Homework 08 One Way ANOVA

Due by 11:59pm, Friday, April 11, 2025

S&DS 230/530/ENV 757

The CSV file `HindiFilm.csv` which you can get HERE contains the 1698 Hindi language movies that released in India across a 13 year period (2005-2017) (thanks to Premkumar, Prashant (2020), "Bollywood Movies data", Mendeley Data, V1, doi: 10.17632/3c57btcxy9.1). Here are the variables:

- Name
- Period (normal or holiday)
- Remake (Yes/No)
- Franchise (Yes/No)
- Genre
- New_Actor (Yes/No)
- New_Director (Yes/No)
- New_Musician (Yes/No - i.e. new music director)
- Lead (Lead actor name)
- Director (Name)
- Musician (Music Director name)
- Screens (how many screens was it shown on)
- Revenue (Indian Rupees)
- Budget (Indian Rupees - usual shorthand is INR)

1) One Way ANOVA of percent return for movies. *(75 pts - 6 pts each except part 1.12 which is 9 pts)*

1.1) Read the data into an object called `movie` (do NOT use the option `as.is = TRUE`). Update this object so that is only contains the columns `Budget`, `Revenue`, `Genre`, and `Name`. In addition, retain only rows that have complete data for these four columns.

Next, create a new column called `pctReturn` that calculates the percentage return (have this be a percentage, NOT a fraction).

Make a table of `Genre` and sort this from high to low. Show these results.

Update `movie` so that it only contains movies in the following genres : "masala", "drama", "fantasy", "love_story", "action".

As per usual, look at the first few rows of the data, get the dimension of the final dataset.

```r
movie <- read.csv(
  'https://raw.githubusercontent.com/jreuning/sds230_data/refs/heads/main/HindiFilm.csv')
movie <- movie[,c('Budget', 'Revenue', 'Genre', 'Name')]
movie <- movie[complete.cases(movie),]
movie['pctReturn'] <- (movie["Revenue"]) / movie["Budget"] * 100
sort(table(movie['Genre']), decreasing=TRUE)
```

```
## Genre
##        drama       comedy     thriller   love_story       action     rom__com
##          639          284          212          133          127           95
##        adult       horror     suspense       masala mythological      fantasy
##           78           53           30           16           14           13
##    animation  documentary
##            3            1
```

```r
movie <- movie[movie$Genre %in% c("masala", "drama", "fantasy",
                                  "love_story", "action"),]
dim(movie)
```

```
## [1] 928   5
```

```r
head(movie, 5)
```

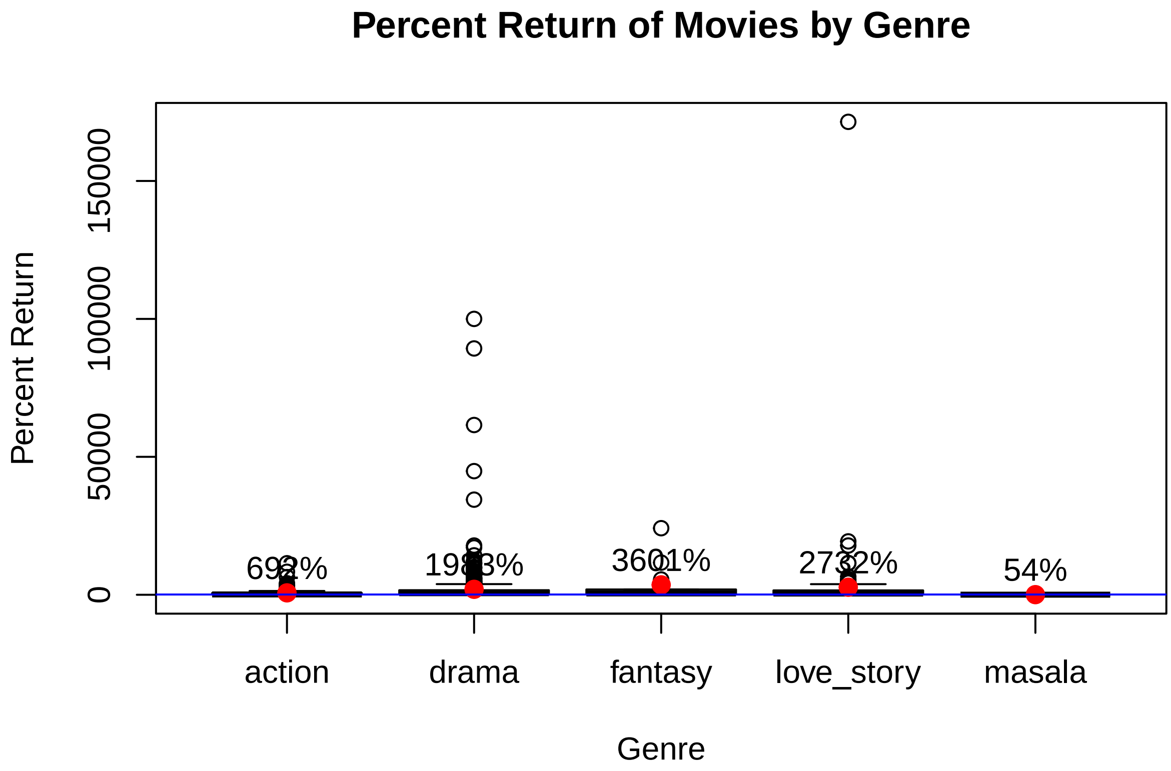
```
##        Budget   Revenue  Genre                    Name  pctReturn
## 2      825000  15000000  drama           Kaccha Limboo 1818.18182
## 4     4500000 210000000  drama              Qaidi Band 4666.66667
## 8      825000  15000000  drama Future To Bright Hai Ji 1818.18182
## 9  1945820000 520000000 action                 Ghajini   26.72395
## 10  875785000 180000000  drama        Taare Zameen Par   20.55299
```

1.2) Make a boxplot of `pctReturn` by `Genre`. Superimpose a red dot at the mean percent return for each Genre; also print this mean rounded to the nearest percent. In addition, add a horizontal dotted blue line at 100 (this represents the place where revenue = budget). Discuss what you observe.

```r
boxplot(pctReturn ~ Genre,
        data=movie,
        main="Percent Return of Movies by Genre",
        xlab="Genre",
        ylab="Percent Return")
means <- aggregate(pctReturn ~ Genre, data = movie, FUN = mean)[['pctReturn']]
paste(round(means), '%', sep="")
```

```
## [1] "692%"  "1983%" "3601%" "2732%" "54%"
```

```r
points(means, col = "red", pch = 19, cex = 1.2)
text(x = c(1:5), y = means + 9000, labels = paste(round(means), '%', sep=''))
abline(h=100, col='blue')
```



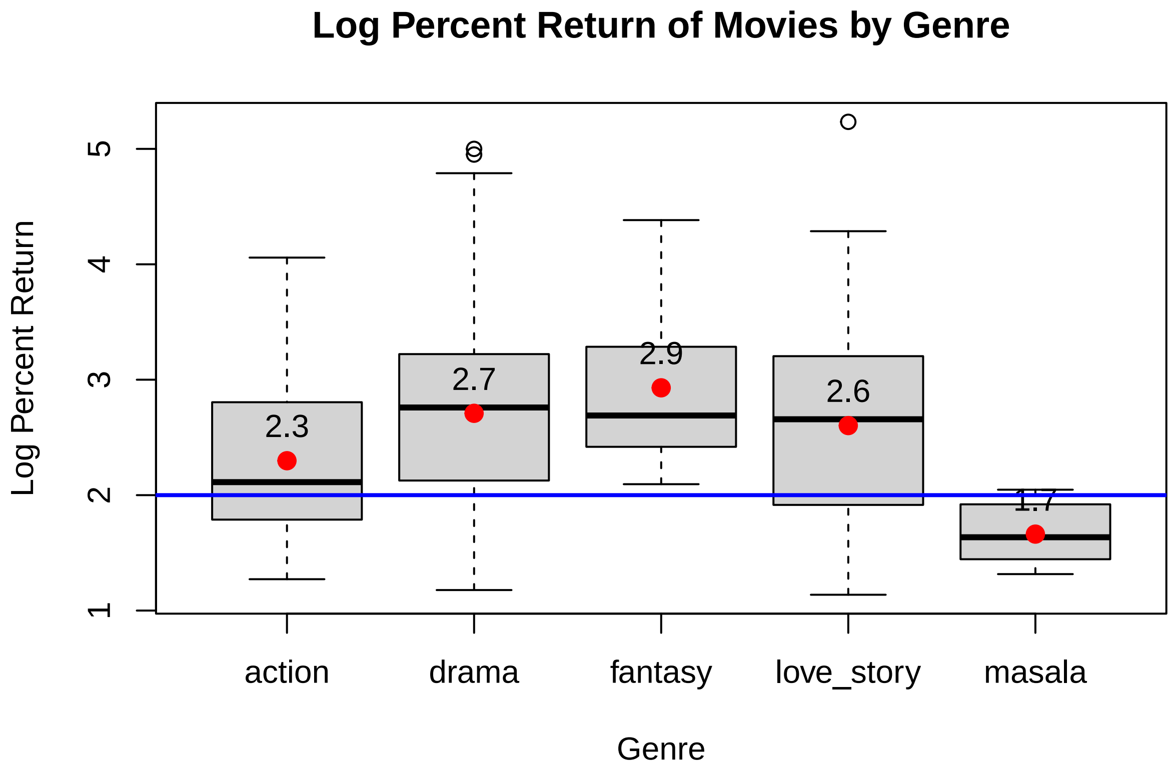
The boxplot is very difficult to analyze because there are some outlier values that cause the boxplot to shrink to the bottom.

1.3) Make a new variable on ‘movie’ called `log10pct` that is log base 10 of percent return (the function you want is `log10()`). Repeat part 1.2, but modify the horizontal line so that it is at the correct value (when rounding mean log10(pct return), round to 1 decimal place). Discuss what you observe.

```
movie["log10pct"] = log10(movie["pctReturn"])
boxplot(log10pct ~ Genre,
        data=movie,
        main="Log Percent Return of Movies by Genre",
        xlab="Genre",
        ylab="Log Percent Return")
means <- aggregate(log10pct ~ Genre, data = movie, FUN = mean)[['log10pct']]
round(means, digits=1)
```

```
## [1] 2.3 2.7 2.9 2.6 1.7
```

```
points(means, col = "red", pch = 19, cex = 1.2)
text(x = c(1:5), y = means + 0.3, labels = round(means, digits=1))
abline(h=2, col='blue', lwd=2)
```



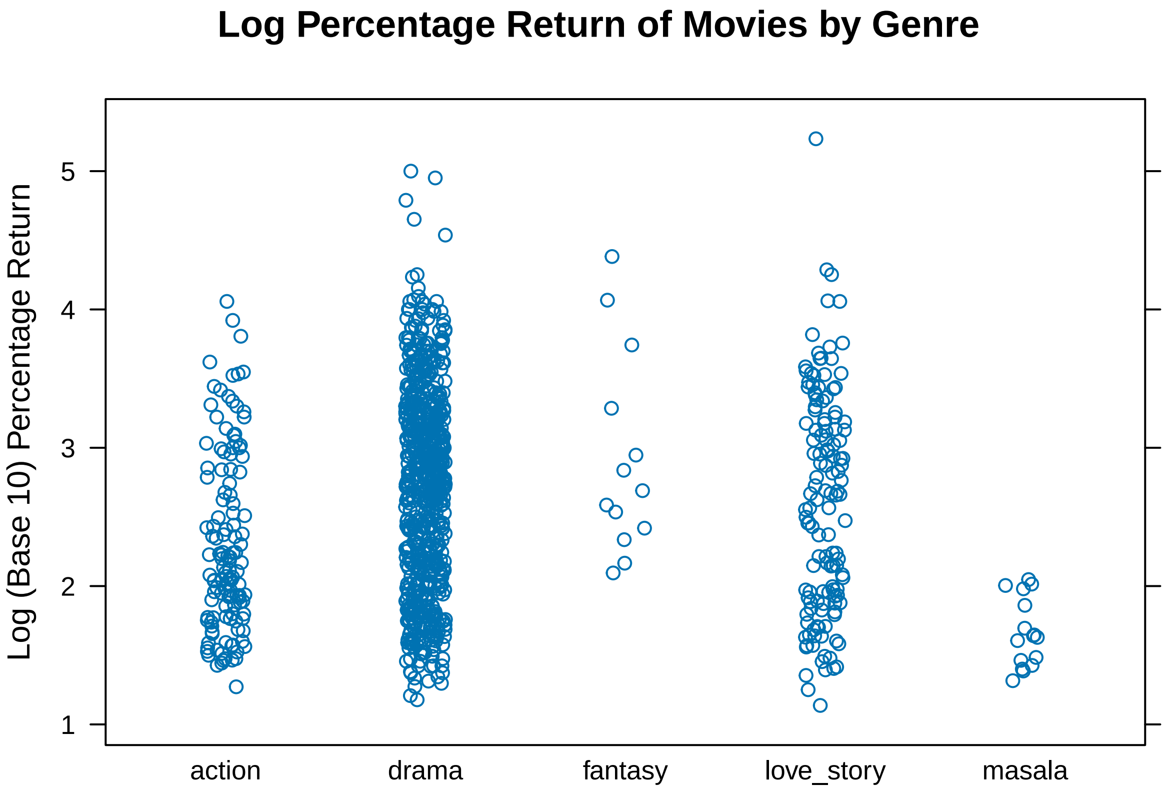
The boxplot looks much better now; we can actually see the means and the IQRs of the log percentage returns. It seems that for every genre besides masala, the movies on average have >100% return.

1.4) Make a `stripplot` of log10pct by Genre. Which plot do you find more instructive - the boxplot or the strip plot (no right answer here, but give a one sentence reason for your opinion)?

```r
library(lattice)
```

```
## Warning: package 'lattice' was built under R version 4.4.3
```

```r
stripplot(movie[["log10pct"]] ~ movie[["Genre"]],
          jitter = .1,
          main = "Log Percentage Return of Movies by Genre",
          Xlab="Genre",
          ylab="Log (Base 10) Percentage Return")
```



I find the boxplot more instructive. The strip plot really highlights the number of movies per genre, which isn't really something that seems too relevant here. Also, it's a little difficult to see what the distribution of the points would be, whereas in the boxplot I can see exactly where the median is and how the 1st and 3rd quartiles compare to it.

1.5) Just using the visual information provided by the boxplot of log10pct, do you think that the distribution of observations inside each Genre is approximately normal? Do you think the standard deviation is approximately the same in each Genre? (Remember that ANOVA assumes we have a normal distribution inside each Genre and that the standard deviation is the same across groups.)

I would say that the distribution of observations inside each Genre is, for the most part, approximately normal. Maybe fantasy and action aren't, but looking at the strip plot as well, they only look slightly off from normally distributed. The standard deviation is definitely not the same across the groups; this is clearly evident in the boxplot with the IQR being much greater for drama than for masala.

1.6) Calculate the sample standard deviation of log10pct for each Genre Calculate the ratio of largest to smallest sample standard deviation. Is it reasonable to assume that the variances are the same across Genres?

```
(sds <- tapply(movie$log10pct, movie$Genre, sd))
```

```
##     action      drama    fantasy love_story     masala
## 0.6441561  0.7210816  0.7319048  0.7944657  0.2487290
```

```
round(max(sds)/min(sds), 1)
```

```
## [1] 3.2
```

It is not reasonable to assume that the variances are the same across Genres since the maximum ratio of sample standard deviations is 3.2, which is greater than 2.

1.7) Use the `aov()` function to compare mean log10pct between Genres. Save your results to an object called `aov1`. Get summary information for `aov1`. Is the mean log10pct return statistically significantly different between Genres? In addition, confirm that the degrees of freedom reported by the test are what you expect (and write a sentence about this).

```
aov1 <- aov(movie$log10pct ~ movie$Genre)
summary(aov1)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## movie$Genre   4   34.2   8.539    16.6 3.79e-13 ***
## Residuals   923  474.7   0.514
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The mean log10pct return is statistically significantly different between Genres with a p value of 3 x 10^-13. The degrees of freedom are what I would expect: between groups is 4, which is the number of groups minus one, and within groups is 923, which is the number of observations (928) minus the number of groups.

1.8) Fit log10pct return based on Genre as a regression model WITHOUT an intercept. Save the results to an object called `mod1`. Following the example code provided in class, calculated confidence intervals for the mean percent return for each Genre and report results to two decimal places. Then make a plot of the resulting confidence intervals using the `plotCI()` function in the `plotrix` library.

You'll notice that some intervals overlap while others do not. What should you conclude from this?

```
mod1 <- lm(movie$log10pct ~ movie$Genre -1)
CIs <- confint(mod1)
round(CIs, 2)
```

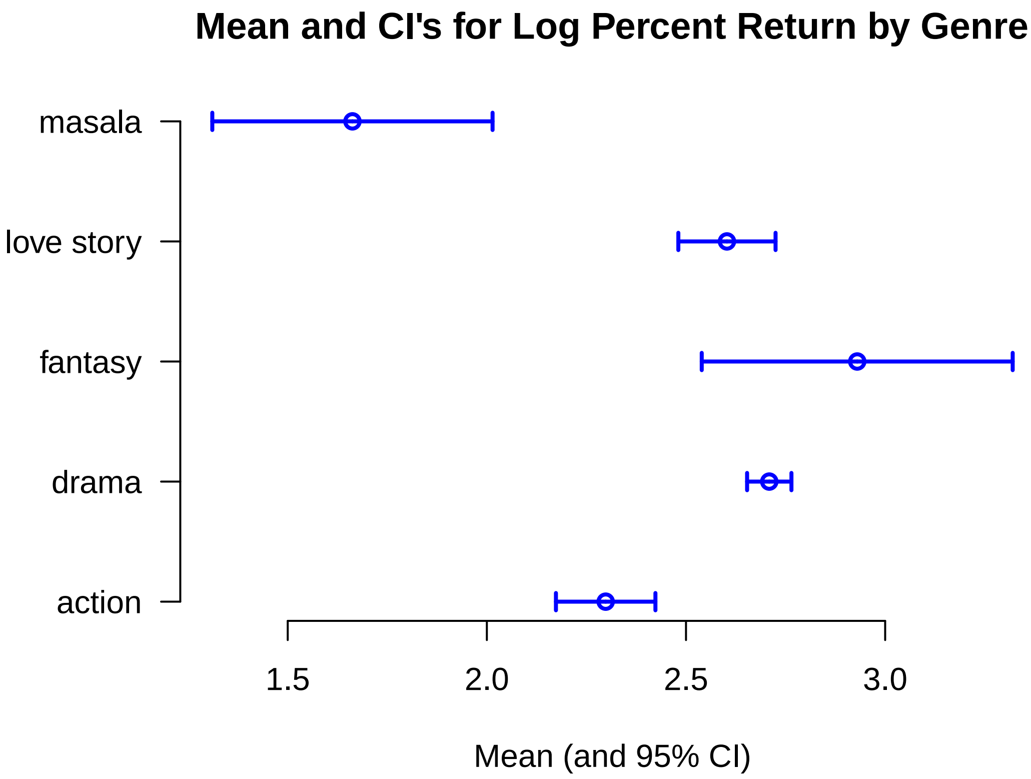
```
##                         2.5 % 97.5 %
## movie$Genreaction        2.17   2.42
## movie$Genredrama         2.65   2.76
## movie$Genrefantasy       2.54   3.32
## movie$Genrelove_story    2.48   2.72
## movie$Genremasala        1.31   2.01
```

```
library(plotrix)

coefs <- coef(mod1)
coefs
```

```
##     movie$Genreaction      movie$Genredrama    movie$Genrefantasy
##              2.298334              2.709028              2.929856
## movie$Genrelove_story     movie$Genremasala
##              2.602779              1.662520
```

```
#Make x margin bigger for names
par(mar=c(5,8,4,2))

#Make plot - err = "x" makes horizontal
```

```r
plotCI(coefs, 1:(length(coefs)),
       ui = CIs[,2],
       li = CIs[,1],
       axes = FALSE,
       err = "x",
       ylab = "",
       xlab = "Mean (and 95% CI)",
       main = "Mean and CI's for Log Percent Return by Genre",
       lwd = 2, col = "blue")

axis(side = 1)

genres = c("action", "drama", "fantasy", "love story", "masala")

#Put emotion labels on
axis(side = 2, at = 1:(length(coefs)), label=genres, las=2)
```



For the genres whose confidence intervals overlap, we can say that there is insufficient evidence to conclude that the mean log percent returns are different between those two genres. For the genres whose confidence intervals do not overlap, however, we can conclude that there is statistically significant evidence that suggests that the mean log percent returns are different between those two genres.

1.9) Use the `pairwise.t.test()` function to calculate Holm's correction for comparing pairs of means. Which pairs of Genres have statistically significantly different percent returns (use alpha = 0.05)?

```r
pairwise.t.test(movie$log10pct, movie$Genre)
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  movie$log10pct and movie$Genre
##
##            action  drama   fantasy love_story
## drama      5.3e-08 -       -       -
## fantasy    0.0103  0.3506  -       -
## love_story 0.0039  0.3506  0.3506  -
## masala     0.0043  1.0e-07 1.8e-05 6.9e-06
##
## P value adjustment method: holm
```

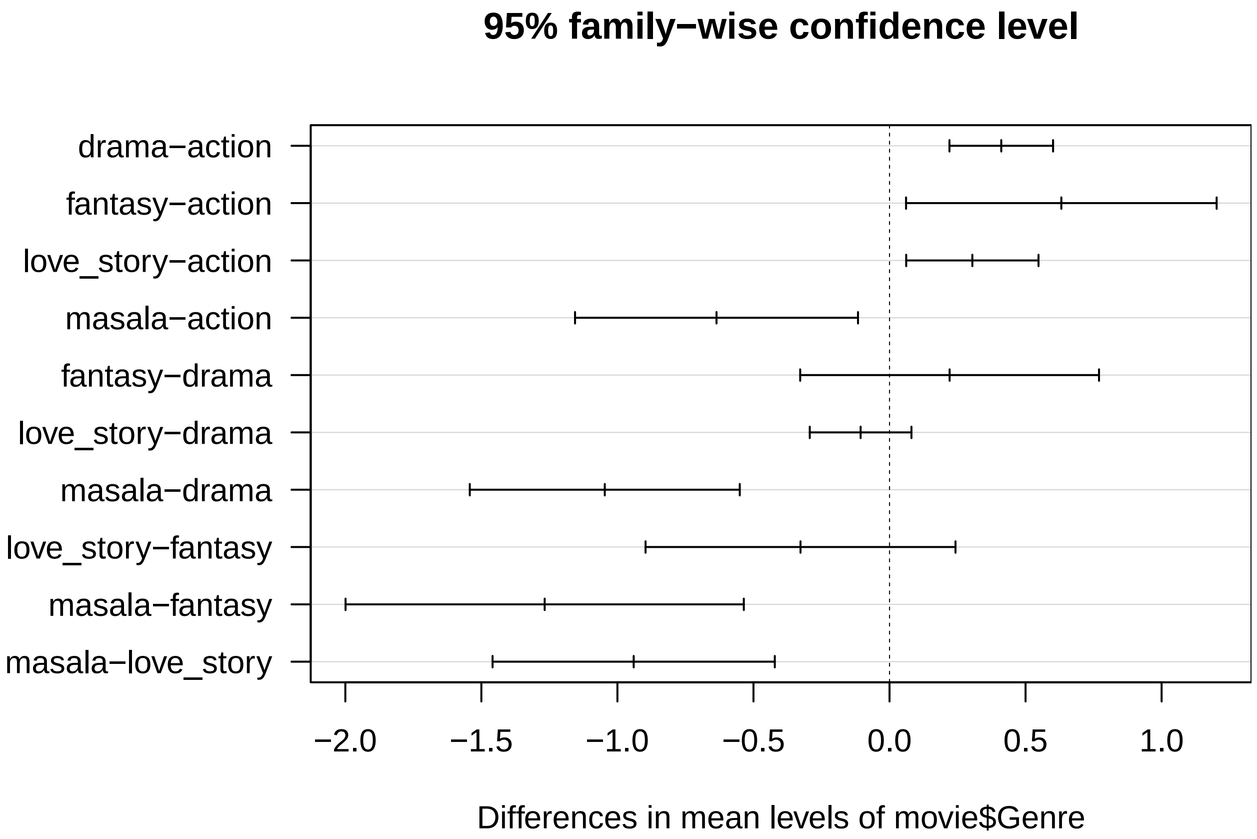
Using a significance level of 0.05, we conclude that the pairs of Genres with statistically significantly different percent returns are drama and action, fantasy and action, love story and action, masala and action, masala and drama, and masala and love story.

1.10) Calculate Tukey simultaneous 95% confidence intervals for differences in mean percent return using the `TukeyHSD()` function. Plot the resulting confidence intervals. Do you reach the same conclusions as when using the Holm correction?

```r
TukeyHSD(aov1)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = movie$log10pct ~ movie$Genre)
##
## $`movie$Genre`
##                          diff         lwr         upr     p adj
## drama-action        0.4106948  0.22026598  0.60112365 0.0000001
## fantasy-action      0.6315223  0.06075232  1.20229222 0.0215328
## love_story-action   0.3044454  0.06126437  0.54762634 0.0058455
## masala-action      -0.6358136 -1.15578179 -0.11584539 0.0076943
## fantasy-drama       0.2208275 -0.32829892  0.76995384 0.8070266
## love_story-drama   -0.1062495 -0.29306070  0.08056179 0.5271769
## masala-drama       -1.0465084 -1.54262167 -0.55039512 0.0000001
## love_story-fantasy -0.3270769 -0.89665012  0.24249628 0.5174215
## masala-fantasy     -1.2673359 -1.99921260 -0.53545911 0.0000252
## masala-love_story  -0.9402589 -1.45891319 -0.42160469 0.0000085
```

```r
par(mar=c(4, 8, 4, 0))
plot(TukeyHSD(aov1), las=1)
```



Yes, the two tests give the same conclusions about the pairwise differences in means.

1.11) Create residual plots appropriate for `aov1`. Does it appear that the residuals have an approximately normal distribution? Is this good or bad? Comment in a sentence or two about the plot of fits versus residuals.

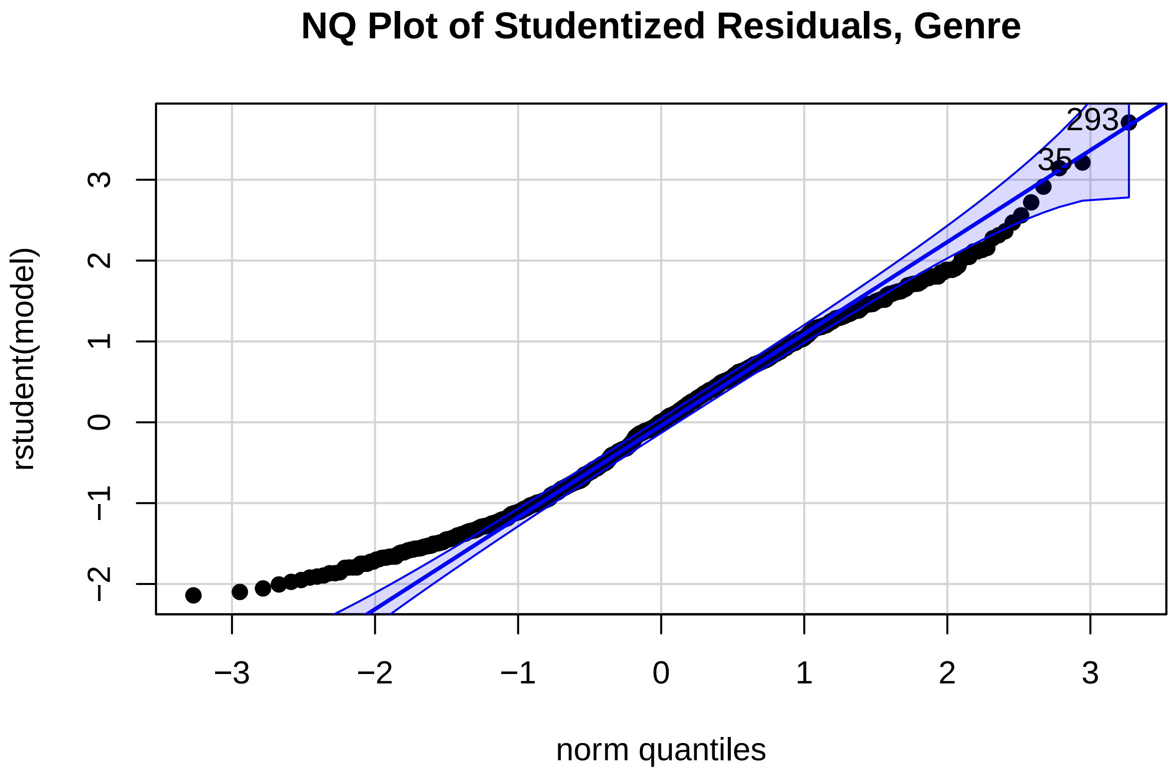
```r
library(leaps)
```

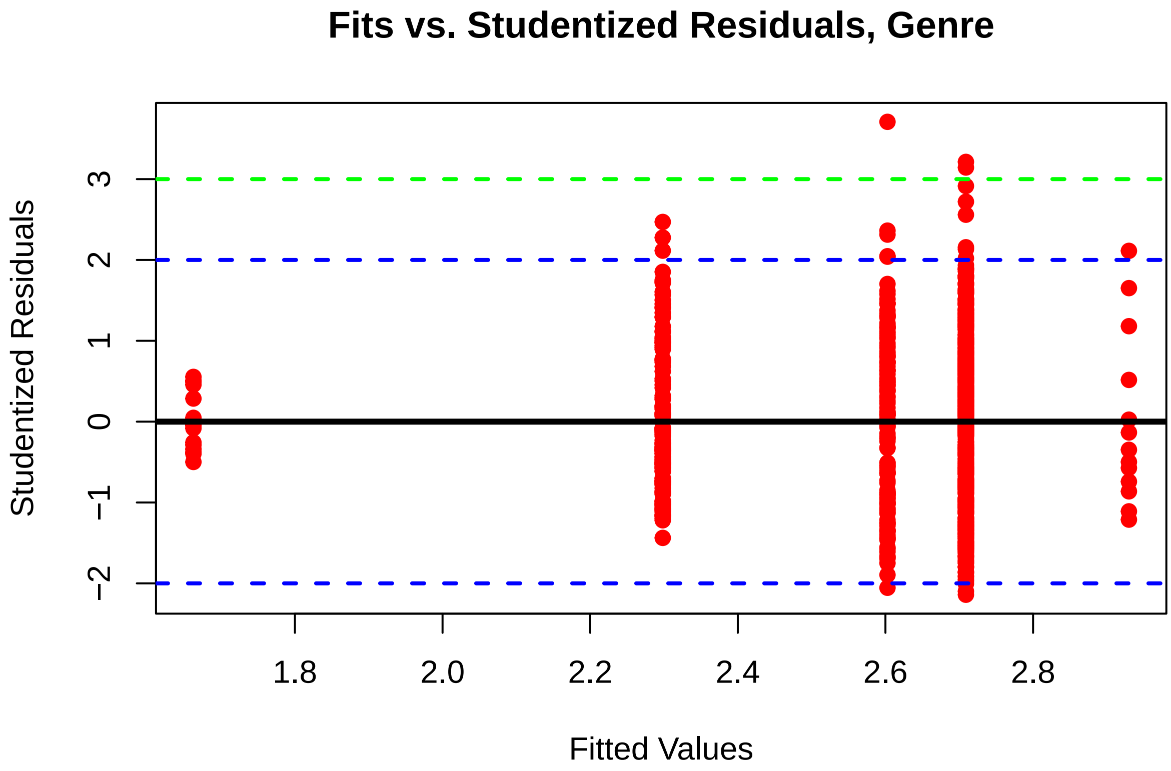
```
## Warning: package 'leaps' was built under R version 4.4.3
```

```r
library(car)
```

```
## Loading required package: carData
```

```r
#Handy residual plots
myResPlots <- function(model, label){

  #Normal quantile plot of studentized residuals
  qqPlot(rstudent(model), pch = 19, main = paste("NQ Plot of Studentized Residuals,", label))

  #plot of fitted vs. studentized residuals
  plot(rstudent(model) ~ model$fitted.values, pch = 19, col = 'red', xlab = "Fitted Values", ylab = "St
     main = paste("Fits vs. Studentized Residuals,", label))
  abline(h = 0, lwd = 3)
  abline(h = c(2,-2), lty = 2, lwd = 2, col="blue")
  abline(h = c(3,-3), lty = 2, lwd = 2, col="green")

}

myResPlots(mod1, label = "Genre")
```





The residuals do not appear to have an approximately normal distribution, which is not good since normally distributed residuals is an assumption we need to make for a proper linear regression. The plot of fits vs residuals appears to have some amount of heteroskedasticity, with the higher fitted values having higher residuals, although this is entirely due to the genre with the low residuals on the left. If we were to remove that genre, then the fits vs studentized residuals would appear to have no signs of heteroskedasticity.

1.12) Just to see what would have happened if we didn't use a log10 transformation, repeat parts 1.7 through 1.11 on the original percent return data (i.e. not log10). Just use one block of code (literally, just copy your code, replace `log10pct` with `pctReturn`).

Write about three sentences commenting on the fit of this model vs. the fit of the previous model. Specifically, comment on whether any pairs of groups changed significance, the overall model significance, and the residual plots.

```
#1.7
aov1 <- aov(movie$pctReturn ~ movie$Genre)
summary(aov1)
```

```
##               Df    Sum Sq  Mean Sq F value Pr(>F)
## movie$Genre    4 3.737e+08 93431520   1.487  0.204
## Residuals    923 5.801e+10 62847159
```

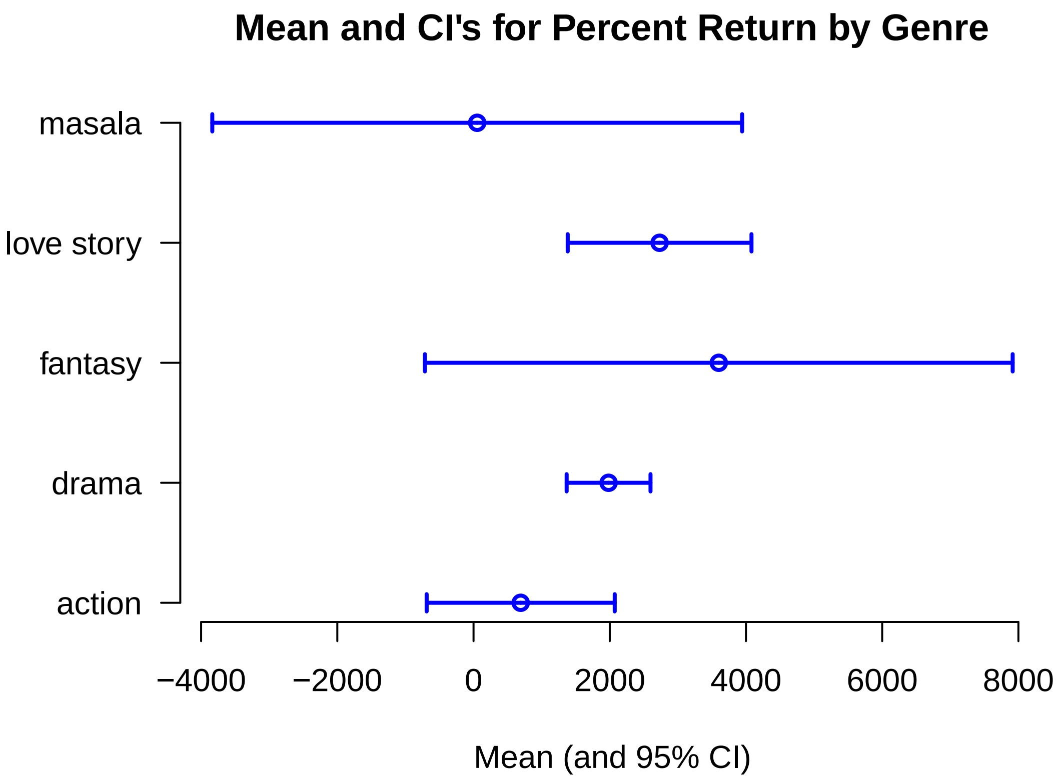
```
#1.8
mod1 <- lm(movie$pctReturn ~ movie$Genre -1)
CIs <- confint(mod1)
round(CIs, 2)
```

```
##                          2.5 %  97.5 %
## movie$Genreaction      -688.17 2072.97
## movie$Genredrama       1367.27 2598.22
## movie$Genrefantasy     -714.30 7915.87
## movie$Genrelove_story  1382.71 4080.85
## movie$Genremasala     -3835.75 3943.37
```

```
library(plotrix)

coefs <- coef(mod1)
coefs
```

```
##     movie$Genreaction      movie$Genredrama    movie$Genrefantasy
##             692.40237            1982.74102            3600.78460
## movie$Genrelove_story     movie$Genremasala
##            2731.77842              53.80833
```

```
#Make x margin bigger for names
par(mar=c(5,8,4,2))

#Make plot - err = "x" makes horizontal
plotCI(coefs, 1:(length(coefs)),
       ui = CIs[,2],
       li = CIs[,1],
       axes = FALSE,
       err = "x",
       ylab = "",
       xlab = "Mean (and 95% CI)",
       main = "Mean and CI's for Percent Return by Genre",
       lwd = 2, col = "blue")

axis(side = 1)

genres = c("action", "drama", "fantasy", "love story", "masala")

#Put emotion labels on
axis(side = 2, at = 1:(length(coefs)), label=genres, las=2)
```



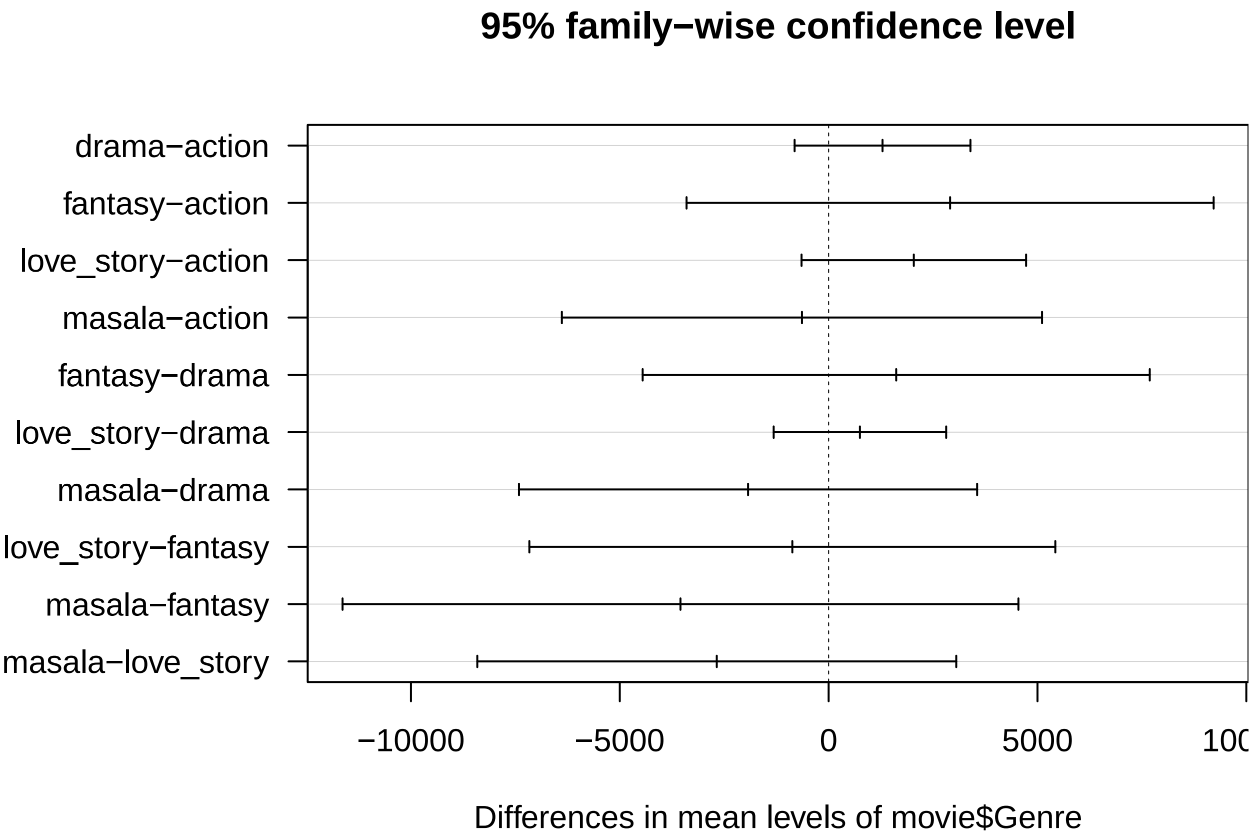
```
#1.9
pairwise.t.test(movie$pctReturn, movie$Genre)
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  movie$pctReturn and movie$Genre
##
##            action drama fantasy love_story
## drama      0.85   -     -       -
## fantasy    1.00   1.00  -       -
## love_story 0.38   1.00  1.00    -
## masala     1.00   1.00  1.00    1.00
##
## P value adjustment method: holm
```

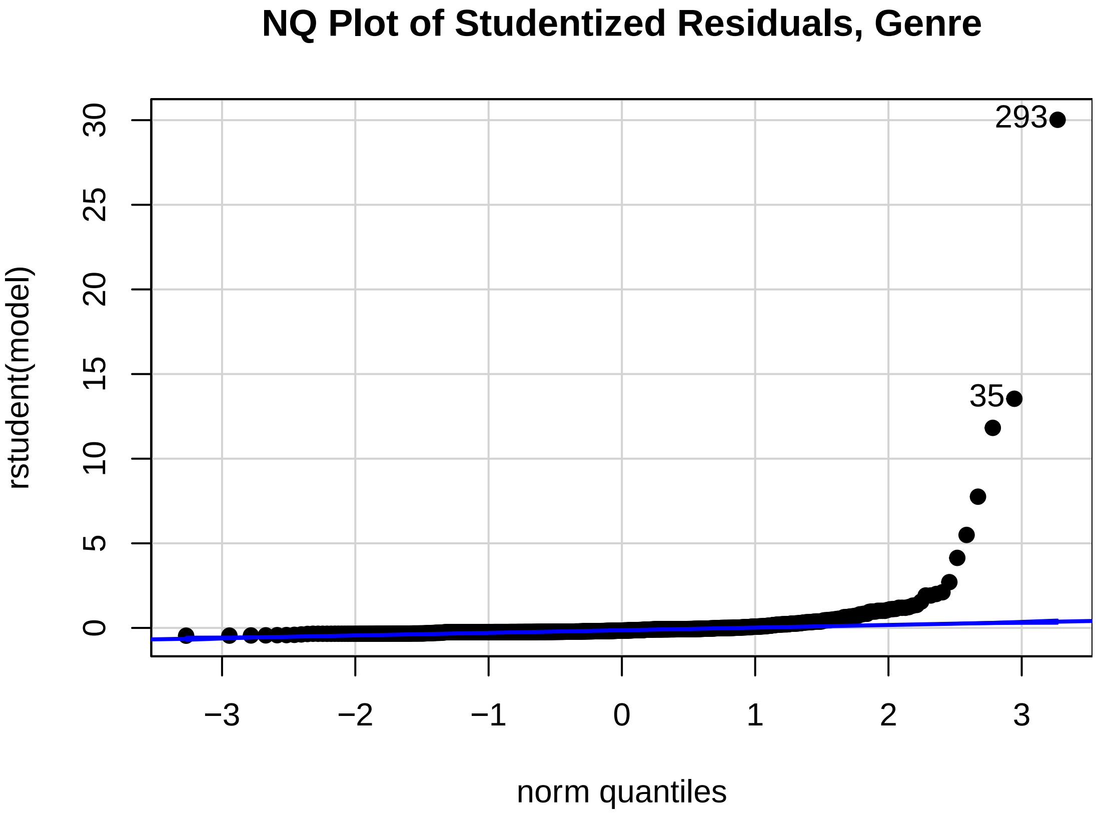
```
#1.10
TukeyHSD(aov1)
```

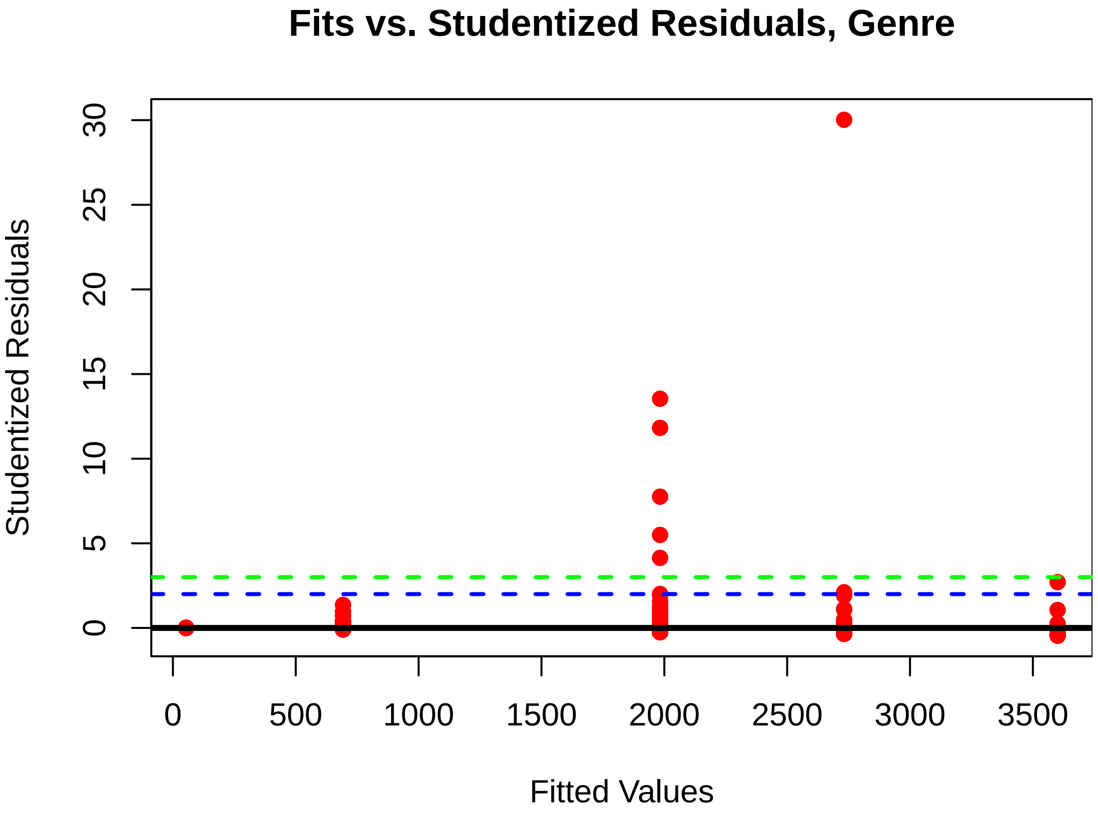
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = movie$pctReturn ~ movie$Genre)
##
## $‘movie$Genre‘
##                          diff         lwr      upr     p adj
## drama-action         1290.3387   -814.7611 3395.438 0.4497932
## fantasy-action       2908.3822  -3401.2066 9217.971 0.7159319
## love_story-action    2039.3761   -648.8734 4727.626 0.2324998
## masala-action        -638.5940  -6386.5939 5109.406 0.9981510
## fantasy-drama        1618.0436  -4452.2860 7688.373 0.9498793
## love_story-drama      749.0374  -1316.0716 2814.146 0.8592871
## masala-drama        -1928.9327  -7413.2277 3555.362 0.8723654
## love_story-fantasy   -869.0062  -7165.3656 5427.353 0.9956975
## masala-fantasy      -3546.9763 -11637.5238 4543.571 0.7524282
## masala-love_story   -2677.9701  -8411.4448 3055.505 0.7057240
```

```
par(mar=c(4, 8, 4, 0))
plot(TukeyHSD(aov1), las=1)
```



```
#1.11
myResPlots(mod1, label = "Genre")
```


NQ Plot of Studentized Residuals, Genre
rstudent(model)
0
5
10
15
20
25
30
−3
−2
−1
0
1
2
3
norm quantiles
293
35



When using percent return instead of log return, the tests conclude that there is insufficient evidence for any of the pairwise comparisons that their means differ. This is true for also the overall model; the ANOVA gave a p value of 0.2. By looking at the normal quantile plot we can very easily see than the residuals are not normally distributed, and the fits vs studentized residuals are definitely not uniformly distributed.

2) Non-parametric tests. *(25 pts - 5pts each)*

2.1) Perform a Bartlett test to see whether variances of percent return are the same across Genres. Then, repeat for `log10pct`. Write one sentence about what you conclude.

```
bartlett.test(movie$pctReturn, movie$Genre)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  movie$pctReturn and movie$Genre
## Bartlett's K-squared = 649.27, df = 4, p-value < 2.2e-16
```

```
bartlett.test(movie$log10pct, movie$Genre)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  movie$log10pct and movie$Genre
## Bartlett's K-squared = 23.962, df = 4, p-value = 8.129e-05
```

Because both p values are below a standard significance level of 0.05, we can conclude that the variances of percent return are different across genres (though in actuality because Bartlett assumes normality we cannot use the Bartlett test here for pctReturn).

2.2) Perform a Levene test to see whether variances of percent return are the same across Genres. Then, repeat for `transpct`. Write one sentence about what you observe (be sure to review the characteristics of Levene’s test).

```
leveneTest(movie$pctReturn, movie$Genre)
```

```
## Warning in leveneTest.default(movie$pctReturn, movie$Genre): movie$Genre
## coerced to factor.
```

```
## Levene’s Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   4  1.3838 0.2376
##       923
```

```
leveneTest(movie$log10pct, movie$Genre)
```

```
## Warning in leveneTest.default(movie$log10pct, movie$Genre): movie$Genre coerced
## to factor.
```

```
## Levene’s Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   4  6.4329 4.167e-05 ***
##       923
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Only the p value for log10pct is below a significance level of 0.05, thus for only log10pct do we conclude that the variance differs across genres.

2.3) Perform Welch’s ANOVA on the original data which you will recall assumes unequal variances but which assumes normal distributions in each group (i.e. compare `pctReturn` by `Genre`). Compare results to regular one-way ANOVA. Is Welch’s ANOVA a good choice here?

```
oneway.test(pctReturn ~ Genre, data = movie)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  pctReturn and Genre
## F = 20.711, num df = 4.000, denom df = 79.718, p-value = 9.859e-12
```

```
summary.aov(aov(pctReturn ~ Genre, data = movie))
```

```
##              Df    Sum Sq  Mean Sq F value Pr(>F)
## Genre         4 3.737e+08 93431520   1.487  0.204
## Residuals   923 5.801e+10 62847159
```

Welch's ANOVA gives a p value under 0.05 whereas one-way ANOVA does not. Welch's ANOVA is not a good choice here, at least for percent returns, since the percent returns are not normally distributed for each group.

2.4) Perform a non-parametric Kruskal Wallis test which recall makes NO assumptions about equal variances or about normality. Compare results to regular one-way ANOVA. Is Kruskal Wallis a good choice here?

```
kruskal.test(pctReturn ~ Genre, data = movie)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  pctReturn by Genre
## Kruskal-Wallis chi-squared = 63.473, df = 4, p-value = 5.397e-13
```

Kruskal Wallis gives a p value under 0.05 whereas one-way ANOVA does not. Kruskal Wallis is a good choice here since the percent returns are not normally distributed for each group and the variance is different across groups.

2.5) Repeat 2.3) and 2.4) on the log10 scale. Discuss your results and compare results to regular ANOVA.

```
oneway.test(log10pct ~ Genre, data = movie)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  log10pct and Genre
## F = 61.337, num df = 4.000, denom df = 55.497, p-value < 2.2e-16
```

```
summary.aov(aov(log10pct ~ Genre, data = movie))
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Genre         4   34.2   8.539    16.6 3.79e-13 ***
## Residuals   923  474.7   0.514
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
kruskal.test(log10pct ~ Genre, data = movie)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  log10pct by Genre
## Kruskal-Wallis chi-squared = 63.473, df = 4, p-value = 5.397e-13
```

All three tests give p values less than 0.05, thus we conclude that the means are difference across genres. ANOVA and Welch's ANOVA are much better here than for just percent returns because the log percent returns are much closer to normally distributed. Welch's ANOVA is better than regular one-way ANOVA here, however, since we found earlier that the variance is still not equal among the genres.

THE END