STAT 382, Project 2

Yelizaveta Semikina

2022-11-23

Task 1

Importing the apps_data.csv and naming it mydata and importing paid_apps.csv and naming it mydata1. Also, converting the Category, Content.Rating and Genre in both data sets to factors.

Task 2

```
##
##  One Sample t-test
##
## data:  mydata$App_Size
## t = 14.652, df = 163, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 0.25
## 92 percent confidence interval:
##  27.94954      Inf
## sample estimates:
## mean of x
##  30.90238
```

H0: mu = 0.25 vs H1: mu > 0.25. P-value is 2.2e-16. p-value is less than significance level of 0.02, thus, reject H0.There is enough evidence that the population mean is greater than 0.25 The confidence interval is 27.94954 > mu. It supports our conclusion because 0.25 is not on included on the interval. We can conduct a t-test despite that data is not normally distributed because. Based on Central Limit Theorem, when the sample size is large and data is skewed, t-test can be still performed based on these requirements.

Task 3

Splitting the data of the Category game is the Category, Game and notgame are Categories Family and Tools.

Equal Variance Test

```
##
##  F test to compare two variances
##
## data:  game and notgame
## F = 7.7201, num df = 40, denom df = 122, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 93 percent confidence interval:
```

```
##   4.963499 12.751215
## sample estimates:
## ratio of variances
##           7.720133
```

p-value (2.2e-16) is less than significance level of 0.07, thus, reject H0, there is significant difference between game and not games categories.

Difference of 2 Means

```
##
##  Welch Two Sample t-test
##
## data:  game and notgame
## t = 1.6701, df = 43.502, p-value = 0.1021
## alternative hypothesis: true difference in means is not equal to 0
## 93 percent confidence interval:
##  -41347.14 777641.39
## sample estimates:
## mean of x mean of y
##  480220.9  112073.8
```

H0: game = notgame vs H1: game != notgame. p-value is 0.1021, it is bigger than significance level of 0.07, thus do not reject H0. There is no significant difference in the means of game and notgame. The confidence interval is -41347.14 < muD < 777641.39, zero is on the interval. Thus, it supports out conclusion that we do not reject H0.

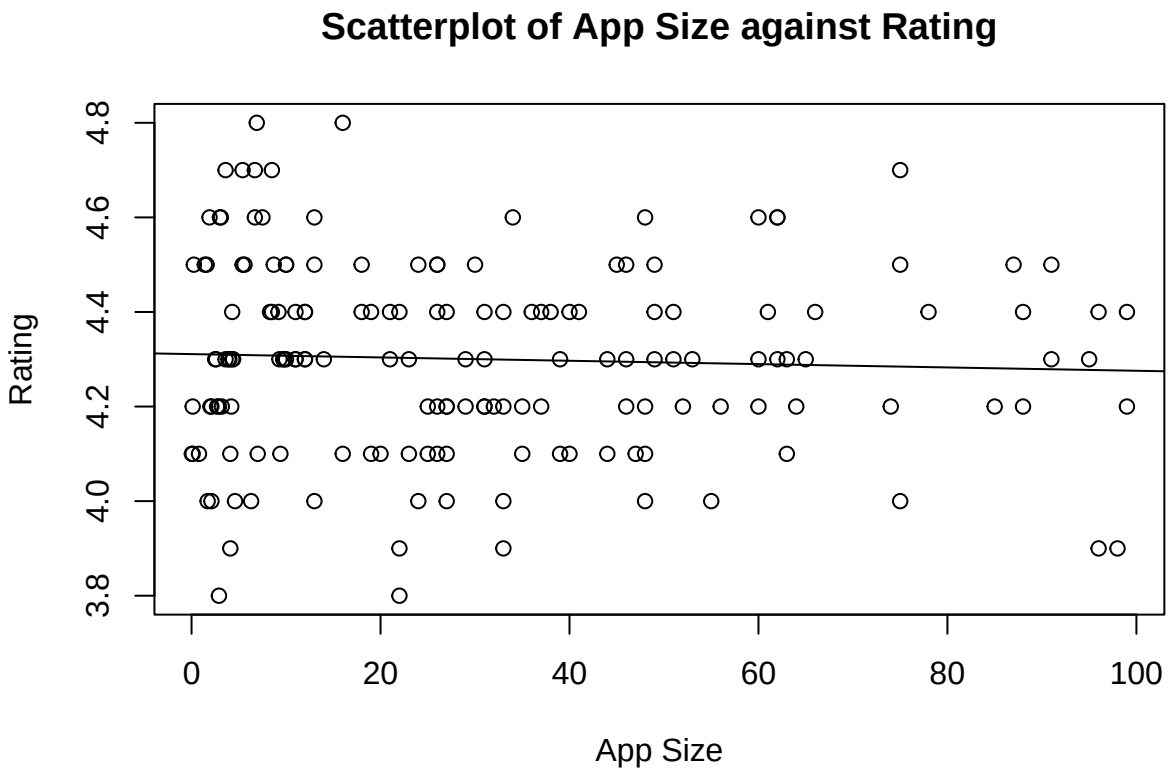
Task 4

```
##
##  Fisher's Exact Test for Count Data
##
## data:  mydata$Content.Rating and mydata$Category
## p-value = 2.314e-09
## alternative hypothesis: two.sided
```

H0: theta = 1 vs H1: theta != 1. p-value is 2.314e-09. P-value is less than significance level of 0.04, thus, reject H0. There is evidence that Content.Rating and Category are not independent.

Task 5

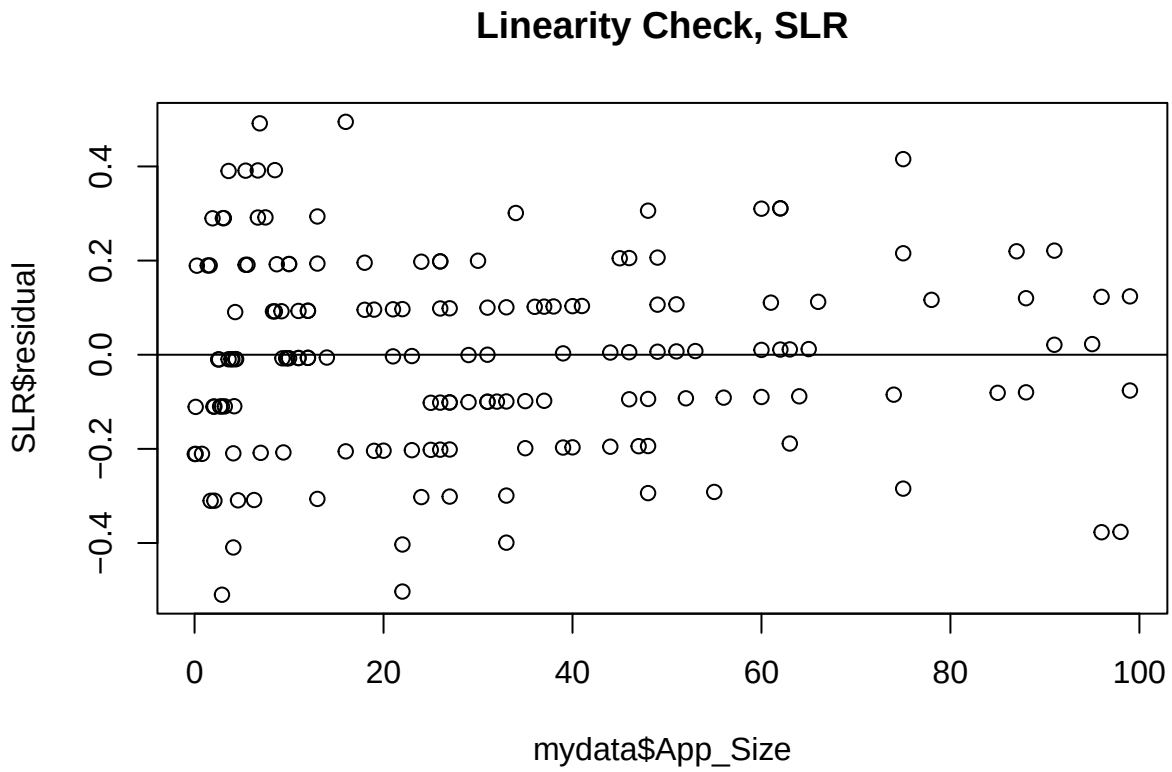
Scatterplot



Pearson

```
##
##  Pearson's product-moment correlation
##
## data:  mydata$Rating and mydata$App_Size
## t = -0.59118, df = 162, p-value = 0.5552
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1982378  0.1076176
## sample estimates:
##        cor
## -0.0463975
```

Linearity Check



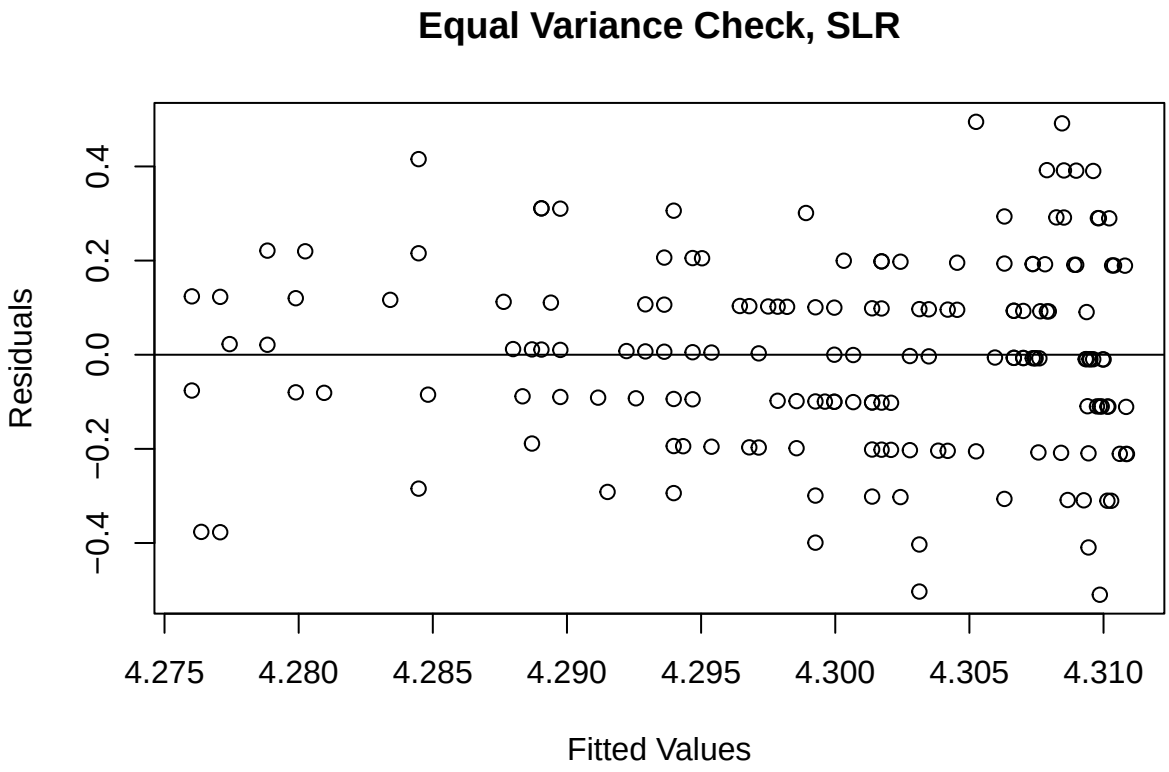
Linear assumption is met because there is no pattern.

Normality Check

```
##
##  Shapiro-Wilk normality test
##
## data:  SLR$residuals
## W = 0.98765, p-value = 0.1587
```

Normality assumption is met because p-value (0.1587) is big compare to significance level of 0.03.

Equal Variance Check



Equal Variance Assumption is met because there is no pattern.

```
##
## Call:
## lm(formula = mydata$Rating ~ mydata$App_Size)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50986 -0.10998 -0.00472  0.11747  0.49475
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.3108829  0.0243315 177.173   <2e-16 ***
## mydata$App_Size -0.0003522  0.0005957  -0.591    0.555
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2038 on 162 degrees of freedom
## Multiple R-squared:  0.002153,   Adjusted R-squared:  -0.004007
## F-statistic: 0.3495 on 1 and 162 DF,  p-value: 0.5552
```

H0: beta1 =0 vs H1: beta2 != 0. p-value is 0.5552. We use F-statistics.p-value is bigger than significance level of 0.03, thus, do not reject H0. There is not enough evidence that App Size is important in explaining some of the variability in Rating.

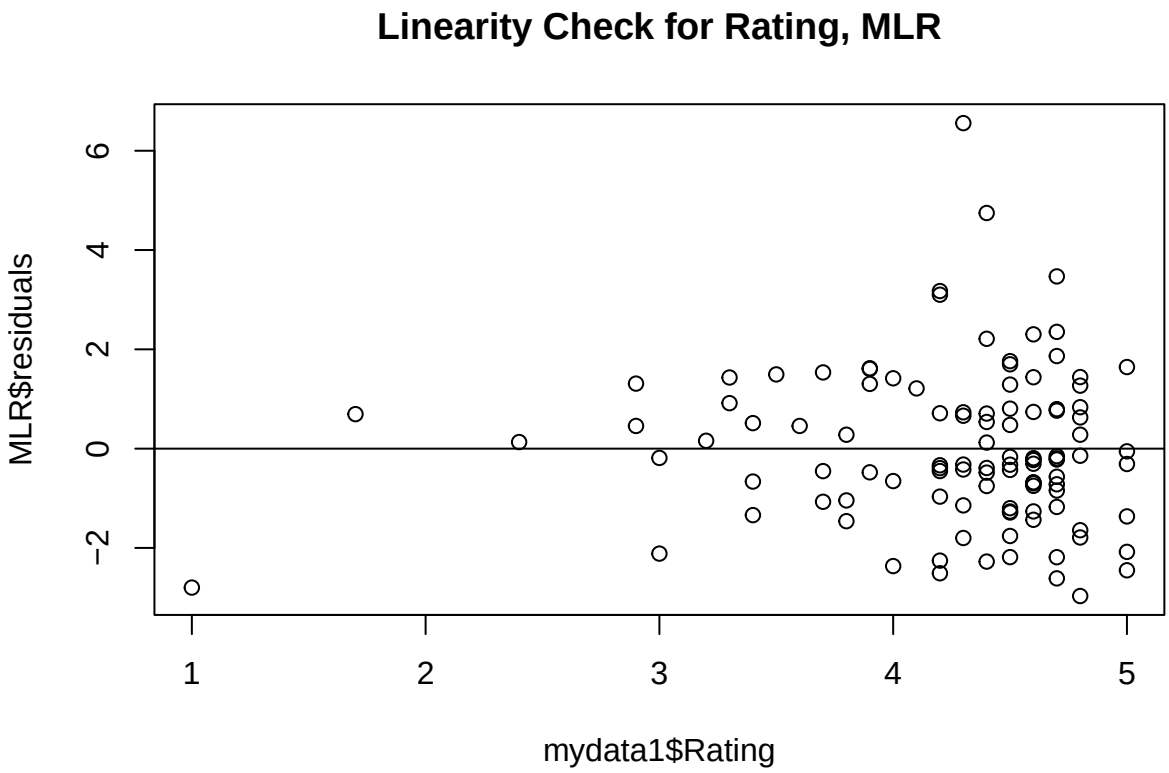
Equation of the Regression Line, and the Value of R2

```
##     (Intercept) mydata$App_Size
##    4.3108828780   -0.0003521695
```

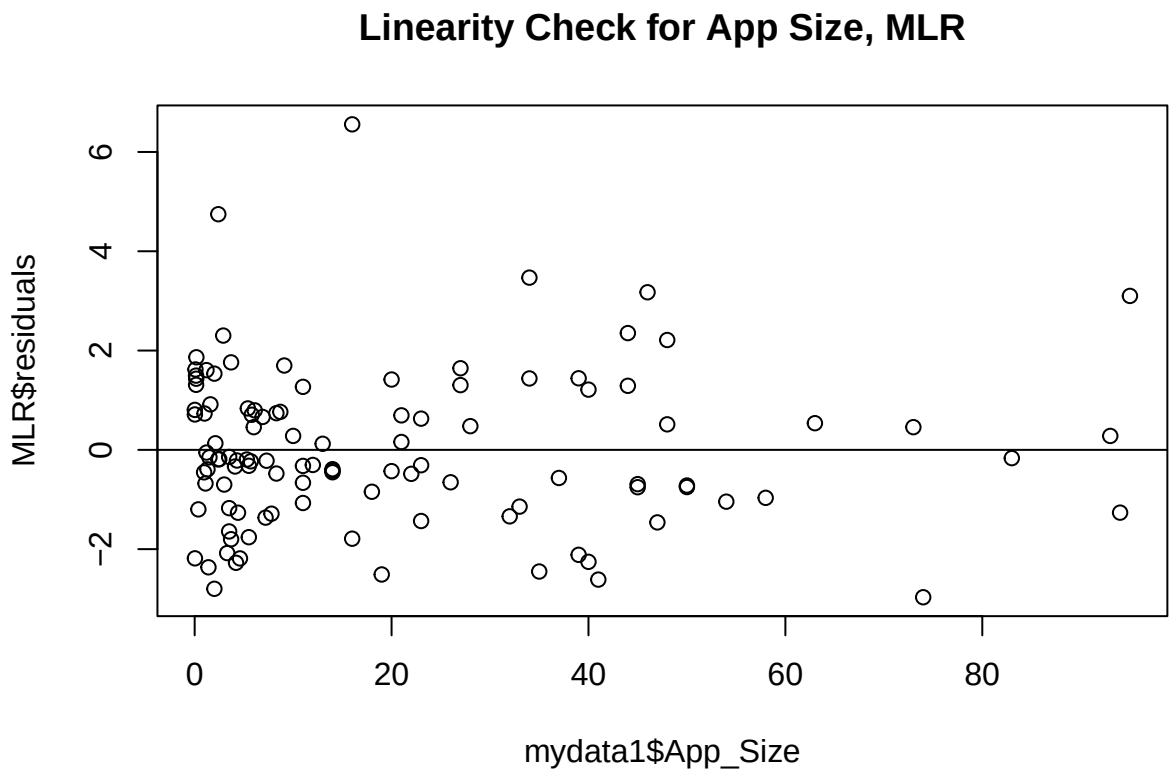
Equation of the Regression Line: yi = 4.31 -0.0003X (4.31 and -0.0003 were found in the SLR$coefficients). R-squared is 0.002153 = 0.21%. R^2 is closer to 0.5, indicating the model is not doing good a job and it does not fit the model well. 0.21% of the variability in Rating is explained by App Size.

Task 6

Linearity Check



Linearity check is not met because there is fanned pattern.



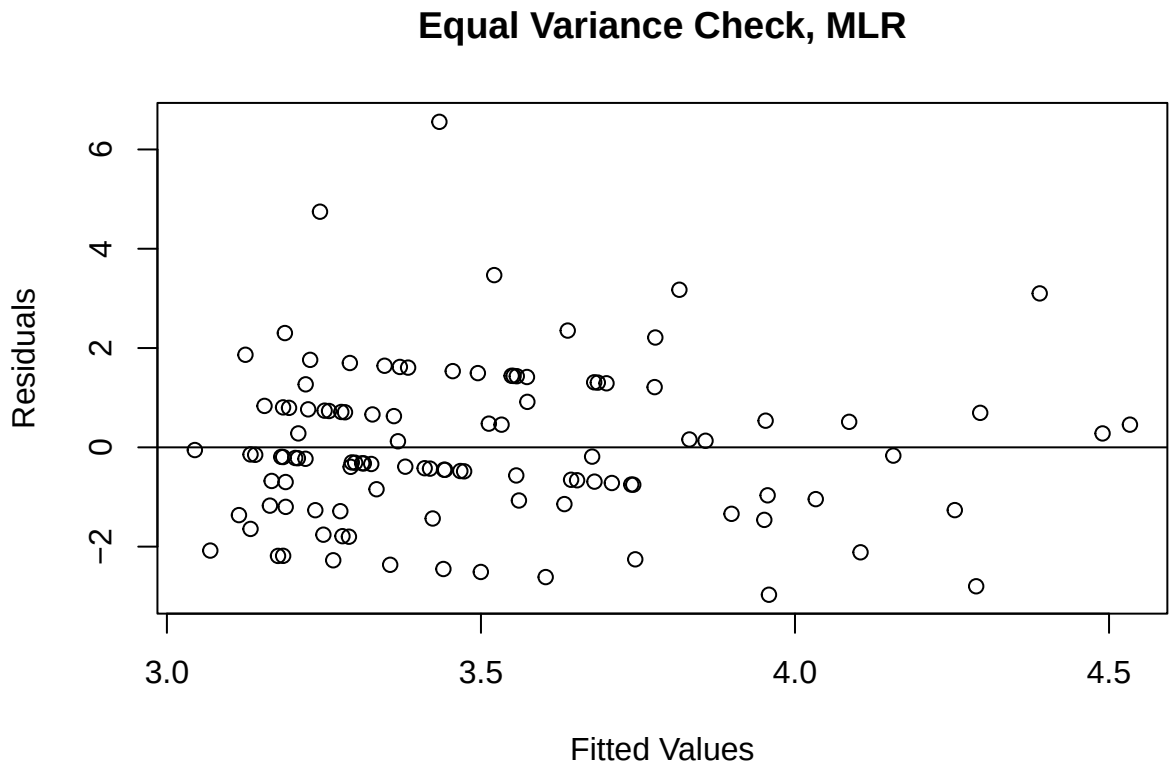
Linearity check is met because there is no pattern.

Normality Check

```
##
##  Shapiro-Wilk normality test
##
## data:  MLR$residuals
## W = 0.95522, p-value = 0.001051
```

The p-value (0.001051) is small compare to significance level of 0.10. It is not normal.

Equal Variance Check



Equal Variance Check is met because there is no pattern.

Independent Variables Check

```
##
## Call:
## lm(formula = mydata1$Price ~ mydata1$Rating + mydata1$App_Size)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9686 -1.0431 -0.2142  0.8050  6.5562
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.573738   1.004706   4.552 1.42e-05 ***
## mydata1$Rating   -0.308662   0.232152  -1.330   0.1865
## mydata1$App_Size  0.011708   0.006692   1.750   0.0831 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.593 on 106 degrees of freedom
## Multiple R-squared:  0.04287,    Adjusted R-squared:  0.02481
## F-statistic: 2.374 on 2 and 106 DF,  p-value: 0.09803
```

H0: beta1 = beta2 = 0 vs H1: at least one betaj != 0. p-value is 0.09803. Compare to significance level of 0.10 p-value is small, thus, Reject H0. There is enough evidence that App Size and Rating explain some of variability in Price. R-squared is 0.04287. Adjusted R-squared is 0.02481.

Determine which independent variables are important

Rating Variable H0: beta1 =0 vs H1: beta1 != 0. p-value for Rating is 0.1865. Compare to significance level of 0.10 p-value is big, thus, do not reject H0. There is not enough evidence that Rating explain some of variability in Price.

App Size Variable H0: beta2 =0 vs H1: beta2 != 0. p-value for App Size is 0.0831. compare to significance level of 0.10 p-value is smaller, thus, reject H0. There is enough evidence that App Size explain some of variability in Price. Thus, App Size variable is an independent variable that is more important than Rating in explaining some variability in Price because it has smaller p-value.

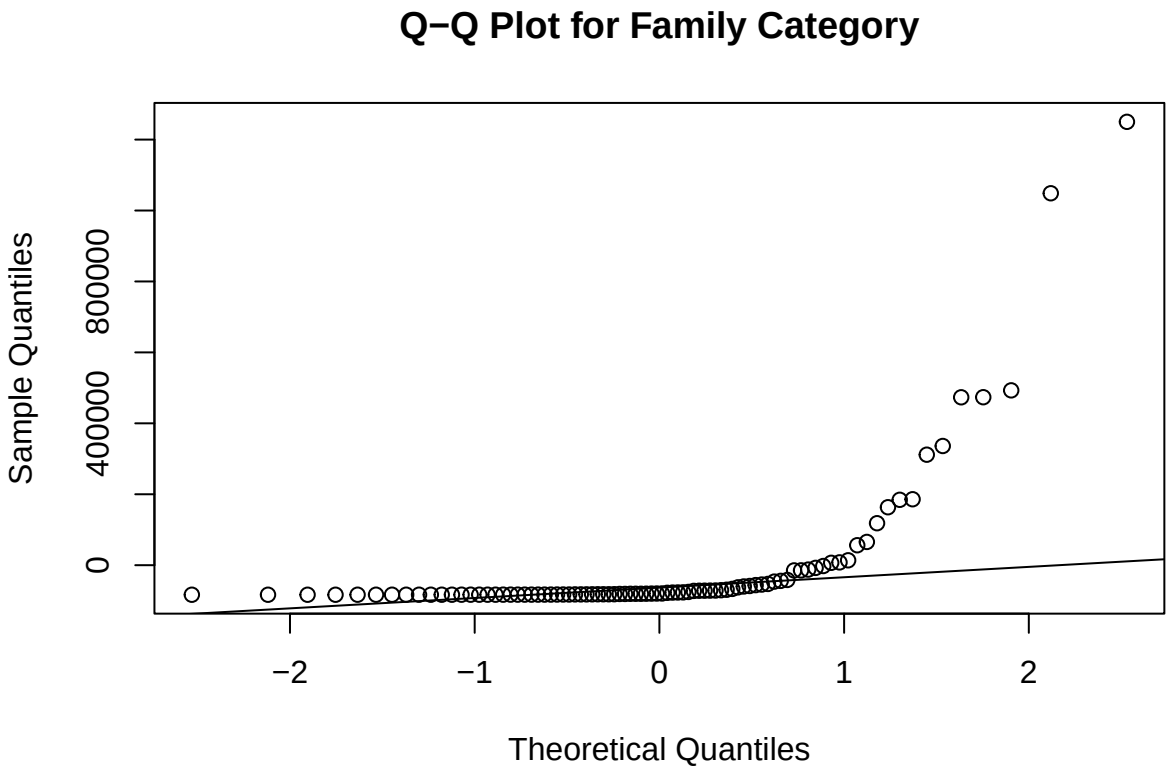
Task 7

```
## FAMILY   GAME  TOOLS
##     88     41     35
```

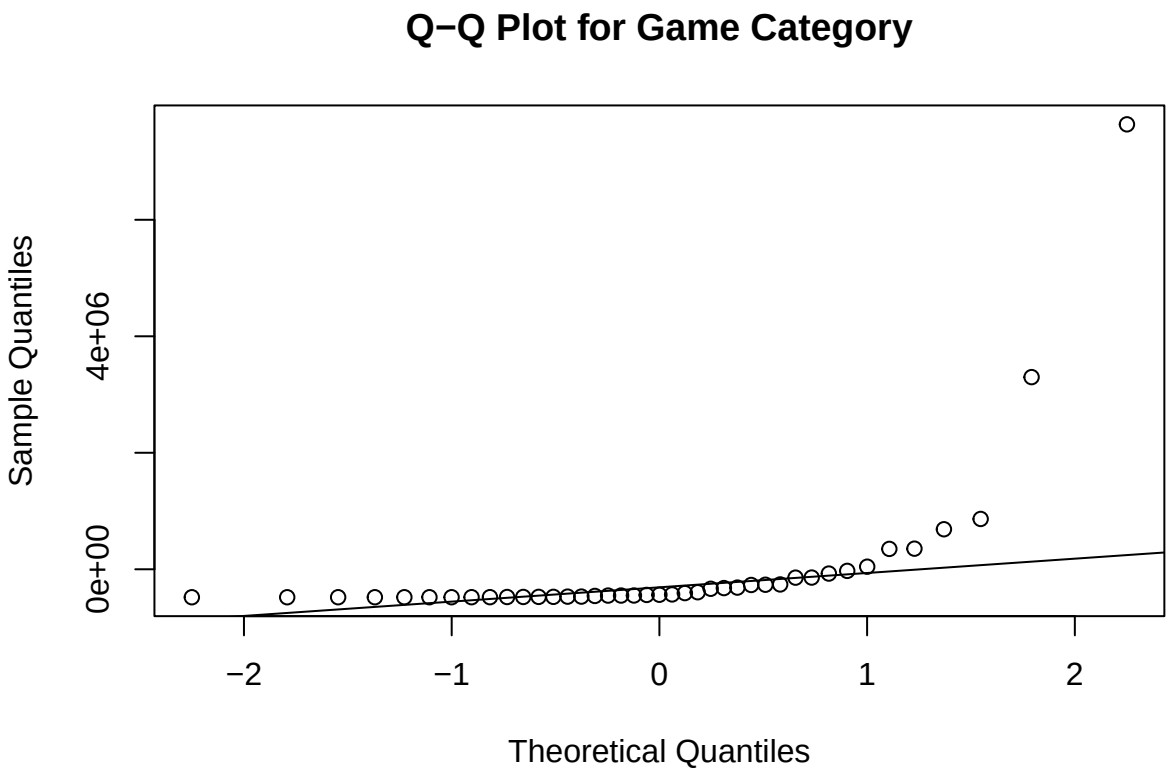
There are three groups and their values are bigeer than 30, thus, it means it is large and we can perform analysis by groups.

Creating a new data set with residuals and splitting the Category Data.

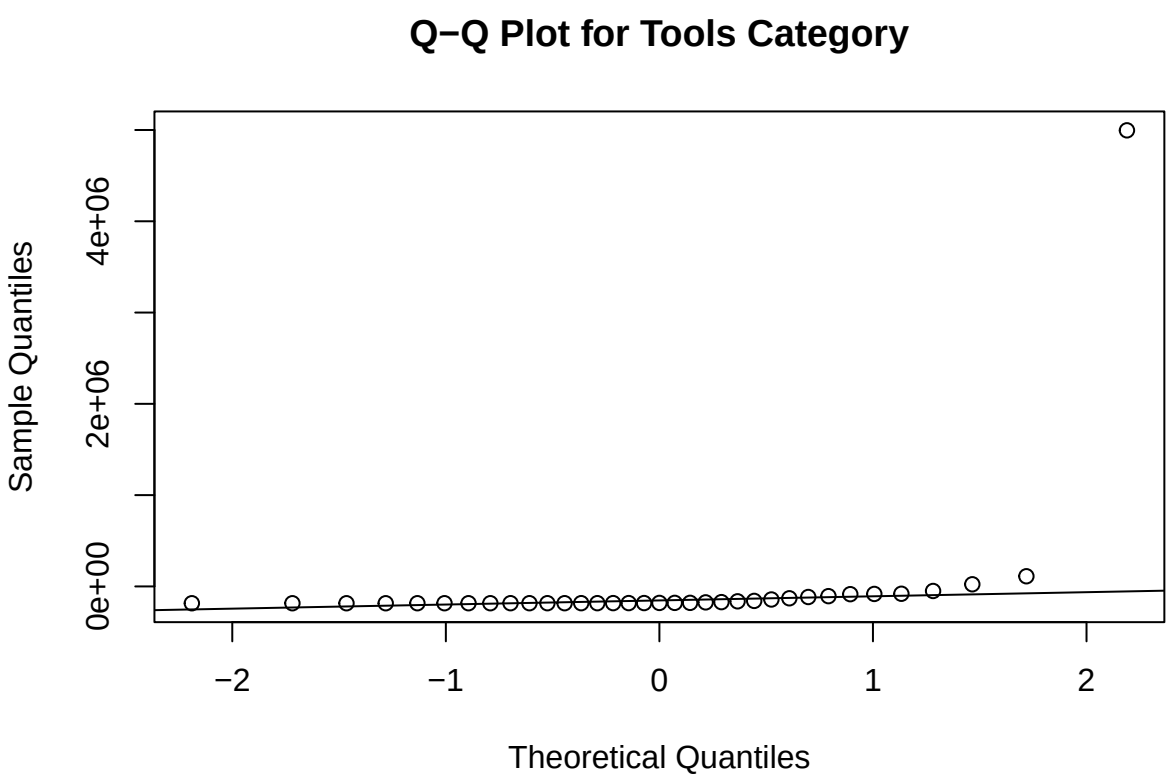
Normality Check



It is not normal because we can see that there are issues in the beginning and ending of the line.



It is not normal because we can see that there are issues in the beginning and ending of the line.



It is not normal because we can see that there are issues in the ending of the line.

```
##
##  Shapiro-Wilk normality test
##
## data:  mydata2fam
## W = 0.43362, p-value < 2.2e-16
```

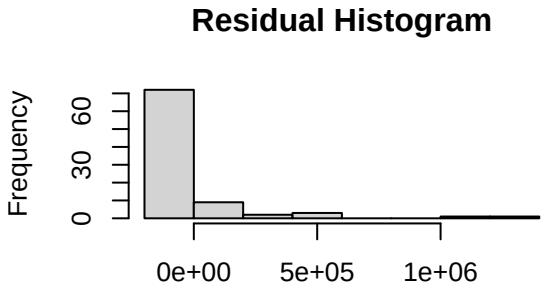
It is not normal because p-value(2.2e-16) is small compare to significance level of 0.04

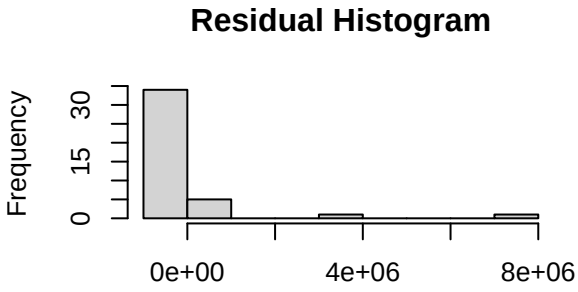
```
##
##  Shapiro-Wilk normality test
##
## data:  mydata2game
## W = 0.37045, p-value = 4.734e-12
```

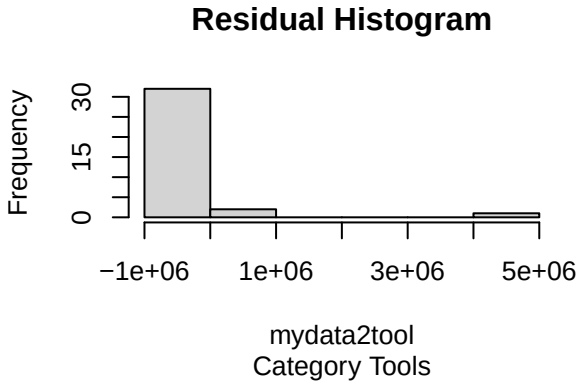
It is not normal because p-value(4.734e-12) is small compare to significance level of 0.04

```
##
##  Shapiro-Wilk normality test
##
## data:  mydata2tool
## W = 0.20486, p-value = 1.435e-12
```

It is not normal because p-value(1.435e-12) is small compare to significance level of 0.04







All of the three histograms are right skewed. It is not normal, thus we can say that normality condition is not met.

Equal Variance Check

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   2  3.2215 0.04248 *
##       161
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value(0.04248) is bigger than significance level of 0.04. Thus, do not reject, there is evidence for equal variance.

H0: muF = muG = muT vs H1: at least one of the means is different. p-value is 0.03758, it is smaller than significance level of 0.04. Thus, Reject H0, there is evidence that at least one of the means is different.

Tukey Test

```
##   Tukey multiple comparisons of means
##     96% family-wise confidence level
##
## Fit: aov(formula = mydata$Reviews ~ mydata$Category)
##
## $`mydata$Category`
```

```
##                   diff       lwr      upr     p adj
## GAME-FAMILY   396907.7   19716.9 774098.5 0.0286411
## TOOLS-FAMILY  101072.9 -297563.3 499709.1 0.8080690
## TOOLS-GAME   -295834.8 -754906.3 163236.7 0.2564046
```

Reject H0 for GAME-FAMILY because the p-value is less than 0.04 significance level and zero is not on the interval.For TOOLS-FAMILY and TOOLS-GAME do not reject H0 because the p-value are bigger than 0.04 significance level and zero is on the intervals. Thus, TOOLS-FAMILY and TOOLS-GAME difference is not significant and for GAME-FAMILY difference is significant.

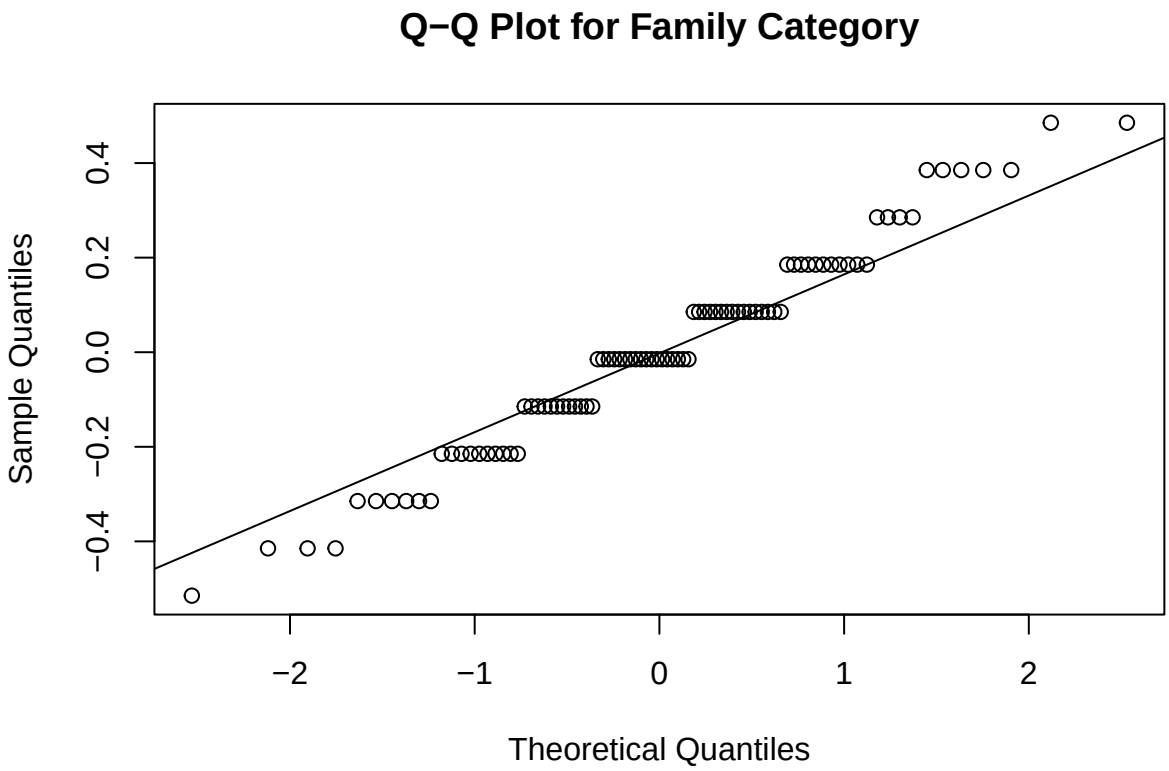
Task 8

```
## FAMILY   GAME  TOOLS
##     88     41     35
```

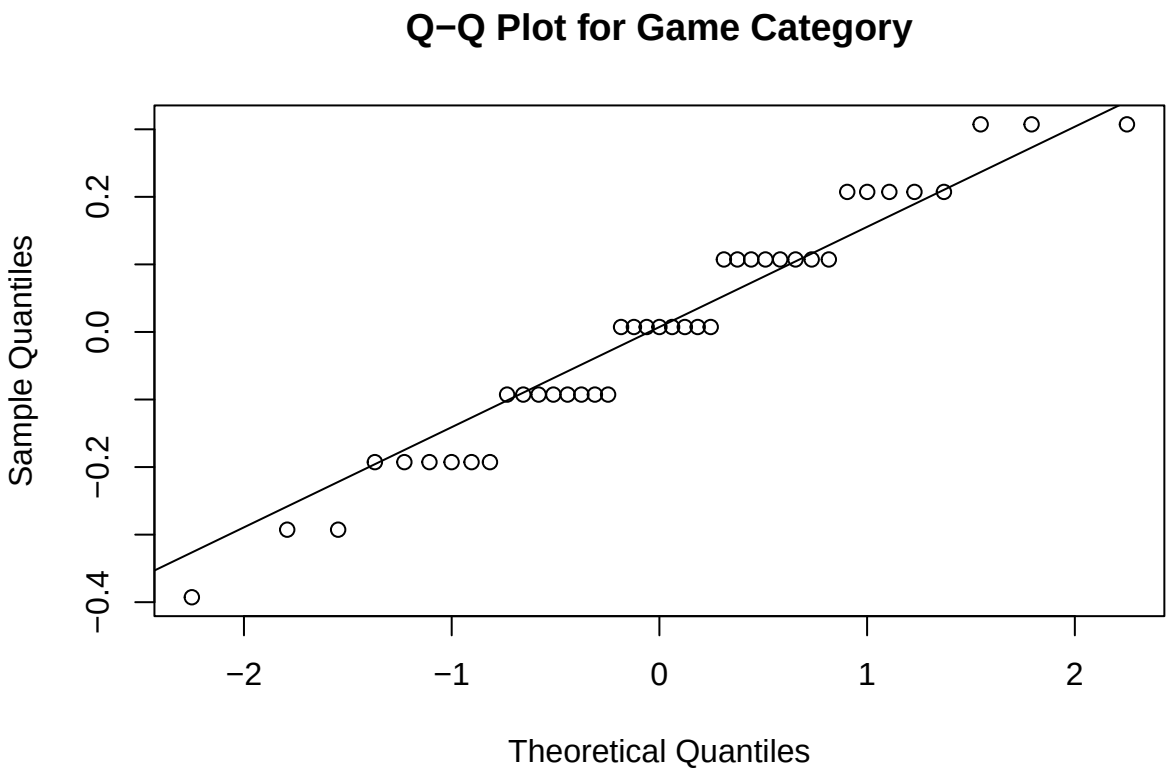
There are three groups and their values are bigeer than 30, thus, it means it is large and we can perform analysis by groups.

Creating a new data set with residuals and splitting the Category Data.

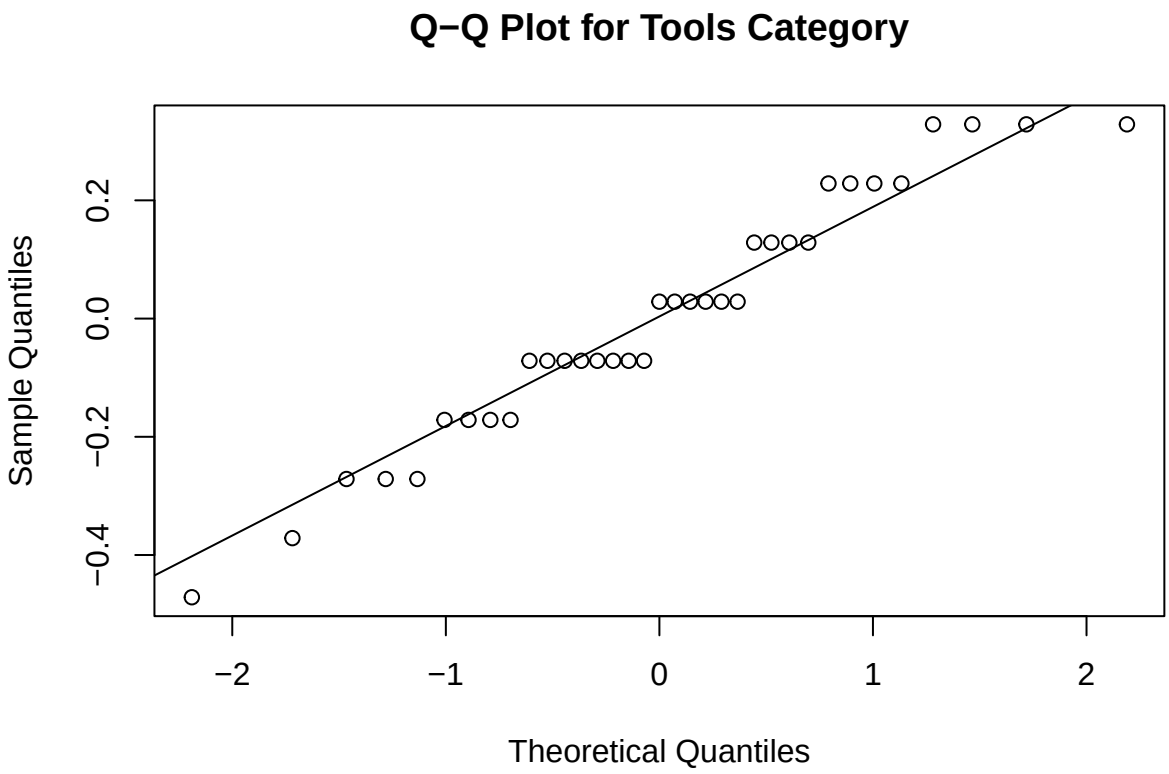
Normality Check



It is not normal because we can see that there are issues all over the line.



It is not normal because we can see that there are issues all over the line.



It is not normal because we can see that there are issues all over the line.

```
##
##  Shapiro-Wilk normality test
##
## data:  mydata3fam
## W = 0.97686, p-value = 0.1166
```

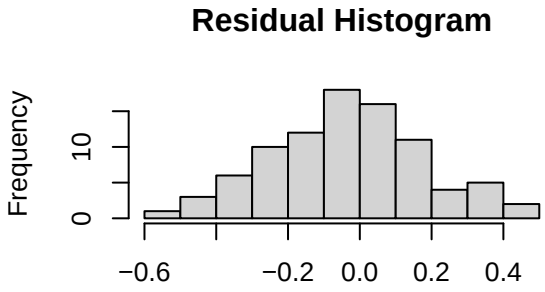
It is normal because p-value(0.1166) is big compare to significance level of 0.03

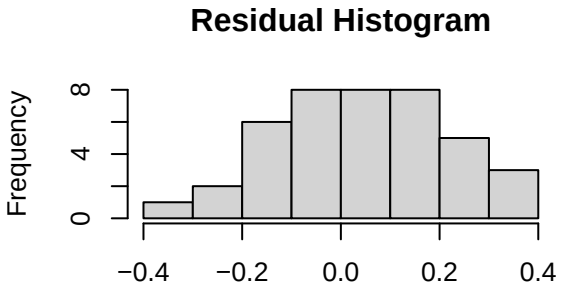
```
##
##  Shapiro-Wilk normality test
##
## data:  mydata3game
## W = 0.96284, p-value = 0.1976
```

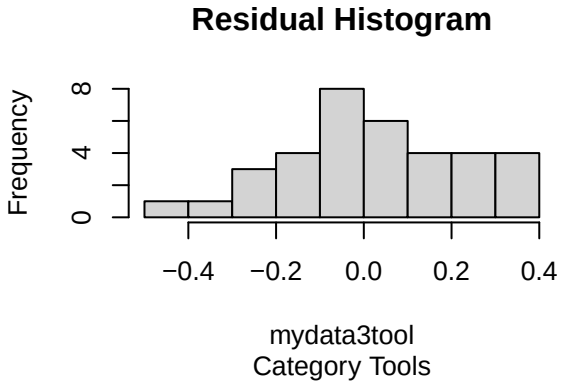
It is normal because p-value(0.1976) is big compare to significance level of 0.03

```
##
##  Shapiro-Wilk normality test
##
## data:  mydata3tool
## W = 0.96074, p-value = 0.2407
```

It is normal because p-value(0.2407) is big compare to significance level of 0.03







All of the three histograms are bell shaped. It is normal.

Equal Variance Check

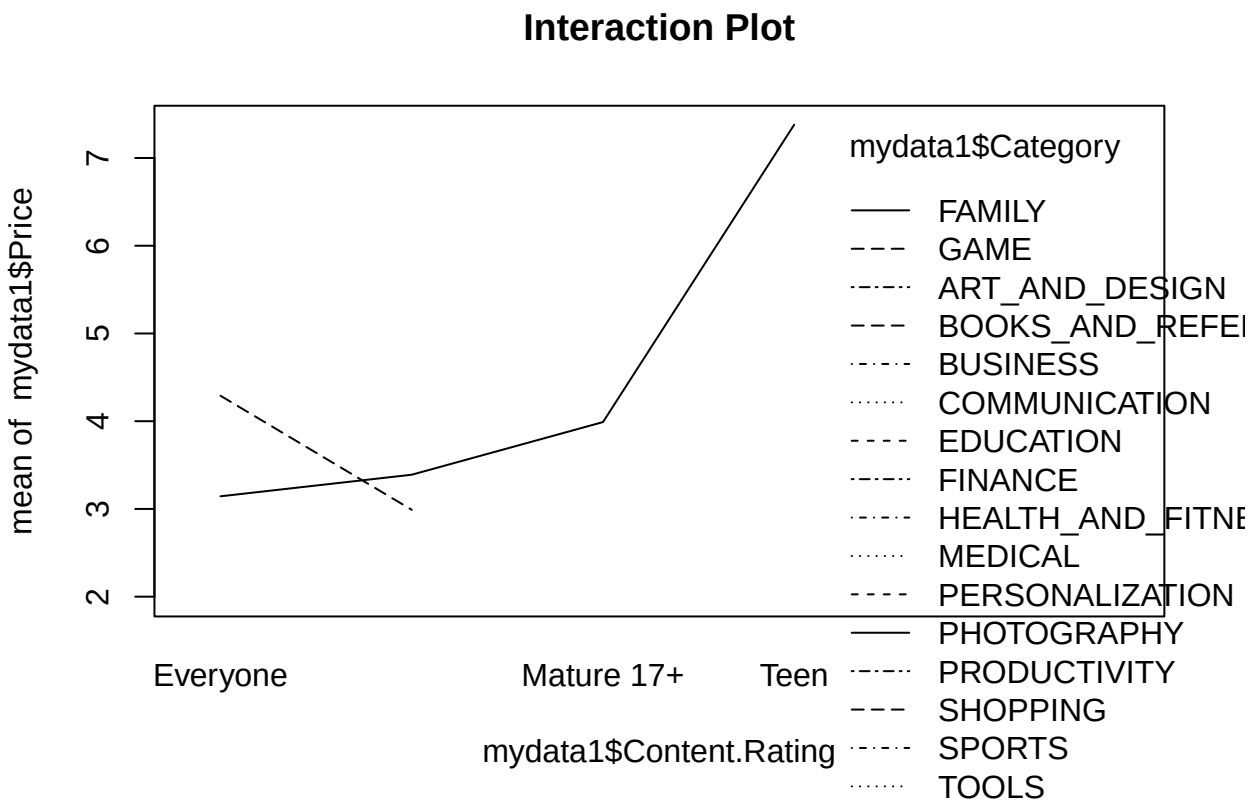
```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   2  0.7488 0.4746
##       161
```

The p-value is 0.4746 which is bigger than significance level of 0.03. Thus, do not reject, there is evidence for equal variance.

H0: muF = muG = muT vs H1: at least one of the means is different. p-value is 0.5493, it is bigger than significance level of 0.03. Thus, do not reject H0, there is not enough evidence that at least one of the means is different.

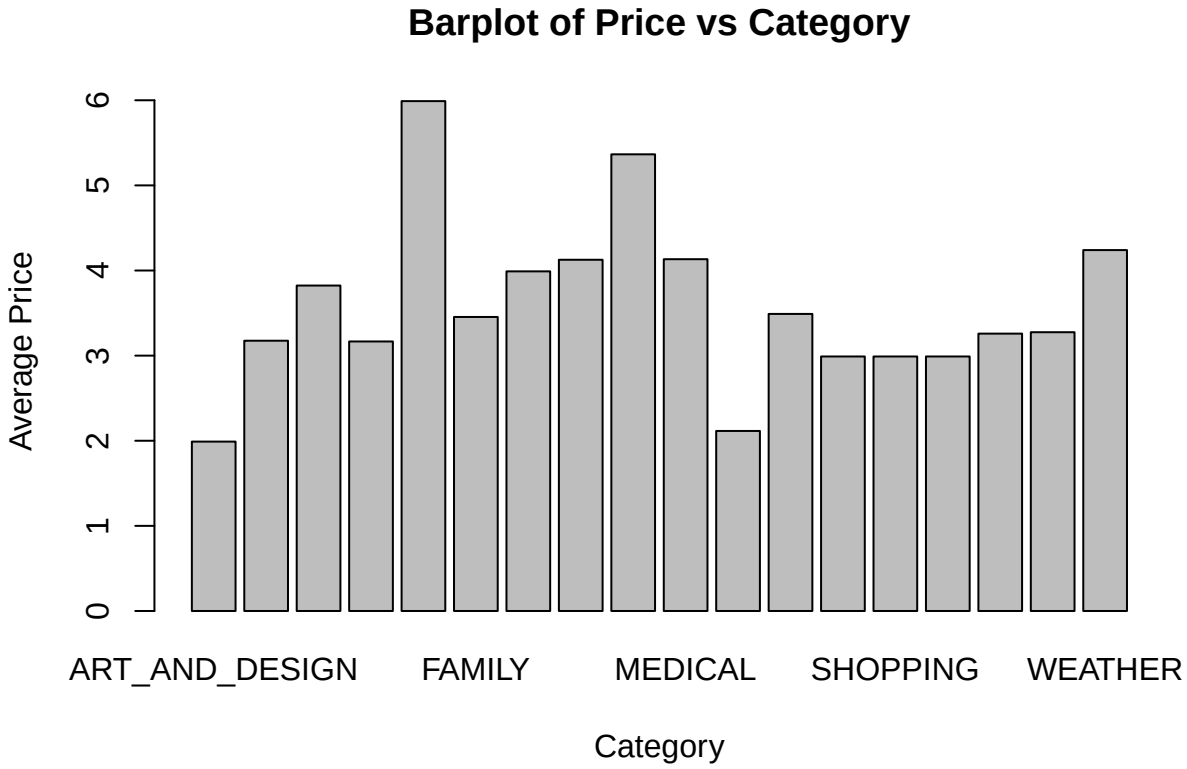
Task 9

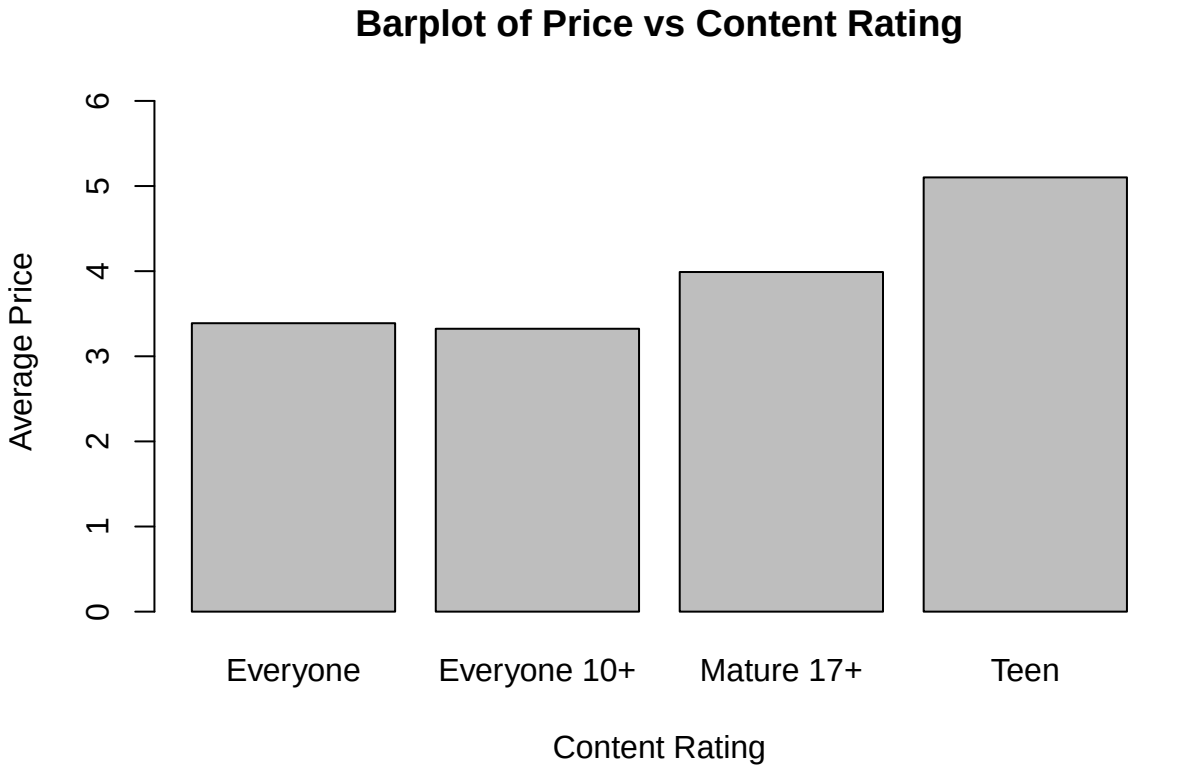
Interaction Plot



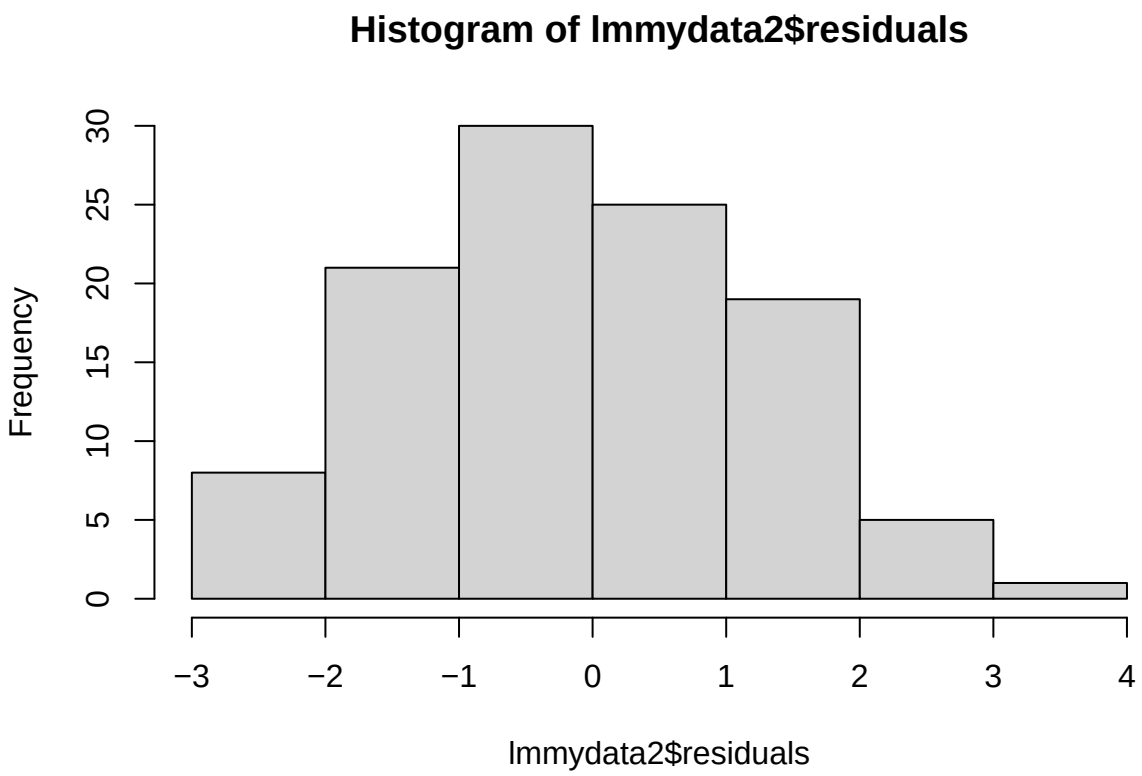
Based on the plot, there is probabaly is significant interaction.

Bar Graphs

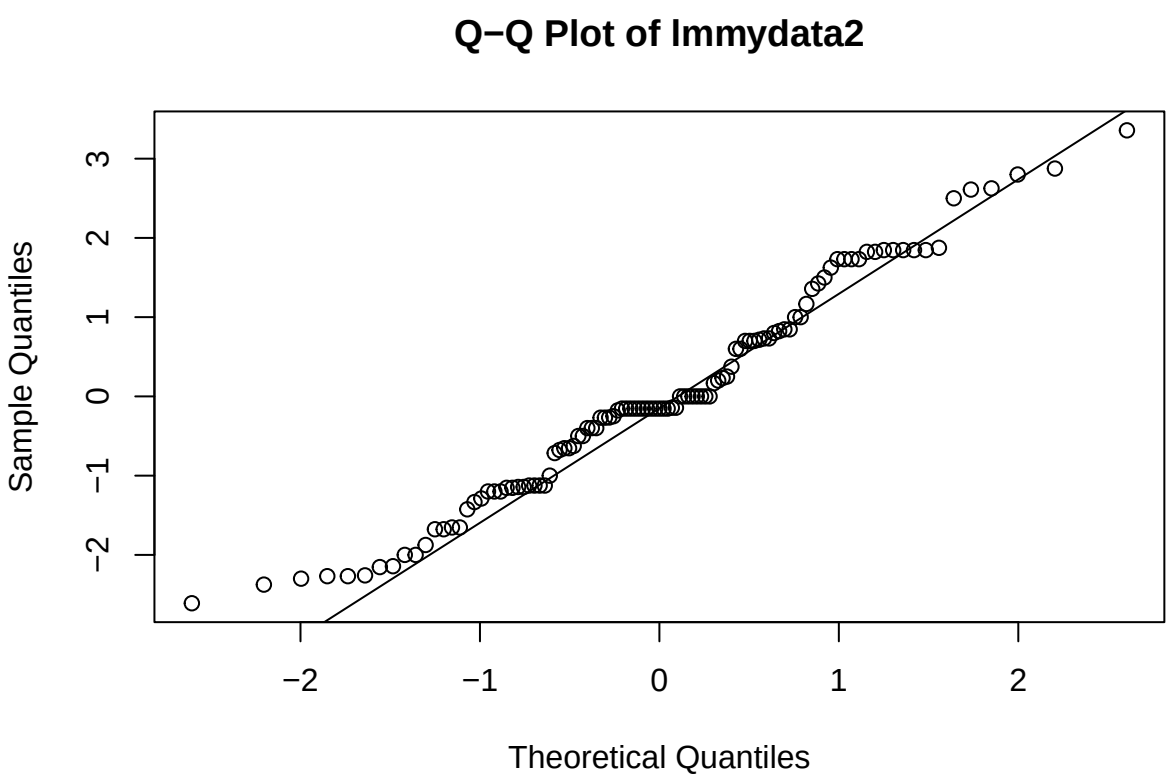

Barplot of Price vs Category
Average Price
0
1
2
3
4
5
6
ART_AND_DESIGN
FAMILY
MEDICAL
SHOPPING
WEATHER
Category


Barplot of Price vs Content Rating
Average Price
0
1
2
3
4
5
6
Everyone
Everyone 10+
Mature 17+
Teen
Content Rating

Normality Check



The histogram looks pretty Bell Shaped. It is normal.

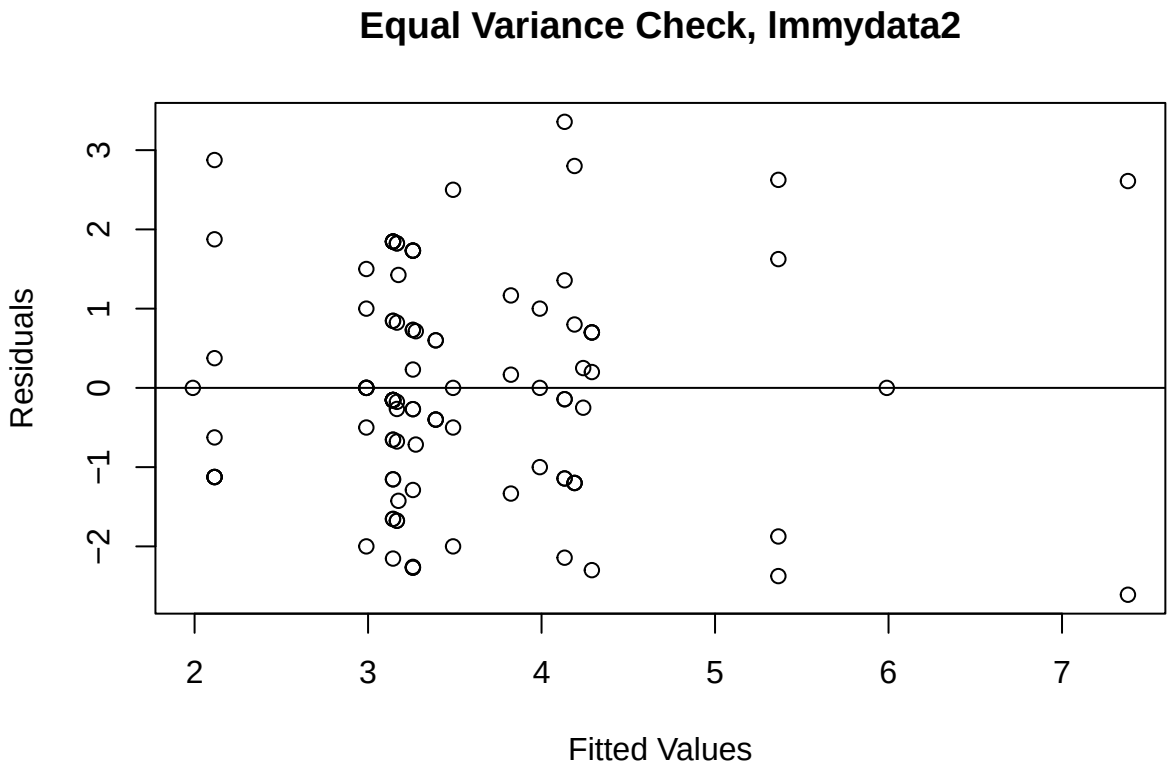


The plot is not normal because there are some issues over the whole line.

```
##
##  Shapiro-Wilk normality test
##
## data:  lmmydata2$residuals
## W = 0.97584, p-value = 0.04457
```

The p-value is 0.04457, it is bigger than significance level of 0.025, thus, it is normal.

Equal Variance Check



Based on the plot, it does not look like there are equal variance because the dots spread differently.

Check if Interactions Significant

```
## Analysis of Variance Table
##
## Response: mydata1$Price
##                                          Df  Sum Sq Mean Sq F value  Pr(>F)
## mydata1$Content.Rating                    3  19.564  6.5214  2.8698 0.04107 *
## mydata1$Category                         17  45.940  2.7024  1.1892 0.29059
## mydata1$Content.Rating:mydata1$Category  2  20.178 10.0892  4.4399 0.01462 *
## Residuals                                86 195.427  2.2724
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

H0: (alphaBeta)ij = 0 vs H1: at least one (alphaBeta)ij != 0. The p-value for interaction is 0.01462, it is small compare to significance level of 0.025, thus, reject H0, there is evidence that interaction is significant and that it is present. We should stop analysis here.