Week 11 Google Scholar

Dana Tomeh

#Rstudio API code

# Libraries

Loading the libraries necessary to scrape data from google scholar and to plot the relationships of the data

knitr::opts\_chunk$set(echo = TRUE)  
library(rvest)

## Loading required package: xml2

library(httr)  
library(ggplot2)  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ tibble 2.1.3 ✓ dplyr 0.8.4  
## ✓ tidyr 1.0.2 ✓ stringr 1.4.0  
## ✓ readr 1.3.1 ✓ forcats 0.4.0  
## ✓ purrr 0.3.3

## ── Conflicts ────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x readr::guess\_encoding() masks rvest::guess\_encoding()  
## x dplyr::lag() masks stats::lag()  
## x purrr::pluck() masks rvest::pluck()

# Data import and cleaning

read\_html allows me to read the web page i am trying to pull information from. html\_nodes pulls all relevant pieces of information, the CSS code was found using the SelectorGadet extension on Chrome. html\_text reads the information from the nodes in as text After pulling the citations, authors, years, and citation counts from the google scholar page of Paul Sackett, apa\_tbl was created that contains article title, authors, citation count and year for all 20 articles.  
Year and citation count were mutated from the class character to numeric using the mutate verb from the tidyverse package so they could be used to calculate a correlation later on.

apa\_html <-read\_html("https://scholar.google.com/citations?hl=en&user=nx\_vfuYAAAAJ")  
apa\_node <- html\_nodes(apa\_html, "#gsc\_a\_b .gsc\_a\_y , #gsc\_a\_b .gsc\_a\_c , .gsc\_a\_at+ .gs\_gray , .gsc\_a\_at")  
apa\_text <- html\_text(apa\_node)  
  
apa\_tbl <- as\_tibble(matrix(apa\_text,nrow=20,ncol=4,byrow=TRUE,  
 dimnames=list(c(),  
 c("Article Title","Authors","CitationCount","Year")))) %>%  
 mutate(Year=as.numeric(Year)) %>%  
 mutate(CitationCount = as.numeric(CitationCount))

#Analysis Finding the correlation between the year and the citation count of an article on the google scholar page for Paul Sackett.

scholar.cor <- cor.test(apa\_tbl$Year, apa\_tbl$`CitationCount`)

The corerlation between the year and the number of times an article has been cited is 0.1837841 with a p value of 0.4379691

# Visualization

A Scatter plot of the year versus the number of citations on the articles on Paul Sackett’s google scholar page was created using ggplot from the ggplot2 package paired with geom\_point(). A regression line is superimposed over the original scatter plot using geom\_smooth() with the method argument as "lm.

ggplot(apa\_tbl, aes(apa\_tbl$Year, apa\_tbl$`CitationCount`)) +   
 geom\_point()+  
 geom\_smooth(method="lm", se=FALSE) +   
 labs(title = "Year of Citation Publication vs Number of times Cited",   
 x="Year of Citation Publication",   
 y="Number of times the article has been Cited")

