```r
# Calculates summary statistics and conducts basic regression analysis to determine
# whether movies which pass the Bechdel test do better or worse at the box office,
# using data from www.bechdeltest.com and www.the-numbers.com

# By Andrew Flowers <andrew.flowers@fivethirtyeight.com>
# See also http://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/

# Install and load required packages
# install.packages(c("gdata", "cwhmisc"))
library(gdata)
library(cwhmisc)

# Load data
rawData<-read.csv("movies.csv", na.strings="#N/A")

# Select movies pre-1990, and format $-denominated data fields
rawData<-rawData[rawData$year>1989,]

# International-only gross profits (which equal total profits minus domestic profits)
rawData$intOnly<-rawData$intgross_2013.-rawData$domgross_2013.

# Return on Investment (ROI) measures
rawData$ROI<-rawData$intgross_2013./rawData$budget_2013. # Total ROI
rawData$ROI1<-rawData$domgross_2013./rawData$budget_2013. # Domestic ROI
rawData$ROI2<-rawData$intOnly/rawData$budget_2013. # International ROI

# Divide movies into FAIL and PASS divisions
failMovies<-rawData[rawData$binary=="FAIL",]
passMovies<-rawData[rawData$binary=="PASS",]

# Include a "generous" category (which includes both "ok" and "dubious" movies)
generous<-rbind(rawData[rawData$clean_test=="ok",],
rawData[rawData$clean_test=="dubious",])

# Print medians: ROI and budget
median(failMovies$ROI, na.rm=T)
median(passMovies$ROI, na.rm=T)
median(rawData$ROI, na.rm=T)

median(failMovies$budget_2013.)
median(passMovies$budget_2013.)
median(rawData$budget_2013.)

# Distributions and logs
hist(rawData$budget_2013.)
hist(log(rawData$budget_2013.))

hist(rawData$intgross_2013.)
hist(log(rawData$intgross_2013.))

hist(rawData$ROI)
hist(log(rawData$ROI))

# Linear regression models

# Movies with higher budgets make more gross revenues
summary(lm(log(intgross_2013.)~log(budget_2013.), data=rawData))

# Bechdel dummy is not significant
summary(lm(log(intgross_2013.)~log(budget_2013.)+factor(binary), data=rawData))

```

```r
# Movies with higher budgets have lower ROI
summary(lm(log(ROI)~log(budget_2013.), data=rawData))

# Bechdel dummy is not significant
summary(lm(log(ROI)~log(budget_2013.)+factor(binary), data=rawData))

# ROI #1 (domestic) used in chart
median(generous$ROI1, na.rm=T)
median(rawData$ROI1[rawData$clean_test=="men"], na.rm=T)
median(rawData$ROI1[rawData$clean_test=="notalk"], na.rm=T)
median(rawData$ROI1[rawData$clean_test=="nowomen"], na.rm=T)

# ROI #2 (international) used in chart
median(generous$ROI2, na.rm=T)
median(rawData$ROI2[rawData$clean_test=="men"], na.rm=T)
median(rawData$ROI2[rawData$clean_test=="notalk"], na.rm=T)
median(rawData$ROI2[rawData$clean_test=="nowomen"], na.rm=T)


```