104 project

1. t.test $conf.int 1.b z- calculate a,b

nomal 1.qqnorm function 2.histgram

3. compare budget, domestic gross and opening weekend for the following three categories

```
data <- readxl::read_xlsx("hollywoodmovies.xlsx")

ind1 <- (data$Genre=="Horror")+0
ind2 <- (data$Genre=="Thriller")+0
ind_1 <- ind1+ind2

ind1 <- (data$Genre=="Animation")+0
ind2 <- (data$Genre=="Fantasy")+0
ind3 <- (data$Genre=="Romance")+0
ind_2 <- ind1+ind2+ind3

ind1 <- (data$Genre=="Action")+0
ind2 <- (data$Genre=="Animation")+0
ind_3<-ind1+ind2

group <- ind_1+2*ind_2+3*ind_3
data$group<-group
data1<-subset(data,data$group != 0)

aov1<-aov(data1$Budget~data1$group)
summary(aov1)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## data1$group   1  51252   51252    24.3 6.84e-06 ***
## Residuals    60 126550    2109                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
aov2<-aov(data1$DomesticGross~data1$group)
summary(aov2)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)  
## data1$group   1  23195   23195   3.878 0.0535 .
## Residuals    60 358869    5981                 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
aov3<-aov(data1$OpeningWeekend~data1$group)
summary(aov3)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## data1$group  1   2263  2263.2    2.43  0.124
## Residuals   60  55892   931.5
```

```
# nonparametric equivalents- Krustal Wallis
kruskal.test(Budget ~ group, data = data1)
```

```
## 
##  Kruskal-Wallis rank sum test
## 
## data:  Budget by group
## Kruskal-Wallis chi-squared = 18.341, df = 2, p-value = 0.000104
```

```
kruskal.test(DomesticGross ~ group, data = data1)
```

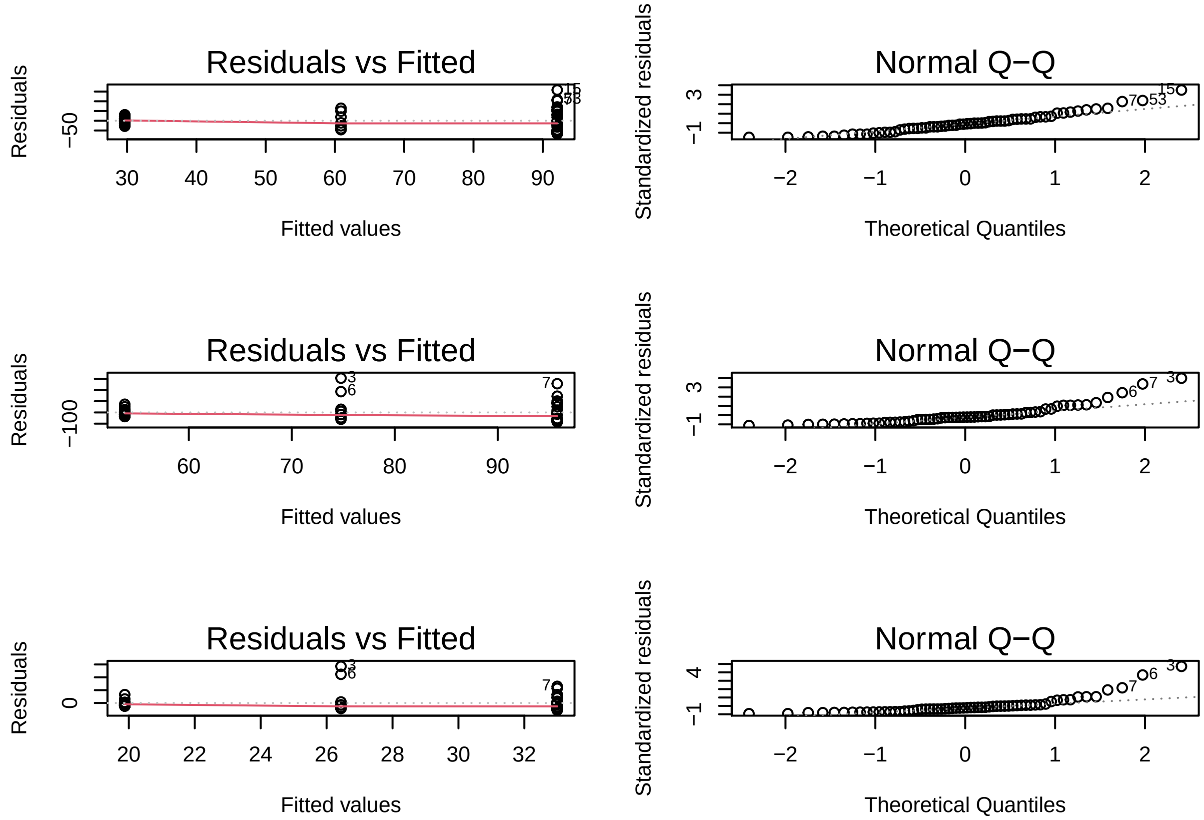
```
## 
##  Kruskal-Wallis rank sum test
## 
## data:  DomesticGross by group
## Kruskal-Wallis chi-squared = 2.215, df = 2, p-value = 0.3304
```

```
kruskal.test(OpeningWeekend ~ group, data = data1)
```

```
## 
##  Kruskal-Wallis rank sum test
## 
## data:  OpeningWeekend by group
## Kruskal-Wallis chi-squared = 2.3109, df = 2, p-value = 0.3149
```

The ANOVA test assumes that: (1) The observations are obtained independently and randomly from the population defined by the factor levels (2) The data of each factor level are normally distributed. (3) These normal populations have a common variance. The residuals versus fits plot can be used to check the homogeneity of variances.

```
par(mfrow=c(3,2))
plot(aov1, c(1,2))
plot(aov2, c(1,2))
plot(aov3, c(1,2))
```



```
# Levene's test to check the homogeneity of variances.
library(car)
```

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

```
leveneTest(Budget ~ factor(group), data = data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value    Pr(>F)
## group  3  12.967 3.343e-07 ***
##       98
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# To check the normality more precisely, we did the Shapiro-Wilk test on the ANOVA residuals (W = 0.942
# Extract the residuals
aov_residuals <- residuals(object = aov1 )
# Run Shapiro-Wilk test
shapiro.test(x = aov_residuals ) # normal
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aov_residuals
## W = 0.94268, p-value = 0.006018
```

```
# 4.
data2 <- subset(data,data$Genre=="Action")
cor(data2$DomesticGross, data2$OpeningWeekend, method = "pearson") #parametric
```

```
## [1] 0.9365815
```

```
cor.test(data2$DomesticGross, data2$OpeningWeekend, method="pearson") #
```

```
##
##  Pearson's product-moment correlation
##
## data:  data2$DomesticGross and data2$OpeningWeekend
## t = 13.887, df = 27, p-value = 8.194e-14
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8680420 0.9700913
## sample estimates:
##       cor
## 0.9365815
```

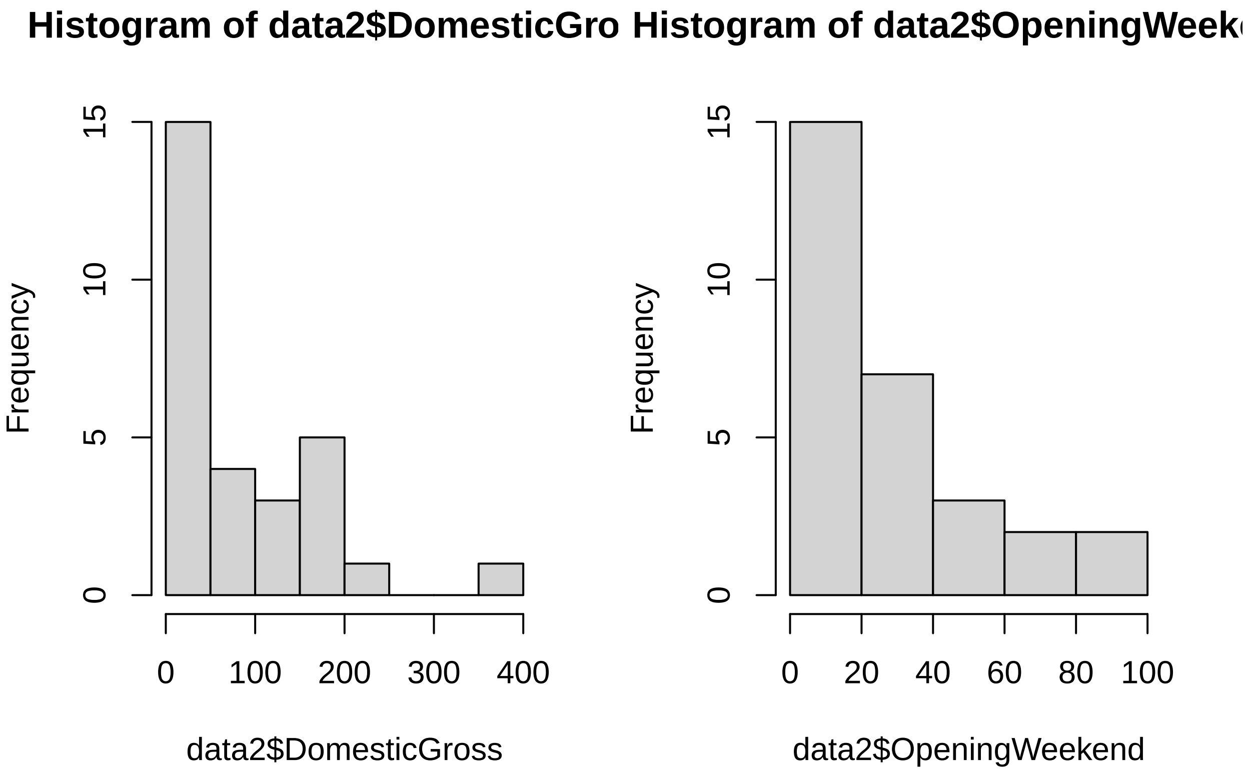
```
cor(data2$DomesticGross, data2$OpeningWeekend, method = "spearman") #nonparametric
```

```
## [1] 0.9438424
```

```
cor.test(data2$DomesticGross, data2$OpeningWeekend, method="spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  data2$DomesticGross and data2$OpeningWeekend
## S = 228, p-value = 8.629e-08
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.9438424
```

```
par(mfrow=c(1,2))
hist(data2$DomesticGross)
hist(data2$OpeningWeekend)
```



```r
plot(data2$DomesticGross, data2$OpeningWeekend)

model_1<-lm(DomesticGross~Budget, data=data2)
summary(model_1)
```

```
##
## Call:
## lm(formula = DomesticGross ~ Budget, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -94.771 -25.277   1.559  22.468 146.711
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -16.3700    16.8170  -0.973    0.339
## Budget        1.1387     0.1514   7.522 4.31e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48.97 on 27 degrees of freedom
## Multiple R-squared:  0.677,  Adjusted R-squared:  0.665
## F-statistic: 56.58 on 1 and 27 DF,  p-value: 4.314e-08
```

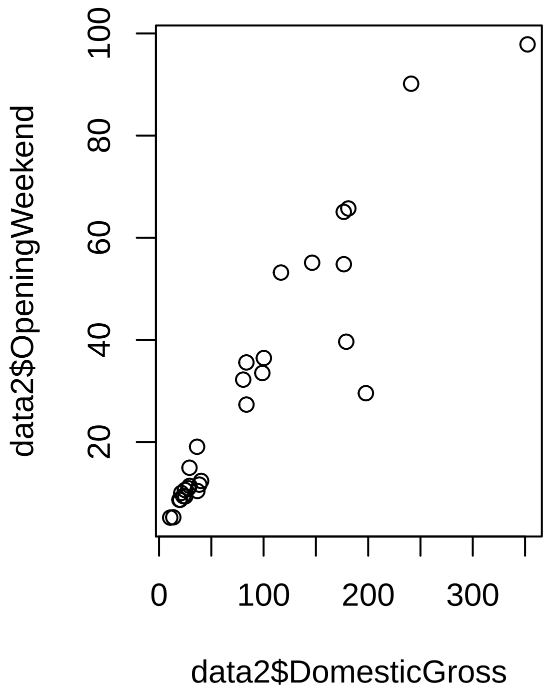
```r
# nonparametric one
library("lmPerm")
model_n<-lmp(DomesticGross~Budget, data=data2)
```

```
## [1] "Settings:  unique SS : numeric variables centered"
```

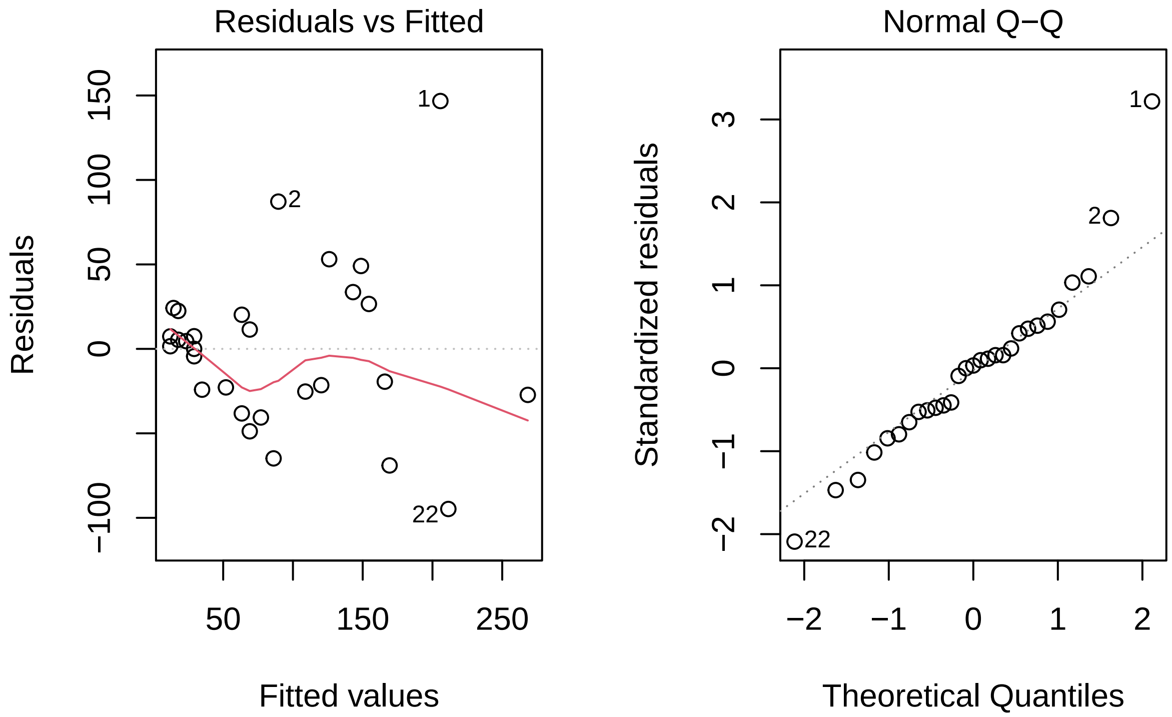
```r
summary(model_n)
```

```
##
## Call:
## lmp(formula = DomesticGross ~ Budget, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -94.771 -25.277   1.559  22.468 146.711
##
## Coefficients:
##        Estimate Iter Pr(Prob)
## Budget    1.139 5000   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 48.97 on 27 degrees of freedom
## Multiple R-Squared: 0.677,   Adjusted R-squared: 0.665
## F-statistic: 56.58 on 1 and 27 DF,  p-value: 4.314e-08
```

```r
par(mfrow=c(1,2))
```

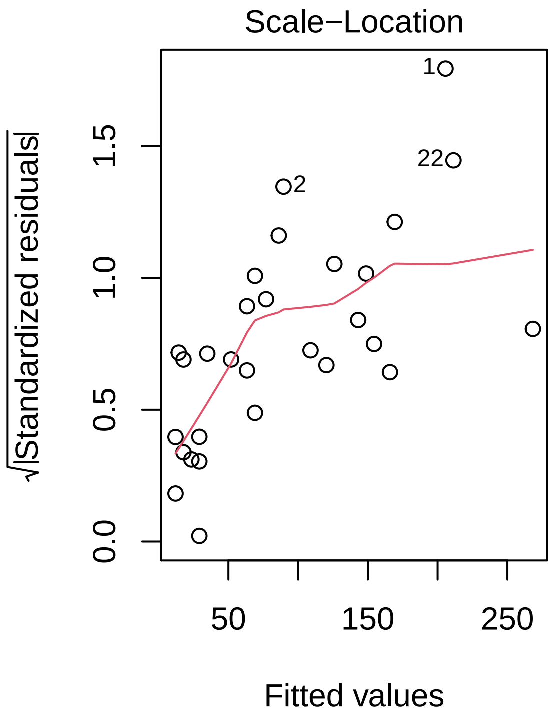


```r
plot(model_l,c(1,2,3))
```



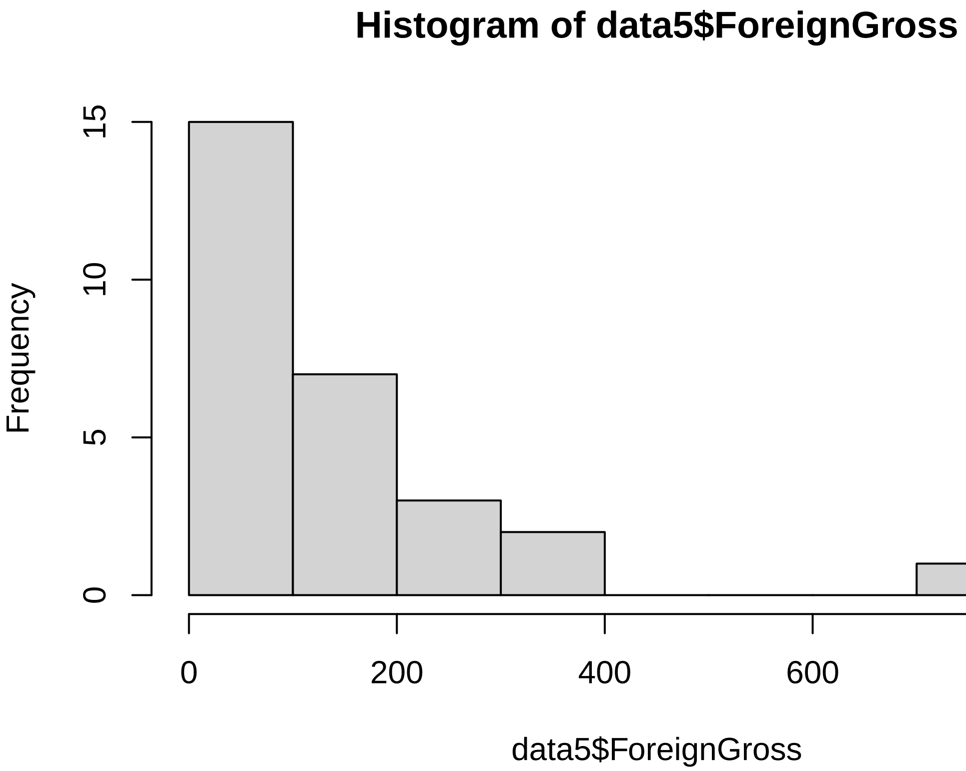
```
aov_residuals <- residuals(object = model_1 )
# Run Shapiro-Wilk test
shapiro.test(x = aov_residuals )
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aov_residuals
## W = 0.95095, p-value = 0.1938
```

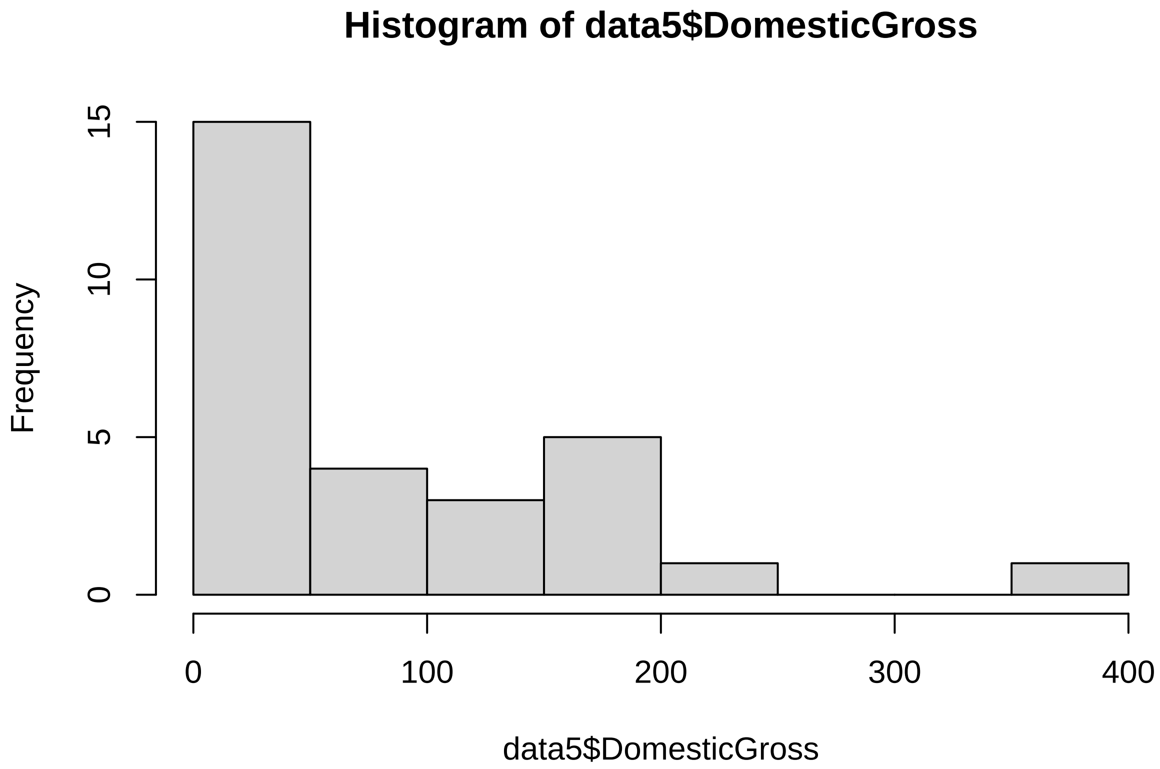


5. paired test: paired t test

```
data5<- subset(data,data$Genre=="Action")
hist(data5$ForeignGross)
```



```
hist(data5$DomesticGross)
```



```r
mean(data5$ForeignGross)
```

```
## [1] 154.5581
```

```r
mean(data5$DomesticGross)
```

```
## [1] 90.04062
```

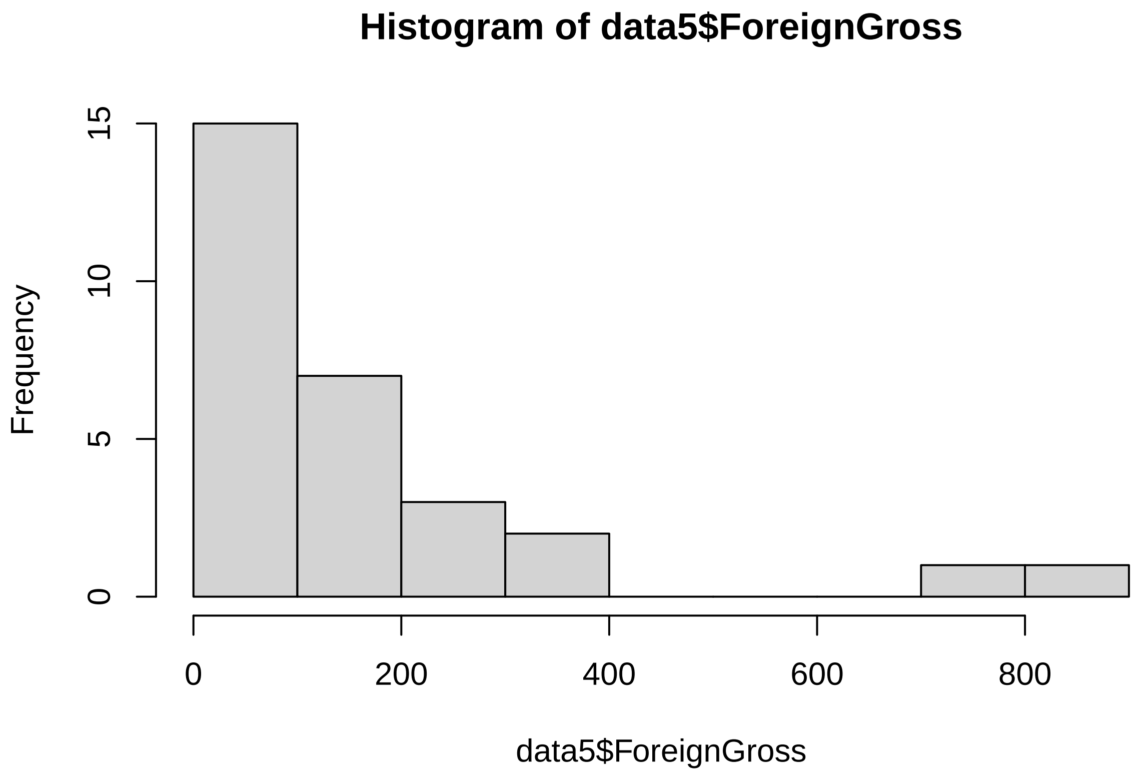
```r
# paired t test
t.test(data5$ForeignGross,data5$DomesticGross,alternative = "greater", paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  data5$ForeignGross and data5$DomesticGross
## t = 2.7269, df = 28, p-value = 0.005454
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  24.26976      Inf
## sample estimates:
## mean of the differences
##                64.51752
```

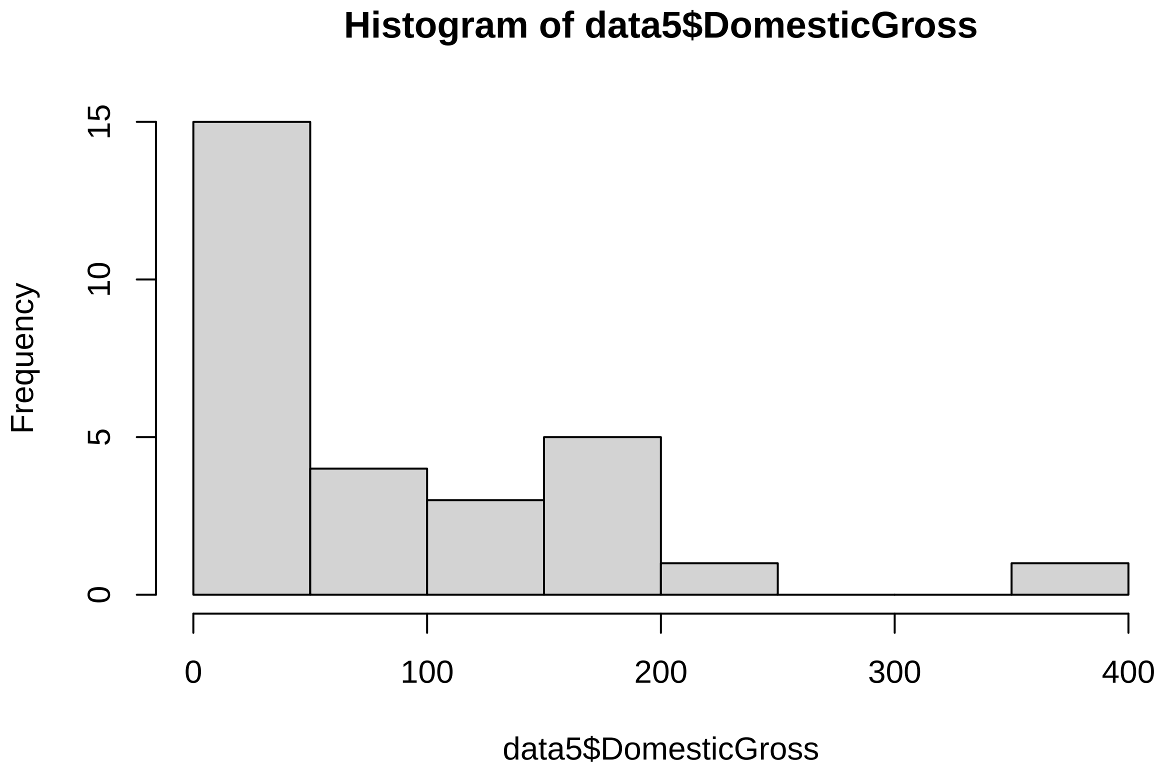
```
# non-parametric equivalent: Wilcoxon signed-rank test
wilcox.test(data5$ForeignGross,data5$DomesticGross,alternative = "greater", paired = TRUE)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  data5$ForeignGross and data5$DomesticGross
## V = 375, p-value = 0.0001661
## alternative hypothesis: true location shift is greater than 0
```

```
# assumption
hist(data5$ForeignGross)
```



```
hist(data5$DomesticGross)
```


Histogram of data5$DomesticGross
Frequency
0
5
10
15
0
100
200
300
400
data5$DomesticGross