

# Assignment 3

Due at 11:59pm on October 15.

You may work in pairs or individually for this assignment. Make sure you join a group in Canvas if you are working in pairs. Turn in this assignment as an HTML or PDF file to ELMS. Make sure to include the R Markdown or Quarto file that was used to generate it. Include the GitHub link for the repository containing these files.

```
library(xml2)
library(rvest)
library(lubridate)
```

Attaching package: 'lubridate'

The following objects are masked from 'package:base':

date, intersect, setdiff, union

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
v dplyr   1.1.4      v readr   2.1.5
v forcats 1.0.0      v stringr 1.5.1
v ggplot2 3.5.1      v tibble  3.2.1
v purrr   1.0.2      v tidyr   1.3.1
```

```
-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter()      masks stats::filter()
x readr::guess_encoding() masks rvest::guess_encoding()
x dplyr::lag()          masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
```

## Web Scraping

In this assignment, your task is to scrape some information from Wikipedia. We start with the following page about Grand Boulevard, a Chicago Community Area.

[https://en.wikipedia.org/wiki/Grand\\_Boulevard,\\_Chicago](https://en.wikipedia.org/wiki/Grand_Boulevard,_Chicago)

The ultimate goal is to gather the table “Historical population” and convert it to a `data.frame`.

As a first step, read in the html page as an R object. Extract the tables from this object (using the `rvest` package) and save the result as a new object. Follow the instructions if there is an error. Use `str()` on this new object – it should be a list. Try to find the position of the “Historical population” in this list since we need it in the next step.

```
base_url <- "https://en.wikipedia.org/wiki/"
chicago_city <- "Grand_Boulevard,_Chicago"

url <- read_html(str_c(base_url, chicago_city))
nds <- html_nodes(url, "table")
# nds <- html_nodes(url, xpath = '//*[contains(concat( " ", @class, " " ), concat( " ", "us-
```

Extract the “Historical population” table from the list and save it as another object. You can use subsetting via `[[...]]` to extract pieces from a list. Print the result.

You will see that the table needs some additional formatting. We only want rows and columns with actual values (I called the table object `pop`).

```
# create a function to extract table from html
## keep what is needed only and rename
html_table_fun <- function(nds_html, chi_city) {

  html_table(nds_html)[[2]] |>
  # remove last row, does not contain data
  filter(row_number() <= n()-1) |>
  # remove empty column
  select(-3) |>
  # give last column a descriptive name
  rename(percent = 3) |>
  # clean numeric values
  mutate(Census = as.numeric(Census),
         Pop. = str_remove(Pop., ","),
         Pop. = as.numeric(Pop.),
         percent = str_remove(percent, "\\%"),
```

```

    percent = na_if(percent, "-"),
    percent = str_replace(percent, "-", "-"),
    percent = as.numeric(percent),
    # add city
    city = str_replace_all(chi_city, "_", " ")
  )
}

```

```
pop <- html_table_fun(nds, chicago_city)
```

## Expanding to More Pages

That's it for this page. However, we may want to repeat this process for other community areas. The Wikipedia page [https://en.wikipedia.org/wiki/Grand\\_Boulevard,\\_Chicago](https://en.wikipedia.org/wiki/Grand_Boulevard,_Chicago) has a section on “Places adjacent to Grand Boulevard, Chicago” at the bottom. Can you find the corresponding table in the list of tables that you created earlier? Extract this table as a new object.

Then, grab the community areas east of Grand Boulevard and save them as a character vector. Print the result.

We want to use this list to create a loop that extracts the population tables from the Wikipedia pages of these places. To make this work and build valid urls, we need to replace empty spaces in the character vector with underscores. This can be done with `gsub()`, or by hand. The resulting vector should look like this: “Oakland,\_Chicago” “Kenwood,\_Chicago” “Hyde\_Park,\_Chicago”

To prepare the loop, we also want to copy our `pop` table and rename it as `pops`. In the loop, we append this table by adding columns from the other community areas.

```

# the table we are looking for is in element 4
# strip only the name to the right
nds_2 <- html_nodes(url, xpath = '//table')

places_adjacent <- html_table(nds_2[[4]]) |>
  select(3) |>
  rename(cities = 1) |>
  filter(str_detect(cities, "Chicago"))

```

Build a small loop to test whether you can build valid urls using the vector of places and pasting each element of it after <https://en.wikipedia.org/wiki/> in a for loop. Calling `url` shows the last url of this loop, which should be [https://en.wikipedia.org/wiki/Hyde\\_Park,\\_Chicago](https://en.wikipedia.org/wiki/Hyde_Park,_Chicago).

```

pop_full <- map_dfr(places_adjacent$cities, function(x) {

  # build the URL
  chi_city = str_replace_all(x, " ", "_")
  print(str_c(base_url, chi_city))

  # read in each page
  url = read_html(str_c(base_url, chi_city))
  nds = html_nodes(url, "table")

  # extract table needed
  dat = html_table_fun(nds, chi_city)

  return(dat)

}) |>
  # add Grand Boul
  full_join(pop)

```

```

[1] "https://en.wikipedia.org/wiki/Oakland,_Chicago"
[1] "https://en.wikipedia.org/wiki/Kenwood,_Chicago"
[1] "https://en.wikipedia.org/wiki/Hyde_Park,_Chicago"

```

Joining with `by = join\_by(Census, Pop., percent, city)`

```
head(pop)
```

```

# A tibble: 6 x 4
  Census   Pop. percent city
  <dbl>   <dbl>   <dbl> <chr>
1  1930  87005     NA  Grand Boulevard, Chicago
2  1940 103256   18.7  Grand Boulevard, Chicago
3  1950 114557   10.9  Grand Boulevard, Chicago
4  1960  80036  -30.1  Grand Boulevard, Chicago
5  1970  80166    0.2  Grand Boulevard, Chicago
6  1980  53741   -33   Grand Boulevard, Chicago

```

Finally, extend the loop and add the code that is needed to grab the population tables from each page. Add columns to the original table pops using `cbind()`.

## Scraping and Analyzing Text Data

Suppose we wanted to take the actual text from the Wikipedia pages instead of just the information in the table. Our goal in this section is to extract the text from the body of the pages, then do some basic text cleaning and analysis.

First, scrape just the text without any of the information in the margins or headers. For example, for “Grand Boulevard”, the text should start with, “**Grand Boulevard** on the [South Side](#) of [Chicago, Illinois](#), is one of the ...”. Make sure all of the text is in one block by using something like the code below (I called my object `description`).

```
# description <- description %>% paste(collapse = ' ')
```

Using a similar loop as in the last section, grab the descriptions of the various communities areas. Make a tibble with two columns: the name of the location and the text describing the location.

Let’s clean the data using `tidytext`. If you have trouble with this section, see the example shown in <https://www.tidytextmining.com/tidytext.html>

```
library(tidytext)
```

Create tokens using `unnest_tokens`. Make sure the data is in one-token-per-row format. Remove any stop words within the data. What are the most common words used overall?

Plot the most common words within each location. What are some of the similarities between the locations? What are some of the differences?