Week 10 Project: Google Scholar

Libraries

```
library(tidyverse)
library(rvest)
```

The tidyverse package is used for data import and cleaning. The rvest package is used to read the html page and find the elements that we want.

Data Import and Cleaning

```
# Download page of Dr. Lynn Eberly
page <- read_html("https://scholar.google.com/citations?user=-vFiWHgAAAAJ&hl=en&oi=ao")

# Find the elements corresponding to article title, author, year, and citations
nodes_titles <- html_nodes(page, "a.gsc_a_at")
nodes_authors <- html_nodes(page, ".gsc_a_at+ .gs_gray")
nodes_years <- html_nodes(page, ".gsc_a_hc")
nodes_citations <- html_nodes(page, ".gsc_a_ac")

# Create a tibble to store above information
profile_tbl <- tibble(title = html_text(nodes_titles),
                      author = html_text(nodes_authors),
                      year = as.numeric(html_text(nodes_years)),
                      citations = as.numeric(html_text(nodes_citations)))
```

The data cleaning process involved retrieving the raw html page, locating the elements of interest (title, authors, year, number of citations), and then storing that information in a tibble.

Analysis

```
# Correlation between year and citation count
year_citations <- cor.test(profile_tbl$year, profile_tbl$citations)
year_citations
```

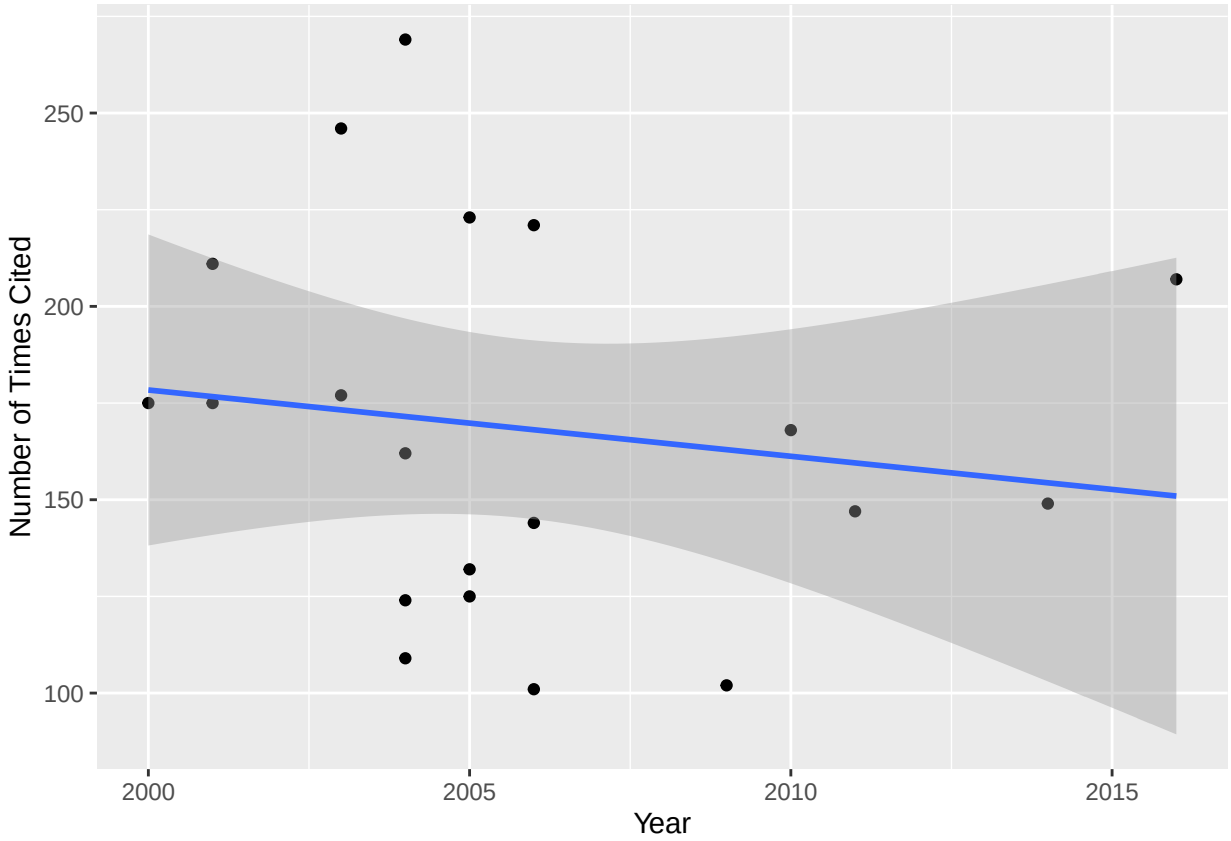
```
##
##  Pearson's product-moment correlation
##
## data:  profile_tbl$year and profile_tbl$citations
## t = -0.63976, df = 18, p-value = 0.5304
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5550070  0.3141415
```

```
## sample estimates:
##        cor
## -0.1491073
```

The correlation between the year and the number of citations was was -0.149 (p=0.53), which is not statistically significant.

Visualization

```
ggplot(profile_tbl, aes(year, citations)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Year", y = "Number of Times Cited")
```



The plot shows the relationship between the year an article was published and the number of times it has been cited. Each point represents an individual article, and the line is the OLS regression line with a 95% confidence band.