# 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 whether an image contains a face or not

```
Response: yes or no

Predictors: nose, eyes, jaw, etc...

Applicational Goal: Predictive because 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

Classifying whether someone will return to a Response: yes or no Predictors: nose, eyes, jaw, etc… Applicational Goal: Predictive because images are being categorized

B. Using a regression model to predict the path of a vehicle Response: Angles in degrees
Predictors: Speed, angle of tires Applicational Goal: Predictive because

Using a regression model to Response:
Predictors: Applicational Goal: Inferential because

Using a regression model to Response:
Predictors: Applicational Goal: Predictive because

C.

## 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)
```

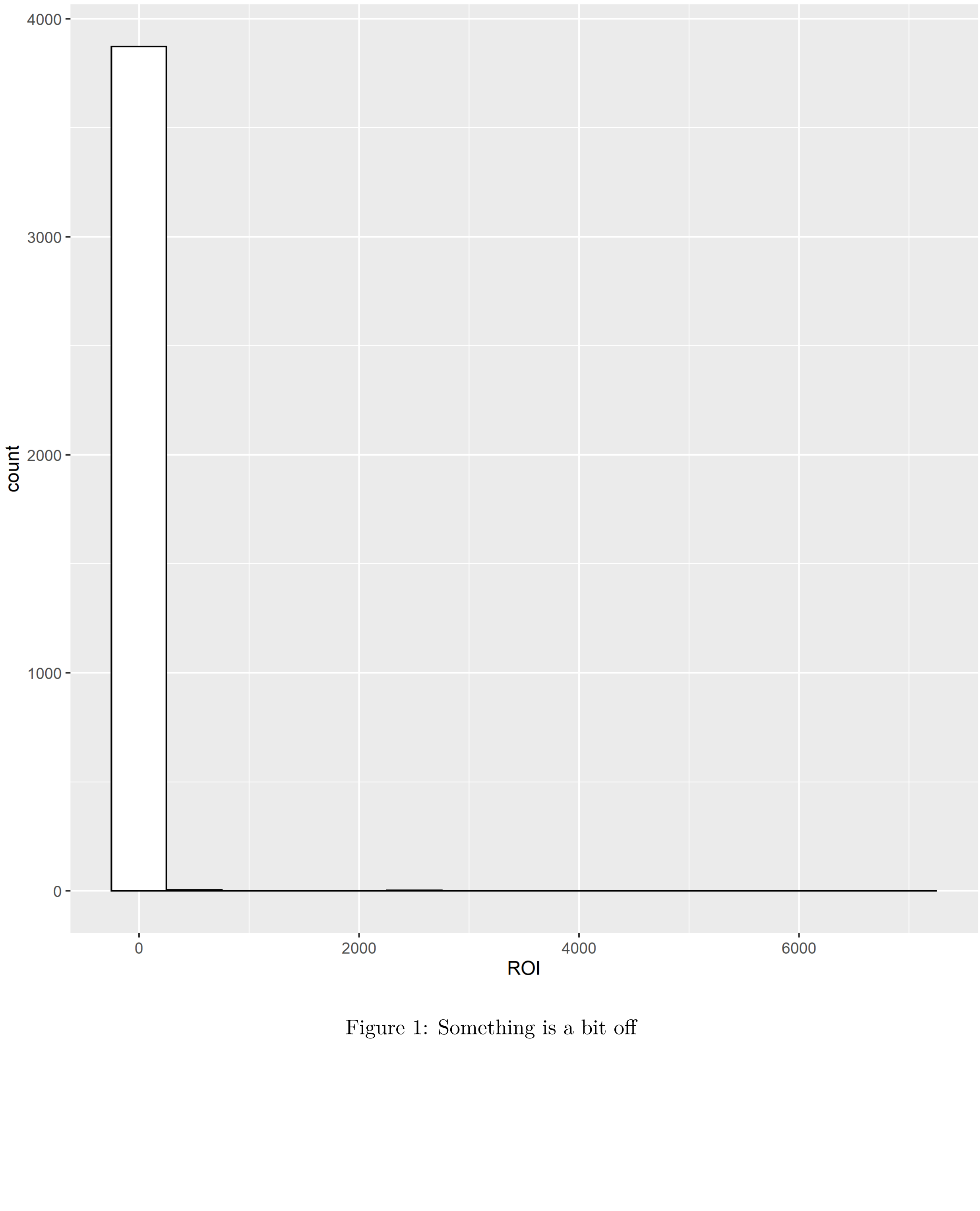


Figure 1: Something is a bit off

### 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)
```

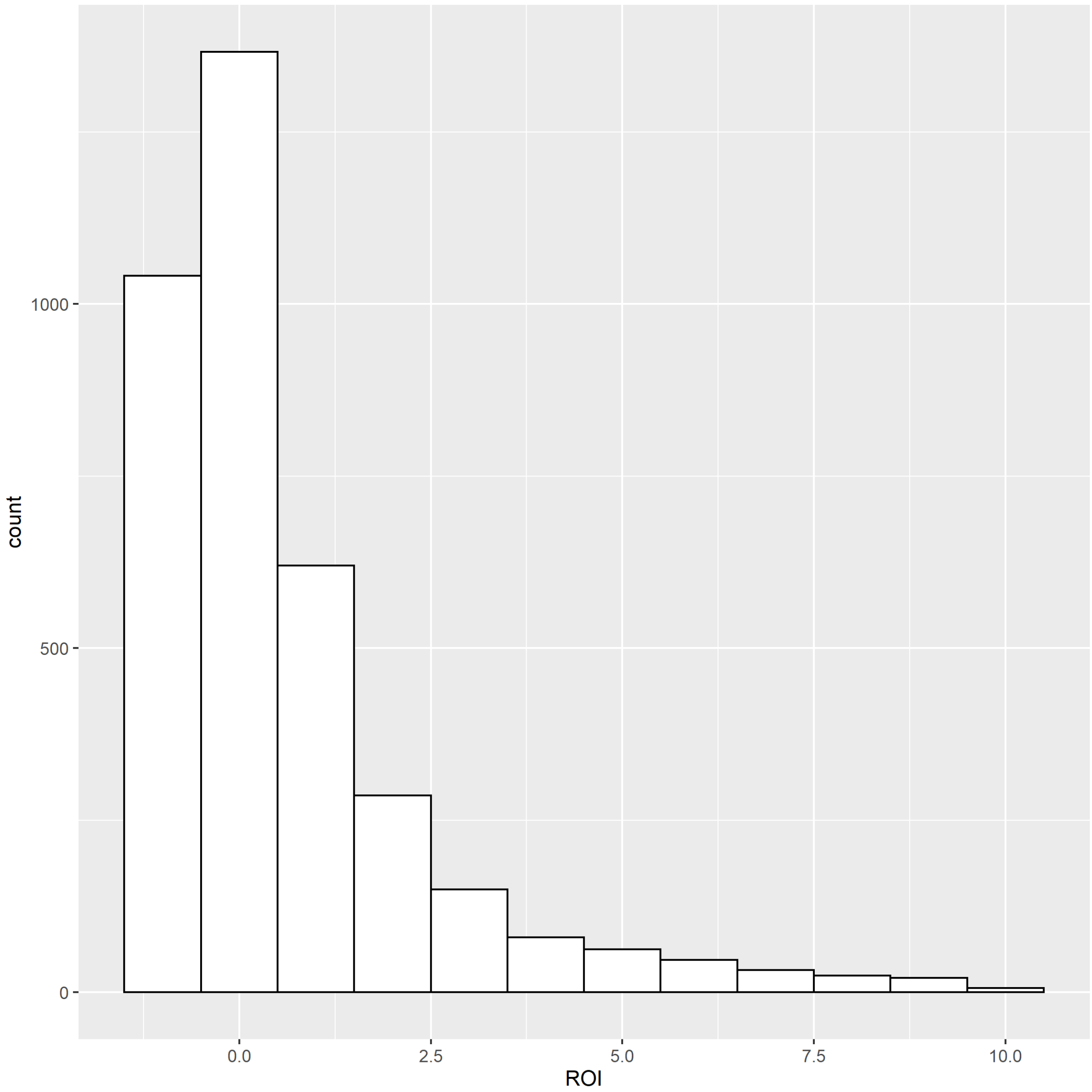


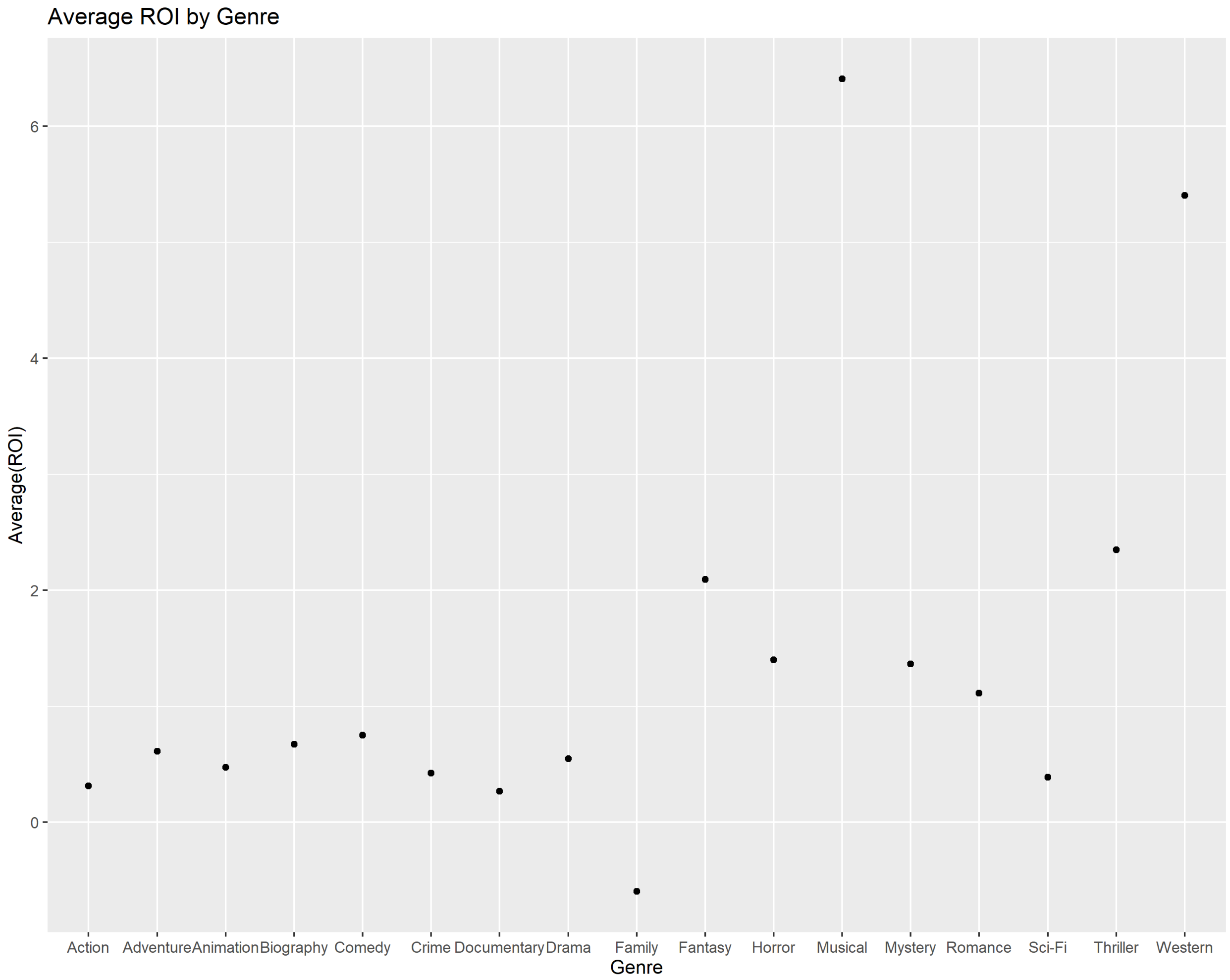
Figure 2: Way better

### 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")
```



## Question 3h

## Question 3i

## Question 3j