

```
1 # Calculates summary statistics and conducts basic regression analysis to determine
2 # whether movies which pass the Bechdel test do better or worse at the box office,
3 # using data from www.bechdeltest.com and www.the-numbers.com
4
5 # By Andrew Flowers <andrew.flowers@fivethirtyeight.com>
6 # See also http://fivethirtyeight.com/features/the-dollar-and-cents-case-against-
  hollywoods-exclusion-of-women/
7
8 # Install and load required packages
9 # install.packages(c("gdata", "cwhmisc"))
10 library(gdata)
11 library(cwhmisc)
12
13 # Load data
14 rawData<-read.csv("movies.csv", na.strings="#N/A")
15
16 # Select movies pre-1990, and format $-denominated data fields
17 rawData<-rawData[rawData$year>1989,]
18
19 # International-only gross profits (which equal total profits minus domestic profits)
20 rawData$intOnly<-rawData$intgross_2013.-rawData$domgross_2013.
21
22 # Return on Investment (ROI) measures
23 rawData$ROI<-rawData$intgross_2013./rawData$budget_2013. # Total ROI
24 rawData$ROI1<-rawData$domgross_2013./rawData$budget_2013. # Domestic ROI
25 rawData$ROI2<-rawData$intOnly/rawData$budget_2013. # International ROI
26
27 # Divide movies into FAIL and PASS divisions
28 failMovies<-rawData[rawData$binary=="FAIL",]
29 passMovies<-rawData[rawData$binary=="PASS",]
30
31 # Include a "generous" category (which includes both "ok" and "dubious" movies)
32 generous<-rbind(rawData[rawData$clean_test=="ok",],
33   rawData[rawData$clean_test=="dubious",])
34
35 # Print medians: ROI and budget
36 median(failMovies$ROI, na.rm=T)
37 median(passMovies$ROI, na.rm=T)
38 median(rawData$ROI, na.rm=T)
39
40 median(failMovies$budget_2013.)
41 median(passMovies$budget_2013.)
42 median(rawData$budget_2013.)
43
44 # Distributions and logs
45 hist(rawData$budget_2013.)
46 hist(log(rawData$budget_2013.))
47
48 hist(rawData$intgross_2013.)
49 hist(log(rawData$intgross_2013.))
50
51 hist(rawData$ROI)
52 hist(log(rawData$ROI))
53
54 # Linear regression models
55
56 # Movies with higher budgets make more gross revenues
57 summary(lm(log(intgross_2013.)~log(budget_2013.), data=rawData))
58
59 # Bechdel dummy is not significant
60 summary(lm(log(intgross_2013.)~log(budget_2013.)+factor(binary), data=rawData))
```

```
61 # Movies with higher budgets have lower ROI
62 summary(lm(log(ROI)~log(budget_2013.), data=rawData))
63
64 # Bechdel dummy is not significant
65 summary(lm(log(ROI)~log(budget_2013.)+factor(binary), data=rawData))
66
67 # ROI #1 (domestic) used in chart
68 median(generous$ROI1, na.rm=T)
69 median(rawData$ROI1[rawData$clean_test=="men"], na.rm=T)
70 median(rawData$ROI1[rawData$clean_test=="notalk"], na.rm=T)
71 median(rawData$ROI1[rawData$clean_test=="nowomen"], na.rm=T)
72
73 # ROI #2 (international) used in chart
74 median(generous$ROI2, na.rm=T)
75 median(rawData$ROI2[rawData$clean_test=="men"], na.rm=T)
76 median(rawData$ROI2[rawData$clean_test=="notalk"], na.rm=T)
77 median(rawData$ROI2[rawData$clean_test=="nowomen"], na.rm=T)
78
79
```

