R Notebook

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------ 
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(httr)
library(rvest)
```

```
## Loading required package: xml2
```

```
##
## Attaching package: 'rvest'
```

```
## The following object is masked from 'package:purrr':
##
##     pluck
```

```
## The following object is masked from 'package:readr':
##
##     guess_encoding
```

Load necessary libraries.

```
# Input the url from which I will be scraping data (Stuart Carr's Google
# Scholar profile)
schol_html <- read_html("https://scholar.google.com/citations?user=gVWAWMOAAAAJ&hl=en&oi=ao")

# Use a function from the rvest package to pull nodes from the previously
# specified URL that matches a given CSS selector
# The CSS selector identifies the title, author, year, and citation count
# of the first 20 articles on the profile identified above
schol_nodes <- html_nodes(schol_html, css = "#gsc_a_b .gs_ibl , .gsc_a_at+ .gs_gray , .gsc_a_at")

# Extract the text from the nodes
node_text <- html_text(schol_nodes)

# From the full text, separate the title, author, year, and citation count
title <- node_text[seq(1, length(node_text), 4)]
author <- node_text[seq(2, length(node_text), 4)]
```

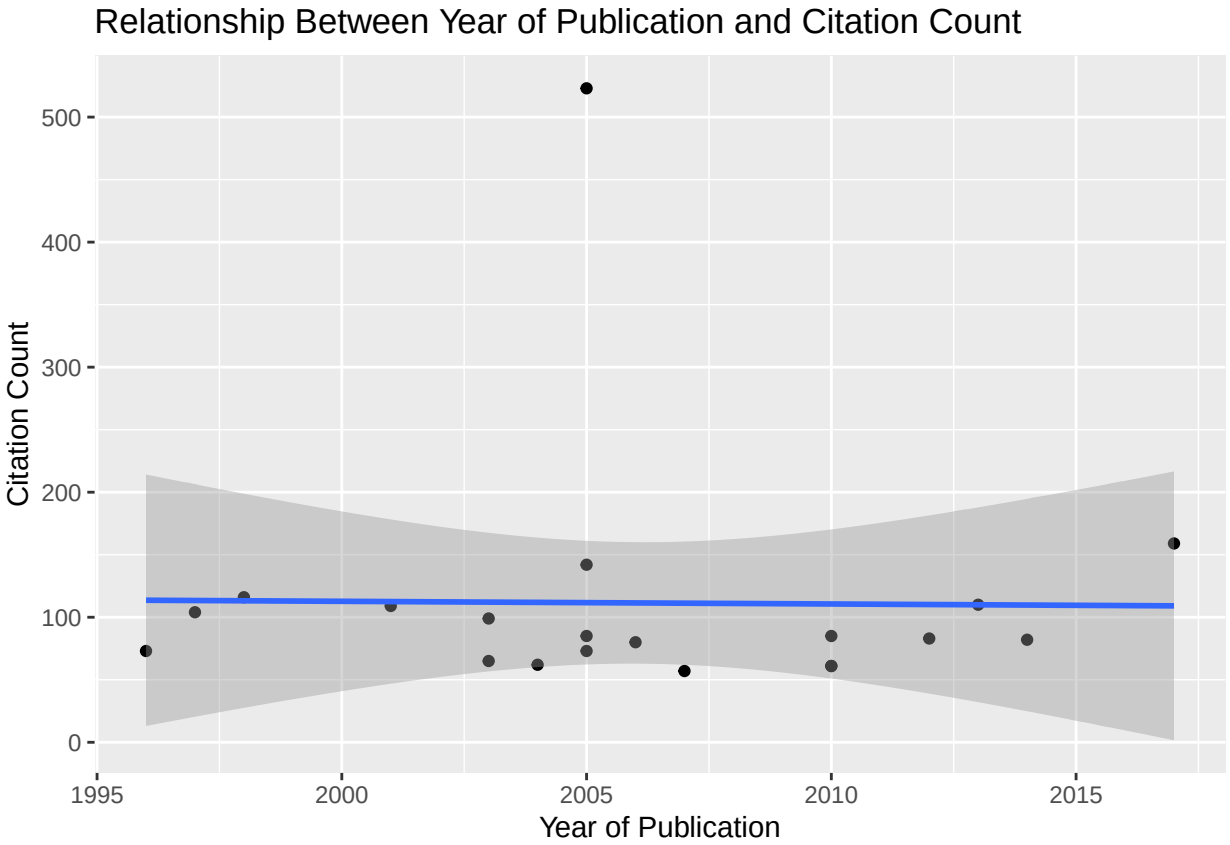
```r
cite <- as.numeric(node_text[seq(3, length(node_text), 4)])
year <- as.numeric(node_text[seq(4, length(node_text), 4)])

# Bind the title, author, year, and citation count vectors into a tibble
profile_tbl <- bind_cols(title = title, authors = author, year = year, citations = cite)
```

```r
profile.cor <- cor.test(profile_tbl$year, profile_tbl$citations)
profile.cor
```

```
##
##  Pearson's product-moment correlation
##
## data:  profile_tbl$year and profile_tbl$citations
## t = -0.051263, df = 18, p-value = 0.9597
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4521852  0.4327525
## sample estimates:
##         cor
## -0.01208203
```

The correlation between when an article (from this specific set of articles) was published and its citation count is -0.012082.

```r
ggplot(profile_tbl, aes(x = year, y = citations)) +
           geom_point() +
    geom_smooth(method = "lm") +
    ggtitle("Relationship Between Year of Publication and Citation Count") +
    xlab("Year of Publication") +
    ylab("Citation Count")
```



This plot shows the relationship between year of publication and citation count, along with the least squares regression line between them.