Problem Set 2

MGSC 310, Fall 2019, Professor Hersh

Elmer Camargo + Nick Trella

Libraries Needed

```
library("tidyverse")
library("ggplot2")
```

Question 1 ISLR Ch.2 Q.2

A. Regression. n(sample) = whatever subset we pick, p(predictors) = the vars

B. Classification. n = 20 similair products, p = success, failure, price, mark budget, comp price, and 10 other vars

C. Regression because output is expected to be a percentage (aka continuous data) Prediction because we are forcasting future percentage change (n = 52, p = % change in [USD/Euro, US Market, British Market, German Market])

Question 2 ISLR Ch.2 Q.4

A. Classifying faces on images, Response: yes or no, Predictors: nose, eyes, jaw, etc… Applicational Goal: Predictive because objects on the images are being categorized

Classifying whether or not to give someone 1 of 3 loan, Response: small, medium, large
Predictors: income, networth, credit history, etc… Applicational Goal: Prediction because output is being categorized into 3 types of loans

Classifying whether someone is a male or female based on previous purchases Response: male or female Predictors: types of purchases, stores of purchases Applicational Goal: Inference because you are exploring the relationship of previous purchases

B. Using a regression model to examine the relationship of marijuana dispensaries and crime in a location Response: Reported crimes in a given location Predictors: Marijuana dispensary locations, historical crime reports in locations Applicational Goal: Inference because you are exploring the relationship between crime and marijuana dispensary within a specified location

Using a regression model to predict a sports teams number of points in a game Response: Points in a game Predictors: Individual player points per game average, defensive stats of the oponent Applicational Goal: Predictive because you are estimating

Using a regression model to predict percent change in a stock Response: Predicted percent change stock Predictors: Previous percent change of stock, media coverage Applicational Goal: Predictive because you are estimating a future variable

C. Using cluster analysis to group businesses together by what they sell Using cluster analysis to group people by income Using cluster analysis to group people interests/facebook likes

Question 3a-b Plotting IMDB's Top 5000 Movies

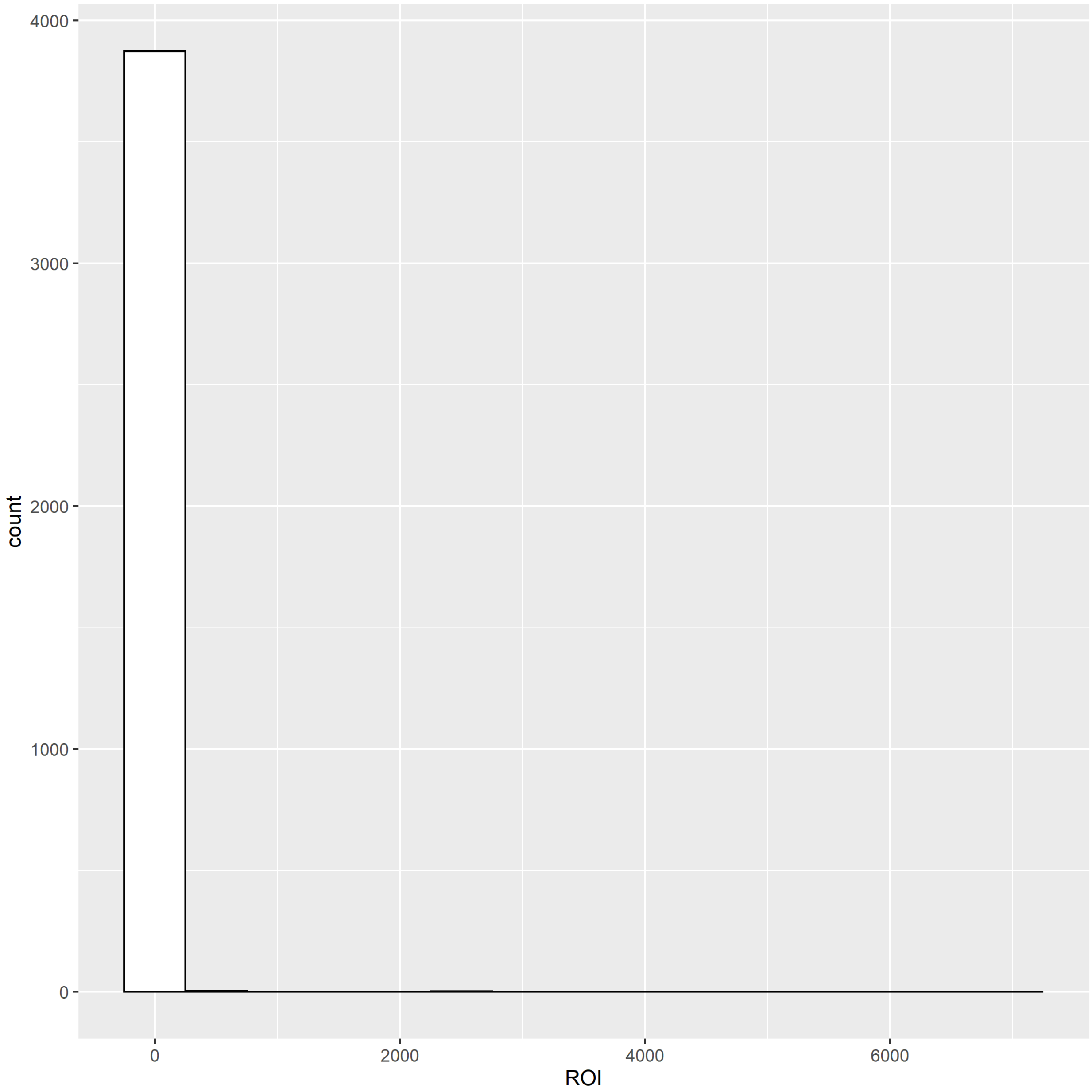
```
movies <-read.csv("data/movie_metadata.csv")

movies <- movies %>% filter(budget<4e+08) #get rid of anomolies
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())
```

Question 3c Profit and ROI

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

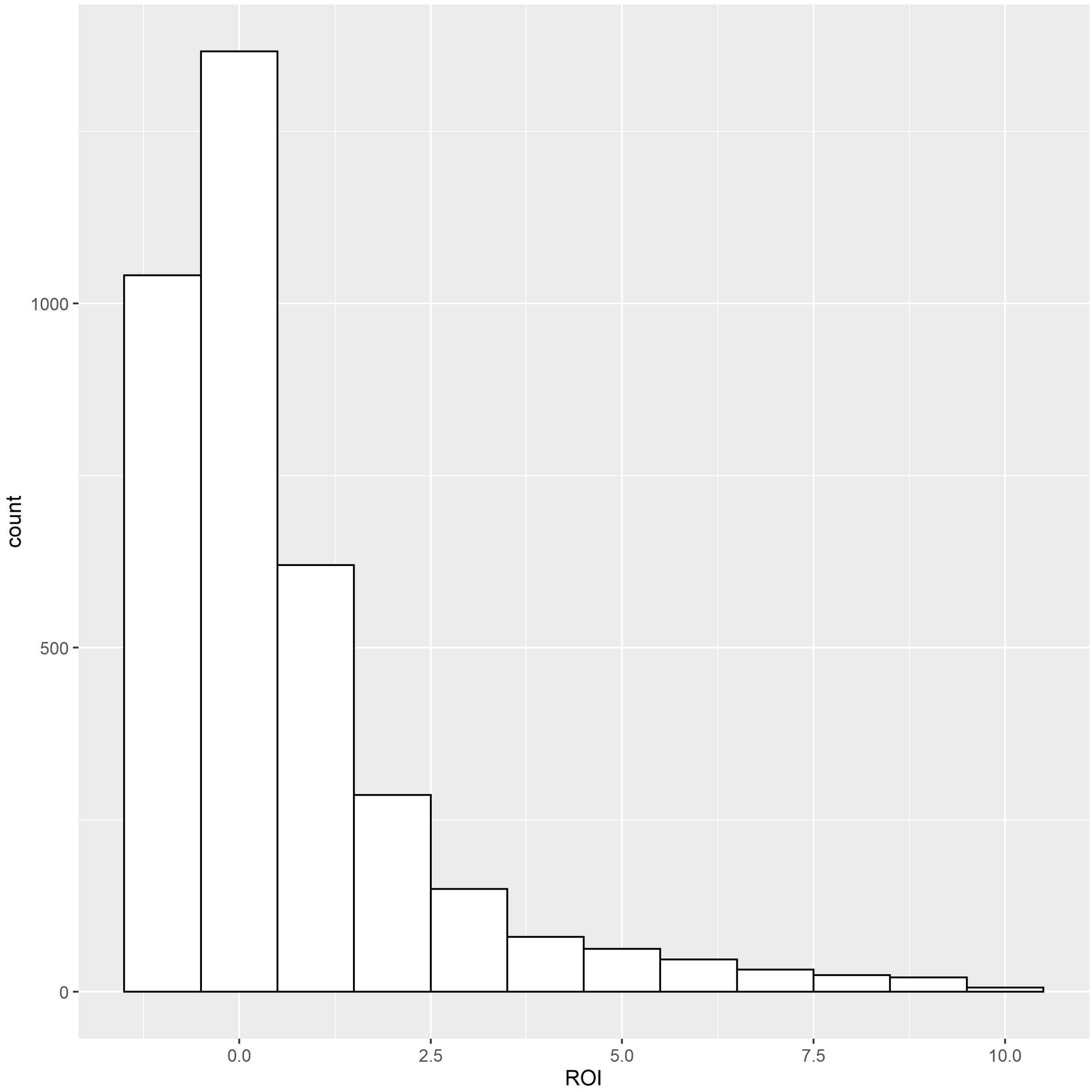
Question 3d Average ROI Plot

```
sum(is.na(movies$ROI))
## [1] 660
movies <- movies %>% drop_na(ROI) #omits NA values in a column
sum(is.na(movies$ROI))
## [1] 0

cat('average ROI is', mean(movies$ROI))
## average ROI is 5.273088

hgp1<-ggplot(movies, aes(x=ROI)) +
  geom_histogram(color="black", fill="white", binwidth = 500)
```


4000
3000
2000
1000
0
count
0
2000
4000
6000
ROI

Question 3e Outliers and Filtering

```
count(movies, vars = ROI > 10)
## # A tibble: 2 x 2
##   vars      n
##   <lgl> <int>
## 1 FALSE  3734
## 2 TRUE    145

movies_filt <- movies %>% filter(ROI < 10) #we want/keep everything < 10

count(movies_filt, vars = ROI > 10)
## # A tibble: 1 x 2
##   vars      n
##   <lgl> <int>
## 1 FALSE  3734

hp2 <- ggplot(data = movies_filt, aes(ROI))+
  geom_histogram(color="black", fill="white", binwidth = 1)
```

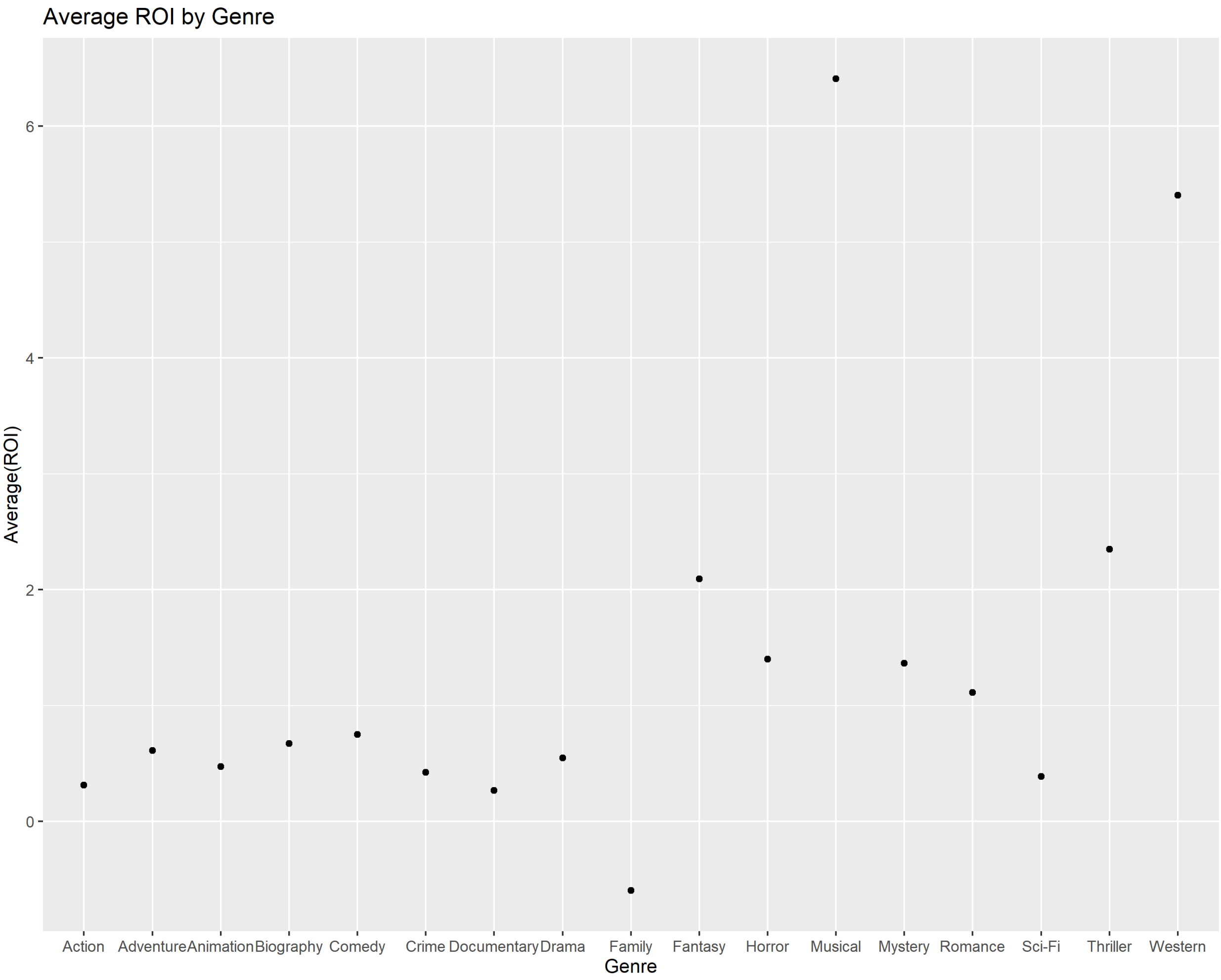

count
1000
500
0
0.0
2.5
5.0
7.5
10.0
ROI

Question 3f Grouping and Summarizing

```
average_roi_bycat <- movies_filt %>%
  group_by(genre_main) %>%
  summarize(mean(ROI))

average_roi_bycat
## # 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

cat("Top 3 Genres: Musical, Western, and Thriller")
## Top 3 Genres: Musical, Western, and Thriller
```

Question 3g

```
genre_meanROI <- average_roi_bycat$`mean(ROI)`
genre <- average_roi_bycat$genre_main

sp1 <- ggplot( data = average_roi_bycat)+
  geom_point(mapping = aes(x = genre, y = genre_meanROI)) +
  labs(x= "Genre",
       y= "Average(ROI)",
       title= "Average ROI by Genre")
```

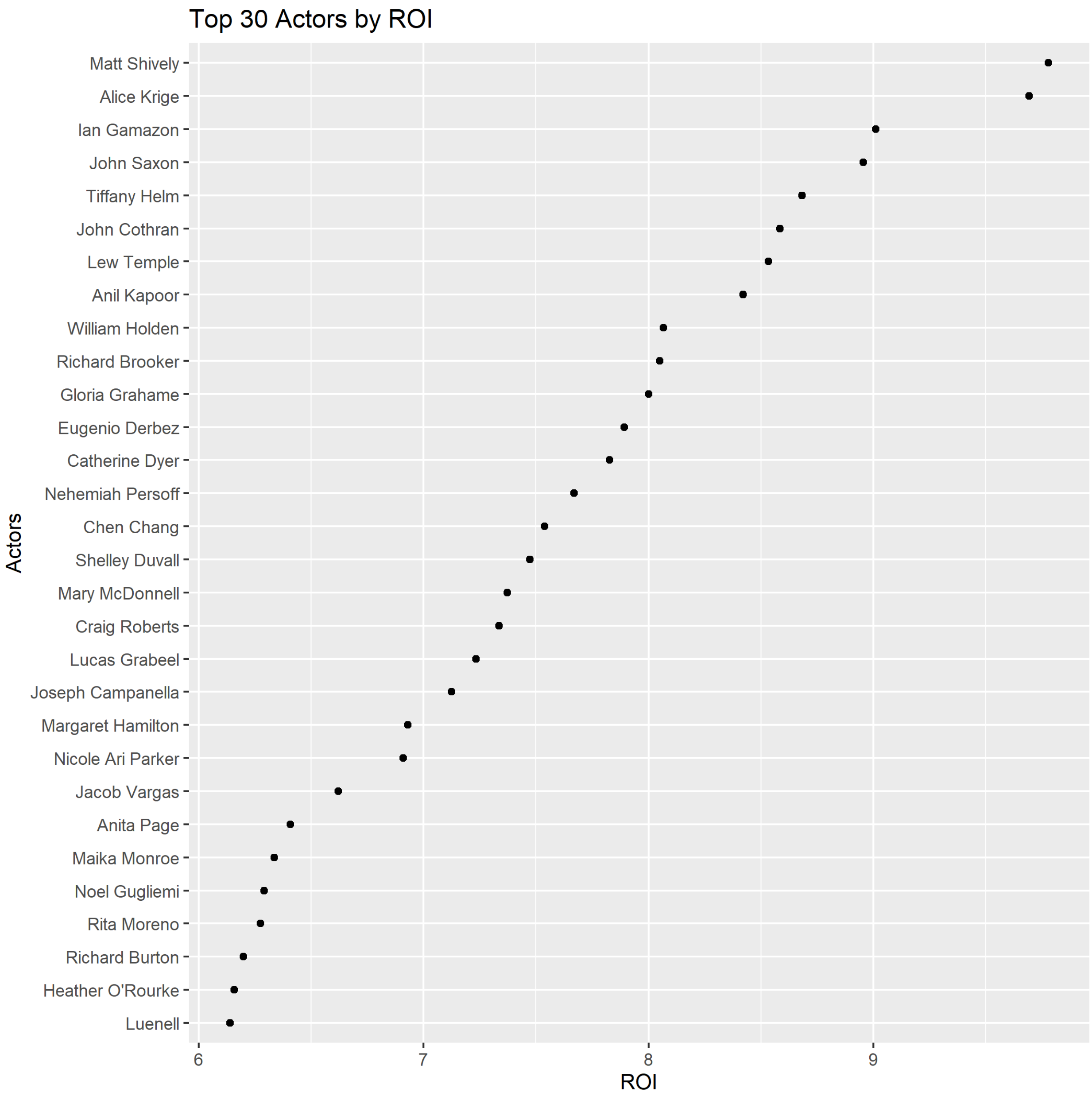

Average ROI by Genre
Average(ROI)
6
4
2
0
Action
Adventure
Animation
Biography
Comedy
Crime
Documentary
Drama
Family
Fantasy
Horror
Musical
Mystery
Romance
Sci-Fi
Thriller
Western
Genre

Question 3h

```
test3 <- group_by(movies_filt,actor_1_name)

df2 <- summarise(test3,mean(ROI),mean(profitM),num_films = n())
x2 <- df2$`mean(ROI)`
df2 <- df2 %>% arrange(desc(x2))


df2 <- df2 %>% slice(1:20)

df2
## # A tibble: 20 x 4
##    actor_1_name      `mean(ROI)` `mean(profitM)` 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
```

Question 3i

```
df3 <- summarise(test3,mean(ROI))
y3 <- df3$actor_1_name
x3 <- df3$`mean(ROI)`
df3 <- df3 %>% arrange(desc(x3))
df3 <- df3 %>% slice(1:30)
y3 <- df3$actor_1_name
x3 <- df3$`mean(ROI)`
df3
## # A tibble: 30 x 2
##    actor_1_name     `mean(ROI)`
##    <fct>                  <dbl>
##  1 Matt Shively            9.78
##  2 Alice Krige             9.69
##  3 Ian Gamazon             9.01
##  4 John Saxon              8.95
##  5 Tiffany Helm           8.68
##  6 John Cothran           8.58
##  7 Lew Temple             8.53
##  8 Anil Kapoor            8.42
##  9 William Holden         8.07
## 10 Richard Brooker        8.05
## # ... with 20 more rows

sp2 <- ggplot(df3 = df3 %>% top_n(30, wt = x3), mapping = aes(x = x3, y = reorder(y3, x3)))+geom_point(
```


Top 30 Actors by ROI
Actors
Matt Shively
Alice Krige
Ian Gamazon
John Saxon
Tiffany Helm
John Cothran
Lew Temple
Anil Kapoor
William Holden
Richard Brooker
Gloria Grahame
Eugenio Derbez
Catherine Dyer
Nehemiah Persoff
Chen Chang
Shelley Duvall
Mary McDonnell
Craig Roberts
Lucas Grabeel
Joseph Campanella
Margaret Hamilton
Nicole Ari Parker
Jacob Vargas
Anita Page
Maika Monroe
Noel Gugliemi
Rita Moreno
Richard Burton
Heather O'Rourke
Luenell
6
7
8
9
ROI

## Question 3j

```
df4 <- summarise(test3,mean(ROI))
y4 <- df4$actor_1_name
x4 <- df4$`mean(ROI)`
df4 <- df4 %>% arrange((x4))
df4 <- df4 %>% slice(30:1)
y4 <- df4$actor_1_name
x4 <- df4$`mean(ROI)`

sp3 <- ggplot(df4 = df4 %>% top_n(30, wt = x4), mapping = aes(x = x4, y = reorder(y4, x4)))+geom_point(
```

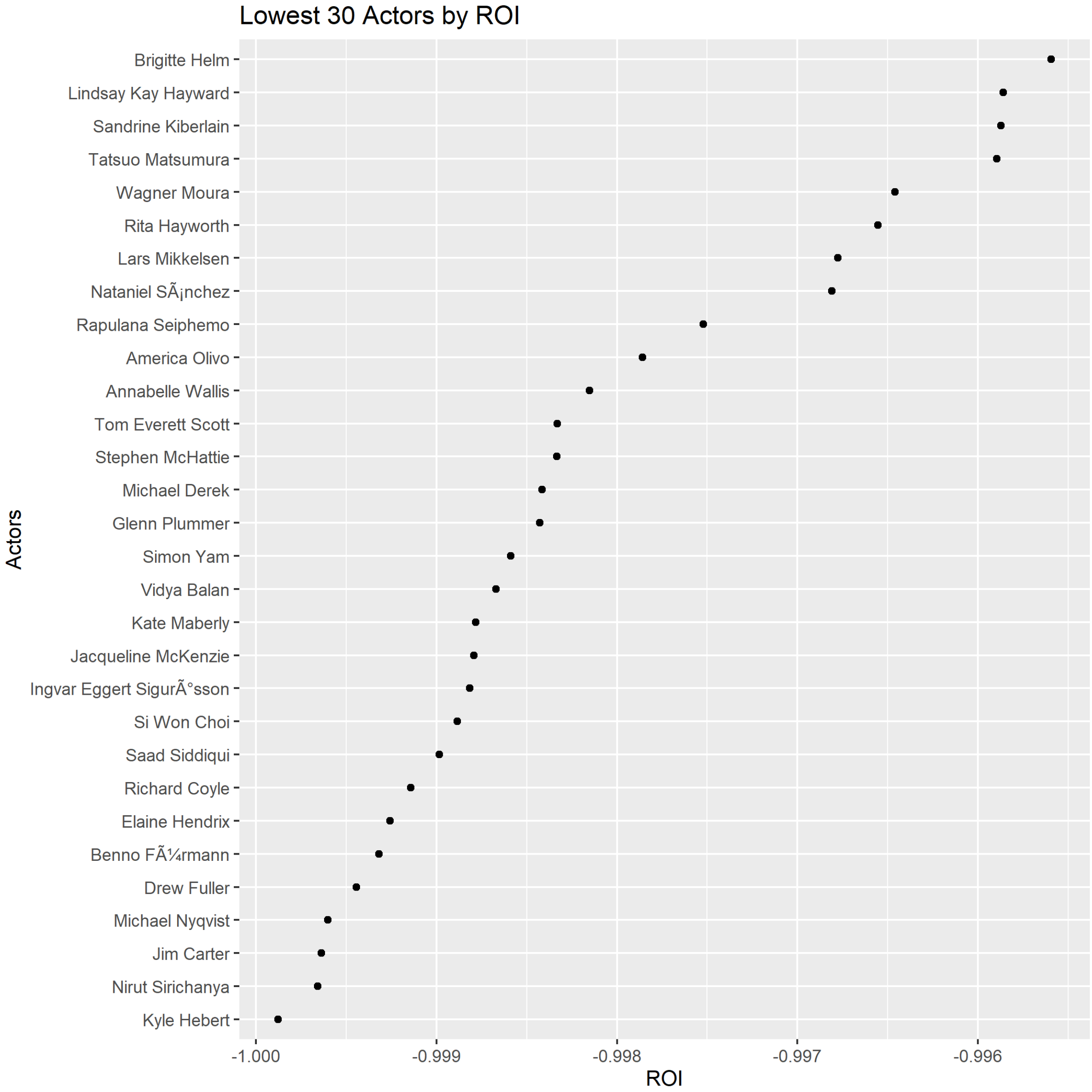


Figure 1: scatterplot 3