Homework 4

Benjamin Jewell - STA141B

General Approach:

For this project I broke the task up into 5 main steps: (1) Get each search results page, (2) Get all the information needed about each question, (3) Get all the information about any answers to that question, (4) Get all the information about any comments to that question, and (5) compile this data into three data frames.

As much data was scrapped from the search results page as possible before diving into each question's own web page. The question web page is passed along the entire script to extract the necessary data for each part. Each page returns the three data frames that are then joined together for the final results.

Data Structure:

I chose to employ a data structure somewhat similar to the SQL data bases in Homework 3, with some minor modifications. Instead of having a PostTypeId to separate questions and answers, I had an answer table and a question table, along with a comments table. Questions have their post ID marked, and answers have the ID of their parent question post marked so one can track the link between these two. Because both questions and answers can have comments, comments have both their parent ID (the question or answer they respond to) marked, as well as the ID of the question whose page they are on.

Here is an attempt at visualizing the data structure using a SQL schema website, I hope this helps.

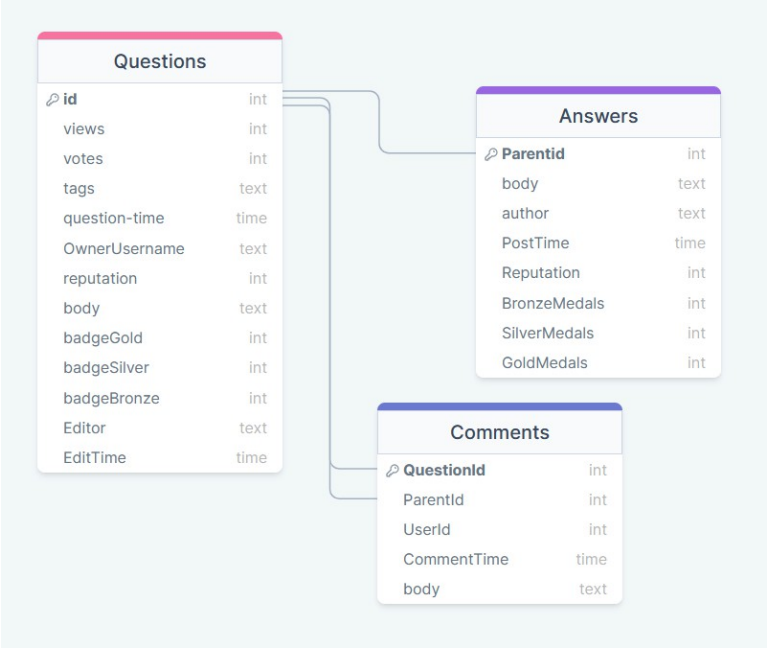


Diagram 1: Schema of output data frames

Detailed Approach:

The general outline above explains the broad steps we used: Parse each search page, go to each question web page, gather the question, answer and comments data, then return to the search page and

go to the next page before beginning this process again. We will now go into a more detailed explanation.

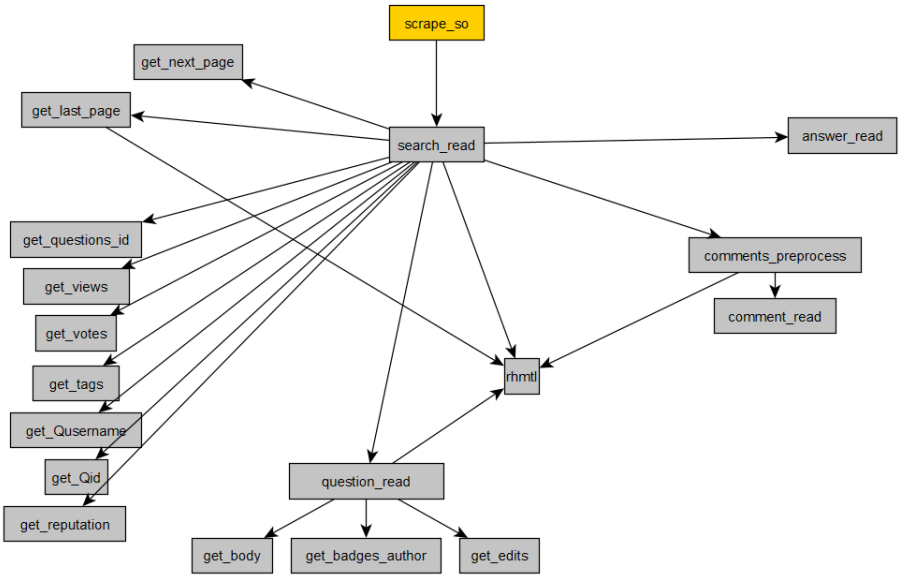


Diagram 2: Overview of functions

Above in Diagram 2 we can see an outline of the functions used in this process, starting with scrape_so. Arrows indicate when a function calls another function. Each function is explained in further detail in the Functions section.

We begin with scrape_so, and an indicated depth we would like to traverse. The start URL of page 1 is passed into search_read where the data is gathered, beginning our process.

search_data takes the given URL and returns the HTML document of that page via rhtml, gathering the URLs for each of the questions on that page. A lot of the information about questions can actually be found on the search page, so as much of that data as possible is gathered here. Using the following functions we can get the respective data: get_question_id gives us the post ID of the question, get_views gives the number of views of the question, get_votes gives the number of votes on the question, get_tags gives all tags on the question, get_Qusername gives the username of the person who posted the question, get_Qid gives the user ID of the poster and get reputation gives the reputation of the poster. As explained in the Functions section this is mostly straight forward, but there are some users missing from the StackOverflow database (which I refer to as “dead users”) and posts that are community wiki posts.

Once this data is constructed we create a data frame of this data, along with empty columns for that data we will have to get from within the question. For this we pass the URL of a question to question_read. Question read gathers the remaining information about the post (post body, their badges, any editors and the time the post was edited). These values are then inserted into the proper row of the data frame and the data frame is returned. The data frame is passed in and out of question_read until it is filled.

From there we can approach getting the data about each answer. Note that whenever we make a web request via rhtml we pause for 1 second to avoid getting rate limited. To increase run time since we already got the HTML document of each question from read_question, that document is passed onto read_answer (and comment_preprocess). read_answer gathers the relevant information as needed, with edge cases for dead users and community wikis. Due to the shortness of this process I did not use excess functions to gather each piece of data, though that might have been cleaner, as my document

was already feeling somewhat cluttered with all the sub functions. Generally gathering all the information on the answers was fast enough I didn't feel the need to break these up for unit testing, as I had grown more comfortable with the xpath and web scrapping process at this point. Do note however there were excessive debug statements throughout this section, many which I removed to make the document readable. We ensure that there is at least one of each value before returning the data frame of these answers to avoid returning empty data frames. Back in search_read we bind these 50 (or less) data frames together.

Comments are also passed the HTML documents gathered from read_question. Originally I was just going to read the comments from this HTML document, however it was discovered thanks to Eric Sun on Piazza that if there too many comments then certain comments are hidden. To get around this a web request can be made to StackOverflow to get the comments from each post (answer or question). As such we have comment_preprocess to extract these HTML documents for each post. We use comment_read to then extract these values and combine them into a data frame. These data frames are then bound together in search_read.

Once we are done gathering questions, answers and comments, search_read then looks at finding the next page URL. It does so by checking the current page versus the n_depth. If our current page is less than the n_depth we simply extract the URL from the 'NEXT PAGE' button via get_next_page and return that URL along with our data frames. If we are on the same page as the n_depth, then we use get_last_page to get the URL of the last page and return it along with our data. If our page number is greater then n_depth we finish our search, only returning our data frames.

Of note, get_next_page simply returns the URL for the next page based on the 'NEXT PAGE' button at the bottom of the search page. However get_last_page cannot simply give us the URL for the largest page number (the last page) and be done. For some reason the last page of StackOverflow does not appear to be stable, and often shifts. Manually clicking on the last page will often reveal that you can go a few pages deeper, however clicking these extra pages will often take you to empty pages and you have to backtrack to find the real last page. However if you refresh the page that number might change, the last page might be before or after the current page you thought to be the last page. The (up to 50) seconds between search_read calling the page and question_read(s) can be enough to throw this off.

As such I had to try to programatically deal with these “mirages”. get_last_page attempts to jump to the last page and return that page. However if we cannot find any question Ids on this page (there are no questions on this page) then it attempts to retreat to a lower page via the 'PREV PAGE' button until it finds a page that has at least one valid question. Unfortunately there often isn't a previous page button to jump onto even. To avoid crashing at this point the program instead jumps to a page I've found to be stable, chosen due to it's proximity to the last page and just because it's a round number: page 9800. While I understand that this isn't actually the last page, it's the compromise I chose to make in my program to allow it to deal with these mirages.

When the data frames and URL are returned to scrape_so we save the data frames and stick the URL back into search_read. Once we are done (having visited the last page), we bind the data frames together and return them as a list.

Verification / Debugging:

A lot of debugging went into this process to match data frame row lengths, data types, etc. Each sub function to get a value has a set of statements I left commented that were giving me the class of the returned objects (usually of the items inside the returned list), the number of returned items (which should usually be 50) as well as just printing out these values. In some cases these might be missing but I can promise you that each time there was a bug I had to go through and check all the values, it just might have been cleaned up accidentally.

In our compiler functions (search_read, comments_read, answers_read, questions_read, etc) there are numerous print statements that allowed me to look through the data types, classes, lengths, dimensions and even just raw values of each variable. I often saved variables into the global variable bank (using `<<-`) so I could see them as they updated and changed, rather than having to work through the print statements.

Any time I found a bug in my sections, I would track down which section it belonged to and found the URL of the post that was causing issues and worked until those bugs were fixed. I specifically also checked posts with lots of answers, posts with no answers, etc to get a good variety of test cases. As well I encountered and often tested dead user posts and community wiki posts due to their *unique* nature.

Unfortunately while I wanted to include several conditional warnings (there are some if you look through the functions below), I realized that the number of edge cases I encountered and the fact that the last page is not necessarily 50 posts would make it more trouble and thus relied on print statements to debug.

At the end as well we graph out all the numeric values we can and look at the distributions to make sure that they're reasonable. Most follow a Poisson distribution as we would expected there to be more users with fewer medals, views, reputation, etc. There are a few values that seem much larger than the rest, but these seem to be coming from the last page which are all very old posts so it makes sense they have lots of views and those users have tons of medals as well. I also noticed when looking at these values and looking at the pages that many users reappear often. I had been looking through the answers data and was worried I was getting duplicates because I saw the same name crop up several times in a row. However when I double checked it turned out this user had just given an answer to like 10 questions on the search page – it seems like there is a good active community here to answer these questions. Thus with users who answer lots of questions they can very high reputation, medals, etc thus giving us some of our outliers.

Functions:

rhtml(url**) –** This function was the way I accesses the web for my entire script. Taking a URL as a string, it uses the `httr` package to get the HTML for that query.

get_question_id(doc**) –** This function fetches the question Id for each question for the search page. It does so by looking under the `questions-summary` section of the HTML, where each unique question ID is stored. Takes the parsed HTML and returns a list of all question Ids.

get_views(doc) – Takes the parsed HTML and returns a list of view counts. Works on the search page, it searches the `question-summary` section for each question, looking under the span for the view count.

get_votes(doc) - Takes the parsed HTML of a search page and returns a list of all vote counts. Similar to get_views it searches the `question-summary` section to retrieve the votes.

get_body(doc) - Takes the parsed HTML of a question's page and returns the body of that page. Most of the actually information about a post can be found under the div where the class contains `post-layout right` and under the `s-prose`.

get_tags(doc) - Takes the parsed HTML of a search page and returns a list of the tags. Due to the fact that the tags are stored in list like `t-r, t-tag1,t-regression` (the tags thus being: r, tag1, regression) we use regular expressions to extract them out of this form. They are then combined into a string with commas between them for easy separation. A lot of work is then done to get them into the right data

type.

get_Qtime(doc) - Takes the parsed HTML of a search page and returns a list of the post time of each tag. This is a usually a rather easy operation, just grabbing the post `relativetime` value and returning it. However there is a rare edge case I found on the last page. One post is “community owned” and does not have a post time. Thus I added a whole for-loop to account for this and insert the time this post was made, which I found by hand.

get_QUsername(doc) - Takes the parsed HTML of a search page and returns a list of the username of the poster of each question. Normally we just look under the question-summary and then s-user-card-link to get the username, however there are a few cases this doesn't work. At least once on the final page there is a post by a user who does not appear to be in the StackOverflow database for some reason, one of these being an 'Eytan'. A whole for-loop is added to account for these 'dead-usernames' as I dubbed them and return the proper value. These “dead users” will continue to be a thorn in our side.

get_Qid(doc) - Takes the parsed HTML of a search page and returns a list of the user ID for each post. This is normally an easy to find value in under @data-user-id. However the “dead users” do not have such an ID, nor does the “Community post(s)” and as such we have to add an extra filtering for-loop to deal with these. These edge cases receive a user Id of 0.

get_reputation(doc) - Takes the parsed HTML of a search page and returns a list of reputations. Normally this is easy to find under the s-user-card—info. However once again our “dead users” and “community post” don't have reputation and we must filter them properly as we did before. As well, reputation is stored two ways: the actual value in a string such as “reputation score 42,401” and an abbreviation such as '42.4k'. It seems that if their score is only 4 characters or less (<9999) then there is no abbreviated score. Sometimes there is also only an abbreviated score. As such we compute a list of both scores, then fill any gaps in the full reputation score list with the abbreviated score.

get_badges(doc) - Takes the parsed HTML of a question page and returns a list of how many Bronze, Silver and Gold badges that user has, each as a separate value. To do so we check the span where the class is `badgecount`, and get the value. If that value is missing we replace it with 0, since not all users have at least 1 of each badge.

get_edits(doc) - Takes the parsed HTML of a question page and returns a list of the editor of each post and the time that post was edited. First we look for an edit time under `show all edits to this post`. If there are no edit times then that means there was no editing happening so NA is returned for the editor and edit time. We also have to account for if the original user edits their post, or if someone else does by checking if the editor has the `@itemprop = author`.

question_read(url, qst_df, i) – This is where the data frame for the question of each search page is compiled. It takes the url of the question page, along with the data frame of all the data that could be scrapped from the search page. `i` is also provided to know what row of the data frame to insert our data into.

It begins by querying the question page for the HTML and parsing it, then sleeping 1 second to avoiding getting blocked by StackOverflow. The body, badges, editor and edit time are all scrapped, then inserted into the appropriate row of the data frame that refers to this question, based on the `i`.

The data frame is then returned, along with the HTML doc object so that we can reuse it and speed up run time when we look for the answer values and the comment values.

answer_read(doc, i) – This is where all data for the answers of each question are gathered. It takes the HTML document of each page which we gathered in question_read() and reuses it, along with the index `i` which is solely here for debugging purposes so that if an error occurs I know what question broke it (there is also a variable that saves the URL of the most recent post read to let me debug).

The body, poster username and time of post for each answer are easy to extract using a simple query. They just look for the appropriate values if there and return them. Note that there aren't always answers.

Similar to how we got the reputation score before we have to extract the correct value from a string. As before we calculate the real and abbreviated score and take the real score, unless it is missing. The badges also follow a similar process as when trying to find them from the questions. We do have to introduce an extra loop fill any missing badges with 0. We treat the number of usernames of answer posters as the canonical number of answers here, since it seems the most reliable.

Then as long as there are any values returned we return a data frame of the answers from this page.

comment_preprocess(qdoc) – This takes the HTML document of a given question page to extract the comments of that page.

This process is particularly difficult, since some comments are hidden and need a click to open. However in inspecting this process I discovered that if you go to https://stackoverflow.com/posts/[post_id]/comments then it will return the data for all comments on that post, even the hidden ones. Since not all pages have comments we have to use some exception handling to avoid a crash. Also extracts the ID of the question page and the parent post Id for each comment. The information for each comment is extracted via **comment_read()**.

Comment_read(doc, q_id, parent_id) – Takes the HTML document for the comments of a post (question or answer), along with the id of that post and the id of the question that post is or is from. The data here is particularly easy to get out, though the time of the comment just requires a minor string trim to get the real time.

get_next_page(doc) – Takes the HTML document of a search page and gets the link to the next page. It does so by getting the link that the NEXT PAGE button would return.

get_last_page(doc) – similar to **get_next_page()**, but finds the page button that precedes the NEXT PAGE button which so happens to be the last page. As explained in the Detailed Approach Section, this function attempts to read the last page. If no question Ids are found on that page it attempts to retreat by reducing the page number until it finds a page that has at least one question Id. If there is no previous page button for it to follow (for example if it overshoots) then it jumps to page 9800 as a compromise to allow the code to not crash.

search_read(url, n_depth**)** – Takes the URL to a search page, along with the depth we want to go to. That page is then parsed and information is extracted. Then if the current page number is equal to the n_depth then we jump to the last page. If we are already on the last page then we return a kill code to stop further recursion.

The HTML for each page is parsed, then information is extracted. Much of the question information is gotten from this page, including the post_id, view, votes, tags, question post time, question poster username, question poster ID and their reputation. The URL for each question is also gathered and passed to **question_read()** to extract the remaining information. This information is bound into a data frame.

The HTML document for each question page is then passed to **answer_read** which will extract the information for all the answers. This information is bound into a data frame. The exact same process occurs for the comments via **comment_read**.

The next page URL is extracted as detailed above and then returned.

scrape_so(n_depth) – This function is our Main function and takes the n_depth as described in search_read.

As long as we do not receive the kill code from search_read this function will call **search_read**, take the questions, answers and comments data frame, save them and then repeat the process with the next page. Once the kill code is received this loop ends and the data frames for each category (question, answer, and comments) is returned and we are done.

Works Cited:

[1] https://stackoverflow.com/questions/32019566/r-xml-parse-for-a-web-address
[2] https://stackoverflow.com/questions/1604471/how-can-i-find-an-element-by-css-class-with-xpath
[3] https://stackoverflow.com/questions/18547410/xpath-with-multiple-contains-on-different-elements
[4] https://stackoverflow.com/questions/11455590/parse-an-xml-file-and-return-an-r-character-vector
[5] https://statisticsglobe.com/concatenate-vector-of-character-strings-in-r
[6] https://stackoverflow.com/questions/34570860/add-nas-to-make-all-list-elements-equal-length
[7] https://www.r-bloggers.com/2020/10/basic-error-handing-in-r-with-trycatch/
[8] https://piazza.com/class/lfxbfh6er6b2jo/post/365

HW04 - Web Scrapping

Benjamin Jewell

2023-06-03

URL: https://stackoverflow.com/questions/tagged/r (https://stackoverflow.com/questions/tagged/r) Pages Example: https://stackoverflow.com/questions/tagged/r?tab=newest&page=3&pagesize=15 (https://stackoverflow.com/questions/tagged/r?tab=newest&page=3&pagesize=15)

Set up

```
library(XML)
library(xml2)
library(httr)
library(RCurl)
library(knitr)
SO_url = 'https://stackoverflow.com/questions/tagged/r'
base_url = 'https://stackoverflow.com'
```

TASKS:

For Each Question:

1 The number of views of the question + search page

1 The number of votes + search page

1 Text of the question –> Title? + question page

1 Tags of question + search page

1 When the question was posted + search page

1 The user/display name of the poster + search page

1 Their reputation + search page

1 How many Gold/Silver/Bronze badges they have + question page

1 Who edited the question, when? + question page

For Each Answer:

1 The text + question page

1 The poster + question page

1 When they posted + question page

1 Their Reputation + question page

1 Their badge info + question page

For ALL Comment:

1 The Text + Sub-comments page

1 Who commented + Sub-comments page

1 When they posted + Sub-comments page

Functions

Read URL

```
#read any html page
rhtml <- function(url){
  Agent = "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/
41.0.2227.0 Safari/537.36"
  z = GET(url, user_agent(Agent))#, verbose(info = FALSE))
  return(htmlParse(z))
}
```

Get the unique ID of each question

```
#Get Question post ID from the search page
get_question_id <- function(doc){
  q_id = getNodeSet(doc, "//div[contains(@id, 'question-summary')]/@data-post-id")
  q_id = lapply(q_id, as.numeric)
  return(q_id)
}
# i = get_question_id(rhtml('https://stackoverflow.com/questions/tagged/r?tab=newest&page=981
2&pagesize=50'))
# unique(lapply(i, typeof))
# unique(lapply(i, class))
# length(i)
# i
```

Get the total number of views

```
#Get the number of views per question from the search page
get_views <- function(doc){
  view_count = getNodeSet(doc, "//div[contains(@id, 'question-summary')]/div/div[contains(@ti
tle, 'views')]/span[not(contains(., 'views'))]/text()")
  view_count = lapply(view_count, function(x) as.character(xmlValue(x)))
  return(view_count)
}
# e = "https://stackoverflow.com/questions/tagged/r?tab=newest&page=9812&pagesize=50"
# v = get_views(rhtml(e))
# unique(lapply(v, typeof))
# unique(lapply(v, class))
# v
```

Get the total number of votes

```
#Get number of votes per question from the search page
get_votes <- function(doc){
  votes_count = getNodeSet(doc, "//div[contains(@id, 'question-summary')]/div/div[contains(@t
itle, 'Score of')]/span[not(contains(., 'vote'))]/text()")
  return(xmlValue(votes_count))
}
# v = get_votes(rhtml('https://stackoverflow.com/questions/tagged/r?tab=newest&page=9812&page
size=50'))
# unique(lapply(v, typeof))
# unique(lapply(v, class))
# v
```

Get the text body

```
#Get the post body text from a question, from a question page
get_body <-function(doc){
  body = getNodeSet(doc, "//div[contains(@class, 'postcell post-layout--right')]/div[contains
(@class, 's-prose ')]")
  return(xmlValue(body))
}
#get_body(rhtml('https://stackoverflow.com/questions/76346497/error-penman-monteith-fao56-ref
erence-crop-et'))
```

Get the tags

```
#Get the tags per post from the search page
get_tags <- function(doc){
  t_str = getNodeSet(doc, "//div[contains(@id, 'question-summary')]/div/div/div[contains(@cla
ss, 'js-tags')]/@class")
  #Not the cleanest way to extrac the tags, but it seems to work
  #[5] collapse string-lists together
  tags = sapply(t_str, function(x) regmatches(x, gregexpr('(?<= t-)[^ ]+', x, perl = TRUE)))
  tags = lapply(tags, function(x) paste(x, collapse = ','))
  tags = lapply(tags, as.character)

  return(as.character(tags))
}

# t = get_tags(rhtml('https://stackoverflow.com/questions/tagged/r?tab=newest&page=9812&pages
ize=50'))
# length(t)
# unique(lapply(t, typeof))
# unique(lapply(t, class))
# t
```

When were the questions posted?

```
#Get the post time per question from the search page
get_Qtime <- function(doc){
  #time = getNodeSet(doc, "//div[contains(@id, 'question-summary')]/div/div/div/time/span/@ti
tle")
  time = getNodeSet(doc, "//div[contains(@id, 'question-summary')]//div/time/span[@class = 'r
elativetime']/@title | //div[contains(@title, 'community owned ')]/text()[1]")

  #debug for when last page is empty - avoids crash
  if (length(time) == 0){return()}

  #debug for when the last page has strange info = a community post
  for (i in 1:length(time)){
  if (typeof(time[[i]]) == "externalptr"){
      time[[i]] = "2009-11-08 07:33:00Z"
    } else {
      time[[i]] = as.character(time[[i]])
    }}

  return(paste(time))
}

# t = get_Qtime(rhtml('https://stackoverflow.com/questions/tagged/r?tab=newest&page=9813&page
size=50'))
# unique(lapply(t, typeof))
# unique(lapply(t, class))
# print(t)
# length(t)
```

Get Username & Id

```
#Get the question poster username from the search page
get_Qusername <- function(doc){
  username = getNodeSet(doc, "//div[contains(@id, 'question-summary')]//div[contains(@class,
's-user-card--link')]/a/text()")
  username = xmlValue(username)

  #Some usernames aren't links anymore, so we get a 2nd set to compare
  usernames_v2 = list()
  deadusername = xmlValue(getNodeSet(doc, "//div[contains(@id, 'question-summary')]//div[cont
ains(@class, 's-user-card--link')]/text()"))
  for (i in 1:length(deadusername)){
    if (nchar(trimws(deadusername[[i]])) > 0){
    usernames_v2 = append(usernames_v2, trimws(deadusername[[i]]))}
  }

  return(as.character(append(username, usernames_v2)))
}

#Get the question ID per post from the search page
get_Qid <- function(doc){
  #includes an option to get community posts
  id = getNodeSet(doc, "//div[contains(@id, 'question-summary')]//a/@data-user-id | //div[con
tains(@id, 'question-summary')]//span/@data-user-id | //div[contains(@id, 'question-summary
')]//div[contains(@title, 'community owned')]/text()[1]")

  #handles when there are "Community Wiki Posts"
  for (i in 1:length(id)){
  if (typeof(id[[i]]) == "externalptr"){
      id[[i]] = "0"
    } else {
      id[[i]] = as.character(id[[i]])
    }}

  return(as.numeric(id))
}

# u = get_Qusername(rhtml('https://stackoverflow.com/questions/tagged/r?tab=newest&page=9812&
pagesize=50'))
# print(unique(lapply(u, typeof)))
# unique(lapply(u, class))
# print(length(u))
# print(u)
#
# u = get_Qid(rhtml('https://stackoverflow.com/questions/tagged/r?tab=newest&page=9812&pagesi
ze=50'))
# print(unique(lapply(u, typeof)))
# unique(lapply(u, class))
# print(length(u))
# print(u)
```

Get reputation

```
#Get the reputation per question from the search page
get_reputation <- function(doc){
  n_users = getNodeSet(doc, "//div[contains(@id, 'question-summary')]//div[@class = 's-user-c
ard--info']//li/span/@title | //div[contains(@id, 'question-summary')]//div[@class = 's-user-
card--info' and count(./ul) = 0]/text()[1]")

  #Turns all nodes that don't have a reputation score blank OR gets Rep Score
  for (i in 1:length(n_users)){
    if (typeof(n_users[[i]]) == "externalptr"){
      n_users[[i]] = trimws(xmlValue(n_users[[i]]))
    } else {
      n_users[[i]] = gsub('reputation score ', '', as.character(n_users[[i]]))
    }
  }

  #Finds rep a 2nd more reliable way to replace any missing from n_user rep
  old_rep = getNodeSet(doc, "//div[contains(@id, 'question-summary')]/div/div/div/div/ul/li/s
pan/text()")

  if (length(n_users) != length(old_rep)){warning(paste('Reputation scores do not match! Nre
p:', length(n_users), 'OldRep:', length(old_rep)))}

  #replaces any missing rep scores with the rep found via old_rep
  for (i in 1:length(old_rep)){
    if (nchar(n_users[[i]]) == 0){n_users[[i]] = xmlValue(old_rep[[i]])}
  }

  n_users = lapply(n_users, function(x) if (nchar(x) == 0){x = 0} else {x=x})

  return(as.character(n_users))
}
# r = get_reputation(rhtml('https://stackoverflow.com/questions/tagged/r?tab=newest&page=9813
&pagesize=50'))
# print(unique(lapply(r, typeof)))
# unique(lapply(t, class))
# print(length(r))
# r
```

Get badges

```
#Gets badges per user from the question page
get_badges_author <- function(doc){
  #bdg = getNodeSet(doc, "//div[@class = 'post-signature owner flex--item']")
  b_bdg = getNodeSet(doc, "//div[@class = 'post-signature owner flex--item']/div[contains(@cl
ass, 'user-info')]//div[contains(@itemprop, 'author')]//span[@class = 'badge3']/ancestor::spa
n[contains(@title, 'badges')]/span[@class = 'badgecount']/text()")

  s_bdg = getNodeSet(doc, "//div[@class = 'post-signature owner flex--item']/div[contains(@cl
ass, 'user-info')]//div[contains(@itemprop, 'author')]//span[@class = 'badge2']/ancestor::spa
n[contains(@title, 'badges')]/span[@class = 'badgecount']/text()")

  g_bdg = getNodeSet(doc, "//div[@class = 'post-signature owner flex--item']/div[contains(@cl
ass, 'user-info')]//div[contains(@itemprop, 'author')]//span[@class = 'badge1']/ancestor::spa
n[contains(@title, 'badges')]/span[@class = 'badgecount']/text()")

  #If a user doesnt have a specific type of badge they get 0
  if (length(b_bdg) < 1){b_bdg = 0} else (b_bdg = xmlValue(b_bdg))
  if (length(s_bdg) < 1){s_bdg = 0} else (s_bdg = xmlValue(s_bdg))
  if (length(g_bdg) < 1){g_bdg = 0} else (g_bdg = xmlValue(g_bdg))
  return(c(b_bdg, s_bdg, g_bdg))
}
#get_badges_author(q_doc)
```

Finding who edited the question & when

```
#Gets the time and editors of a post from a question page
get_edits <- function(doc){
  edit_time = getNodeSet(doc, "//div[@class = 'user-action-time']/a[contains(@title, 'show al
l edits to this post')]/span/@title")
  #print(edit_time)

  #Is there an edit?
  if (length(edit_time) > 0){
    #Is the editor the original author?
    author = getNodeSet(doc, "//div[@class = 'post-signature flex--item']//div[@itemprop = 'a
uthor']")
    #If the author edited this:
    if (length(author) > 0){
      #the original author
      editor = xmlValue(getNodeSet(doc, "//div[@class = 'post-signature owner flex--item']//d
iv[@class = 'user-details']/a/text()"))
    } else {
      #There is another user as our editor, AKA not the author
      editor = xmlValue(getNodeSet(doc, "//div[@class = 'post-signature flex--item']//div[@cl
ass = 'user-details']/a/text()"))
    }
  } else {
    editor = NA
    edit_time = NA
  }

  return(list(editor, paste(edit_time)))
}
#get_edits(q_doc)
```

Question Page Reading

```
#Reads a given question URL
question_read <- function(url, qst_df, i){
  Sys.sleep(1) # Ensures we don't query SO too quickly
  qdoc = rhtml(url)

  #pastes the URL for debugging purposes
  print(paste('Current URL:', url))

  #Extract data from the question page
  body = get_body(qdoc)
  auth_badges = get_badges_author(qdoc)
  edit_gen_info = get_edits(qdoc)
  editor = edit_gen_info[[1]]
  edit_time = paste(edit_gen_info[[2]], collapse =', ')

  #print(c(length(body),))

  #Insert the data into the empty data frame
  qst_df$body[i] = body
  qst_df$badgeGold[i] = auth_badges[[3]]
  qst_df$badgeSilver[i] = auth_badges[[2]]
  qst_df$badgeBronze[i] = auth_badges[[1]]
  qst_df$Editor[i] = editor
  qst_df$EditTime[i] = edit_time

  #print(qst_df)

  return(list(qst_df, qdoc))
}

#question_read('https://stackoverflow.com/questions/tagged/r?tab=newest&page=9810&pagesize=50
', 1, 1)
```

Read the answers from each page

Minor Issue: There are two values for the Reputation, one is abbreviated once it is above like 4 characters long so we have to extract the reputation from a @title value instead. However when the Reputation is like 4 characters or less that @title value is no longe present so we have to get the normal value.

```
#Reads a question page and returns it
answer_read <- function(doc, i){
    #saves the link for debugging purposes
    why <<- getNodeSet(doc, "//head/link[@rel = 'canonical']/@href")
    #print(why)

    # --- The text ---
    a_text = getNodeSet(doc, "//div[@id = 'answers']/div[contains(@class, 'js-answer')]//div
[@itemprop = 'text']")
    a_text = lapply(a_text, function(x) as.character(xmlValue(x)))
    #print(paste('Text Length :', length(a_text)))


    # ---The poster ---
    a_poster = getNodeSet(doc, "//div[@id = 'answers']/div[contains(@class, 'js-answer')]//di
v[@class = 'user-details' and @itemprop = 'author']/a/text() |//div[@id = 'answers']/div[cont
ains(@class, 'js-answer')]//div[@class = 'user-details']/span[@class = 'community-wiki']/text
()")
    a_poster = lapply(a_poster, function(x) trimws(xmlValue(x)))

    #Certain users arent in the SO database so we handle them here
    a_dead_poster = getNodeSet(doc, "//div[@id = 'answers']/div[contains(@class, 'js-answer
')]//div[@class = 'user-details' and @itemprop = 'author']/text()[1]")
    a_dead_poster = lapply(a_dead_poster, function(x) trimws(xmlValue(x)))
    #replace any missing values with the dead author

    if (length(a_dead_poster) < 1){a_dead_poster = c('')} #avoids crashing the loop when ther
e are 0 answers
    for (i in 1:length(a_dead_poster)){
      if (nchar(a_dead_poster[[i]]) > 0){
        a_poster = append(a_poster, a_dead_poster[[i]], after = (i - 1))
      }}


    # --- When they posted ---
    a_time = getNodeSet(doc, "//div[@id = 'answers']/div[contains(@class, 'js-answer')]//div
[@class = 'user-action-time']/span/@title")
    a_time = lapply(a_time, as.character)
    #Deals with community wiki posts
    community_time = getNodeSet(doc, "//div[@id = 'answers']/div[contains(@class, 'js-answer
')]//div[@class = 'user-details']/span[@class = 'community-wiki']/@title")
    community_time = lapply(community_time, function(x) gsub('(This post is community owned a
s of )?(. Votes do not generate reputation, and it can be edited by users with 100 rep)?',
'', as.character(x)))
    community_time = lapply(community_time, function(x) gsub('at', '', x))

    c = 1
    if (length(community_time) > 0){for (i in 1:length(a_poster)){
      if (a_poster[[i]] == 'community wiki'){
        a_time = append(a_time, community_time[[1]], after = (i - 1))
        c = c + 1
```

```
      }
    }}


    # --- Their Reputation ---
    #Version One
    a_rep = getNodeSet(doc, "//div[@id = 'answers']/div[contains(@class, 'js-answer')]//div[@
class = 'user-details' and @itemprop = 'author']/div[@class = '-flair']/span[@class = 'reputa
tion-score']/@title | //div[@id = 'answers']/div[contains(@class, 'js-answer')]//div[@class =
'user-details']/span[@class = 'community-wiki']/text()")
    for (i in 1:length(a_rep)){
      if (typeof(a_rep[[i]]) == 'externalptr'){
        a_rep[[i]] = 'reputation score 0'}
      }
    #Reputation score is saved in a string for some reason so we pull out the number
    a_rep = lapply(a_rep, function(x) as.numeric(gsub('(reputation score )?,?', '', x)))

    #Version Two, as per same situation as get_rep()
    a_rep2 = getNodeSet(doc, "//div[@id = 'answers']/div[contains(@class, 'js-answer')]//div
[@class = 'user-details' and @itemprop = 'author']/div[@class = '-flair']/span[@class = 'repu
tation-score']/text()  | //div[@id = 'answers']/div[contains(@class, 'js-answer')]//div[@clas
s = 'user-details']/span[@class = 'community-wiki']/text()")
    # print(paste('REP 2:', a_rep2)
    a_rep2 = lapply(a_rep2, function(x) gsub('(reputation score )?,?', '', xmlValue(x)))
    # print(a_rep2)
    #if there is no value from Version 1, then we use the Version 2 value instead
    if (length(a_text) > 0){
    for (i in 1:length(a_rep)){if (is.na(a_rep[[i]]) == TRUE){
      a_rep[[i]] = a_rep2[[i]]}}}

    #Handles dead users Reputation
    a_dead_rep = getNodeSet(doc, "//div[@id = 'answers']/div[contains(@class, 'js-answer')]//
div[@class = 'user-details' and @itemprop = 'author']/text()[1]")
    a_dead_rep = lapply(a_dead_rep, function(x) trimws(xmlValue(x)))
    if (length(a_dead_rep) < 1){a_dead_rep = c('')} #avoids crashing the loop when there are
0 answers
    for (i in 1:length(a_dead_rep)){
      if (nchar(a_dead_rep[[i]]) > 0){
        a_rep= append(a_rep, 0, after = (i - 1))
      }}


    # --- Their badge info ---

    #bronze badges
    a_bronze = getNodeSet(doc, "//div[@id = 'answers']/div[contains(@class, 'js-answer')]//di
v[@class = 'user-details' and @itemprop = 'author']/div[@class = '-flair']/span[contains(@tit
le, 'bronze badges')]/@title")
    a_bronze = lapply(a_bronze, function(x) as.numeric(gsub(' bronze badges', '', x)))
    if (length(a_bronze) < length(a_poster)){
      a_bronze = append(a_bronze, rep(0, length(a_poster) - length(a_bronze)))}
```

```
    #silver badges
    a_silver = getNodeSet(doc, "//div[@id = 'answers']/div[contains(@class, 'js-answer')]//di
v[@class = 'user-details' and @itemprop = 'author']/div[@class = '-flair']/span[contains(@tit
le, 'silver badges')]/@title")
    a_silver = lapply(a_silver, function(x) as.numeric(gsub(' silver badges', '', x)))
    if (length(a_silver) < length(a_poster)){
      a_silver = append(a_silver, rep(0, length(a_poster) - length(a_silver)))}

    #gold badges
    a_gold = getNodeSet(doc, "//div[@id = 'answers']/div[contains(@class, 'js-answer')]//div
[@class = 'user-details' and @itemprop = 'author']/div[@class = '-flair']/span[contains(@titl
e, 'gold badges')]/@title")
    a_gold = lapply(a_gold, function(x) as.numeric(gsub(' gold badges', '', x)))
    if (length(a_gold) < length(a_poster)){
      a_gold = append(a_gold, rep(0, length(a_poster) - length(a_gold)))}



    # --- question id ---
    #for reference's sake we get the question ID to link them
    a_id = rep(as.numeric(getNodeSet(doc, "//div[@class = 'question js-question']/@data-quest
ionid")), length(a_poster))

    #debug statement
    # print(c('AiD', length(a_id), 'text', length(a_text), 'poster', length(a_poster), 'time
', length(a_time), 'rep', length(a_rep), 'brnz',  length(a_bronze), 'slvr', length(a_silver),
'Gold', length(a_gold)))
    #print(a_poster)

    if (!(0 %in% c(length(a_id), length(a_text), length(a_poster), length(a_time), length(a_r
ep), length(a_bronze), length(a_silver), length(a_gold)))){
    a_df = data.frame(a_id, unlist(a_text), unlist(a_poster), unlist(a_time), unlist(a_rep),
unlist(a_bronze), unlist(a_silver), unlist(a_gold))
    colnames(a_df) = c('ParentId', 'body', 'author', 'PostTime', 'Reputation', 'BronzeMedals
', 'SilverMedals', 'GoldMedals')
      return(a_df)
    } else {return(FALSE)}
}
# a = answer_read(rhtml('https://stackoverflow.com/questions/2564258/plot-two-graphs-in-a-sam
e-plot'),1)
# lapply(a, length)
# print(a)
# as.data.frame(a)
```

Read Comments

```
comment_preprocess <- function(qdoc){
  #Gather all question+answer IDs into a singular list, allowing us to get all comments - eve
n the hidden ones!
  q_id = as.numeric(getNodeSet(qdoc, "//div/@data-questionid"))
  a_ids = lapply(getNodeSet(qdoc, "//div/@data-answerid"), as.numeric)
  comment_ids = append(q_id, a_ids)

  comment_data = list()
  for (i in 1:length(comment_ids)){
    Sys.sleep(1)
    comment_url = paste('https://stackoverflow.com/posts/', comment_ids[[i]], '/comments', se
p='')

    #[7] Exception handling when there are no comments
    comment_data[[i]] = tryCatch({
        comment_read(rhtml(comment_url), q_id, comment_ids[[i]])
      }, error = function(e){})
    }

  return(do.call("rbind", comment_data))
}

#Actually read the comment
comment_read <- function(doc, q_id, parent_id){
  # Text
  c_text = xmlValue(getNodeSet(doc, "//li[contains(@class, 'comment js-comment')]//span[@clas
s='comment-copy']"))

  # Comment User
  c_user = xmlValue(getNodeSet(doc, "//a[contains(@class, 'comment-user')]/text()"))

  # Time
  c_time = getNodeSet(doc, "//li[contains(@class, 'comment js-comment')]//span[@class = 'comm
ent-date']//span/@title")
  c_time = lapply(c_time, function(x) substr(as.character(x), 0, 20))

  c_df = data.frame(list(list(rep(q_id, length(c_user))), list(rep(parent_id, length(c_use
r))), c_user, unlist(c_time), c_text))
  colnames(c_df) <- c('QuestionId', 'ParentId', 'UserId', 'CommentTime', 'body')
  return(c_df)
}
#c = rhtml('https://stackoverflow.com/questions/2851327/combine-a-list-of-data-frames-into-on
e-data-frame-by-row')
#comment_preprocess(c)

# c = rhtml('https://stackoverflow.com/posts/2851327/comments')
# print(c)
# comment_read(c, 20, 22)
```

Get next page functions

```
#Gets the URL for the next page
get_next_page <- function(doc){
  return(as.character(getNodeSet(doc, "//div[@id = 'mainbar']//a[@rel ='next']/@href")))
}

#Gets the URL for the "Last" page
get_last_page <- function(doc){
  #Adding extra loop to deal with the 'Last Page Mirage'
  last_pg = as.character(getNodeSet(doc, "//div[@id = 'mainbar']//a[@rel = 'next']/preceding-
sibling::a[1]/@href"))
  #print(last_pg)

  #return('/questions/tagged/r?tab=newest&page=9800&pagesize=50')

  #If the last page is empty and has no value, retreat to a known page that works.
  #This might not be exactly the last page, but since the "last page" is broken
  #I chose a number that is stable
  if (length(last_pg) < 1){
    print('Last page given broken! Retreating to stable backup at page 9800')
    return('/questions/tagged/r?tab=newest&page=9800&pagesize=50')
  }

  #Here we attempt to "retreat" and lower the page number to find a stable page
  #only works if there are page buttons
  mirage = TRUE
  #try to query that page, see how many post ids there are
  while (mirage == TRUE){
    #print(last_pg)
    Sys.sleep(1)
    mdoc = get_question_id(rhtml(paste(base_url, paste(last_pg), sep='')))
    if (length(mdoc) > 1){
      return(last_pg)
      mirage = FALSE
    } else {
      print('Mirage detected! Retreating!')
        last_page = paste(base_url, paste(as.character(getNodeSet(doc, "//div[@id = 'mainbar
']//a[@rel ='prev']/@href"))), sep='')
      }
  }

  return(last_pg)
}

get_last_page(rhtml('https://stackoverflow.com/questions/tagged/r?tab=newest&page=9820&pagesi
ze=50'))
```

```
## [1] "Mirage detected! Retreating!"
## [1] "Mirage detected! Retreating!"
## [1] "Mirage detected! Retreating!"
```

```
## [1] "/questions/tagged/r?tab=newest&page=9822&pagesize=50"
```

Read Generic Search Page

```
#Gathers all the information for a given page
search_read <- function(url, n_depth){
    doc = rhtml(url)
    #return(doc)

    #Get all URLs on the page
    #[2] Searching for specific values
    #[3] How to use 'AND contains'
    qlinks = getNodeSet(doc, "//div[contains(@id, 'question-summary')]/div/h3/a/@href")

    #Gathers the data from the search page
    #[4] Coercing xml values into lists
    id = get_question_id(doc)
    views = get_views(doc)
    votes = get_votes(doc)
    body_txt = get_body(doc)
    tags = get_tags(doc)
    question_times = get_Qtime(doc)
    Qusers = get_Qusername(doc)
    Qids = get_Qid(doc)
    reps = get_reputation(doc)

    # print(paste('ID', unique(lapply(id, class))))
    # print(paste('Views', unique(lapply(views, typeof))))
    # print(paste('votes', unique(lapply(id, typeof))))
    # print(paste('question_times', unique(lapply(tags, typeof))))
    # print(paste('Users', unique(lapply(Qusers, typeof))))
    # print(paste('Qids', unique(lapply(Qids, typeof))))
    # print(paste('Rep', unique(lapply(reps, typeof))))
    #
    # print(c('id', length(id), 'view', length(views), 'votes', length(votes), 'tags', length
(tags), 'qtime', length(question_times), 'qusers', length(Qusers), 'qIDs', length(Qids), 'rep
', length(reps)))

    #print(reps)

    q_df_inc = data.frame(unlist(id), unlist(views), unlist(votes), unlist(tags), question_ti
mes, Qusers, Qids, reps, list(rep('', length(id))), list(rep('', length(id))),list(rep('', le
ngth(id))), list(rep('', length(id))), list(rep('', length(id))), list(rep('', length(id))))

    colnames(q_df_inc) <- c('id', 'views', 'votes', 'tags', 'question-time', 'OwnerUsername',
'ownerId', 'reputation', 'body', 'badgeGold', 'badgeSilver', 'badgeBronze', 'Editor', 'EditTi
me')
    #print(q_df_inc)

    #Queries each Question, updates the data frame with the missing infor about the poste
    #Then returns the HTML of that page so we don't have to query it again - not the cleanest
solution
    q_html_objs = list()
    for (i in 1:length(qlinks)){
    q_query = question_read(paste(base_url, paste(qlinks[[i]]), sep=''),
```

```
              q_df_inc,
              i)
q_df_inc = q_query[[1]]
q_html_objs[[i]] = q_query[[2]]
}
#print(q_df_inc)

# --- Answers ---
j = 1
ans_df_list = list()
#Takes each Question Page HTML obj and parses through to get the answer
for (i in 1:length(q_html_objs)){
  a_query = answer_read(q_html_objs[[i]], i)

  if (length(a_query) > 1){
    #print(length(a_query))
    ans_df_list[[j]] = a_query
    j = j + 1
  }}

a_df_inc = do.call("rbind", ans_df_list)
#print(a_df_inc)


# --- Comments ---
k = 1
com_df_list = list()
for (i in 1:length(qlinks)){
  c_query = comment_preprocess(q_html_objs[[i]])

  if (length(c_query) > 1){
    com_df_list[[k]] = c_query
    k = k + 1
  }}

c_df = do.call("rbind", com_df_list)
#print(c_df)

#Go to the next page of the search to n
#if page_depth == n, then go to last page

#get page depth from URL using REGEX
page_depth = as.numeric(regmatches(url, gregexpr('(?<=page=)[0-9]+', url, perl = TRUE)))

#define the next page
if (page_depth == n_depth){
  #Return link to the last page
  next_url = get_last_page(doc)
  #next_url = TRUE
} else if (page_depth > n_depth){
  #Once we are on the last page send the kill command back to the main module
```

```
      next_url = TRUE
    }else {
    #Else return link to the next page
      next_url = get_next_page(doc)
    }

    return(list(q_df_inc, a_df_inc, c_df, next_url))
}

# init_url = 'https://stackoverflow.com/questions/tagged/r?tab=newest&page=9812&pagesize=50'
# init_url = 'https://stackoverflow.com/questions/tagged/r?tab=newest&page=1&pagesize=50'
# a = search_read(init_url, n_depth = 3)
# qq = a[[1]]
# qa = a[[2]]
# qc = a[[3]]
```

MAIN FUNCTION

This is the function that we actually call to run our script.

```
#Our main function, given a start page and a depth, it gathers data from that many
#search pages + last and returns a list of dataframes
scrape_so <- function(n_depth){
  input_url = 'https://stackoverflow.com/questions/tagged/r?tab=newest&page=1&pagesize=50'
  scrping = TRUE
  questions_list = list()
  answers_list = list()
  comments_list = list()

  d = 1
  while (scrping == TRUE){
    search_data <<- search_read(input_url, n_depth)
    # print(paste("SEARCH DATA LENGTH:", length(search_data)))
    # print(c(typeof(search_data[[1]]), print(dim(search_data[[1]]))))
    # print(c(typeof(search_data[[2]]), print(dim(search_data[[2]]))))
    # print(c(typeof(search_data[[3]]), print(dim(search_data[[3]]))))

    questions_list[[d]] = search_data[[1]]
    answers_list[[d]] = search_data[[2]]
    comments_list[[d]] = search_data[[3]]
    next_url = search_data[[4]]
    d = d + 1

    if (next_url == TRUE){scrping = FALSE} #kill the loop once we are done

    input_url = paste(base_url, paste(next_url), sep='')
    print(paste('INPUT URL:', input_url))
  }

  questions_list <<- questions_list
  answers_list <<- answers_list
  comments_list <<- comments_list

  # print('Shape of Questions:')
  # print(lapply(questions_list, dim))
  # print('Shape of Answers:')
  # print(lapply(answers_list, dim))
  # print('Shape of Comments:')
  # print(lapply(comments_list, dim))

  q_df <<- do.call('rbind', questions_list)
  a_df <<- do.call('rbind', answers_list)
  c_df <<- do.call('rbind', comments_list)

  # print('experimental bind')
  # exp_df <<- rbind(questions_list[[1]], questions_list[[2]])
  # print(exp_df)

  # print(typeof(q_df))
  # print(typeof(a_df))
  # print(typeof(c_df))
```

```
  final_df = list()
  final_df[[1]] = q_df
  final_df[[2]] = a_df
  final_df[[3]] = c_df
  return(final_df)
}

final_dataframes = scrape_so(3)
```

```
## [1] "Current URL: https://stackoverflow.com/questions/76398634/r-check-model-for-model-summaries-as-list-objects-and-6-plots-in-1-row"
## [1] "Current URL: https://stackoverflow.com/questions/76398607/why-is-my-fasterize-function-making-a-raster-filled-with-0s-for-my-presence-abse"
## [1] "Current URL: https://stackoverflow.com/questions/76398561/sequential-count-of-values-within-factor-level-ignoring-nas"
## [1] "Current URL: https://stackoverflow.com/questions/76398512/how-can-i-use-r-pdftools-and-stringr-to-extract-the-authors-name-from-the-first"
## [1] "Current URL: https://stackoverflow.com/questions/76398472/replace-multiple-patterns-by-start-and-end-index-in-r"
## [1] "Current URL: https://stackoverflow.com/questions/76398388/missing-one-line-in-ggplot2-line-graph-and-highlighting-particular-data-in-line"
## [1] "Current URL: https://stackoverflow.com/questions/76398208/why-monthly-mean-raster-output-seem-to-get-multiplied-by-10000"
## [1] "Current URL: https://stackoverflow.com/questions/76398196/analyze-an-entire-sheet-and-display-the-results-in-a-table"
## [1] "Current URL: https://stackoverflow.com/questions/76397984/is-there-an-r-function-that-would-produce-results-similar-to-scipy-s-gaussian-fi"
## [1] "Current URL: https://stackoverflow.com/questions/76397812/seurat-findclusters-seems-to-freeze-after-one-iteration"
## [1] "Current URL: https://stackoverflow.com/questions/76397805/r-matching-a-matrix-to-a-dataframe"
## [1] "Current URL: https://stackoverflow.com/questions/76397627/ggplot-reorder-within-order-month"
## [1] "Current URL: https://stackoverflow.com/questions/76397512/ggplot-ordering-legends-with-guides-changes-continuous-legend-to-discrete"
## [1] "Current URL: https://stackoverflow.com/questions/76397435/how-can-i-put-the-iterable-numbers-into-a-new-column-each-iteration"
## [1] "Current URL: https://stackoverflow.com/questions/76397278/how-to-add-multiple-column-headers-that-span-specified-columns-to-a-dt-table-ren"
## [1] "Current URL: https://stackoverflow.com/questions/76397124/python-or-r-packages-to-draw-2d-diagrams-of-intermolecular-interactions"
## [1] "Current URL: https://stackoverflow.com/questions/76396915/aggregating-model-summaries-from-list-of-models"
## [1] "Current URL: https://stackoverflow.com/questions/76396811/using-call-for-c-pkolmogorov2x-in-r-package-submitted-to-cran"
## [1] "Current URL: https://stackoverflow.com/questions/76396563/how-to-run-a-geographically-weighted-logistic-regression-in-r"
## [1] "Current URL: https://stackoverflow.com/questions/76396425/how-to-fix-the-subscript-out-of-bounds-error-when-converting-nc-to-text-in-r-u"
## [1] "Current URL: https://stackoverflow.com/questions/76396351/how-to-stitch-merge-two-rasterstack-map-overlapping-each-other"
## [1] "Current URL: https://stackoverflow.com/questions/76396349/is-there-a-r-function-to-make-pairwise-correlation-tests-between-2-different-sub"
## [1] "Current URL: https://stackoverflow.com/questions/76396323/spgwr-error-new-data-matrix-rows-mismatch"
## [1] "Current URL: https://stackoverflow.com/questions/76396291/error-encountered-when-trying-to-plot-individual-trees-from-cforest-forest-usi"
## [1] "Current URL: https://stackoverflow.com/questions/76396233/error-in-the-mapnames-operator-is-invalid-for-atomic-vectors"
## [1] "Current URL: https://stackoverflow.com/questions/76396165/how-to-create-a-spatial-gri
```

```
dlines-using-latitude-and-longitude-in-r"
## [1] "Current URL: https://stackoverflow.com/questions/76395825/efficient-way-to-change-the
-class-of-several-matrices-in-r"
## [1] "Current URL: https://stackoverflow.com/questions/76395807/error-in-st-kriging-of-unev
en-spacetime-data-the-leading-minor-of-order-2-is-n"
## [1] "Current URL: https://stackoverflow.com/questions/76395682/how-to-merge-two-data-frame
s-by-first-4-digits-of-mergeable-values-only"
## [1] "Current URL: https://stackoverflow.com/questions/76395579/error-publishing-flexdashbo
ard-as-shiny-app-in-r"
## [1] "Current URL: https://stackoverflow.com/questions/76395522/how-to-reshape-data-from-lo
ng-to-wide-format-in-r"
## [1] "Current URL: https://stackoverflow.com/questions/76395332/extract-iteratively-data-in
-subfolders-within-folder-into-googledrive"
## [1] "Current URL: https://stackoverflow.com/questions/76395254/r-efficient-way-to-apply-a-
match-function-on-every-element-of-vector"
## [1] "Current URL: https://stackoverflow.com/questions/76395142/including-the-title-page-an
d-back-page-in-a-quarto-book-in-pdf-format"
## [1] "Current URL: https://stackoverflow.com/questions/76395098/dotplot-of-enrichgo-results
-with-all-of-the-ontology-terms-on-same-plot-for-comp"
## [1] "Current URL: https://stackoverflow.com/questions/76394809/create-a-variable-filtering
-for-all-rows-which-are-based-on-a-numerical-value-wi"
## [1] "Current URL: https://stackoverflow.com/questions/76394793/prior-posterior-predictive-
distributions-for-an-unconstrained-constrained-ppo-mo"
## [1] "Current URL: https://stackoverflow.com/questions/76394787/how-to-label-levels-in-r"
## [1] "Current URL: https://stackoverflow.com/questions/76394767/is-there-a-way-to-add-an-ex
ternal-term-to-the-bekk-garch-process-in-r"
## [1] "Current URL: https://stackoverflow.com/questions/76394709/reshaping-data-frame-with-m
utiple-value-columns-from-wide-to-long"
## [1] "Current URL: https://stackoverflow.com/questions/76394485/grouping-data-with-conditio
n-in-r"
## [1] "Current URL: https://stackoverflow.com/questions/76394461/r-comparing-distance-calcul
ations"
## [1] "Current URL: https://stackoverflow.com/questions/76394455/function-with-shiny-survey"
## [1] "Current URL: https://stackoverflow.com/questions/76394451/why-is-sub-in-r-not-recogni
zing-the-pattern-that-i-provided-in-r"
## [1] "Current URL: https://stackoverflow.com/questions/76394390/how-to-replace-a-string-def
ined-by-starting-and-ending-index-by-another-string-i"
## [1] "Current URL: https://stackoverflow.com/questions/76394371/create-columns-based-on-the
-values-in-another-data-frame"
## [1] "Current URL: https://stackoverflow.com/questions/76394359/simulating-a-spatio-tempora
l-gaussian-process"
## [1] "Current URL: https://stackoverflow.com/questions/76394250/negative-gvif1-2df-values"
## [1] "Current URL: https://stackoverflow.com/questions/76394080/single-season-occupancy-mod
eling-in-unmarked-how-to-create-a-model-averaged-pred"
## [1] "Current URL: https://stackoverflow.com/questions/76394018/node-inconsistent-with-pare
nts"
## [1] "INPUT URL: https://stackoverflow.com/questions/tagged/r?tab=newest&page=2&pagesize=5
0"
## [1] "Current URL: https://stackoverflow.com/questions/76393934/how-do-add-sign-of-rows-in-
one-column-to-rows-in-new-column-in-r"
## [1] "Current URL: https://stackoverflow.com/questions/76393814/complete-missing-non-overla
```

```
pping-date-ranges"
## [1] "Current URL: https://stackoverflow.com/questions/76393742/using-partykit-to-predict-s
urvival-probabilities"
## [1] "Current URL: https://stackoverflow.com/questions/76393686/how-to-set-external-variabl
e-or-program-for-rstudio"
## [1] "Current URL: https://stackoverflow.com/questions/76393636/interpretation-of-roc-curve
-curving-early"
## [1] "Current URL: https://stackoverflow.com/questions/76393577/how-can-i-find-the-length-o
f-row-wise-set-differences-between-two-list-columns-u"
## [1] "Current URL: https://stackoverflow.com/questions/76393401/switching-columns-within-a-
geom-col-plot-in-r"
## [1] "Current URL: https://stackoverflow.com/questions/76393378/how-to-replace-multiple-occ
urrences-of-a-pattern-in-r"
## [1] "Current URL: https://stackoverflow.com/questions/76393250/bslibvalue-box-within-bs4da
sh-displays-not-as-intended"
## [1] "Current URL: https://stackoverflow.com/questions/76393244/how-can-i-disregard-the-the
me-for-1-slide-in-xaringan"
## [1] "Current URL: https://stackoverflow.com/questions/76393217/how-to-plot-multiple-plots-
in-r-for-different-variables"
## [1] "Current URL: https://stackoverflow.com/questions/76393207/rselenium-timeout-error-whi
le-trying-to-connect-to-server-running-on-docker-usi"
## [1] "Current URL: https://stackoverflow.com/questions/76392914/removing-unwanted-special-e
tc-and-html-a8-etc-characters-in-r"
## [1] "Current URL: https://stackoverflow.com/questions/76392902/how-to-launch-second-shiny-
modal-based-on-an-event-in-the-main-shiny-app-ui-and"
## [1] "Current URL: https://stackoverflow.com/questions/76392893/how-to-format-negative-numb
ers-with-parenthesis-with-openxlsx-in-r"
## [1] "Current URL: https://stackoverflow.com/questions/76392884/complex-conditional-df-subs
etting-with-nested-for-loops-in-r"
## [1] "Current URL: https://stackoverflow.com/questions/76392840/how-can-i-conditionally-ass
ign-values-to-a-column-in-data-table-using-a-function"
## [1] "Current URL: https://stackoverflow.com/questions/76392784/how-do-i-automatically-plot
-overlapping-curves-with-ggplot2"
## [1] "Current URL: https://stackoverflow.com/questions/76392725/setting-dynamic-yaxis-label
s-ggplot"
## [1] "Current URL: https://stackoverflow.com/questions/76392709/create-a-custom-legend-for-
facet-grid-variables"
## [1] "Current URL: https://stackoverflow.com/questions/76392631/r-using-group-by-for-all-va
lues"
## [1] "Current URL: https://stackoverflow.com/questions/76392610/how-do-i-fix-errors-when-re
naming-observations-in-r-for-filtering-purposes"
## [1] "Current URL: https://stackoverflow.com/questions/76392410/how-can-i-make-movies-from-
the-movies-dataset-with-multiple-genres-only-have-1"
## [1] "Current URL: https://stackoverflow.com/questions/76392409/how-can-i-add-the-overall-h
istogram-in-a-grouped-histogram-plot"
## [1] "Current URL: https://stackoverflow.com/questions/76392386/how-do-i-add-a-var-constrai
nt-to-portfolioanalytics-optimization-function-in-r"
## [1] "Current URL: https://stackoverflow.com/questions/76392367/write-to-google-sheet-skipp
ing-1st-row"
## [1] "Current URL: https://stackoverflow.com/questions/76392347/mixed-model-ancova-when-var
iables-are-not-of-the-same-length"
```

```
## [1] "Current URL: https://stackoverflow.com/questions/76392326/rmarkdown-how-to-use-result
s-asis-and-fig-show-hold-together"
## [1] "Current URL: https://stackoverflow.com/questions/76392259/how-do-i-plot-my-data-throu
gh-in-discrete-time-bins"
## [1] "Current URL: https://stackoverflow.com/questions/76392228/lmer-gives-estimates-for-no
n-existing-variables-after-equation-with-interaction"
## [1] "Current URL: https://stackoverflow.com/questions/76392177/how-to-assign-unique-observ
ations-to-values-within-a-group"
## [1] "Current URL: https://stackoverflow.com/questions/76392149/linear-programming-problem-
r-using-lpsolveapi"
## [1] "Current URL: https://stackoverflow.com/questions/76392055/how-to-match-multiple-occur
rences-of-strings-given-a-start-and-end-pattern-in-r"
## [1] "Current URL: https://stackoverflow.com/questions/76391960/r-studio-add-row-with-dynam
ic-field-name"
## [1] "Current URL: https://stackoverflow.com/questions/76391906/how-to-pass-an-unquoted-col
umn-name-to-a-custom-summary-function-min-in-r"
## [1] "Current URL: https://stackoverflow.com/questions/76391856/r-argument-is-not-a-charact
er-vector-enc2utf8path"
## [1] "Current URL: https://stackoverflow.com/questions/76391852/recreate-hist-binning-in-gg
plot2-with-geom-histogram"
## [1] "Current URL: https://stackoverflow.com/questions/76391835/format-graph-ggcuminc-cumul
ative-incidence"
## [1] "Current URL: https://stackoverflow.com/questions/76391830/three-way-interaction-in-re
gression-does-not-show-the-interaction-terms-and-esti"
## [1] "Current URL: https://stackoverflow.com/questions/76391751/if-match-between-column-id-
in-two-different-datasets-then-create-a-new-dataset"
## [1] "Current URL: https://stackoverflow.com/questions/76391676/r-how-to-mutate-strings-int
o-icons-within-one-cell"
## [1] "Current URL: https://stackoverflow.com/questions/76391552/how-to-extract-a-string-tha
t-starts-and-ends-with-a-specific-pattern-in-r"
## [1] "Current URL: https://stackoverflow.com/questions/76391482/in-ggplot2-print-an-express
ion-in-the-facet-wrap"
## [1] "Current URL: https://stackoverflow.com/questions/76391426/what-directory-do-i-use-to-
install-libraries-for-a-container-image-with-r-for-an"
## [1] "Current URL: https://stackoverflow.com/questions/76391151/how-program-to-join-2-diffe
rent-tables-based-on-which-one-has-the-highest-number"
## [1] "Current URL: https://stackoverflow.com/questions/76391100/r-add-current-row-value-of-
adjacent-column-with-previous-row-value-of-current"
## [1] "Current URL: https://stackoverflow.com/questions/76391093/r-analysis-of-number-of-mat
ure-animals"
## [1] "Current URL: https://stackoverflow.com/questions/76390879/how-can-i-edit-shiny-server
s-sockjs-websocket-to-add-content-type-headers-to-it"
## [1] "Current URL: https://stackoverflow.com/questions/76390817/is-it-possible-to-conduct-a
nova-with-a-frequency-variable"
## [1] "Current URL: https://stackoverflow.com/questions/76390798/undefined-columns-selected-
when-subsetting-in-r"
## [1] "INPUT URL: https://stackoverflow.com/questions/tagged/r?tab=newest&page=3&pagesize=5
0"
## [1] "Current URL: https://stackoverflow.com/questions/76390501/is-there-a-way-to-untar-a-f
ile-and-merge-all-the-extracted-files-into-one-for-ef"
## [1] "Current URL: https://stackoverflow.com/questions/76390478/import-decimal-interval-of-
```

```
months-with-lubridate-to-convert-into-days"
## [1] "Current URL: https://stackoverflow.com/questions/76390426/how-to-change-the-y-axis-to
-scientific-annotation-problem-with-characters"
## [1] "Current URL: https://stackoverflow.com/questions/76390374/web-crawling-with-naver-cli
ent-id-why-is-my-id-invalid-and-how-can-i-fix-it"
## [1] "Current URL: https://stackoverflow.com/questions/76390329/change-x-or-y-position-of-d
ensity-plot"
## [1] "Current URL: https://stackoverflow.com/questions/76390299/r-tapply-how-to-use-index-n
ames-as-a-fun-additional-argument"
## [1] "Current URL: https://stackoverflow.com/questions/76390196/error-in-seq-defaultfrom-1-
to-total-window-1-by-absstep"
## [1] "Current URL: https://stackoverflow.com/questions/76389748/how-do-i-calculate-the-omeg
a-values-for-my-bifactor-model"
## [1] "Current URL: https://stackoverflow.com/questions/76389710/how-do-i-apply-conditional-
statemtents-to-create-a-new-raster-layer-based-on-3-g"
## [1] "Current URL: https://stackoverflow.com/questions/76389697/clean-data-in-r-from-image"
## [1] "Current URL: https://stackoverflow.com/questions/76389613/r-append-matrix-rows-if-con
dition-is-met"
## [1] "Current URL: https://stackoverflow.com/questions/76389515/retrieve-every-value-betwee
n-an-alphanumeric-range-in-r-using-ifelse"
## [1] "Current URL: https://stackoverflow.com/questions/76389510/add-a-column-with-a-string-
value-based-on-other-column-values-r"
## [1] "Current URL: https://stackoverflow.com/questions/76389468/how-to-subtract-values-betw
een-months-for-each-group-and-each-year-separately-in"
## [1] "Current URL: https://stackoverflow.com/questions/76389386/r-function-that-summarize-r
ows-grouped-by-but-disregards-duplicate-strings-and"
## [1] "Current URL: https://stackoverflow.com/questions/76389243/is-it-possible-to-bypass-th
e-9-argument-limit-in-commandargs-for-r-programs"
## [1] "Current URL: https://stackoverflow.com/questions/76389056/how-to-achieve-k-anonymity-
in-r-changing-python-to-r-function"
## [1] "Current URL: https://stackoverflow.com/questions/76389050/cleaning-data-in-r-by-using
-a-reference-date"
## [1] "Current URL: https://stackoverflow.com/questions/76389028/is-there-a-way-to-add-secti
on-labels-to-the-x-axis-of-a-plot-using-ggplot2"
## [1] "Current URL: https://stackoverflow.com/questions/76388858/change-the-size-and-orienta
tion-of-legend-title-while-plotting-raster"
## [1] "Current URL: https://stackoverflow.com/questions/76388339/r-data-table-lost-rows-afte
r-order"
## [1] "Current URL: https://stackoverflow.com/questions/76388303/how-can-i-get-the-primary-k
ey-of-a-selected-option-from-a-dataframe-based-select"
## [1] "Current URL: https://stackoverflow.com/questions/76388156/customise-and-order-legends
-in-survival-analysis-in-r"
## [1] "Current URL: https://stackoverflow.com/questions/76388032/convert-data-frame-with-cha
racter-column-to-data-frame-with-integer-column"
## [1] "Current URL: https://stackoverflow.com/questions/76387964/h2o-mean-residual-deviance-
for-poisson-family"
## [1] "Current URL: https://stackoverflow.com/questions/76387937/rearranging-a-dataframe-cre
ation-of-new-columns-and-pivot-to-wide-format-based"
## [1] "Current URL: https://stackoverflow.com/questions/76387472/tidyrpivot-longer-with-dupl
icate-problems-with-no-apparent-duplicate-column"
## [1] "Current URL: https://stackoverflow.com/questions/76387441/obtaining-predicted-values-
```

```
of-the-outcome-variable-using-margins-command"
## [1] "Current URL: https://stackoverflow.com/questions/76387313/expectation-step-in-gaussia
n-mixture-model-for-matrix-data-not-producing-proper"
## [1] "Current URL: https://stackoverflow.com/questions/76387295/error-in-installing-glmnet-
4-1-7-in-r-4-2-3-in-rhel-7-9"
## [1] "Current URL: https://stackoverflow.com/questions/76387189/fatal-error-relating-to-inc
lude-s-h-when-installing-r-scalop-package"
## [1] "Current URL: https://stackoverflow.com/questions/76387115/ggplot2-dodge-overlapping-w
hen-y-axis-is-not-count-frequency"
## [1] "Current URL: https://stackoverflow.com/questions/76387096/why-are-the-fitting-results
-of-the-arima-and-glm-function-different"
## [1] "Current URL: https://stackoverflow.com/questions/76386943/incompatible-shapes-1024-3-
vs-1024-1024-in-neural-network-training"
## [1] "Current URL: https://stackoverflow.com/questions/76386904/how-to-merge-columns-and-ro
ws-in-r-so-that-i-can-calculate-the-average-number-o"
## [1] "Current URL: https://stackoverflow.com/questions/76386871/how-do-i-automatically-set-
a-custom-path-for-r-packages-when-using-ms-visual-stu"
## [1] "Current URL: https://stackoverflow.com/questions/76386843/how-to-use-dplyr-verbs-rena
me-and-mutate-in-a-same-chunk"
## [1] "Current URL: https://stackoverflow.com/questions/76386840/why-is-remove-punct-not-rem
oving-apostrophes-when-tokenizing-a-corpus-in-quanted"
## [1] "Current URL: https://stackoverflow.com/questions/76386786/after-stat-and-glue-do-not-
work-together"
## [1] "Current URL: https://stackoverflow.com/questions/76386732/how-do-i-stop-highlighted-r
-code-runs-twice"
## [1] "Current URL: https://stackoverflow.com/questions/76386620/fisher-exact-using-tidyr-or
-dplyr-approach"
## [1] "Current URL: https://stackoverflow.com/questions/76386429/how-can-i-download-pdfs-fro
m-a-website-that-stores-them-on-aws-using-rvest-in-r"
## [1] "Current URL: https://stackoverflow.com/questions/76386409/smoothing-splines-is-there-
a-way-to-automatically-add-s-to-each-column-name"
## [1] "Current URL: https://stackoverflow.com/questions/76386402/compatibility-issue-between
-dplyr-versions-in-rstudio-error-with-mutate-functio"
## [1] "Current URL: https://stackoverflow.com/questions/76386356/how-can-i-optimize-this-pat
tern-matching-function-to-handle-large-data-sets-in-r"
## [1] "Current URL: https://stackoverflow.com/questions/76386347/how-can-i-convert-a-data-ta
ble-to-its-logarithm-1-base-2-form-in-r-while-ign"
## [1] "Current URL: https://stackoverflow.com/questions/76386335/how-do-i-collapse-data-fram
e-rows-if-values-are-identical-in-r"
## [1] "Current URL: https://stackoverflow.com/questions/76386256/export-a-plotly-graph-in-r-
to-powerpoint"
## [1] "Current URL: https://stackoverflow.com/questions/76386162/is-there-a-way-to-merge-two
-mutated-dataframes-together"
## [1] "Current URL: https://stackoverflow.com/questions/76386109/how-can-i-control-for-multi
ple-categorical-variables-in-my-regression-analysis-o"
## [1] "INPUT URL: https://stackoverflow.com/questions/tagged/r?tab=newest&page=9815&pagesize
=50"
```

```
## Warning in get_reputation(doc): Reputation scores do not match! Nrep: 50
## OldRep: 45
```

```
## [1] "Current URL: https://stackoverflow.com/questions/1658032/draw-hyperplane-in-r"
## [1] "Current URL: https://stackoverflow.com/questions/1656026/why-doesnt-r-add-the-title-a
t-the-top-of-the-page"
## [1] "Current URL: https://stackoverflow.com/questions/1655792/r-lag-over-missing-data"
## [1] "Current URL: https://stackoverflow.com/questions/1655454/how-do-you-make-a-new-datase
t-given-a-set-of-vectors"
## [1] "Current URL: https://stackoverflow.com/questions/1653710/r-analogous-to-sql-inner-joi
n-selection"
## [1] "Current URL: https://stackoverflow.com/questions/1653271/how-do-you-find-the-median-o
f-2-columns-using-r"
## [1] "Current URL: https://stackoverflow.com/questions/1652522/rbind-dataframes-in-a-list-o
f-lists"
## [1] "Current URL: https://stackoverflow.com/questions/1649503/strange-problem-with-rpy2"
## [1] "Current URL: https://stackoverflow.com/questions/1647236/matching-strings-across-colu
mns-in-r"
## [1] "Current URL: https://stackoverflow.com/questions/1644661/add-a-vertical-line-with-dif
ferent-intercept-for-each-panel-in-ggplot2"
## [1] "Current URL: https://stackoverflow.com/questions/1642201/your-experiences-with-matlab
-f-r-for-data-analysis-and-modeling-algorithms"
## [1] "Current URL: https://stackoverflow.com/questions/1642119/consensus-tree-or-bootstrap-
proportions-from-multiple-hclust-objects"
## [1] "Current URL: https://stackoverflow.com/questions/1641488/lattice-problems-lattice-obj
ects-coming-from-jags-but-device-cant-be-set"
## [1] "Current URL: https://stackoverflow.com/questions/1635278/saving-a-data-frame-as-a-bin
ary-file"
## [1] "Current URL: https://stackoverflow.com/questions/1632772/unseen-factor-levels-when-ap
pending-new-records-with-unseen-string-values-to-a-d"
## [1] "Current URL: https://stackoverflow.com/questions/1630724/can-i-gracefully-include-for
matted-sql-strings-in-an-r-script"
## [1] "Current URL: https://stackoverflow.com/questions/1628383/finding-the-minimum-differen
ce-between-each-element-of-one-vector-and-another-ve"
## [1] "Current URL: https://stackoverflow.com/questions/1622797/debugging-littler-rscripts"
## [1] "Current URL: https://stackoverflow.com/questions/1622419/all-the-directions-perpendic
ular-to-hyperplane-through-p-data-points"
## [1] "Current URL: https://stackoverflow.com/questions/1621848/looping-through-a-column-in-
r"
## [1] "Current URL: https://stackoverflow.com/questions/1617061/include-levels-of-zero-count
-in-result-of-table"
## [1] "Current URL: https://stackoverflow.com/questions/1616983/building-r-packages-using-al
ternate-gcc"
## [1] "Current URL: https://stackoverflow.com/questions/1614889/how-do-you-know-which-functi
ons-in-r-are-flagged-for-debugging"
## [1] "Current URL: https://stackoverflow.com/questions/1614331/write-plot-text-binary-into-
variable"
## [1] "Current URL: https://stackoverflow.com/questions/1608130/equivalent-of-throw-in-r"
## [1] "Current URL: https://stackoverflow.com/questions/1607413/using-summary-lm-function-in
-rapache"
## [1] "Current URL: https://stackoverflow.com/questions/1594121/how-do-i-best-simulate-an-ar
bitrary-univariate-random-variate-using-its-probabil"
## [1] "Current URL: https://stackoverflow.com/questions/1586744/complex-object-initializatio
n-scope-issues-with-nested-functions"
```

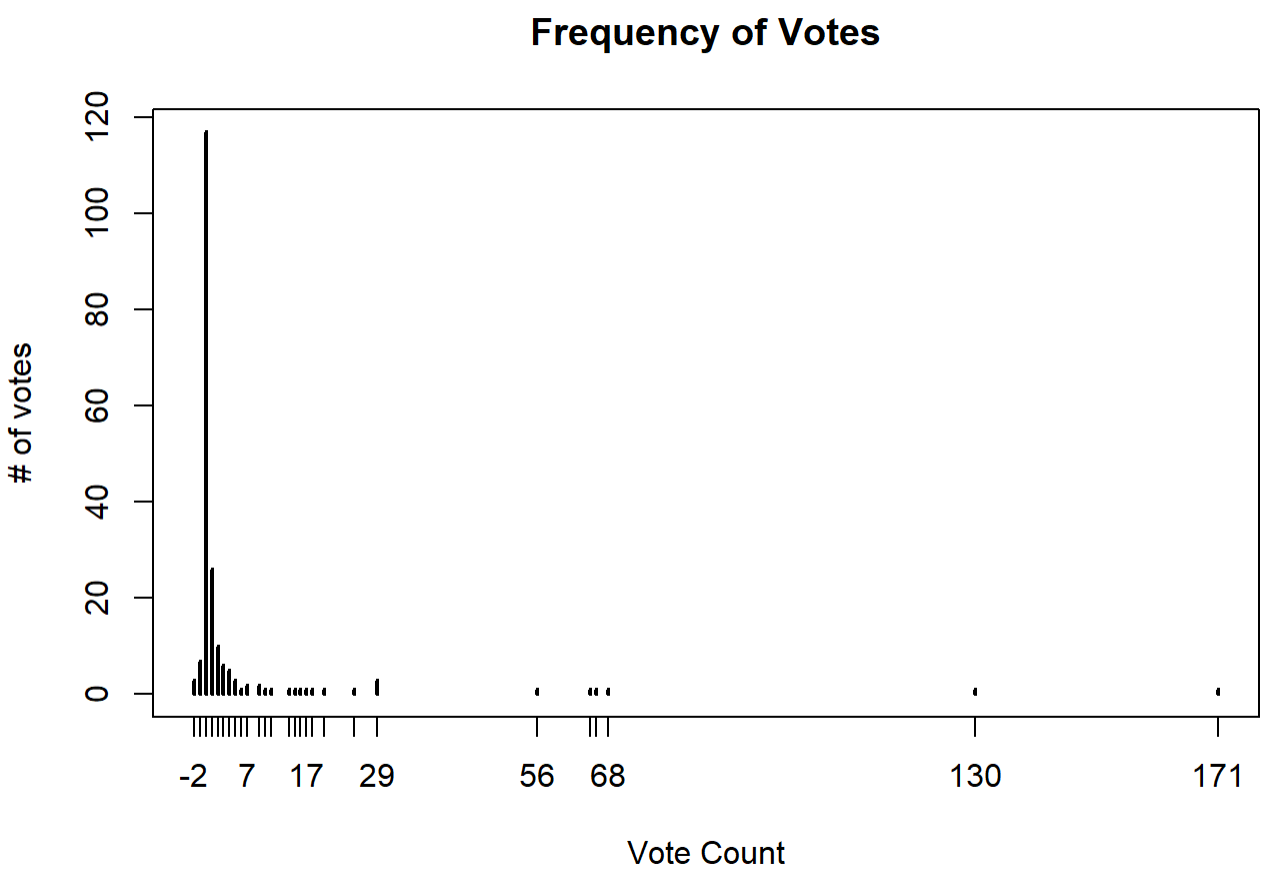
```
## [1] "Current URL: https://stackoverflow.com/questions/1581232/add-commas-into-number-for-o
utput"
## [1] "Current URL: https://stackoverflow.com/questions/1580308/r2html-number-formatting"
## [1] "Current URL: https://stackoverflow.com/questions/1577480/r-occurrence-times-binary-se
quence"
## [1] "Current URL: https://stackoverflow.com/questions/1576201/how-do-i-find-peak-values-ro
w-numbers"
## [1] "Current URL: https://stackoverflow.com/questions/1576075/kohonen-som-maps-in-r-tutori
al"
## [1] "Current URL: https://stackoverflow.com/questions/1570379/adding-stat-smooth-in-to-onl
y-1-facet-in-ggplot2"
## [1] "Current URL: https://stackoverflow.com/questions/1570263/alternatives-to-using-text-t
o-adding-text-to-a-plot"
## [1] "Current URL: https://stackoverflow.com/questions/1570050/cross-platform-zip-file-crea
tion"
## [1] "Current URL: https://stackoverflow.com/questions/1568511/how-do-i-sort-one-vector-bas
ed-on-values-of-another"
## [1] "Current URL: https://stackoverflow.com/questions/1567718/getting-a-function-name-as-a
-string"
## [1] "Current URL: https://stackoverflow.com/questions/1563961/how-to-find-top-n-of-records
-in-a-column-of-a-dataframe-using-r"
## [1] "Current URL: https://stackoverflow.com/questions/1562124/how-can-i-merge-many-data-fr
ames-from-csv-files-when-the-id-column-is-implied"
## [1] "Current URL: https://stackoverflow.com/questions/1560397/how-to-view-the-contents-of-
parsed-r-functions"
## [1] "Current URL: https://stackoverflow.com/questions/1559724/can-you-use-the-lapply-funct
ion-to-alter-the-value-of-input"
## [1] "Current URL: https://stackoverflow.com/questions/1557137/double-for-loops-in-r"
## [1] "Current URL: https://stackoverflow.com/questions/1554942/placement-of-axis-labels-at-
minor-breaks-with-ggplot2"
## [1] "Current URL: https://stackoverflow.com/questions/1552438/working-with-data-frames-in-
r-using-sas-code-to-describe-what-i-wantr"
## [1] "Current URL: https://stackoverflow.com/questions/1551554/constrained-least-squares"
## [1] "Current URL: https://stackoverflow.com/questions/1548913/time-series-in-r"
## [1] "Current URL: https://stackoverflow.com/questions/1545591/how-to-make-a-ggplot-line-pl
ot-with-different-color-segments-conditional-on-di"
## [1] "Current URL: https://stackoverflow.com/questions/1545302/multiple-data-points-in-one-
r-ggplot2-plot"
## [1] "Current URL: https://stackoverflow.com/questions/1544907/melt-to-two-variable-column
s"
## [1] "INPUT URL: https://stackoverflow.comTRUE"
```

```
# write.csv(final_dataframes[[1]], 'E:\\College\\UC Davis\\STA141B\\HW4\\questions.csv')
# write.csv(final_dataframes[[2]], 'E:\\College\\UC Davis\\STA141B\\HW4\\answers.csv')
# write.csv(final_dataframes[[3]], 'E:\\College\\UC Davis\\STA141B\\HW4\\comments.csv')
```

Analysis graphs for questions:

```
q_df = final_dataframes[[1]]
a_df = final_dataframes[[2]]
c_df = final_dataframes[[3]]

#Plot votes
plot(table(q_df$votes), main = 'Frequency of Votes', ylab = '# of votes', xlab = 'Vote Count
')
```

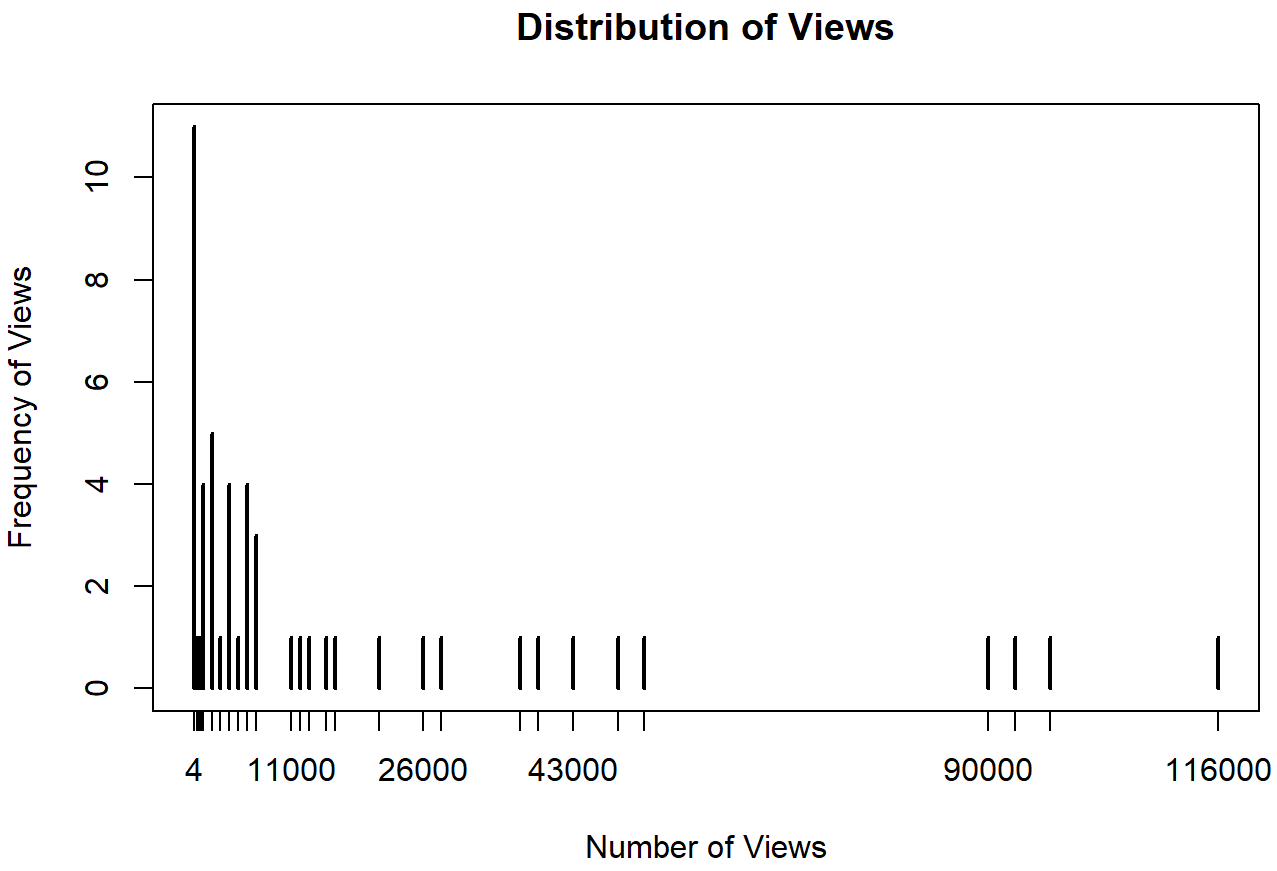



```
kable(table(q_df$votes), col.names = c('Vote Count', 'Frequency'))
```

Vote Count	Frequency
-1	7
-2	3
0	117
1	26
10	1
11	1
130	1

Vote Count	Frequency
14	1
15	1
16	1
17	1
171	1
18	1
2	10
20	1
25	1
29	3
3	6
4	5
5	3
56	1
6	1
65	1
66	1
68	1
7	2
9	2

```
#Fix the abbreviations of the Views
q_df$views = sapply(q_df$views, function(x) if (grepl('k', x) == TRUE){x = as.numeric(substr
(x, 1, nchar(x) - 1)) * 1000} else {x = as.numeric(x)})
plot(table(q_df$views), main = 'Distribution of Views', ylab = 'Frequency of Views', xlab = '
Number of Views')
```




```
kable(table(q_df$views), col.names = c('View Count', 'Frequency'))
```

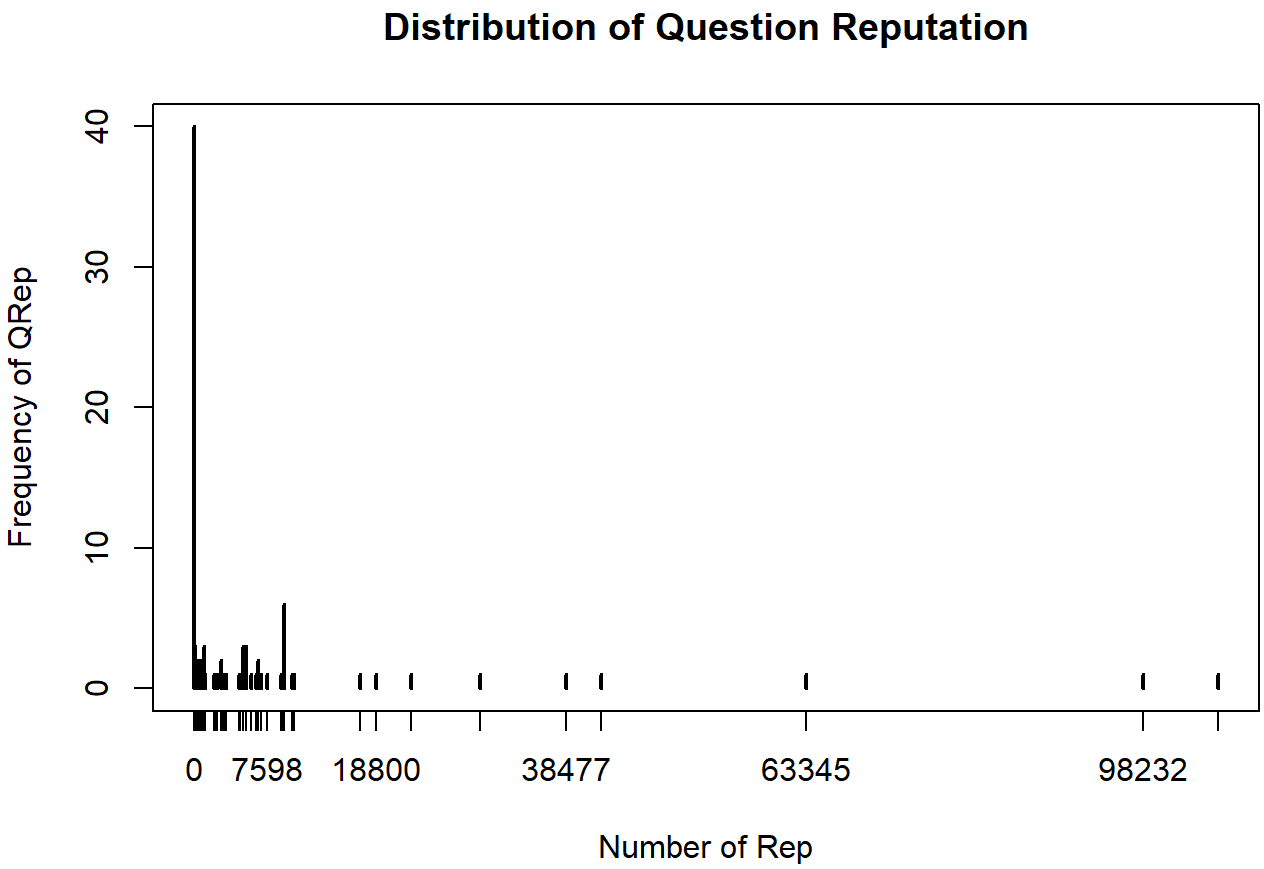
View Count	Frequency
4	1
6	1
7	2
8	1
9	1
10	3
11	5
12	1
13	5
14	5
15	5
16	4

View Count	Frequency
17	4
18	3
19	4
20	4
21	4
22	3
23	4
24	3
25	10
26	3
27	2
28	4
29	4
30	1
31	5
32	2
33	11
34	8
35	1
36	1
37	6
38	4
39	4
40	1
41	2
42	1
43	1
44	3
45	1
46	1

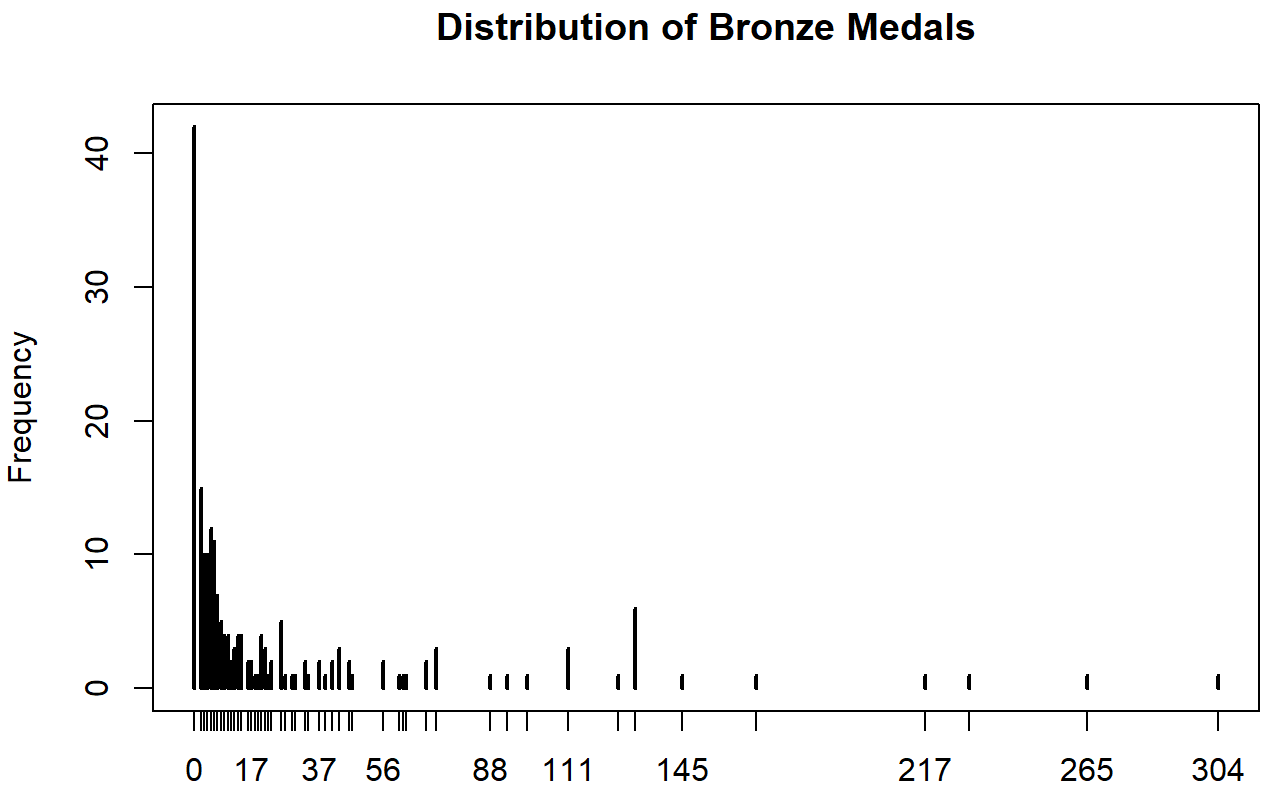
View Count	Frequency
47	2
50	2
52	1
55	2
56	1
57	1
58	1
64	1
378	1
380	1
391	1
451	1
548	1
579	1
588	1
650	1
706	1
849	1
998	1
1000	4
2000	5
3000	1
4000	4
5000	1
6000	4
7000	3
11000	1
12000	1
13000	1
15000	1

View Count	Frequency
16000	1
21000	1
26000	1
28000	1
37000	1
39000	1
43000	1
48000	1
51000	1
90000	1
93000	1
97000	1
116000	1

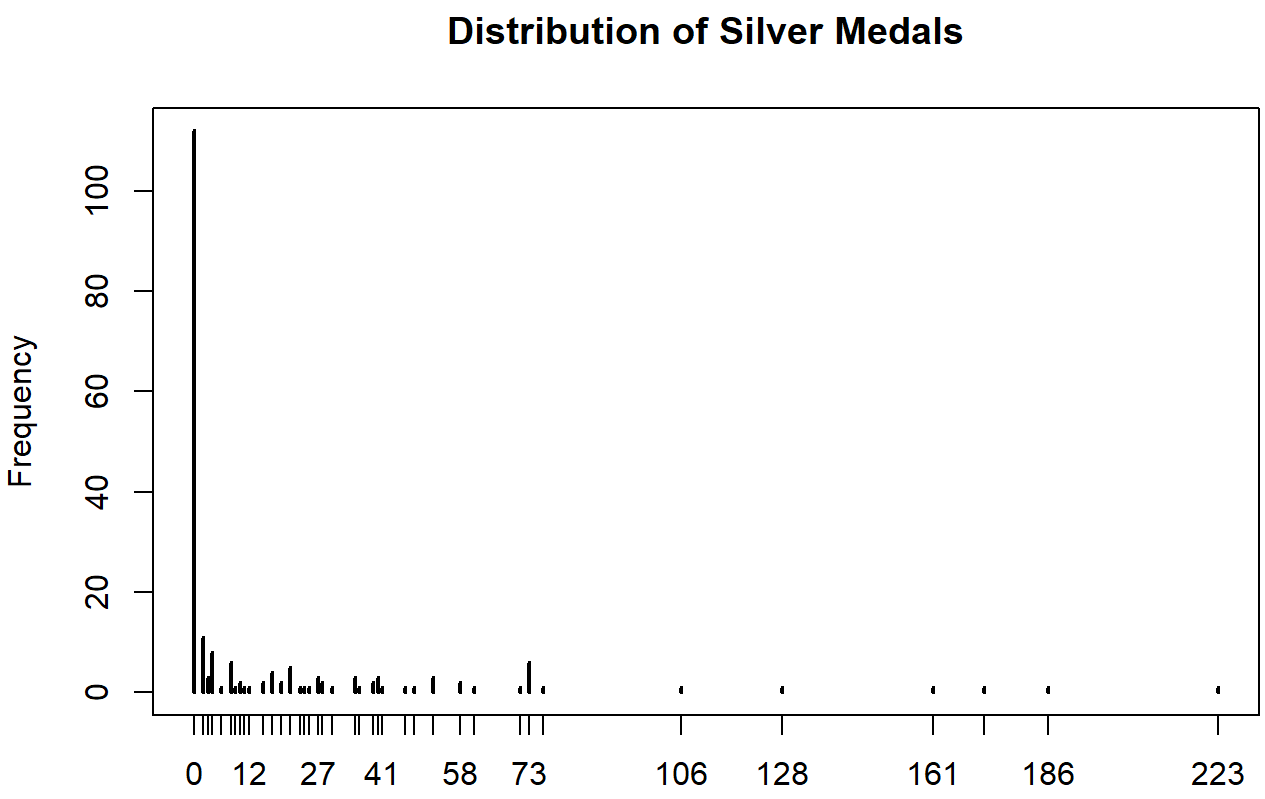
```
q_df$reputation = sapply(q_df$reputation, function(x) if (grepl('k', x) == TRUE){x = as.numer
ic(substr(x, 1, nchar(x) - 1)) * 1000} else {x = x})
q_df$reputation = sapply(q_df$reputation, function(x) if (nchar(x) < 1){x = 0} else {as.numer
ic(gsub(',', '', x))})
plot(table(q_df$reputation), main = 'Distribution of Question Reputation', ylab = 'Frequency
of QRep', xlab = 'Number of Rep')
```



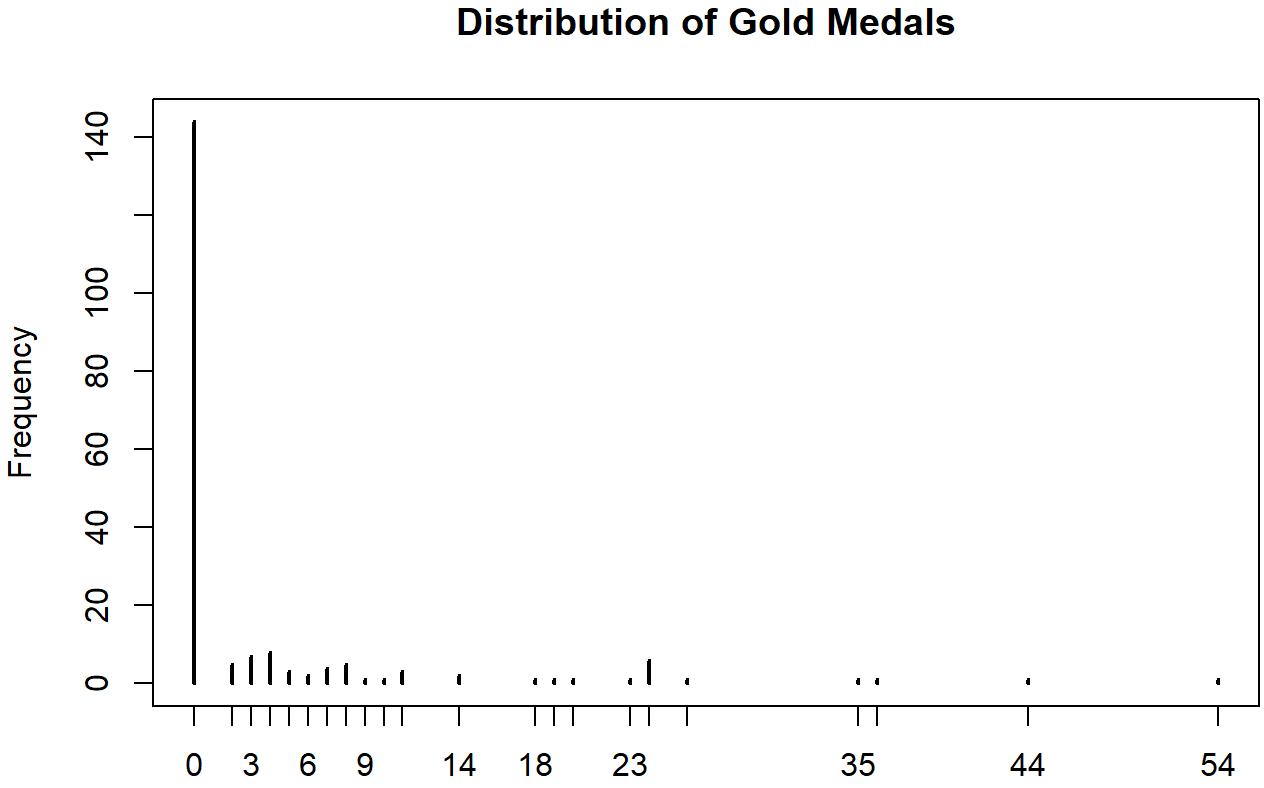

```
plot(table(q_df$badgeBronze), main = 'Distribution of Bronze Medals', ylab = 'Frequency')
```




```
plot(table(q_df$badgeSilver), main = 'Distribution of Silver Medals', ylab = 'Frequency')
```

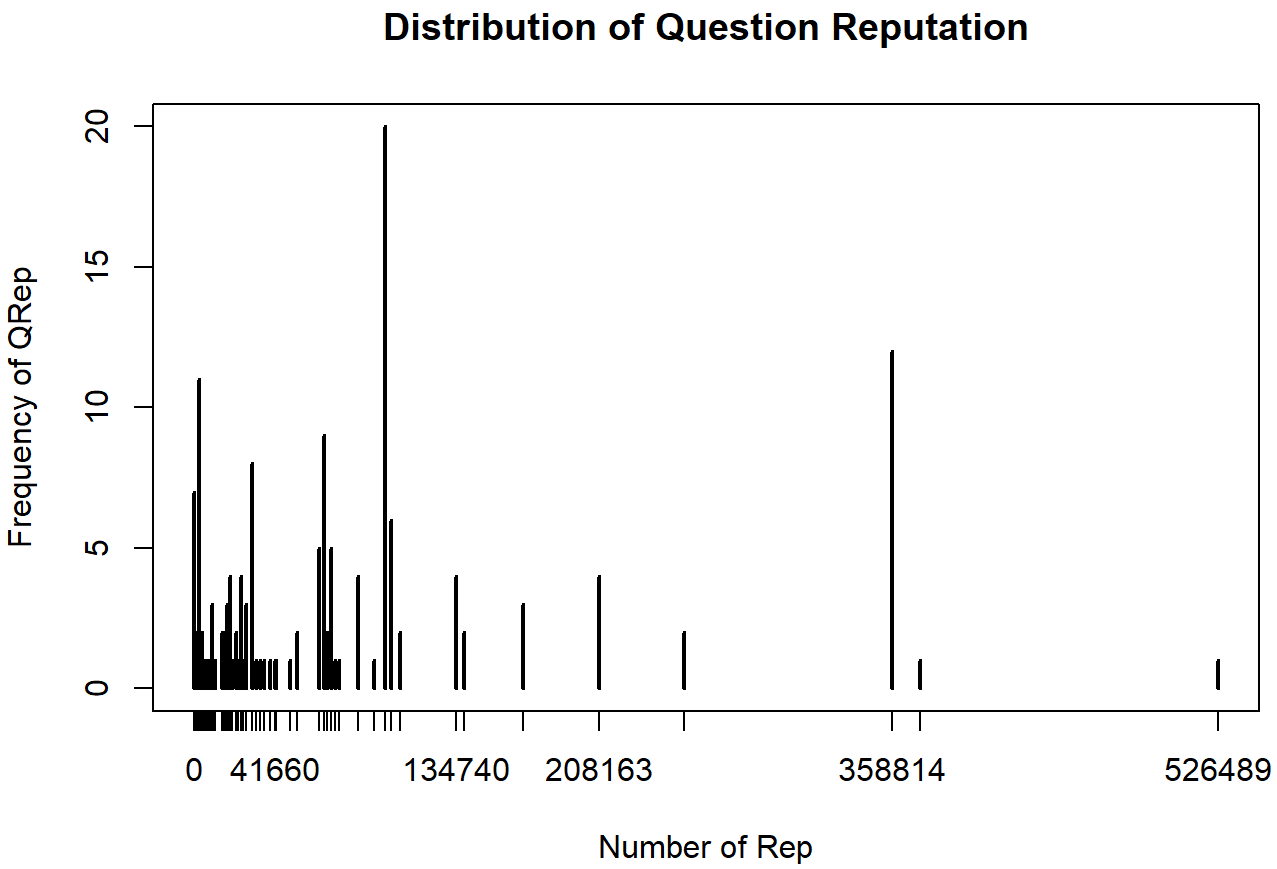



```
plot(table(q_df$badgeGold), main = 'Distribution of Gold Medals', ylab = 'Frequency')
```

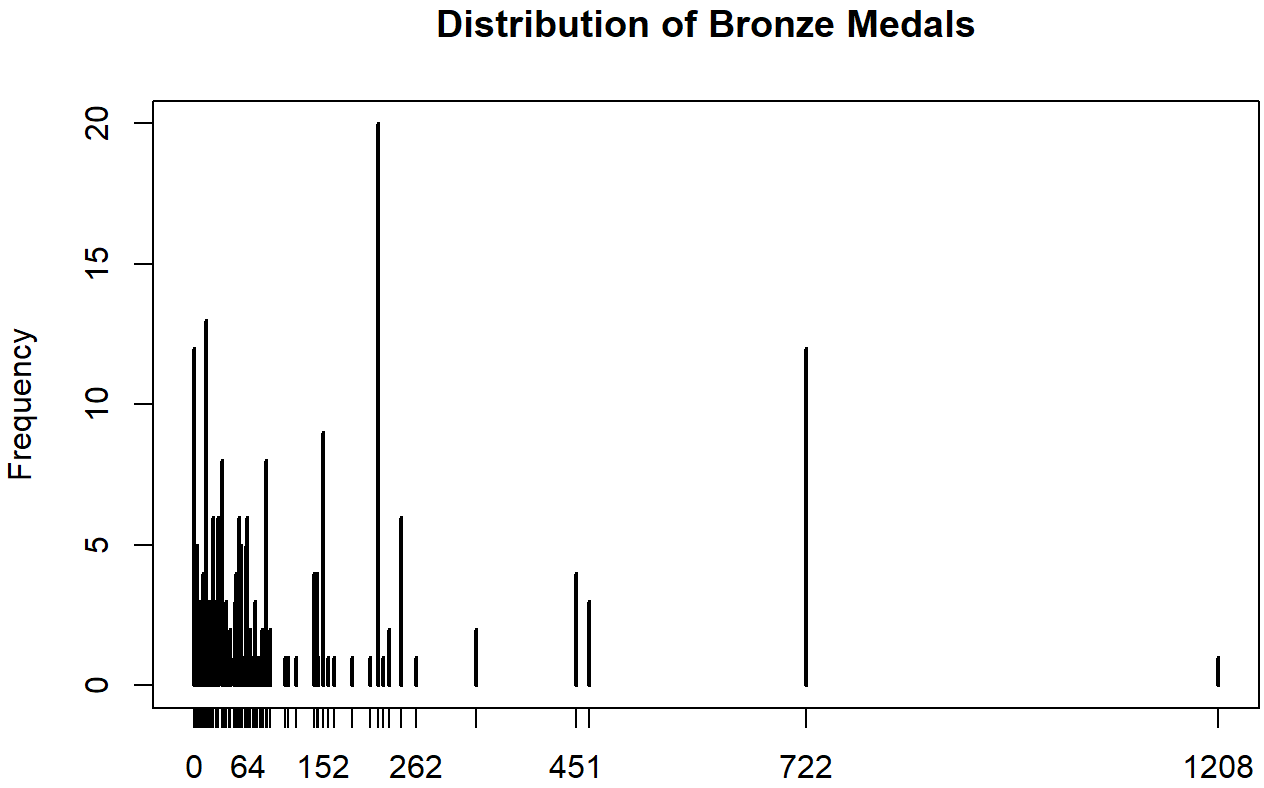


Answers Analysis

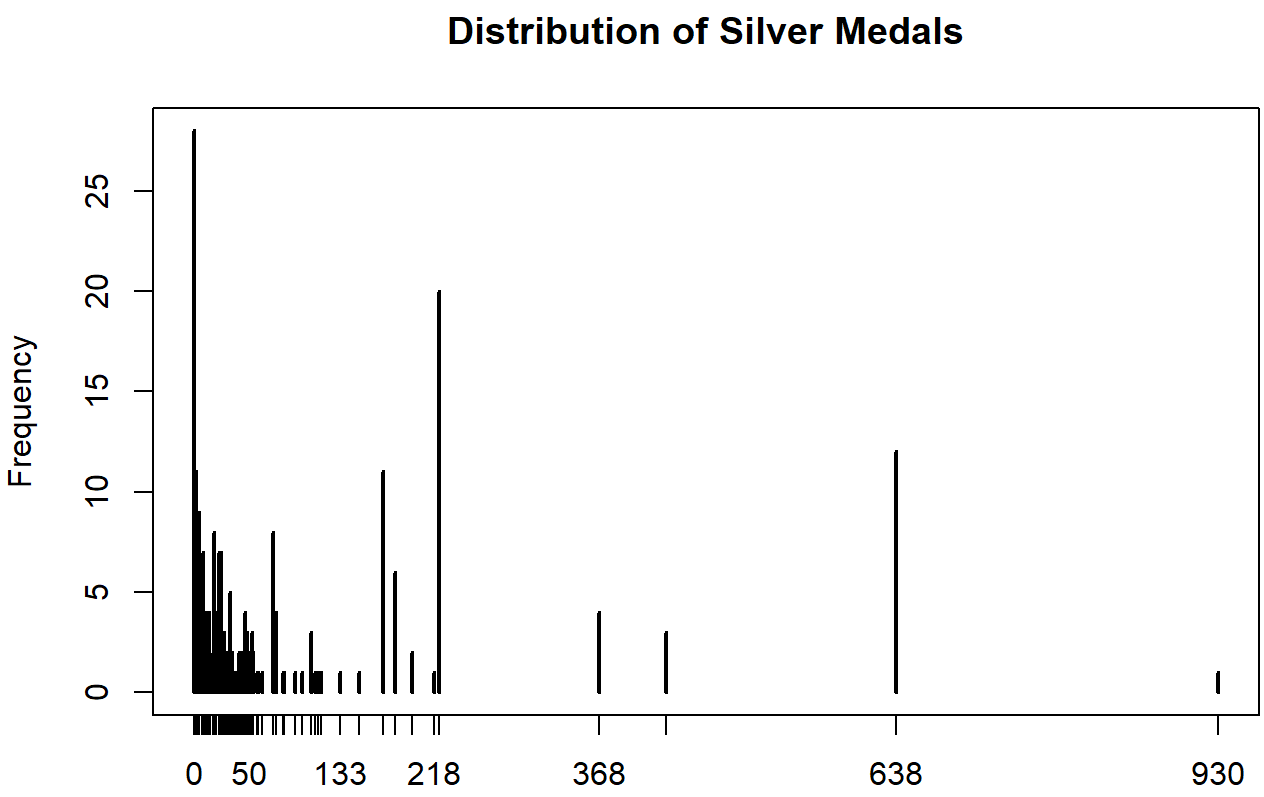
```
a_df$Reputation = sapply(a_df$Reputation, function(x) if (grepl('k', x) == TRUE){x = as.numer
ic(substr(x, 1, nchar(x) - 1)) * 1000} else {x = x})
a_df$Reputation = sapply(a_df$Reputation, function(x) if (nchar(x) < 1){x = 0} else {as.numer
ic(gsub(',', '', x))})
plot(table(a_df$Reputation), main = 'Distribution of Question Reputation', ylab = 'Frequency
of QRep', xlab = 'Number of Rep')
```



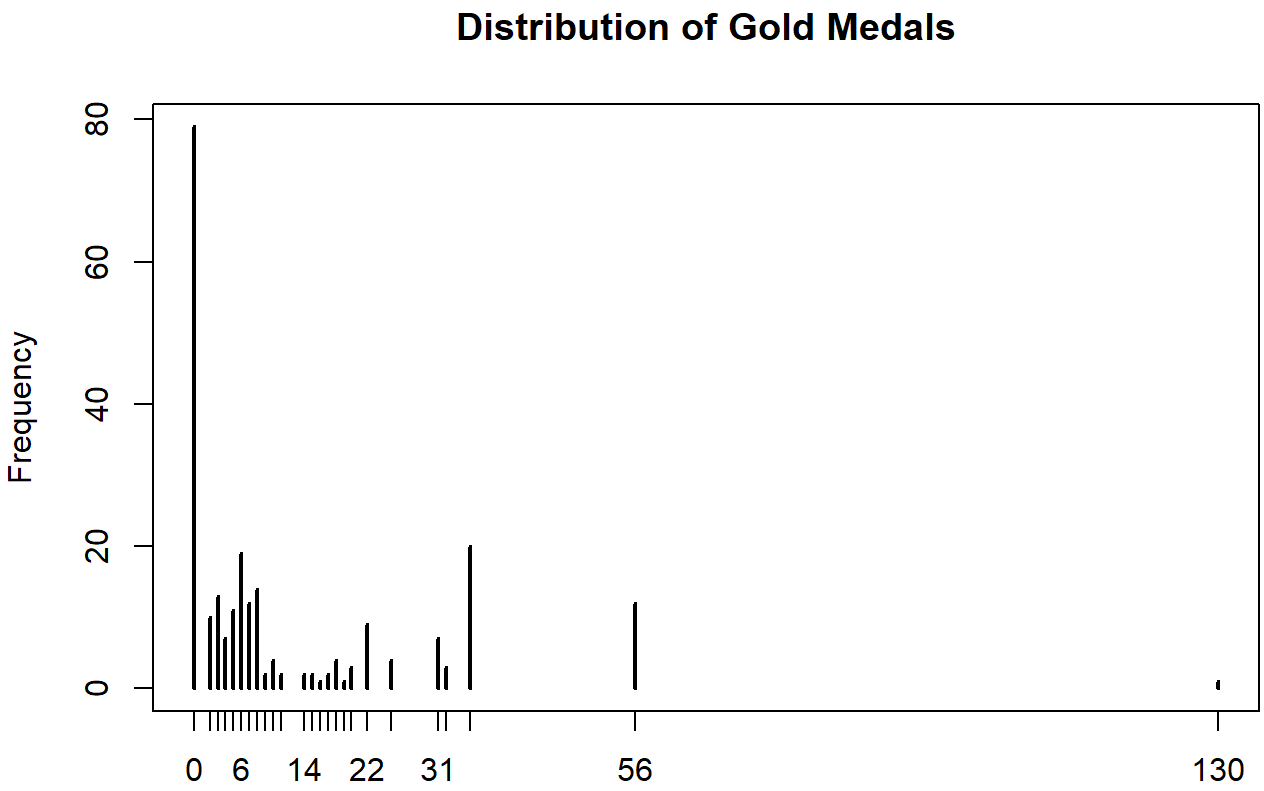

```
plot(table(a_df$BronzeMedals), main = 'Distribution of Bronze Medals', ylab = 'Frequency')
```




```
plot(table(a_df$SilverMedals), main = 'Distribution of Silver Medals', ylab = 'Frequency')
```




```
plot(table(a_df$GoldMedals), main = 'Distribution of Gold Medals', ylab = 'Frequency')
```



Print out tables

```
head(q_df)
```

```
##         id views votes                                tags
## 1 76398634     6     0 r,plot,linear-regression,diagnostics
## 2 76398607     4     0             r,gis,raster,rasterizing
## 3 76398561    14     0                  r,dplyr,sequential
## 4 76398512    10     0              r,text,stringr,pdftools
## 5 76398472    25     1               r,regex,string,replace
## 6 76398388    15     0                           r,ggplot2
##          question-time OwnerUsername  ownerId reputation
## 1 2023-06-04 01:33:11Z     bison2178  3135514        713
## 2 2023-06-04 01:15:59Z       procyon 21996869          1
## 3 2023-06-04 00:48:01Z       ghaines 16675997          1
## 4 2023-06-04 00:18:22Z   user6542495  6542495         11
## 5 2023-06-03 23:59:13Z        Adrian  3391549       9287
## 6 2023-06-03 23:19:39Z       esteban 16466230        102
##
body
## 1
\r\n                \r\nI'm using the check_model function within the performance library in
r to check the assumptions of a simple linear regression model using lm.\nI am fitting 3 mode
ls on different subsets of the data.\ndata("mtcars")\nhead("mtcars")\n\ntable(mtcars$cyl, use
NA = "ifany")\n\nfoo <- mtcars %>%\n          group_by(cyl) %>%\n             nest() %>%\n
mutate(model= map(data, lm(mpg ~ hp + wt, data = .))\n\nfoo %>%\n    {map(.$model, summary)}\
n\nI want check_model to plot all the 6 diagnostic plots per model in one row. Which means 3
rows, 6 plots in each row, using the model summaries which is stored as a list object.\ncheck
_model(foo$model..cylinder 4)...all 6 plots in 1 row\ncheck_model(foo$model..cylinder 6)...al
l 6 plots in 1 row\ncheck_model(foo$model..cylinder 8)...all 6 plots in 1 row\n\nAny suggesti
ons for how to accomplish this is much appreciated. Thanks.\n
## 2 \r\n                \r\nI'm trying to produce a presence/absence raster from a set of co
ordinates. I've turned the coordinates into a polygon, but when I use fasterize it produces a
copy of the template shape filled with 0s.\nI think there's a problem with the line reading i
n the presence data (which is just a column of 1s named "yes") because there isn't an error a
nd a raster is made, but I don't know how to fix it and I'm new to working with spatial dat
a.\nThe code I'm using is:\nparaster <- fasterize(paframe, cropbgEU,\n
field = "yes",\n                         background = 0) %>% \n  mask(cropbgEU)\n\nThese are t
he details for the polygon paframe, and the template raster cropbgEU:\n> paframe \n\nSimple f
eature collection with 2103 features and 1 field\nGeometry type: POLYGON\nDimension:     XY\n
Bounding box:  xmin: 1.576116 ymin: 41.1864 xmax: 28.48484 ymax: 70.51157\nGeodetic CRS:  WGS
84\nFirst 10 features:\n        yes                       geometry\n1         1 POLYGON ((17.
85235 59.86206...\n2         1 POLYGON ((15.26105 59.28842...\n3         1 POLYGON ((6.511416
52.38192...\n4         1 POLYGON ((22.4303 60.42065,...\n5         1 POLYGON ((20.73348 50.34
674...\n6         1 POLYGON ((21.05561 55.90442...\n7         1 POLYGON ((25.47543 61.041
7,...\n8         1 POLYGON ((24.82326 60.19986...\n9         1 POLYGON ((13.92114 53.0886
7...\n10        1 POLYGON ((20.99356 56.31544...\n\n> cropbgEU\n\nclass      : RasterLayer \n
dimensions : 4260, 4620, 19681200  (nrow, ncol, ncell)\nresolution : 0.008333333, 0.008333333
(x, y)\nextent     : -10, 28.5, 36, 71.5  (xmin, xmax, ymin, ymax)\ncrs        : +proj=longla
t +datum=WGS84 +no_defs \nsource     : r_tmp_2023-06-04_002524_22764_76822.grd \nnames      :
layer \nvalues     : -11.09583, 19.84583  (min, max)\n\nThe extent and CRS line up as far as
I can tell, which seemed to be the problem in a similar question where the raster was just NA
s. There was a different question using the rasterize function, but I don't know how to apply
that to fasterize: calling the data with field = paframe@data[,"yes"] returns the error:\nErr
or in h(simpleError(msg, call)) : \n  error in evaluating the argument 'x' in selecting a met
```

hod for function 'mask': trying to get slot "data" from an object (class "sf") that is not an S4 object \n\nI'm not sure if the problem could be further back when creating the polygon, but the code was:\npaframe <- presence %>% \n filter(yes %in% 1) %>% \n st_as_sf(coords=c("decimalLongitude","decimalLatitude"),\n   crs = 4326) %>% \n # convert to an sf polygon object by buffering the points\n st_buffer(1) \n\nWith the original presence dataframe looking like:\n> head(presence)\n   decimalLatitude decimalLongitude yes\n26   59.86205   17.85235   1\n31   59.28843   15.26104   1\n95   52.38192   6.51143   1\n128   60.42064   22.43029   1\n189   50.34675   20.73349   1\n190   55.90442   21.05563   1\n\nFor reference, I'm following the instructions in the blog post here by Amy Whitehead.\nIf anyone knows how to fix this, I would really appreciate it :)\n
3
\r\n   \r\nI have mark-recapture data in long form, and I want a column counting the number of times each individual has been seen at the time of each observation.\nhere is an example of the sort of data that I have:\ndat<-tibble(ID=c("A","A","A","A","A","B","B","B","B","B"),\n   period=c("Aug.2012","Jun.2013","Aug.2013","Jun.2014","Aug.2014","Aug.2012","Jun.2013","Aug.2013","Jun.2014","Aug.2014"),\n   length=c(12,NA,NA,15,19,NA,3,6,10,NA))\ndat$sample.event<-rep(1:5,dim(dat)[1]/5)\n\n# A tibble: 10 Ã\u0097 4\n   ID   period   length sample.event\n   <chr> <chr>   <dbl>   <int>\n 1 A   Aug.2012   12   1\n 2 A   Jun.2013   NA   2\n 3 A   Aug.2013   NA   3\n 4 A   Jun.2014   15   4\n 5 A   Aug.2014   19   5\n 6 B   Aug.2012   NA   1\n 7 B   Jun.2013   3   2\n 8 B   Aug.2013   6   3\n 9 B   Jun.2014   10   4\n10 B   Aug.2014   NA   5\n\nso I want a new column called ind.obs counting each time an individual has been seen, like this, but I want to keep the rows of the dataframe where the individual was not seen:\n   ID   period length sample.event ind.obs\n1   A Aug.2012   12   1   1\n2   A Jun.2013   NA   2   NA\n3   A Aug.2013   NA   3   NA\n4   A Jun.2014   15   4   2\n5   A Aug.2014   19   5   3\n6   B Aug.2012   NA   1   NA\n7   B Jun.2013   3   2   1\n8   B Aug.2013   6   3   2\n9   B Jun.2014   10   4   3\n10 B Aug.2014   NA   5   NA\n\nThis seems like it should be possible using dplyr, but I can't figure it out.\nI have tried:\ndat%>%group_by(ID)%>%\n drop_na(length) %>%\n mutate(ind.obs=sequence(n()))\n\n ID   period   length sample.event ind.obs\n <chr> <chr>   <dbl>   <int>   <int>\n1 A   Aug.2012   12   1   1\n2 A   Jun.2014   15   4   2\n3 A   Aug.2014   19   5   3\n4 B   Jun.2013   3   2   1\n5 B   Aug.2013   6   3   2\n6 B   Jun.2014   10   4   3\n\nBut as you can see, this completely removes the rows without observations.\nI've also tried this, but get an error:\ndat%>%group_by(ID)%>%mutate(ind.obs=sequence(n(na.rm=T)))\n\nError in `mutate()`:\nâ\u0084¹ In argument: `ind.obs = sequence(n(na.rm = T))`.\nâ\u0084¹ In group 1: `ID = "A"`.\nCaused by error in `n()`:\n! unused argument (na.rm = T)\n\nWould appreciate any tips for resolving this, thanks\n
4
\r\n   \r\nI'm trying to extract a line of text from the first page of each multi-page PDF file in a list of PDFs. I'm trying to get the text into a dataframe so I can extract the author of each PDF, which is on the first page and the same word precedes the author in every single document.\nI found the resource below by Packt Publishing that gets very close to what I'm trying to do, but when I implement the for loop (I just copied and pasted and plugged in my object names), R throws this error:\nFor loop:\ntext_df <- data.frame(matrix(ncol=2, nrow=0))\ncolnames(text_df) <- c("pdf title", "text")\n\nfor (i in 1:length(vector)){\n print(i)\n pdf_text(paste("folder/", vector[i],sep = "")) %>% \n   strsplit("\\n")-> document_text\n data.frame("pdf title" = gsub(x =vector[i],pattern = ".pdf", replacement = ""), \n "text" = document_text, stringsAsFactors = FALSE) -> document\n colnames(document) <- c("pdf title", "text")\n text_df <- rbind(text_df,document) \n}\n\n\nError in (function (..., row.n

```
ames = NULL, check.rows = FALSE, check.names = TRUE,  :  arguments imply differing number of rows: 50, 60, 11\nCould someone help me understand what this error means? Could someone direct me to other resources that accomplish what I'm trying to do? Thank you in advance!\nResource:\nhttps://www.r-bloggers.com/2018/01/how-to-extract-data-from-a-pdf-file-with-r/\n
## 5
\r\n                \r\nmystring <- c("code IS (384333)\\n   AND parse = TURE \\n ) \\n \n \\n code IS (43343344)\\n ) some information here\n              code IS (23423422) ) and more information")\n\nI have a string where I want to replace characters from index 1-40 with "orange" and characters from index 61-81 with "blue".\n1. Attempt with gsubfn\nlibrary(gsubfn)\ntoreplace <- list(replace1 = "orange", replace2 = "blue")\nnames(toreplace) <- c(substr(mystring, 1, 40), substr(mystring, 61, 81))\n\nI got the following error:\n> gsubfn(paste(names(toreplace), collapse = "|"), toreplace, mystring)\nError in structure(.External(.C_dotTcl,
...), class = "tclObj") : \n  [tcl] couldn't compile regular expression pattern: parentheses () not balanced.\n\n2. Attempt with gsub:\npat = paste(c(substr(mystring, 1, 40), substr(mystring, 61, 81)), collapse='|')\nreplace = paste("orange", "blue", collapse = "|")\n> gsub(pat, replace, mystring)\n[1] "code IS (384333)\\n   AND parse = TURE \\n ) \\n \\n              \\n code IS (43343344)\\n ) some information here\\n              code IS (23423422) ) and more information"\n\nBut I did not get the desired output. In fact, it didn't seem to have replaced anything.\nThe desired output is\n"orange \\n \n              \\n blue some information here\n              code IS (23423422) ) and more information"\n\n
## 6
\r\n                \r\nI'm working with some regional level data about household internet access in Latinamerica and the Caribbean, I made a df in order to use ggplot2 to do some graphic representations of my data and I encountered some problems. This is my commented code:\nlibrary(ggplot2)\nlibrary(tidyverse)\nlibrary(scales)\nlibrary(ggrepel)\nlibrary(gghighlight)\n\n#Doing the dataframe with the data I want to visualize\ndf <- data.frame(\n  'PaÃs' = c('Argentina', 'MÃ©xico', 'Colombia', 'PerÃº', 'Brasil'),\n  "2017" = c(75.9, 50.7, NA_real_, 28.2, 61.0),\n  "2018" = c(80.3, 52.5, 52.7, 29.8, 67.0),\n  "2019" = c(82.9, 55.8, 51.9, 35.9, 71.0),\n  "2020" = c(90.0, 59.9, 58.1, 38.7, 83.0),\n  "2021" = c(90.4, 66.4, 61.6, 48.7, 82.0)\n)\ndf\n\n#There is no regional level data, so I create a mean of every column that would represent the regional mean\nlac_row<-apply(df[,2:6], 2, mean, na.rm=TRUE)\n\n#I paste the new row in the original df adding "LAC" in the "PaÃs" col\n\ndf<-rbind(df, c("LAC", as.numeric(lac_row)))\n\n#Remove X\nrownames(df) <- gsub('^X', '', rownames(df))\n\n#In order to get a proper tibble to use with ggplot2 I use pivot longer to transpose my df into a longer format\ndf_long <- df %>%\n  pivot_longer(cols = -PaÃs, names_to = "AÃ±o", values_to = "Porcentaje") %>%\n  mutate(Year = as.numeric(str_remove(AÃ±o, "X")))\n\nstr(df_long)\n\n\n# Create the ggplot\nggplot(df_long, aes(x = AÃ±o, y = as.numeric(Porcentaje), color = PaÃs, group = PaÃs)) +\n  geom_line(size = 1) +\n    labs(title = "Porcentaje de hogares con acceso a internet en PaÃses de LatinoamÃ©rica",\n       x = "AÃ±os",\n       y = "Porcentaje",\n       color = "PaÃs") +\n  scale_y_continuous(labels = scales::percent_format(scale = 1)) +\n  scale_x_discrete(labels = c("2017", "2018", "2019", "2020", "2021"))+\n  geom_text_repel(aes(label = scales::percent(as.numeric(Porcentaje), scale = 1, size = 1)),\n            nudge_y = 0.5,\n            show.legend = FALSE,\n            size = 3)+\n  geom_point(size= 2.5)+theme_classic()\n\nI got this warning messages:\nWarning messages:\n1: Removed 2 rows containing missing values (`geom_line()`). \n2: Removed 6 rows containing missing values (`geom_text_repel()`). \n3: Removed 6 rows containing missing values (`geom_point()`).\n\nAnd this is my output:\n\nAs you can see, everything seems fine except for the LAC data, which is not shown properly, I would like its line to be drawn, and, additional to that, I would like that line to be slightly bold than the others (since is the regional mean).\nHope you can help, thanks in advance.\n
##   badgeGold badgeSilver badgeBronze  Editor             EditTime
## 1         0           8          21    <NA>                   NA
```

```
## 2         0           0           0    <NA>                   NA
## 3         0           0           2    <NA>                   NA
## 4         0           0           2    <NA>                   NA
## 5        24          73         131  Adrian 2023-06-04 01:15:56Z
## 6         0           0           8 esteban 2023-06-04 01:53:37Z
```

```
head(a_df)
```

```
##   ParentId
## 1 76398561
## 2 76398472
## 3 76397805
## 4 76397627
## 5 76397512
## 6 76397435
##
body
## 1
\r\nI bet there's something sharper, but:\ndat %>%\n  mutate(flag = cumsum(!is.na(length)),\n
ind.obs = if_else(flag != lag(flag, default = 0), flag, NA),\n         .by = ID)\n\nResult\n#
A tibble: 10 Ã\u0097 6\n   ID    period   length sample.event  flag ind.obs\n   <chr> <chr>
<dbl>        <int> <int>   <int>\n 1 A     Aug.2012     12            1     1       1\n 2 A
Jun.2013     NA            2     1      NA\n 3 A     Aug.2013     NA            3     1
NA\n 4 A     Jun.2014     15            4     2       2\n 5 A     Aug.2014     19
5     3       3\n 6 B     Aug.2012     NA            1     0      NA\n 7 B     Jun.2013
3            2     1       1\n 8 B     Aug.2013      6            3     2       2\n 9 B     J
un.2014     10            4     3       3\n10 B     Aug.2014     NA            5     3      N
A\n\n
## 2
\r\nYou can do a nested sub.\nsub(substr(mystring, 1, 40), "orange", \n    sub(substr(mystrin
g, 61, 81), "blue", mystring, fixed = T), fixed = T)\n[1] "orange \\n \\n              \\n bl
ue some information here\\n              code IS (23423422) ) and more information"\n\n
## 3 \r\nwhat about:\nlibrary(geosphere)\nlibrary(dplyr)\n\nthe_distances <- \n    expand.gri
d(name_1 = df_1$name_1, name_2 = df_2$name_2) |> \n    left_join(df_1 |> mutate(coords_1 = cb
ind(lon, lat)), by = 'name_1') |>\n    left_join(df_2 |> mutate(coords_2 = cbind(lon, lat)),
by = 'name_2') |> \n    rowwise() |>\n    mutate(hav_dist = distHaversine(coords_1, coords_
2)) |>\n    select(c(starts_with('name_'), hav_dist))\n\n## > the_distances |> head()\n## # A
tibble: 6 x 3\n## # Rowwise: \n##   name_1 name_2  hav_dist\n##   <chr>  <chr>      <dbl>\n##
1 john   matthew    1564.\n## 2 david  matthew    1903.\n## 3 alex   matthew    2028.\n##
...\n\nthe_distances |>\n    group_by(name_1) |>\n    summarize(min = min(hav_dist),\n
ave = mean(hav_dist), \n              max = max(hav_dist)\n             )\n\n## # A tibble:
15 x 4\n##    name_1    min   ave   max\n##    <chr>   <dbl> <dbl> <dbl>\n##  1 alex     354.
1361. 3108.\n##  2 chris   1477. 2607. 3599.\n##  3 david    302. 1519. 2678.\n##  4 henry
1289. 2541. 3935.\n##  5 john     880. 1862. 2701.\n## ...\n\n
## 4
\r\n               \r\nWe can use fct_inorder here:\nggplot will order x axis alphapeticall
y. To get the order in your table use fct_inorder:\nlibrary(ggplot2)\nlibrary(forcats)\nlibra
ry(dplyr)\n\nde %>% \n  mutate(Mes = fct_inorder(Mes)) %>% \n  ggplot(aes(fill = Cidade, y =
Leitura, x = Mes)) +\n  geom_bar(position = 'dodge', stat = 'identity')\n\n\n
## 5
\r\n               \r\nIt's simply\np + guides(size = guide_legend(order = 1), \n
color = guide_colorbar(order = 2))\n\n\n
## 6
\r\nR is vectorized and this can be done without loops at all.\nNote that nums <- seq_len(3)
is an alternative way of creating the vector nums.\nnums <- 1:3\nrow <- setNames(nums, paste0
("Column", nums))\nas.data.frame(t(row))\n#>   Column1 Column2 Column3\n#> 1       1       2
3\n\nCreated on 2023-06-03 with reprex v2.0.2\n
##          author             PostTime Reputation BronzeMedals SilverMedals
## 1    Jon Spring 2023-06-04 00:53:49Z      52995           52           35
```

```
## 2      benson23 2023-06-04 01:46:39Z      15250           38           18
## 3           I_O 2023-06-03 21:03:32Z       2755           14            2
## 4        TarJae 2023-06-03 19:08:42Z      70408           62           18
## 5 Allan Cameron 2023-06-03 18:46:52Z     138808           80           42
## 6  Rui Barradas 2023-06-03 18:24:05Z      68610           63           32
##   GoldMedals
## 1          4
## 2          8
## 3          0
## 4          6
## 5          7
## 6          8
```

```
head(c_df)
```

```
##   QuestionId ParentId       UserId          CommentTime
## 1   76398388 76398388      r2evans 2023-06-03 23:26:23Z
## 2   76398388 76398388      r2evans 2023-06-03 23:28:23Z
## 3   76398196 76398196 Leroy Tyrone 2023-06-03 22:48:20Z
## 4   76397984 76397984   Ben Bolker 2023-06-03 20:49:01Z
## 5   76397984 76397984  user2554330 2023-06-03 21:58:02Z
## 6   76397984 76397984      JStorey 2023-06-03 22:41:07Z
##
body
## 1 Looking, but (1) never do df<-rbind(df, c("LAC", as.numeric(lac_row))), you are corrupti
ng your data. You are trying to fix it with as.numeric(.) later, but you should not be doing
that in the first place. It's better to rbind a frame (or list), names much match. (Consider
add_row or bind_rows as alternatives.)
## 2
When I try your code, I see the LAC line, i.stack.imgur.com/wMKd1.png
## 3                                                    Welcome to SO Rossy. Please review H
ow to make a great R reproducible example and update your question accordingly. While the SO
community are very helpful, you will increase your chances of getting an answer if your make
your question easier to answer. Thanks
## 4                                                                                      C
an you please give us some more information, including (1) a link to the documentation of gau
ssian_filter(); (2) a minimal reproducible example; (3) some specific examples (with code) of
what you've tried that hasn't worked?
## 5
Showing the output of the Python code on a simple dataset would be extremely helpful.
## 6
Apologies, I should have included some examples. I've updated my answer. Thanks!
```

[1] https://stackoverflow.com/questions/32019566/r-xml-parse-for-a-web-address (https://stackoverflow.com/questions/32019566/r-xml-parse-for-a-web-address)

[2] https://stackoverflow.com/questions/1604471/how-can-i-find-an-element-by-css-class-with-xpath (https://stackoverflow.com/questions/1604471/how-can-i-find-an-element-by-css-class-with-xpath)

[3] https://stackoverflow.com/questions/18547410/xpath-with-multiple-contains-on-different-elements

(https://stackoverflow.com/questions/18547410/xpath-with-multiple-contains-on-different-elements)

[4] https://stackoverflow.com/questions/11455590/parse-an-xml-file-and-return-an-r-character-vector (https://stackoverflow.com/questions/11455590/parse-an-xml-file-and-return-an-r-character-vector)

[5] https://statisticsglobe.com/concatenate-vector-of-character-strings-in-r (https://statisticsglobe.com/concatenate-vector-of-character-strings-in-r)

[6] https://stackoverflow.com/questions/34570860/add-nas-to-make-all-list-elements-equal-length (https://stackoverflow.com/questions/34570860/add-nas-to-make-all-list-elements-equal-length) [

7] https://www.r-bloggers.com/2020/10/basic-error-handing-in-r-with-trycatch/ (https://www.r-bloggers.com/2020/10/basic-error-handing-in-r-with-trycatch/)