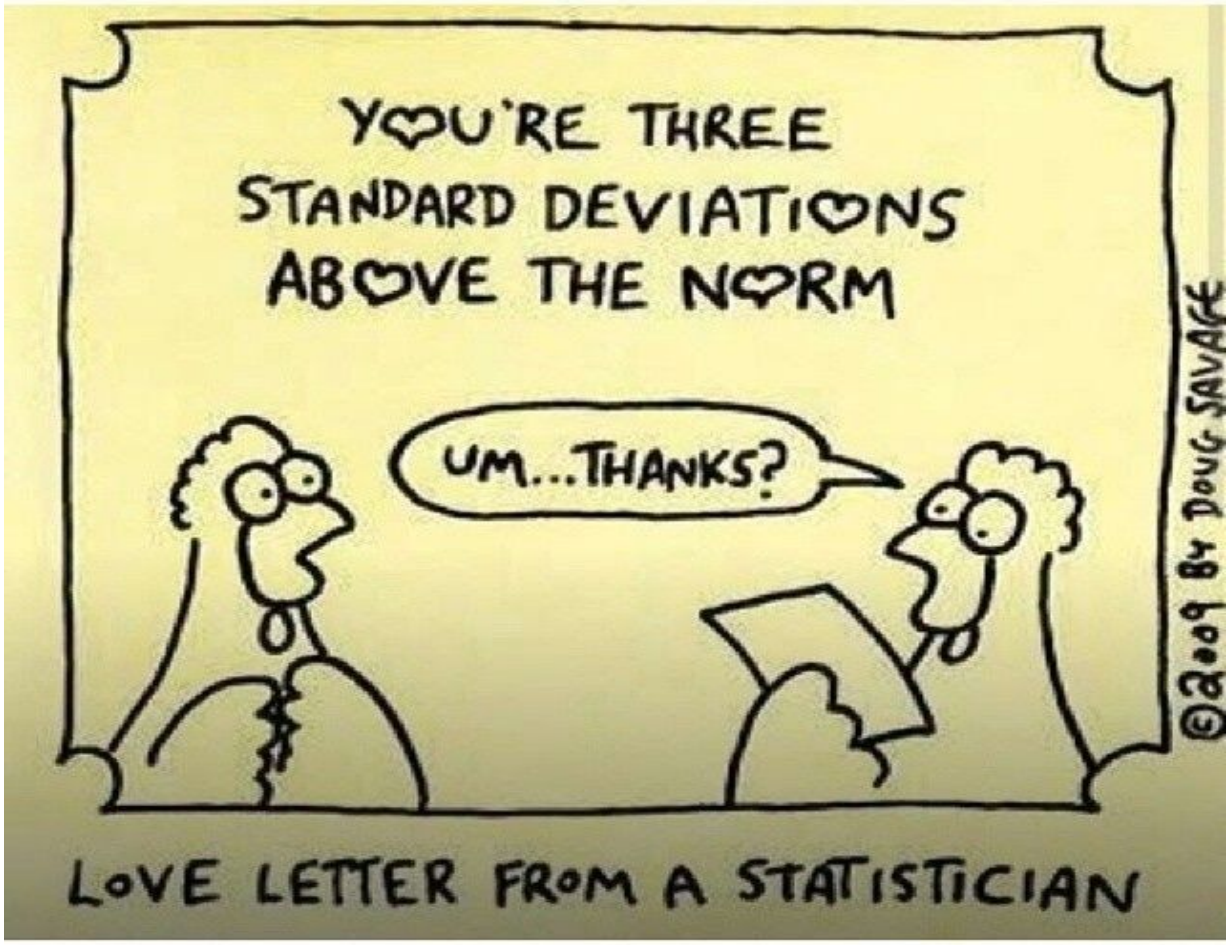
STA 104 Final Project



Zenghui Yu 914405982

Yutian Yang 915327218

1. Abstract

This project will discuss the mean, medians for different variables and their associated 95% confidence interval and compare variables across genres by using **both parametric and non-parametric analysis methods**.

Our project contains three main parts:

A **methods/results** section which Present different methods analysis implemented during the computation and the corresponding results in details.

A **Discussion** which elaborates each interpretation of the following result for selecting t-test and ANOVA F-test as parametric method; Binomial test, Wilcoxon Rank SUm test, Kruskal Wallis test as non-parametric method.

An **Appendix** that contains all computer code used and evidence that we actually used the computer code.

2. Introduction

The project is based on the dataset *hollywoodmovie.xls* which is about information for 136 movies released from Hollywood in 2011 with 136 observations. The dataset contains 14 variables-Movie,LeadStudio,RottenTomatoes,story, genre, TheaterOpen week, BOAverageOpenWeek, DomesticGross, ForeignGross, WorldGross, Budget, Profitability, Openingweek.

For our project, we will focus more on the variables “Budget”, “DomesticGross”, and “TheatersOpenWeek” for different genres because those three are the main influence.

3. Methods & Results

We first eliminated all the NAs in Budget, DomesticGross, OpeningWeekend for all Genres, and calculate confidence interval of means between each two of the categories. We then split Comedy and Drama from the

dataset and created a new separate data frame. After that, we conducted both anova test and non-parametric test to analyze differences between both genres' Budget, DomesticGross, and OpeningWeekend data. We continued to compare Budget, DomesticGross, and OpeningWeekened between horror/thriller, animation/fantasy, and action/adventure using both anova analysis and non-parametric test.

After that, we focused on action movies and calculated both pearson and spearsman rank correlation to check if action movie's DomesticGross and OpeningWeekend are correlated. Finally, we compared action movie's DomesticGross and ForeignGross to check if it's earning higher inside or outside of the United States.

4. Discussion

We used which function to find out all possible index that contain NAs within Budget, DomesticGross, OpeningWeekend across all genres. After eliminating NAs, we then have a new data frame that contains all valid data.

We first found out the means of Budget = 54.07308, DomesticGross = 63.26781, OpeningWeekend = 20.65766 across all genres. We also conducted 95% confidence intervals for the means of the three variables. We have (45.65901, 62.48715) for 95% CI of Budget mean, (51.43363, 75.10200) for 95% CI of DomesticGross mean, and (16.36125, 24.95407) for 95% CI of OpeningWeekend mean. We then calculated median for these three variables and get results of Budget median= 37.0000, DomesticGross median = 37.5370, OpeningWeekend median = 13.2385 across all genres. After that we conducted 95% binomial confidence intervals for the medians of the three variables. We have (27, 50) for 95% CI of Budget mean, (24.800, 66.626) for 95% CI of DomesticGross mean, and (9.352, 19.501) for 95% CI of OpeningWeekend mean. Comparing both types of confidence intervals, mean CIs are narrower comparing to median CIs. However, according to the confidence interval, means and medians are further apart suggesting data is not normal or skewed.

We then specifically compared Comedy genre and Drama genre using both two sample t-test and wilcoxon rank sum test after sorting both genres' Budget, DomesticGross, and OpeningWeekend data. For comparison of Budget of these two genres, the two sample test suggests that the average comedy budget is greater than the average drama budget since the p-value = 0.009037 is less than alpha = 0.05 and we have a 95 percent confidence interval of difference in average Budget equals to (3.806198 25.111751). For non-parametric test, it suggests the same result. The p-value is 0.02591 and we have confidence interval of difference in median budget equals to (0.9999482, 24.9999753) which suggests the median comedy budget is greater than the median drama budget. Similarly, we got the results of the two sample test and nonparametric test of DomesticGross and OpeningWeekend for comedy and drama. The two sample test for DomesticGross suggests that there is no significant difference between the average comedy domestic gross and the average drama domestic gross since the p-value = 0.07623 is greater than alpha = 0.05 and we have a 95 percent confidence interval of difference in average DomesticGross equals to (-2.787612, 53.519582). For non-parametric test, it suggests that the median comedy domestic gross is greater than the median drama domestic gross since the p-value is 0.02296 and we have confidence interval of difference in median budget equals to (1.996, 38.896). The two sample test for OpeningWeekend suggests that the average comedy opening weekend is greater than the average drama opening weekend since the p-value = 0.01041 is less than alpha = 0.05 and we have a 95 percent confidence interval of difference in average Budget equals to (2.500665, 17.577822). For non-parametric test, it suggests the same result. Since the p-value is 0.00519 and we have confidence interval of difference in median budget equals to (2.354 13.223), it suggests the median comedy opening weekend is greater than the median drama opening weekend. Since the sample is small and potentially violated the normality assumption (according to previous conclusion), we decided that using non-parametric test is more suitable in this situation.

We then compared Budget, DomesticGross, and OpeningWeekend of movies that are similar (horror/thriller, animation/fantasy, action/adventure). We ran both anova test and permutation F test to analysis the dataset. For horror vs. thriller, both tests suggest that there is no significant difference between average(median) horror and thriller budget,domestic gross, and opening weekend since both test's p-value is greater than 0.05. For animation vs. fantasy, again, both tests suggest that there is no significant difference between average(median) animation and fantasy budget,domestic gross, and opening weekend since both test's p-value is greater than 0.05. However, since there is no adventure movies after we eliminate all the NAs, we cannot compare these factors if each factor contains only one level. In general, we decided that non-parametric test is more accurate in the case since the sample size is relatively small, and some of the hypothesis test violated the assumption that the expected value of all residuals should equals to 0.

We then focused on action movies and trying to analysis that if it's DomesticGross and OpeningWeekend are correlated. We calculated both Pearson and Spearman Rank correlation and got results of 0.9389937 and 0.6751947. Both correlation suggests that these two factors are positively correlated. We then conducted a regression analysis along with a permutation F-test. Both tests gave back small p-value (less than 0.05), which indicates that there is a significant difference between action movies average (median) domestic gross and opening weekend. However, the Pearson correlation evaluates the linear relationship between two continuous variables, which assumes normality, and the Spearman correlation coefficient is based on the ranked values for each variable rather than the raw data. Since the linear relationship requires normality, in this case, the action movie dataset is relatively small (30 obs) and it is not normal based on the plot of the anova test (Residual vs Fitted plot and normal QQ plot), so it is more suitable to use Spearman Rank correlation and non-parametric test.

Finally, we were still focusing action movies but trying to analyze its domestic gross and foreign gross. The anova analysis suggests that there is a significant difference between the average domestic gross of action movies and average foreign gross of action movies. However, the permutation F-test suggests that there is no significant difference between the median domestic gross and median foreign gross.

5. Appendix

```
library(readxl)
hollywoodmovies <- read_excel("C:/Users/charl/Downloads/hollywoodmovies.xls")


hollywoodmovies = hollywoodmovies[-which(is.na(hollywoodmovies$Budget)),]
hollywoodmovies = hollywoodmovies[-which(is.na(hollywoodmovies$DomesticGross)),]
hollywoodmovies = hollywoodmovies[-which(is.na(hollywoodmovies$OpeningWeekend)),]

Budget_mean = mean(hollywoodmovies$Budget)
DomestricGross_mean = mean(hollywoodmovies$DomesticGross)
Openingweekend_mean = mean(hollywoodmovies$OpeningWeekend)
c(Budget_mean, DomestricGross_mean, Openingweekend_mean)

Budget_sd = sd(hollywoodmovies$Budget)
DomestricGross_sd = sd(hollywoodmovies$DomesticGross)
Openingweekend_sd = sd(hollywoodmovies$OpeningWeekend)
n = length(hollywoodmovies$Budget)

Budget_error = qnorm(0.975)*Budget_sd/sqrt(n)
DomestricGross_error = qnorm(0.975)*DomestricGross_sd/sqrt(n)
Openingweekend_error = qnorm(0.975)*Openingweekend_sd/sqrt(n)

Budget_mean_CI = c((Budget_mean-Budget_error),(Budget_mean+Budget_error))
Budget_mean_CI

DomestricGross_mean_CI = c((DomestricGross_mean-DomestricGross_error),(DomestricGross_mean+Domestr
icGross_error))
DomestricGross_mean_CI

Openingweekend_mean_CI = c((Openingweekend_mean-Openingweekend_error),(Openingweekend_mean+Opening
weekend_error))
Openingweekend_mean_CI

Budget_median = median(hollywoodmovies$Budget)
DomestricGross_median = median(hollywoodmovies$DomesticGross)
Openingweekend_median = median(hollywoodmovies$OpeningWeekend)
c(Budget_median, DomestricGross_median, Openingweekend_median)

a = -1.96*sqrt(n)+0.5*(n)
b = 1.96*sqrt(n)+0.5*(n)+1

sorted_hollywoodmovies = hollywoodmovies[order(hollywoodmovies$Budget),]
Budget_median_CI = c(sorted_hollywoodmovies$Budget[round(a)],sorted_hollywoodmovies$Budget[round
(b)])
Budget_median_CI

sorted_hollywoodmovies = hollywoodmovies[order(hollywoodmovies$DomesticGross),]
DomesticGross_median_CI = c(sorted_hollywoodmovies$DomesticGross[round(a)],sorted_hollywoodmovies
$DomesticGross[round(b)])
DomesticGross_median_CI

sorted_hollywoodmovies = hollywoodmovies[order(hollywoodmovies$OpeningWeekend),]
Openingweekend_median_CI = c(sorted_hollywoodmovies$OpeningWeekend[round(a)],sorted_hollywoodmovie
```

```
s$OpeningWeekend[round(b)])
Openingweekend_median_CI

c(Budget_mean_CI, Budget_median_CI)
c(DomestricGross_mean_CI, DomesticGross_median_CI)
c(Openingweekend_mean_CI, Openingweekend_median_CI)
hollywoodmovies_comedy_idx = which(hollywoodmovies$Genre == "Comedy")
hollywoodmovies_comedy = hollywoodmovies[hollywoodmovies_comedy_idx,]
hollywoodmovies_drama_idx = which(hollywoodmovies$Genre == "Drama")
hollywoodmovies_drama = hollywoodmovies[hollywoodmovies_drama_idx,]



t.test(hollywoodmovies_comedy$Budget, hollywoodmovies_drama$Budget)
wilcox.test(hollywoodmovies_comedy$Budget, hollywoodmovies_drama$Budget,
            alternative="two.sided",
            correct=TRUE,
            conf.int=TRUE,
            conf.level=0.95)

t.test(hollywoodmovies_comedy$DomesticGross, hollywoodmovies_drama$DomesticGross)
wilcox.test(hollywoodmovies_comedy$DomesticGross, hollywoodmovies_drama$DomesticGross,
            alternative="two.sided",
            correct=TRUE,
            conf.int=TRUE,
            conf.level=0.95)

t.test(hollywoodmovies_comedy$OpeningWeekend, hollywoodmovies_drama$OpeningWeekend)
wilcox.test(hollywoodmovies_comedy$OpeningWeekend, hollywoodmovies_drama$OpeningWeekend,
            alternative="two.sided",
            correct=TRUE,
            conf.int=TRUE,
            conf.level=0.95)

#seperate horror and thriller from dataset
hollywoodmovies_horror_idx = which(hollywoodmovies$Genre == "Horror")
hollywoodmovies_horror = hollywoodmovies[hollywoodmovies_horror_idx,]
hollywoodmovies_thriller_idx = which(hollywoodmovies$Genre == "Thriller")
hollywoodmovies_thriller = hollywoodmovies[hollywoodmovies_thriller_idx,]
hollywoodmovies_comparison1 = rbind(hollywoodmovies_horror, hollywoodmovies_thriller)

# horror vs thriller budget
model1 = lm(Budget~Genre,  data = hollywoodmovies_comparison1)
anova.table1 = anova(model1)
plot(aov(model1))
anova.table1
summary(anova.table1)

kruskal.test(Genre~Budget, data = hollywoodmovies_comparison1)

outcome=c(hollywoodmovies_horror$Budget, hollywoodmovies_thriller$Budget)
treat=c(rep(1, length(hollywoodmovies_horror$Budget)), rep(2, length(hollywoodmovies_thriller$Budge
t)))
diffobs=mean(hollywoodmovies_horror$Budget)-mean(hollywoodmovies_thriller$Budget)
```

```
d=c()
p=c()

for(i in 1:5000){
  permut=sample(outcome)
  d[i]=mean(permut[treat==1])-mean(permut[treat==2])
  p[i]=(d[i]>=diffobs)+0
  }
pvalue=sum(p)/5000
pvalue

# horror vs thriller DomestricGross
model2 = lm(DomesticGross~Genre, data = hollywoodmovies_comparison1)
anova.table2 = anova(model2)
plot(aov(model2))
anova.table2
summary(anova.table2)

kruskal.test(Genre~DomesticGross,data = hollywoodmovies_comparison1)

outcome=c(hollywoodmovies_horror$DomesticGross,hollywoodmovies_thriller$DomesticGross)
treat=c(rep(1,length(hollywoodmovies_horror$DomesticGross)),rep(2,length(hollywoodmovies_thriller
$DomesticGross)))
diffobs=mean(hollywoodmovies_horror$DomesticGross)-mean(hollywoodmovies_thriller$DomesticGross)
d=c()
p=c()

for(i in 1:5000){
  permut=sample(outcome)
  d[i]=mean(permut[treat==1])-mean(permut[treat==2])
  p[i]=(d[i]>=diffobs)+0
  }
pvalue=sum(p)/5000
pvalue

# horror vs thriller OpeningWeekend
model3 = lm(OpeningWeekend~Genre, data = hollywoodmovies_comparison1)
anova.table3 = anova(model3)
plot(aov(model3))
anova.table3
summary(anova.table3)

kruskal.test(Genre~OpeningWeekend,data = hollywoodmovies_comparison1)

outcome=c(hollywoodmovies_horror$OpeningWeekend,hollywoodmovies_thriller$OpeningWeekend)
treat=c(rep(1,length(hollywoodmovies_horror$OpeningWeekend)),rep(2,length(hollywoodmovies_thriller
$OpeningWeekend)))
diffobs=mean(hollywoodmovies_horror$OpeningWeekend)-mean(hollywoodmovies_thriller$OpeningWeekend)
d=c()
p=c()

for(i in 1:5000){
  permut=sample(outcome)
  d[i]=mean(permut[treat==1])-mean(permut[treat==2])
```

```
  p[i]=(d[i]>=diffobs)+0
  }
pvalue=sum(p)/5000
pvalue


#seperate animation and fantasy
hollywoodmovies_animation_idx = which(hollywoodmovies$Genre == "Animation")
hollywoodmovies_animation = hollywoodmovies[hollywoodmovies_animation_idx,]
hollywoodmovies_fantasy_idx = which(hollywoodmovies$Genre == "Fantasy")
hollywoodmovies_fantasy = hollywoodmovies[hollywoodmovies_fantasy_idx,]
hollywoodmovies_comparison2 = rbind(hollywoodmovies_animation, hollywoodmovies_fantasy)

#animation vs fantasy budget
model4 = lm(Budget~Genre, data = hollywoodmovies_comparison2)
anova.table4 = anova(model4)
plot(aov(model4))
anova.table4
summary(anova.table4)

kruskal.test(Genre~Budget, data = hollywoodmovies_comparison2)

outcome=c(hollywoodmovies_animation$Budget, hollywoodmovies_fantasy$Budget)
treat=c(rep(1, length(hollywoodmovies_animation$Budget)), rep(2, length(hollywoodmovies_fantasy$Budge
t)))
diffobs=mean(hollywoodmovies_animation$Budget)-mean(hollywoodmovies_fantasy$Budget)
d=c()
p=c()

for(i in 1:5000){
  permut=sample(outcome)
  d[i]=mean(permut[treat==1])-mean(permut[treat==2])
  p[i]=(d[i]>=diffobs)+0
  }
pvalue=sum(p)/5000
pvalue

#animation vs fantasy DomesticGross
model5 = lm(DomesticGross~Genre, data = hollywoodmovies_comparison2)
anova.table5 = anova(model5)
plot(aov(model5))
anova.table5
summary(anova.table5)

kruskal.test(Genre~DomesticGross, data = hollywoodmovies_comparison2)

outcome=c(hollywoodmovies_animation$DomesticGross, hollywoodmovies_fantasy$DomesticGross)
treat=c(rep(1, length(hollywoodmovies_animation$DomesticGross)), rep(2, length(hollywoodmovies_fantas
y$DomesticGross)))
diffobs=mean(hollywoodmovies_animation$DomesticGross)-mean(hollywoodmovies_fantasy$DomesticGross)
d=c()
p=c()

for(i in 1:5000){
```

```
  permut=sample(outcome)
  d[i]=mean(permut[treat==1])-mean(permut[treat==2])
  p[i]=(d[i]>=diffobs)+0
  }
pvalue=sum(p)/5000
pvalue


#animation vs fantasy OpeningWeekend
model6 = lm(OpeningWeekend~Genre, data = hollywoodmovies_comparison2)
anova.table6 = anova(model6)
plot(aov(model6))
anova.table6
summary(anova.table6)

kruskal.test(Genre~OpeningWeekend,data = hollywoodmovies_comparison2)

outcome=c(hollywoodmovies_animation$OpeningWeekend,hollywoodmovies_fantasy$OpeningWeekend)
treat=c(rep(1,length(hollywoodmovies_animation$OpeningWeekend)),rep(2,length(hollywoodmovies_fanta
sy$OpeningWeekend)))
diffobs=mean(hollywoodmovies_animation$OpeningWeekend)-mean(hollywoodmovies_fantasy$OpeningWeeken
d)
d=c()
p=c()

for(i in 1:5000){
  permut=sample(outcome)
  d[i]=mean(permut[treat==1])-mean(permut[treat==2])
  p[i]=(d[i]>=diffobs)+0
  }
pvalue=sum(p)/5000
pvalue


#seperate action and adventure
hollywoodmovies_action_idx = which(hollywoodmovies$Genre == "Action")
hollywoodmovies_action = hollywoodmovies[hollywoodmovies_action_idx,]
hollywoodmovies_adventure_idx = which(hollywoodmovies$Genre == "Adventure")
hollywoodmovies_adventure = hollywoodmovies[hollywoodmovies_adventure_idx,]
hollywoodmovies_comparison3 = rbind(hollywoodmovies_action,hollywoodmovies_adventure)

#action vs adventure budget

mean(hollywoodmovies_comparison3$Budget)
sd(hollywoodmovies_comparison3$Budget)

#only one factor with one level

outcome=c(hollywoodmovies_action$Budget,hollywoodmovies_adventure$Budget)
treat=c(rep(1,length(hollywoodmovies_action$Budget)),rep(2,length(hollywoodmovies_adventure$Budge
t)))
diffobs=mean(hollywoodmovies_action$Budget)-mean(hollywoodmovies_adventure$Budget)
d=c()
p=c()
```

```
for(i in 1:1000){
  permut=sample(outcome)
  d[i]=mean(permut[treat==1])-mean(permut[treat==2])
  p[i]=(d[i]>=diffobs)+0
  }
pvalue=sum(p)/1000
pvalue

#Only one level so no p-value

#action vs adventure DomesticGross
mean(hollywoodmovies_comparison3$DomesticGross)
sd(hollywoodmovies_comparison3$DomesticGross)

#only one factor with one level

outcome=c(hollywoodmovies_action$DomesticGross,hollywoodmovies_adventure$DomesticGross)
treat=c(rep(1,length(hollywoodmovies_action$DomesticGross)),rep(2,length(hollywoodmovies_adventure
$DomesticGross)))
diffobs=mean(hollywoodmovies_action$DomesticGross)-mean(hollywoodmovies_adventure$DomesticGross)
d=c()
p=c()

for(i in 1:1000){
  permut=sample(outcome)
  d[i]=mean(permut[treat==1])-mean(permut[treat==2])
  p[i]=(d[i]>=diffobs)+0
  }
pvalue=sum(p)/1000
pvalue

#Only one level so no p-value


#action vs adventure OpeningWeekend
mean(hollywoodmovies_comparison3$OpeningWeekend)
sd(hollywoodmovies_comparison3$OpeningWeekend)

#only one factor with one level

outcome=c(hollywoodmovies_action$OpeningWeekend,hollywoodmovies_adventure$OpeningWeekend)
treat=c(rep(1,length(hollywoodmovies_action$OpeningWeekend)),rep(2,length(hollywoodmovies_adventur
e$OpeningWeekend)))
diffobs=mean(hollywoodmovies_action$OpeningWeekend)-mean(hollywoodmovies_adventure$OpeningWeekend)
d=c()
p=c()

for(i in 1:1000){
  permut=sample(outcome)
  d[i]=mean(permut[treat==1])-mean(permut[treat==2])
  p[i]=(d[i]>=diffobs)+0
  }
pvalue=sum(p)/1000
```

```
pvalue

#Only one level so no p-value

pearson <- sum((hollywoodmovies_action$DomesticGross-mean(hollywoodmovies_action$DomesticGross))*
(hollywoodmovies_action$OpeningWeekend-mean(hollywoodmovies_action$OpeningWeekend)))/
  sqrt(sum((hollywoodmovies_action$DomesticGross-mean(hollywoodmovies_action$DomesticGross))^2)*su
m((hollywoodmovies_action$OpeningWeekend-mean(hollywoodmovies_action$OpeningWeekend))^2))
pearson

DomesticGross <- order(hollywoodmovies_action$DomesticGross)
OpeningWeekend <- order(hollywoodmovies_action$OpeningWeekend)

spearman <- sum((DomesticGross-mean(DomesticGross))*(OpeningWeekend-mean(OpeningWeekend)))/
  sqrt(sum((DomesticGross-mean(DomesticGross))^2)*sum((OpeningWeekend-mean(OpeningWeekend))^2))
spearman

#The Pearson correlation evaluates the linear relationship between two continuous variables, which
assumes normality. The Spearman correlation coefficient is based on the ranked values for each var
iable rather than the raw data.

model_action = lm(DomesticGross~OpeningWeekend, data = hollywoodmovies_action)
plot(aov(model_action))
summary(aov(model_action))

kruskal.test(DomesticGross~OpeningWeekend,data = hollywoodmovies_action)

outcome=c(hollywoodmovies_action$DomesticGross,hollywoodmovies_action$OpeningWeekend)
treat=c(rep(1,length(hollywoodmovies_action$DomesticGross)),rep(2,length(hollywoodmovies_action$Op
eningWeekend)))
diffobs=mean(hollywoodmovies_action$DomesticGross)-mean(hollywoodmovies_action$OpeningWeekend)
d=c()
p=c()

for(i in 1:5000){
  permut=sample(outcome)
  d[i]=mean(permut[treat==1])-mean(permut[treat==2])
  p[i]=(d[i]>=diffobs)+0
  }
pvalue=sum(p)/5000
pvalue

#Since the action movie dataset is relatively small (30 obs) and it is not normal based on the plo
t of the anova test (Residual vs Fitted plot and normal QQ plot), so it is more suitable to use no
n-parametric test.
model_action2 = lm(DomesticGross~ForeignGross, data = hollywoodmovies_action)
plot(aov(model_action2))
summary(aov(model_action2))

kruskal.test(DomesticGross~ForeignGross,data = hollywoodmovies_action)

outcome=c(hollywoodmovies_action$DomesticGross,hollywoodmovies_action$ForeignGross)
treat=c(rep(1,length(hollywoodmovies_action$DomesticGross)),rep(2,length(hollywoodmovies_action$Fo
reignGross)))
```

```
diffobs=mean(hollywoodmovies_action$ForeignGross)-mean(hollywoodmovies_action$ForeignGross)
d=c()
p=c()

for(i in 1:5000){
  permut=sample(outcome)
  d[i]=mean(permut[treat==1])-mean(permut[treat==2])
  p[i]=(d[i]>=diffobs)+0
  }
pvalue=sum(p)/5000
pvalue

#Since the action movie dataset is relatively small (30 obs) and it is not normal based on the plo
t of the anova test (Residual vs Fitted plot and normal QQ plot), so it is more suitable to use no
n-parametric test. According to both kruskal test and permutation test, both tests suggest that th
ere is no significant difference between Domestic and Foreign Gross.
```