Basic Harvesting

William Lovejoy

7/6/2022

```
library(rvest)
library(scales)
library(dplyr)
library(reshape2)
library(tidyverse)
```

We'll load in these packages to scrape and handle our data.

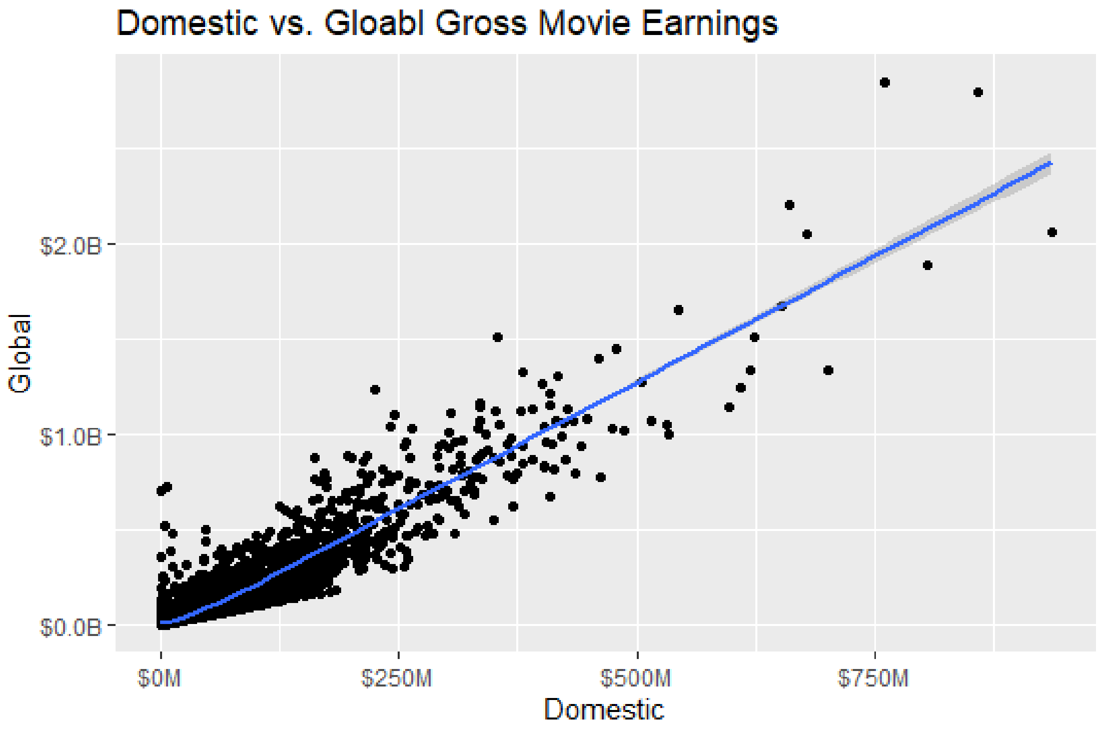
```
base_webpage <- read_html("https://www.the-numbers.com/movie/budgets/all")
new_urls <- "https://www.the-numbers.com/movie/budgets/all/%s"

table_base <- html_table(base_webpage)[[1]] %>%
  as_tibble(.name_repair = "unique")
table_base

## # A tibble: 100 x 6
##     ...1 ReleaseDate  Movie        ProductionBudget DomesticGross
WorldwideGross
##    <int> <chr>        <chr>        <chr>            <chr>         <chr>
##  1     1 Apr 23, 2019 Avengers: E~ $400,000,000     $858,373,000
$2,797,800,564
##  2     2 May 20, 2011 Pirates of ~ $379,000,000     $241,071,802
$1,045,713,802
##  3     3 Apr 22, 2015 Avengers: A~ $365,000,000     $459,005,868
$1,395,316,979
##  4     4 Dec 16, 2015 Star Wars E~ $306,000,000     $936,662,225
$2,064,615,817
##  5     5 Apr 25, 2018 Avengers: I~ $300,000,000     $678,815,482
$2,048,359,754
##  6     6 May 24, 2007 Pirates of ~ $300,000,000     $309,420,425
$960,996,492
##  7     7 Nov 13, 2017 Justice Lea~ $300,000,000     $229,024,295
$655,945,209
##  8     8 Oct 6, 2015  Spectre      $300,000,000     $200,074,175
$879,500,760
##  9     9 Jul 13, 2023 Mission: Im~ $290,000,000     $0            $0
## 10    10 Dec 18, 2019 Star Wars: ~ $275,000,000     $515,202,542
$1,072,848,487
## # ... with 90 more rows

table_new <- data.frame()
df <- data.frame()
```

```
i <- 101

while(i < 5502) {
  new_webpage <- read_html(sprintf(new_urls, i))
  table_new <- html_table(new_webpage)[[1]] %>%
    as_tibble(.name_repair = "unique")
  df <- rbind(df, table_new)
  i = i + 100
}

movies <- rbind(table_base, df)
glimpse(movies)

## Rows: 5,600
## Columns: 6
## $ ...1             <chr> "1", "2", "3", "4", "5", "6", "7", "8", "9",
"10", "1~
## $ ReleaseDate      <chr> "Apr 23, 2019", "May 20, 2011", "Apr 22, 2015",
"Dec ~
## $ Movie            <chr> "Avengers: Endgame", "Pirates of the Caribbean:
On St~
## $ ProductionBudget <chr> "$400,000,000", "$379,000,000", "$365,000,000",
"$306~
## $ DomesticGross    <chr> "$858,373,000", "$241,071,802", "$459,005,868",
"$936~
## $ WorldwideGross   <chr> "$2,797,800,564", "$1,045,713,802",
"$1,395,316,979",~
```

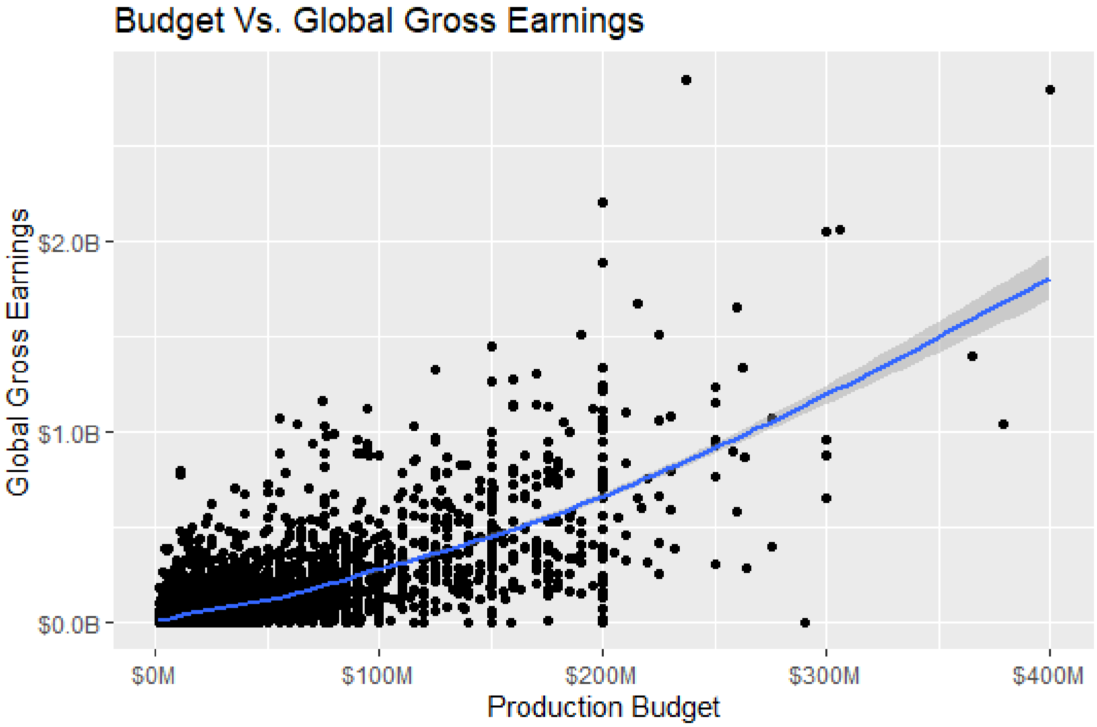
This part iterates 5501 times to extract the data from the rest of the webpages

```
movies$ProductionBudget <- gsub("\\$|,", "", movies$ProductionBudget)
movies$DomesticGross <- gsub("\\$|,", "", movies$DomesticGross)
movies$WorldwideGross <- gsub("\\$|,", "", movies$WorldwideGross)
movies$ReleaseDate <- gsub(",", "", movies$ReleaseDate)

movies$ReleaseDate <- as.Date(as.character(movies$ReleaseDate), format = "%b
%d %Y")
movies <- movies %>%
  mutate_at(c("ProductionBudget", "DomesticGross", "WorldwideGross"),
as.numeric)

glimpse(movies)

## Rows: 5,600
## Columns: 6
## $ ...1             <chr> "1", "2", "3", "4", "5", "6", "7", "8", "9",
"10", "1~
## $ ReleaseDate      <date> 2019-04-23, 2011-05-20, 2015-04-22, 2015-12-16,
2018~
## $ Movie            <chr> "Avengers: Endgame", "Pirates of the Caribbean:
```

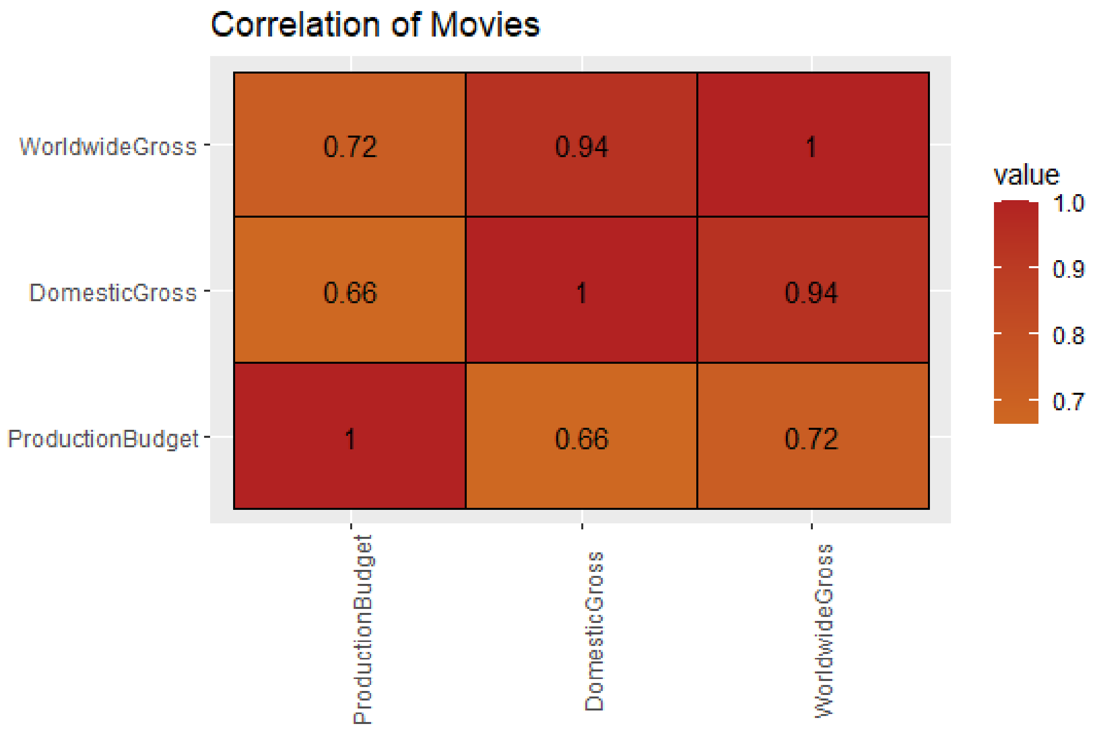
```
On St~
## $ ProductionBudget <dbl> 400000000, 379000000, 365000000, 306000000,
300000000~
## $ DomesticGross    <dbl> 858373000, 241071802, 459005868, 936662225,
678815482~
## $ WorldwideGross   <dbl> 2797800564, 1045713802, 1395316979, 2064615817,
20483~

write.csv(movies,
          "C:\\Users\\William
Lovejoy\\Documents\\Codes\\R\\DataScience\\movies.csv",
          row.names = FALSE)

movies %>%
  ggplot(aes(x = DomesticGross, y = WorldwideGross)) + geom_point() +
  geom_smooth() + scale_x_continuous(labels = label_number(prefix = "$",
                                                           suffix = "M",
                                                           scale = 1 / 1e6))
+
  scale_y_continuous(labels = label_number(prefix = "$", suffix = "B",
                                           scale = 1 / 1e9)) +
  labs(title = "Domestic vs. Gloabl Gross Movie Earnings", x = "Domestic",
       y = "Global")
```



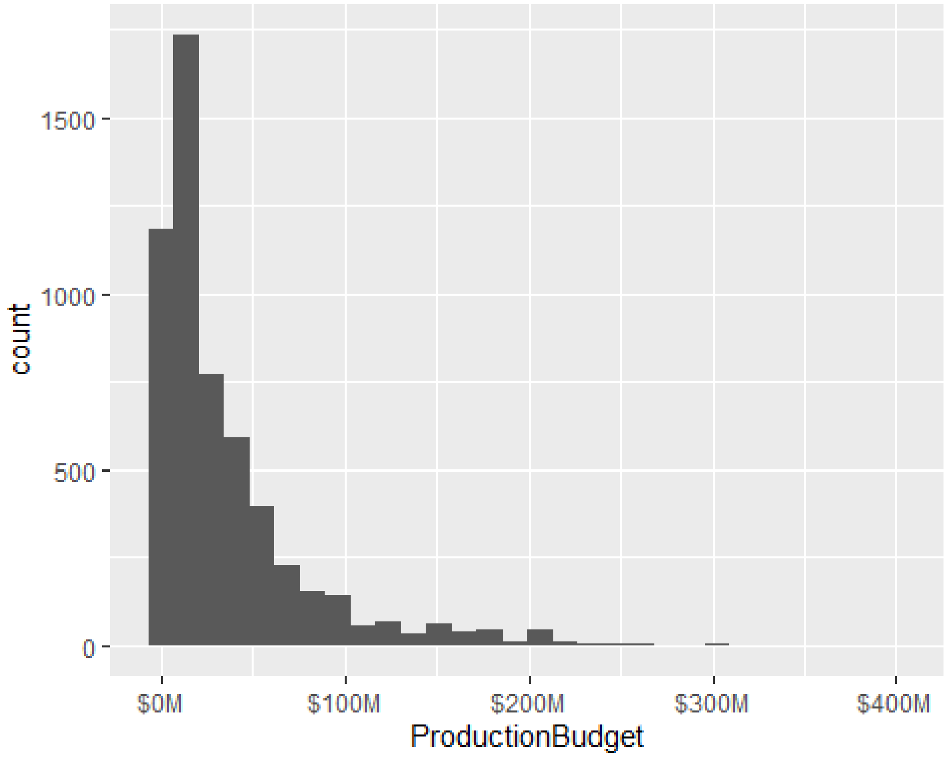
```
movies %>%
  ggplot(aes(x = ProductionBudget, y = WorldwideGross)) + geom_point()+
  geom_smooth() + scale_x_continuous(labels = label_number(prefix = "$",
                                                           suffix = "M",
                                                           scale = 1 / 1e6))
+
    scale_y_continuous(labels = label_number(prefix = "$", suffix = "B",
                                           scale = 1 / 1e9)) +
  labs(title = "Budget Vs. Global Gross Earnings", x = "Production Budget",
       y = "Global Gross Earnings")
```



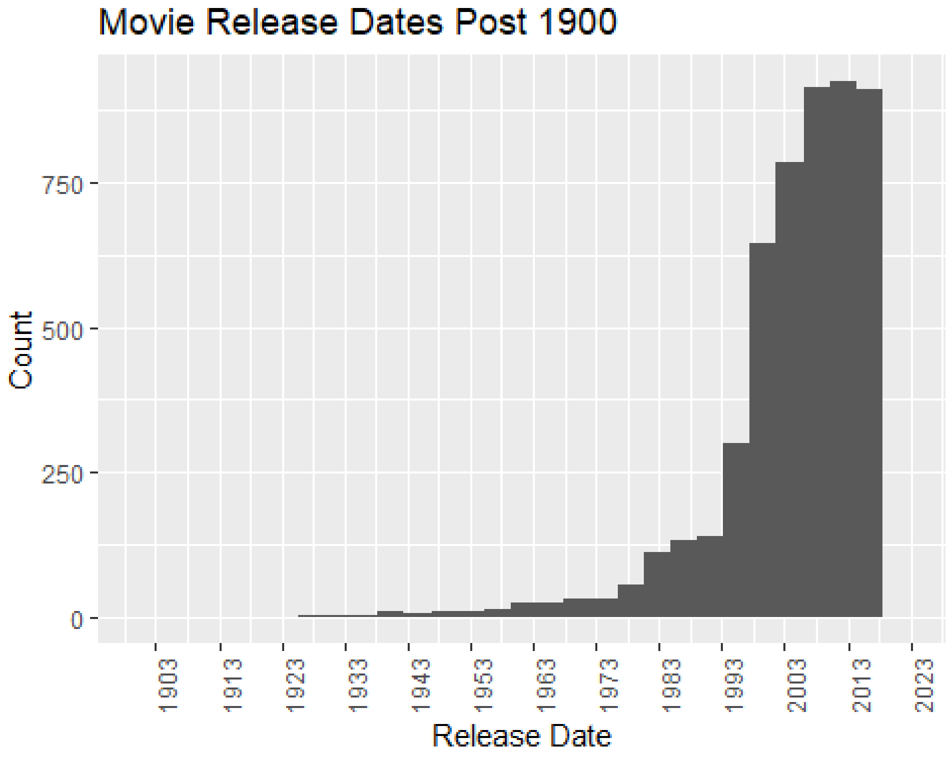
```
correlated <- movies[, -c(1:3)] %>%
  cor() %>%
  melt()

ggplot(correlated, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile(color = "black") +
  scale_fill_gradient2(low = "antiquewhite", mid = "gold", high =
"firebrick") +
  geom_text(aes(label = round(value, 2)), size = 4) +
  theme(axis.text.x = element_text(angle = 90), axis.title = element_blank())
+
  labs(title = "Correlation of Movies")
```



```
movies %>%
  ggplot(aes(x = ProductionBudget)) + geom_histogram() +
  scale_x_continuous(labels = label_number(prefix = "$", suffix = "M",
                                            scale = 1 / 1e6))
```



```r
movies %>%
  ggplot(aes(x = ReleaseDate)) + geom_histogram() +
  scale_x_date(limit = c(as.Date("1900-01-01"), as.Date("2022-07-06")),
               date_labels = "%Y", breaks = date_breaks("10 years")) +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(title = "Movie Release Dates Post 1900", x = "Release Date", y =
"Count")
```


Movie Release Dates Post 1900
Count
Release Date