dat$publisher <- str_replace_all(dat$publisher, "wolters kluwer.*", "wolters kluwer")
dat$publisher <- str_replace_all(dat$publisher, "springer.*", "springer")
dat$publisher <- str_replace_all(dat$publisher, "nature.*", "nature")
dat$publisher <- str_replace_all(dat$publisher, "bmj.*", "bmj")
```

```
dat$publisher <- str_replace_all(dat$publisher, "cold spring harbor.*", "cold spring harbor")
dat$publisher <- str_replace_all(dat$publisher, "acs.*", "acs")
dat$publisher <- str_replace_all(dat$publisher, "asbmb.*", "asbmb")
dat$publisher <- str_replace_all(dat$publisher, "dartmouth.*", "dartmouth")
dat$publisher <- str_replace_all(dat$publisher, "cambridge.*", "cambridge")
dat$publisher <- str_replace_all(dat$publisher, "oxford.*", "oxford")
dat$publisher <- str_replace_all(dat$publisher, "plos.*", "plos")
dat$publisher <- str_replace_all(dat$publisher, "rsc.*", "rsc")
dat$publisher <- str_replace_all(dat$publisher, "portland.*", "portland")
dat$publisher <- str_replace_all(dat$publisher, "&", "and")
dat$publisher <- str_replace_all(dat$publisher, "taylor and francis.*", "taylor and francis")
dat$publisher <- str_replace_all(dat$publisher, "bentham.*", "bentham")
dat$publisher <- str_replace_all(dat$publisher, "national academy of sciences.*",
                                 "national academy of sciences")
dat$publisher <- str_replace_all(dat$publisher, "camdus.*", "camdus")
dat$publisher <- str_replace_all(dat$publisher, "cenveo.*", "cenveo")
dat$publisher <- str_replace_all(dat$publisher, "mary ann liebert.*", "mary ann liebert")
dat$publisher <- str_replace_all(dat$publisher, "impact.*", "impact")
dat$publisher <- str_replace_all(dat$publisher, "frontiers.*", "frontiers")
dat$publisher <- str_replace_all(dat$publisher, "future medicine.*", "future science")
dat$publisher <- str_replace_all(dat$publisher, "hindawi.*", "hindawi")
dat$publisher <- str_replace_all(dat$publisher, "informa healthcare.*", "informa healthcare")
dat$publisher <- str_replace_all(dat$publisher, "mit press.*", "mit press")
dat$publisher <- str_replace_all(dat$publisher, "humana press.*", "humana press")
dat$publisher <- str_replace_all(dat$publisher, "pubmed.*", "pubmed")
```

There will also be some name corrections - for example npg is 'nature publishing group' and so should be under 'nature'.

```
dat$publisher <- str_replace_all(dat$publisher, "npg.*", "nature")
dat$publisher <- str_replace_all(dat$publisher, "oup.*", "oxford")
dat$publisher <- str_replace_all(dat$publisher, ".*royal society.*", "rsc")
dat$publisher <- str_replace_all(dat$publisher, "public library of science.*", "plos")
dat$publisher <- str_replace_all(dat$publisher,
                                 ".*american society for biochemistry and molecular biology.*",
                                 "asbmb")
dat$publisher <- str_replace_all(dat$publisher,
                                 "federation of american societies for experimental biology.*",
                                 "faseb")
dat$publisher <- str_replace_all(dat$publisher,
                                 "federation of american society of experimental biology.*", "faseb")
dat$publisher <- str_replace_all(dat$publisher, "american chemical society.*", "acs")
dat$publisher <- str_replace_all(dat$publisher, "tandf", "taylor and francis")
dat$publisher <- str_replace_all(dat$publisher, "international union of crystallography.*", "iucr")
dat$publisher <- str_replace_all(dat$publisher, ".*american society for microbiology.*", "asm")
dat$publisher <- str_replace_all(dat$publisher, "biomed central.*", "bmc")
dat$publisher <- str_replace_all(dat$publisher, "british medical journal", "bmj")
dat$publisher <- str_replace_all(dat$publisher, ".*american society of microbiology.*", "asm")
dat$publisher <- str_replace_all(dat$publisher, "of general microbiology", "for general microbiology")
dat$publisher <- str_replace_all(dat$publisher, "of leukocyte", "for leukocyte")
dat$publisher <- str_replace_all(dat$publisher, "publisher society", "society")
dat$publisher <- str_replace_all(dat$publisher, "company of biologist$", "company of biologists")
dat$publisher <- str_replace_all(dat$publisher, "american psychiatric publishing",
                                 "american psychiatric association")
```

```
dat$publisher <- str_replace_all(dat$publisher, "institute of physics", "iop publishing")
```

Remove newline characters.

```
dat$publisher <- str_remove_all(dat$publisher, "\r\n")
dat$publisher <- str_trim(dat$publisher, side = "both")
dat$publisher <- str_remove_all(dat$publisher, "the ")
```

Let's see what we have.

```
levels(as.factor(dat$publisher))
```

```
##  [1] "acs"
##  [2] "aga institute"
##  [3] "ambsb"
##  [4] "american association of immunologists"
##  [5] "american college of chest physicians"
##  [6] "american physiological society"
##  [7] "american psychiatric association"
##  [8] "american psychological association"
##  [9] "american public health association"
## [10] "american society for investigative pathology"
## [11] "american society for nutrition"
## [12] "american society of hematology"
## [13] "american society of pediatrics"
## [14] "american speech-language-hearing association"
## [15] "asbmb"
## [16] "asm"
## [17] "association for research in vision and ophthalmology"
## [18] "bentham"
## [19] "berghahn"
## [20] "biochem journal"
## [21] "biophysical society"
## [22] "bioscientifica"
## [23] "bmc"
## [24] "bmj"
## [25] "boulevard"
## [26] "brill"
## [27] "cambridge"
## [28] "camdus"
## [29] "cell press"
## [30] "cenveo"
## [31] "coaction"
## [32] "cold spring harbor"
## [33] "company of biologists"
## [34] "copyright clearance center"
## [35] "cshlp"
## [36] "cup"
## [37] "dartmouth"
## [38] "elsevier"
## [39] "endocrine society"
## [40] "european respiratory society"
## [41] "european society of endocrinology"
## [42] "faseb"
## [43] "federation of american society of experimental biology"
```

```
## [44] "ferrata storti foundation"
## [45] "frontiers"
## [46] "future science"
## [47] "hindawi"
## [48] "humana press"
## [49] "impact"
## [50] "informa healthcare"
## [51] "international aids society"
## [52] "international union against tuberculosis and lung disease"
## [53] "iop publishing"
## [54] "ios press"
## [55] "iucr"
## [56] "ivyspring international publisher"
## [57] "j med internet research"
## [58] "johns hopkins university press"
## [59] "journal of american physiological proceedings of national academy of sciences"
## [60] "journal of visualized experiments"
## [61] "jove"
## [62] "jscimed central"
## [63] "karger"
## [64] "landes bioscience"
## [65] "lww"
## [66] "mary ann liebert"
## [67] "mdpi"
## [68] "mit press"
## [69] "national academy of sciences"
## [70] "nature"
## [71] "open access reg ltd"
## [72] "optical society of america"
## [73] "oxford"
## [74] "palgrave macmillan"
## [75] "pion"
## [76] "plos"
## [77] "pnas"
## [78] "policy press"
## [79] "portland"
## [80] "public.service.co.uk"
## [81] "pubmed"
## [82] "research media ltd"
## [83] "royal college of psychiatrists"
## [84] "rsc"
## [85] "sage"
## [86] "sciedu press"
## [87] "scientific research publishing"
## [88] "sheridan press"
## [89] "society for endocrinologyy"
## [90] "society for general microbiology"
## [91] "society for leukocyte biology"
## [92] "society for neuroscience"
## [93] "society for publication of acta dermato-venereologica"
## [94] "springer"
## [95] "taylor and francis"
## [96] "transcript verlag"
## [97] "university of basque country press"
```

```
## [98] "wiley"
## [99] "wolters kluwer"
```

```
length(levels(as.factor(dat$publisher)))
```

```
## [1] 99
```

We now have 99 publishers. I am going to call this 'good enough' to move on.

Question 2: What is the mean cost of publishing for the top 3 most popular publishers?

Which publishers are the most popular to publish with?

```
dat %>% group_by(publisher) %>% summarize(mean = mean(cost), n=n()) %>% arrange(desc(n))
```

```
## # A tibble: 99 x 3
##    publisher  mean     n
##    <chr>     <dbl> <int>
##  1 elsevier  2436.   409
##  2 plos      1139.   307
##  3 wiley     2009.   270
##  4 oxford    1850.   167
##  5 bmc       1343.    95
##  6 springer  2024.    94
##  7 nature    2673.    81
##  8 asbmb     1385.    73
##  9 bmj       2075.    58
## 10 acs       1252.    34
## # ... with 89 more rows
```

What is the mean cost of publishing for the top 3 most popular publishers?

```
dat %>%
  filter(publisher == "elsevier"| publisher == "plos"| publisher == "wiley") %>%
  summarize(mean = mean(cost))
```

```
## # A tibble: 1 x 1
##    mean
##   <dbl>
## 1 1915.
```

Question 3: What is the number of publications by PLOS One in this dataset?

Let's start by converting the journal titles to lowercase. Then we can grab everything with plos to see what we are dealing with.

```
dat$journal <- tolower(dat$journal)

filter(dat, grepl("plos", dat$journal)) %>%
  select(journal)
```

```
## # A tibble: 298 x 1
##    journal
##    <chr>
##  1 plos
##  2 plos
##  3 plos
```

```
##  4 plos
##  5 plos  computational biology
##  6 plos  one
##  7 plos 1
##  8 plos 1
##  9 plos 1
## 10 plos 1
## # ... with 288 more rows
```

We should probably check for 'public library of science one' listed. Looks like there is 1.

```
filter(dat, grepl("public", dat$journal)) %>%
  select(journal) %>%
  tail()
```

```
## # A tibble: 6 x 1
##   journal
##   <chr>
## 1 public health nutrition
## 2 public health nutrition
## 3 public library of science
## 4 public library of science one
## 5 public service review
## 6 zoonoses and public health
```

It looks like our variations are "plos one", "plos one", plosone","public library of science one" and "plos 1". Instead of doing data-cleaning for the journal column, we are just going to grab all of the rows in our data frame that contain one of these versions of PLOS One, and then count the number of those rows. It's regex time!!

```
filter(dat, grepl("p(los[[:space:]]*(one|1)|ublic library of science one)", dat$journal)) %>%
  nrow(.)
```

```
## [1] 208
```

Another, less clunky way.

```
filter(dat, grepl("[^b]\\s?(1|one)$", dat$journal)) %>% select(journal)%>% nrow(.)
```

```
## [1] 208
```

That is our answer!

Question 4: Convert sterling to CAD. What is the median cost of publishing with Elsevier in CAD?

```
dat %>%
  filter(publisher == "elsevier") %>%
  summarize(median_cad = median(cost)*1.79)
```

```
## # A tibble: 1 x 1
##   median_cad
##        <dbl>
## 1      4198.
```

Question 5: Annotate your data cleaning efforts and answers to these questions in an .Rmd file. Knit your final answers to pdf.

Congratulate yourself of how professional your document looks.