Problem Set 2

MGSC 310, Fall 2019, Professor Hersh (BEST PROFESSOR EVER!!!)

Geoffrey Hughes

9/13/2019

```r
library("tidyverse")
```

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

## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.3     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()
```

Question 1) ISLR Ch. 2, Problem 2

a.

- This is a regression problem because we are trying to predict quantitative numbers using the given data (seeing which factors affect CEO by a numerical amount), as opposed to trying to predict some categorical classification labels.
- In this problem we are interested in the **inference**, because we are trying to define the relationship between the outcome (CEO Salary) and all of the premises/inputs (record profit, number of employees, industry) as they change.
- Here, we have the top 500 firms (n = 500), and record profit, number of employees, industry, and CEO salary as variables (p = 4).

b.

- This is a categorical problem, as we are classifying the outcome as either a *success* or *failure*.
- Here we are interested in **prediction**, as we are using the inputs to predict an output that is unknown at this time (new product).
- Data is collected on 20 products (n = 20), with their variables being [whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables] so (p = 14).

c.

- Predicting % change in the USD/Euro exchange rate is a regression problem, because we are trying to quantify a specific number, not a categorical label.
- This one is right in the wording! We are focusing on **prediction**, as we are trying to predict the outcome, given a history of % changes in the global economy.
- We are using weekly data gathered over all of 2012, and since we have 52 weeks in a year, (n = 52). Our variables consist of [% change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market] a total of 4 variables (p = 4).

Question 2) ISLR Ch. 2, Problem 4

a. Classification examples:

- Predicting, based on (predictors) income, marital status, # of children, which (response) tax bracket you fall into (say there are 5: 12%, 17%, 23%, 30%, and 40% of income). This is focusing on **prediction**, as we are using these three variables (p = 3) to infer someone's tax bracket.
- Predicting, based on (predictors) T-Cell level, X-Rays, and immune response to find out whether someone has HIV or does not. This is focusing on **prediction**, as health care professionals want to determine a patient's HIV status based on these variables.
- Say we have 300 Olympic Gold medalists, 300 silver, and 300 bronze (response). We want to figure out which variables have the most influence on their ranking. So we measure each medalist's lung capacity, resting heart rate, BMI, and sleeping habits (predictors). This is focusing on **inference** because we want to find how each variable (x) affects the categorical placement. (Let's find how much influence each variable has for the Gold medalists, in contrast to the other medalists.)

b. Regression examples:

- We want to predict a student's final grade [0, 100] % (response) using the variables of the student's avg hours of sleep, avg hours of studying a week, and hours played of World of Warcraft (predictors). This is focusing on **prediction**, as we are using the variables to predict the outcome. Regression is useful here because we are trying to predict a specific numeric score (quantitative)!
- We want to find out which variable most significantly affects affects a person's longest relationship in years (response). Using the variables (predictors) income, education status, and age. This is focusing on **inference**, because we are trying to decide which of those p = 3 variables most influence their longest relationship.
- Say we want to determine a dog's lifespan (response). We use the variables # times walked a week, breed, type of food, and income of owner (predictors). This is focusing on **prediction** and should be modeled regressivly, since we want to determine a specific age (say, in dog years) based on the given variables as input.

c. Cluster analysis examples:

- If we want to determine the relationship between a person's alcohol use and their income, age, weight, and mental health.
- If we want to determmine the relationship between people who read a lot and those who don't, depending on education and free time.
- If we want to determine groups of people who will complete, partially complete, or not complete this assignment based on their current classes, course load, and drinking habits.

Question 3)

a.

```
movies <- read.csv("/Users/geoffreyhughes/Documents/MGSC_310/MGSC310/Datasets/movie_metadata.csv")
```

b. Filter out unreasonably large budgets; create new variables with mutate()

```r
library("tidyverse")
movies <- movies %>% filter(budget < 4e+08)
movies <- movies %>% mutate(genre_main = unlist(map(strsplit(as.character(movies$genres),
  "\\|"), 1)), grossM = gross/1e+06, budgetM = budget/1e+06)
movies <- movies %>% mutate(genre_main = factor(genre_main) %>%
  fct_drop())
```

c. Use mutate() to generate profitM and ROI (profit / budget)

```r
movies <- movies %>% mutate(profitM = grossM - budgetM)
movies <- movies %>% mutate(ROI = profitM / budgetM)
names(movies)
```

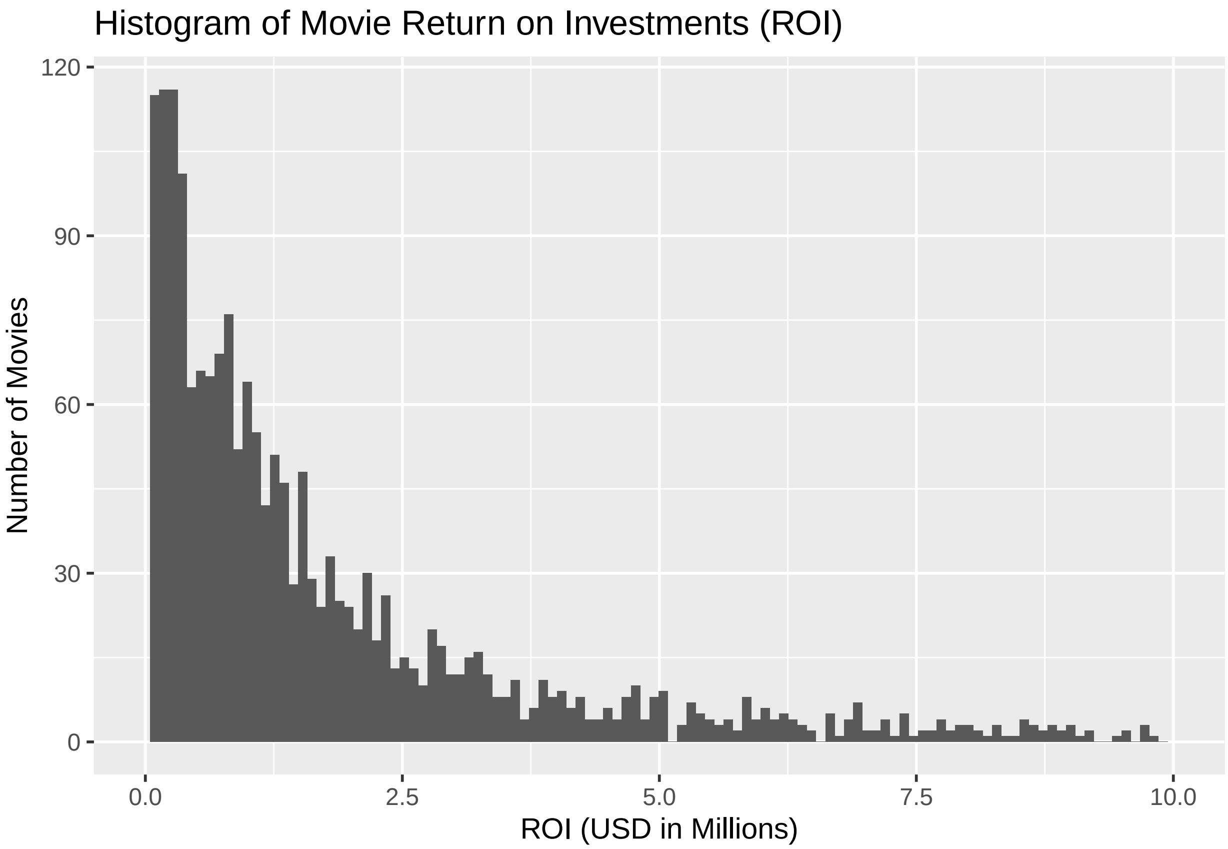
```
##  [1] "color"                     "director_name"
##  [3] "num_critic_for_reviews"    "duration"
##  [5] "director_facebook_likes"   "actor_3_facebook_likes"
##  [7] "actor_2_name"              "actor_1_facebook_likes"
##  [9] "gross"                     "genres"
## [11] "actor_1_name"              "movie_title"
## [13] "num_voted_users"           "cast_total_facebook_likes"
## [15] "actor_3_name"              "facenumber_in_poster"
## [17] "plot_keywords"             "movie_imdb_link"
## [19] "num_user_for_reviews"      "language"
## [21] "country"                   "content_rating"
## [23] "budget"                    "title_year"
## [25] "actor_2_facebook_likes"    "imdb_score"
## [27] "aspect_ratio"              "movie_facebook_likes"
## [29] "genre_main"                "grossM"
## [31] "budgetM"                   "profitM"
## [33] "ROI"
```

d. Average ROI and Histogram

```r
ggplot(data = movies, aes(x = ROI)) +
  geom_histogram(binwidth = 0.09) +
  labs(x = "ROI (USD in Millions)", y = "Number of Movies", title = "Histogram of Movie Return on Inves
  xlim(0, 10)
```

```
## Warning: Removed 2633 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



e. Count and Filter out movies with ROI > 10m

```
count(movies, ROI > 10)
```

```
## # A tibble: 3 x 2
##   `ROI > 10`     n
##   <lgl>      <int>
## 1 FALSE       3734
## 2 TRUE         145
## 3 NA           660
```

```
movies <- filter(movies, movies$ROI < 10)
count(movies, ROI > 10)
```

```
## # A tibble: 1 x 2
##   `ROI > 10`     n
##   <lgl>      <int>
## 1 FALSE       3734
```

As shown above, there are 145 instances where ROI > 10. Filtering them out now.

f. Group movies by Genre, and Summarize them - which have highest ROI?

```
groupby_genre_ROI <- movies %>% group_by(genre_main) %>%
  summarize(mean(ROI))
groupby_genre_ROI
```

```
## # A tibble: 17 x 2
##    genre_main  `mean(ROI)`
##    <fct>             <dbl>
##  1 Action            0.315
##  2 Adventure         0.612
##  3 Animation         0.475
##  4 Biography         0.673
##  5 Comedy            0.750
##  6 Crime             0.423
##  7 Documentary       0.268
##  8 Drama             0.548
##  9 Family           -0.597
## 10 Fantasy           2.09
## 11 Horror            1.40
## 12 Musical           6.41
## 13 Mystery           1.37
## 14 Romance           1.11
## 15 Sci-Fi            0.389
## 16 Thriller          2.35
## 17 Western           5.40
```

It appears Musicals, Westerns, and Fantasy movies have the highest ROIs.

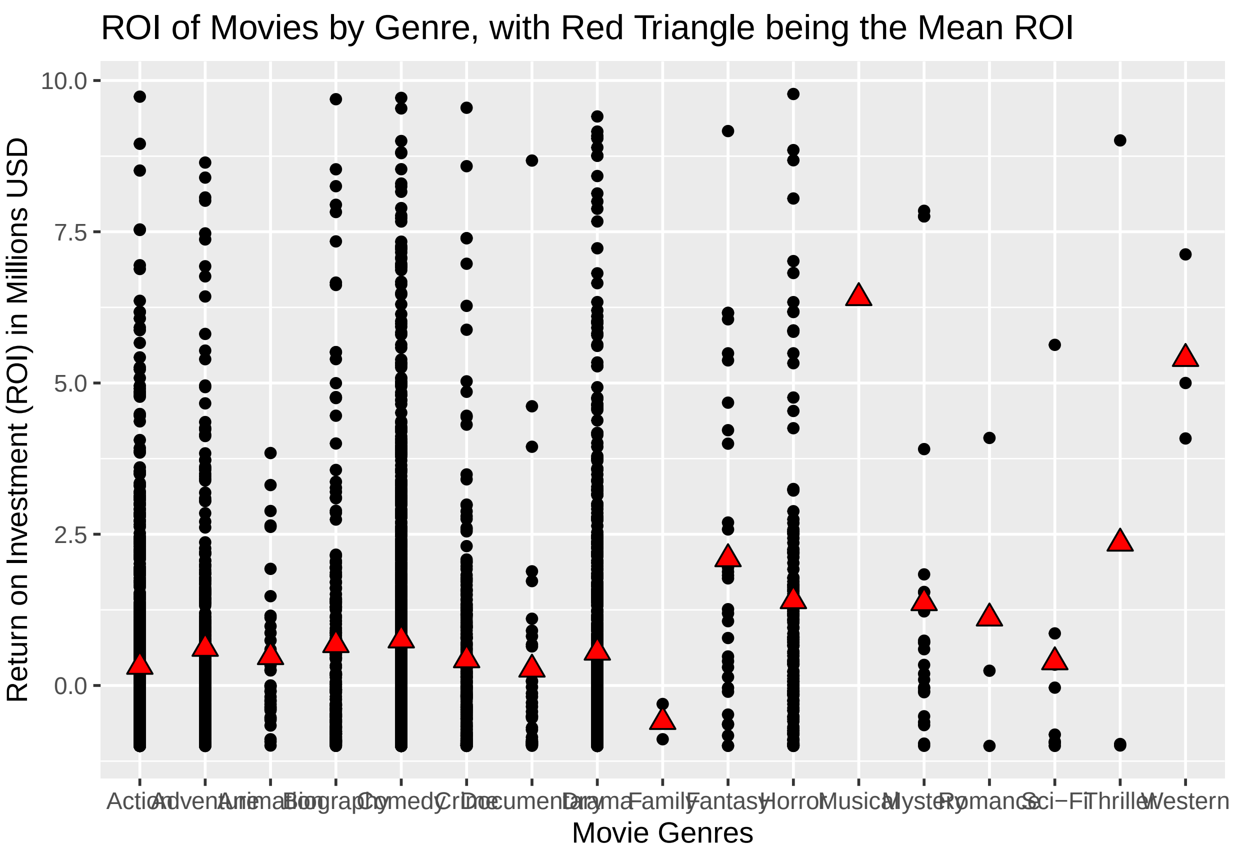
g. Plot graph the Mean ROIs for each Genre

```
library('ggridges')
```

```
##
## Attaching package: 'ggridges'

## The following object is masked from 'package:ggplot2':
##
##     scale_discrete_manual
```

```
ggplot(movies, aes(x = genre_main, y = ROI)) +
  geom_point() +
  labs(x = "Movie Genres",
       y = "Return on Investment (ROI) in Millions USD",
       title = "ROI of Movies by Genre, with Red Triangle being the Mean ROI") +
    stat_summary(
     geom = "point",
     fun.y = "mean",
     col = "black",
     size = 3,
     shape = 24,
     fill = "red"
  )
```


ROI of Movies by Genre, with Red Triangle being the Mean ROI

h. Find Actors with top ROIs

```
actors_ROI <- movies %>% group_by(actor_1_name) %>%
  summarize(actor_mean_ROI = mean(ROI),
            actor_mean_profit = mean(profitM),
            num_films = n())
actors_ROI <- actors_ROI %>% arrange(desc(actor_mean_ROI))
actors_ROI
```

```
## # A tibble: 1,429 x 4
##    actor_1_name    actor_mean_ROI actor_mean_profit num_films
##    <fct>                    <dbl>             <dbl>     <int>
##  1 Matt Shively              9.78           48.9            1
##  2 Alice Krige               9.69           53.3            1
##  3 Ian Gamazon               9.01            0.0631         1
##  4 John Saxon                8.95           40.3            1
##  5 Tiffany Helm              8.68           19.1            1
##  6 John Cothran              8.58           51.5            1
##  7 Lew Temple                8.53           17.1            1
##  8 Anil Kapoor               8.42          126.             1
##  9 William Holden            8.07           24.2            1
## 10 Richard Brooker           8.05           32.2            1
## # ... with 1,419 more rows
```

```
actors_ROI <- actors_ROI %>% slice(1:20)
actors_ROI
```

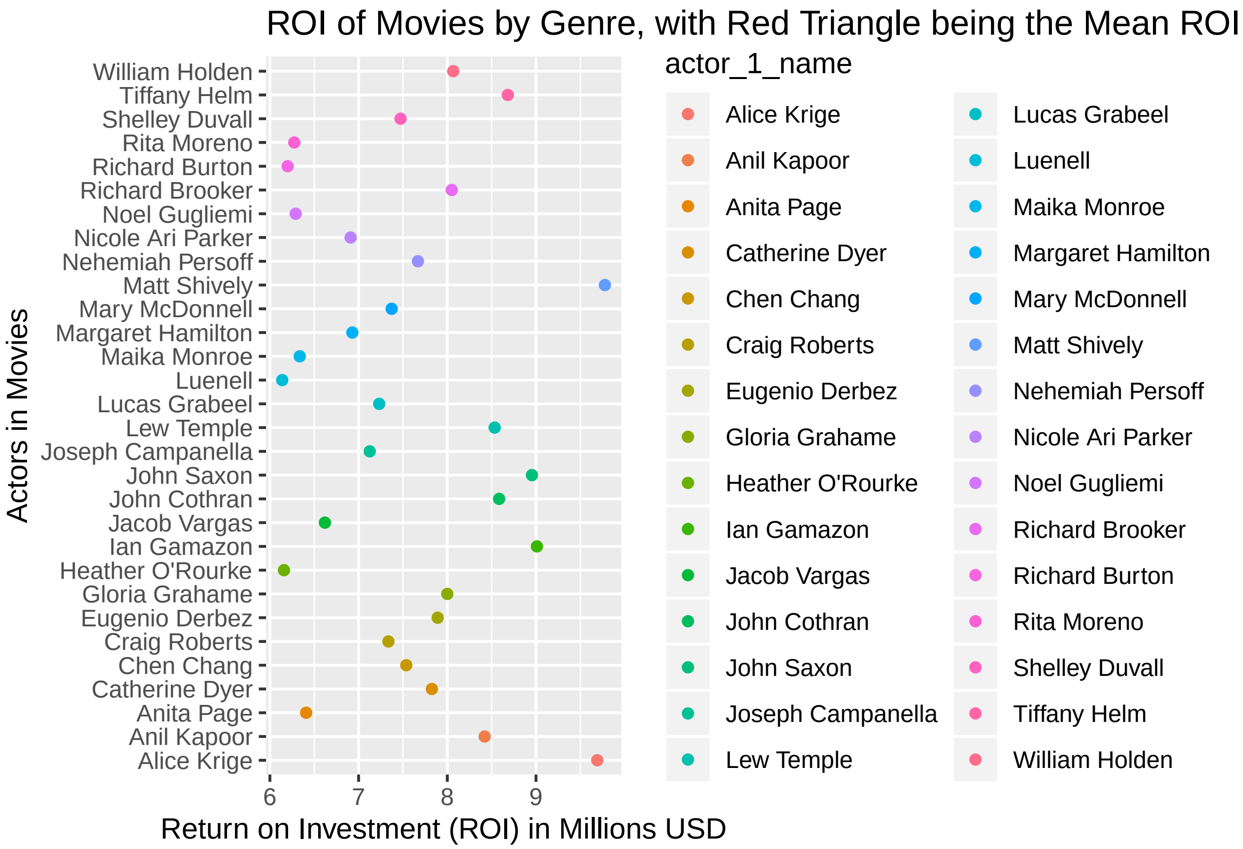
```
## # A tibble: 20 x 4
##    actor_1_name      actor_mean_ROI actor_mean_profit num_films
##    <fct>                      <dbl>             <dbl>     <int>
##  1 Matt Shively                9.78           48.9            1
##  2 Alice Krige                 9.69           53.3            1
##  3 Ian Gamazon                 9.01            0.0631         1
##  4 John Saxon                  8.95           40.3            1
##  5 Tiffany Helm                8.68           19.1            1
##  6 John Cothran                8.58           51.5            1
##  7 Lew Temple                  8.53           17.1            1
##  8 Anil Kapoor                 8.42          126.             1
##  9 William Holden              8.07           24.2            1
## 10 Richard Brooker             8.05           32.2            1
## 11 Gloria Grahame              8              32              1
## 12 Eugenio Derbez              7.89           39.5            1
## 13 Catherine Dyer              7.83          227.             1
## 14 Nehemiah Persoff            7.67           22.1            1
## 15 Chen Chang                  7.54          113.             1
## 16 Shelley Duvall              7.47           37.4            1
## 17 Mary McDonnell              7.37          162.             1
## 18 Craig Roberts               7.34          132.             1
## 19 Lucas Grabeel               7.23           79.6            1
## 20 Joseph Campanella           7.13            0.214          1
```

```
actors_ROI <- movies %>% group_by(actor_1_name) %>%
  summarize(actor_mean_ROI = mean(ROI),
            actor_mean_profit = mean(profitM),
            num_films = n())
actors_ROI <- actors_ROI %>% arrange(desc(actor_mean_ROI))
```

Finally!!! It appears Matt Shively, Alice Krige, and Ian Gamazon have the highest average ROIs.

i. Plot actors with the 30 highest ROIs

```
actors_ROI <- actors_ROI %>% slice(1:30)
ggplot(actors_ROI, aes(x = actor_mean_ROI, y = actor_1_name, fill = actor_1_name, color = actor_1_name)
  geom_point() +
  labs(x = "Return on Investment (ROI) in Millions USD",
       y = "Actors in Movies",
       title = "ROI of Movies by Genre, with Red Triangle being the Mean ROI")
```

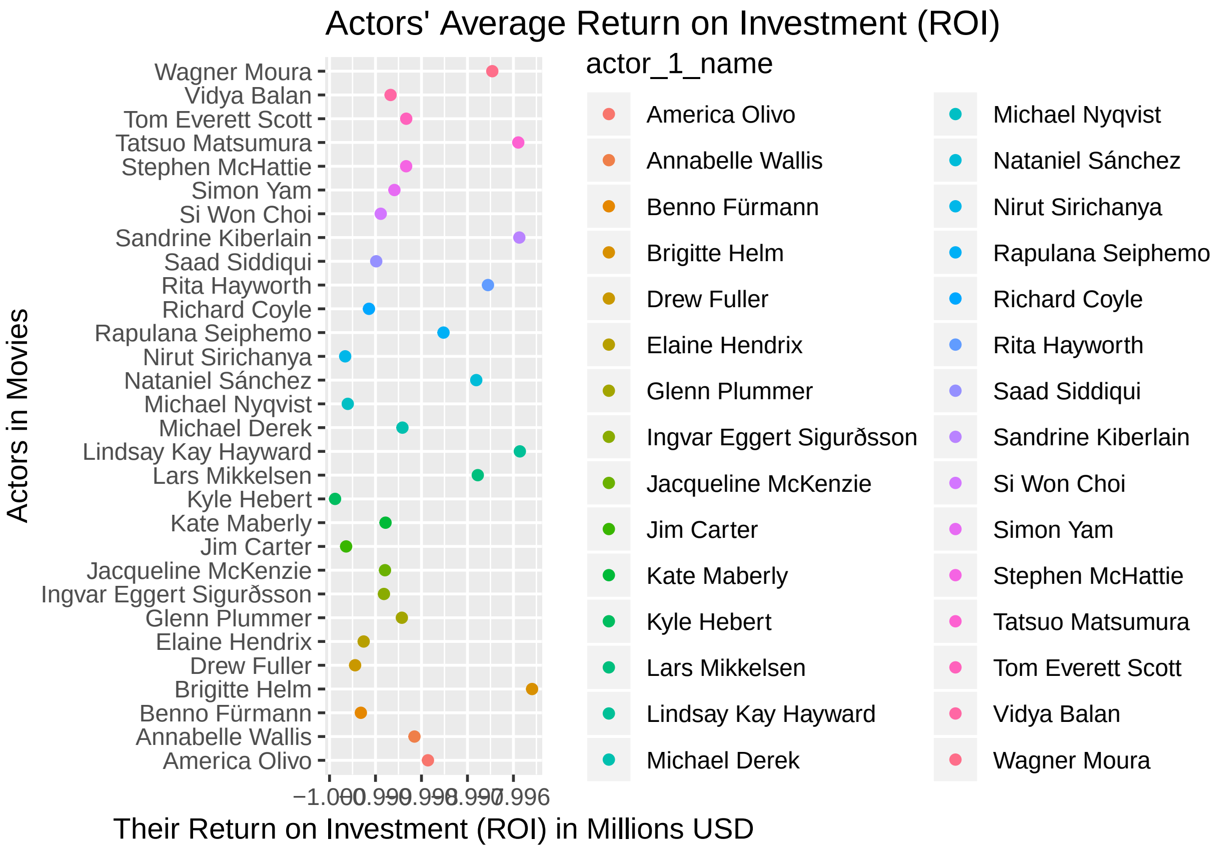


```
actors_ROI <- movies %>% group_by(actor_1_name) %>%
  summarize(actor_mean_ROI = mean(ROI),
            actor_mean_profit = mean(profitM),
            num_films = n())


actors_ROI <- actors_ROI %>% arrange(actor_mean_ROI)

actors_ROI <- actors_ROI %>% slice(1:30)
actors_ROI <- actors_ROI %>% arrange(actor_mean_ROI)

ggplot(actors_ROI, aes(x = actor_mean_ROI, y = actor_1_name, fill = actor_1_name, color = actor_1_name)
  geom_point() +
  labs(x = "Their Return on Investment (ROI) in Millions USD",
       y = "Actors in Movies",
       title = "Actors' Average Return on Investment (ROI)")
```


Actors' Average Return on Investment (ROI)
Actors in Movies
Wagner Moura
Vidya Balan
Tom Everett Scott
Tatsuo Matsumura
Stephen McHattie
Simon Yam
Si Won Choi
Sandrine Kiberlain
Saad Siddiqui
Rita Hayworth
Richard Coyle
Rapulana Seiphemo
Nirut Sirichanya
Nataniel Sánchez
Michael Nyqvist
Michael Derek
Lindsay Kay Hayward
Lars Mikkelsen
Kyle Hebert
Kate Maberly
Jim Carter
Jacqueline McKenzie
Ingvar Eggert Sigurðsson
Glenn Plummer
Elaine Hendrix
Drew Fuller
Brigitte Helm
Benno Fürmann
Annabelle Wallis
America Olivo
−1.000
−0.999
−0.998
−0.997
−0.996
Their Return on Investment (ROI) in Millions USD
actor_1_name
America Olivo
Annabelle Wallis
Benno Fürmann
Brigitte Helm
Drew Fuller
Elaine Hendrix
Glenn Plummer
Ingvar Eggert Sigurðsson
Jacqueline McKenzie
Jim Carter
Kate Maberly
Kyle Hebert
Lars Mikkelsen
Lindsay Kay Hayward
Michael Derek
Michael Nyqvist
Nataniel Sánchez
Nirut Sirichanya
Rapulana Seiphemo
Richard Coyle
Rita Hayworth
Saad Siddiqui
Sandrine Kiberlain
Si Won Choi
Simon Yam
Stephen McHattie
Tatsuo Matsumura
Tom Everett Scott
Vidya Balan
Wagner Moura

WOW! gg, wp