PS12

Dilip Nikhil Francies

2023-12-07

```
getwd()
```

```
## [1] "C:/Users/dilip/OneDrive - Indiana University/Stats/PS12"
```

Question 1: Brother vs Sister

Lets read the data as vectors:

```
sister <- c(69, 64, 65, 63, 65, 62, 65, 64, 66, 59, 62)
brother <- c(71, 68, 66, 67, 70, 71, 70, 73, 72, 65, 66)
```

Question 1a: Sample Coeffecient of Determination : Rsquared

```
lin_model <- lm(brother ~ sister)
rSquared <- summary(lin_model)$r.squared
rSquared
```

```
## [1] 0.3114251
```

Question 1b:

alpha = 0.05

Lets find the p-value for the hypothesis that the slope in the linear regression model is 0, and the alternate hypothesis that it is not.

```
summary(lin_model)
```

```
##
## Call:
## lm(formula = brother ~ sister)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5909 -1.2273 -0.9545  1.1136  4.0000
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  31.1818    18.7584   1.662   0.1308
## sister        0.5909     0.2929   2.018   0.0744 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.379 on 9 degrees of freedom
## Multiple R-squared:  0.3114, Adjusted R-squared:  0.2349
```

```
## F-statistic:  4.07 on 1 and 9 DF,  p-value: 0.07442
```

With the p-value of around 0.07442, we can safely conclude that the data is compatible with the null hypothesis that knowing sisters height is not enough to predict brothers height.

Question 1c:

```
# Confidence interval for the slope
conf_interval <- confint(lin_model, level = 0.90)
conf_interval
```

```
##                      5 %      95 %
## (Intercept) -3.20446954 65.568106
## sister       0.05401643  1.127802
```

The 90% confidence interval for the slope is 0.054 to 1.127.

Question 2: Anxiety and Exams

Read the given data:

```
exam_anxiety <- read.csv("examanxiety.txt",sep = "\t")
head(exam_anxiety,2)
```

```
##   Code Revise Exam Anxiety Gender
## 1    1      4   40  86.298   Male
## 2    2     11   65  88.716 Female
```

Split the data:

```
anxiety_male <- exam_anxiety$Anxiety[exam_anxiety$Gender =="Male"]
anxiety_feamale <- exam_anxiety$Anxiety[exam_anxiety$Gender =="Female"]
```

Question 2a: Lets perform two sample test with the following hypothesis:

Null : There is no mean difference in anxiety levels between males and females ie. muFemales = muMales
Alternate hypotheses: There is difference betweeen anxiety levels between males and females ie muFemales != muMales

```
result <- t.test(anxiety_male,anxiety_feamale,var.equal =FALSE)
print(result)
```

```
##
##  Welch Two Sample t-test
##
## data:  anxiety_male and anxiety_feamale
## t = -0.32961, df = 100.41, p-value = 0.7424
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.147444  5.110827
## sample estimates:
## mean of x mean of y
##  74.38373  75.40204
```

From the above results, its clear that there is no difference as the p-value obtained is not tiny.

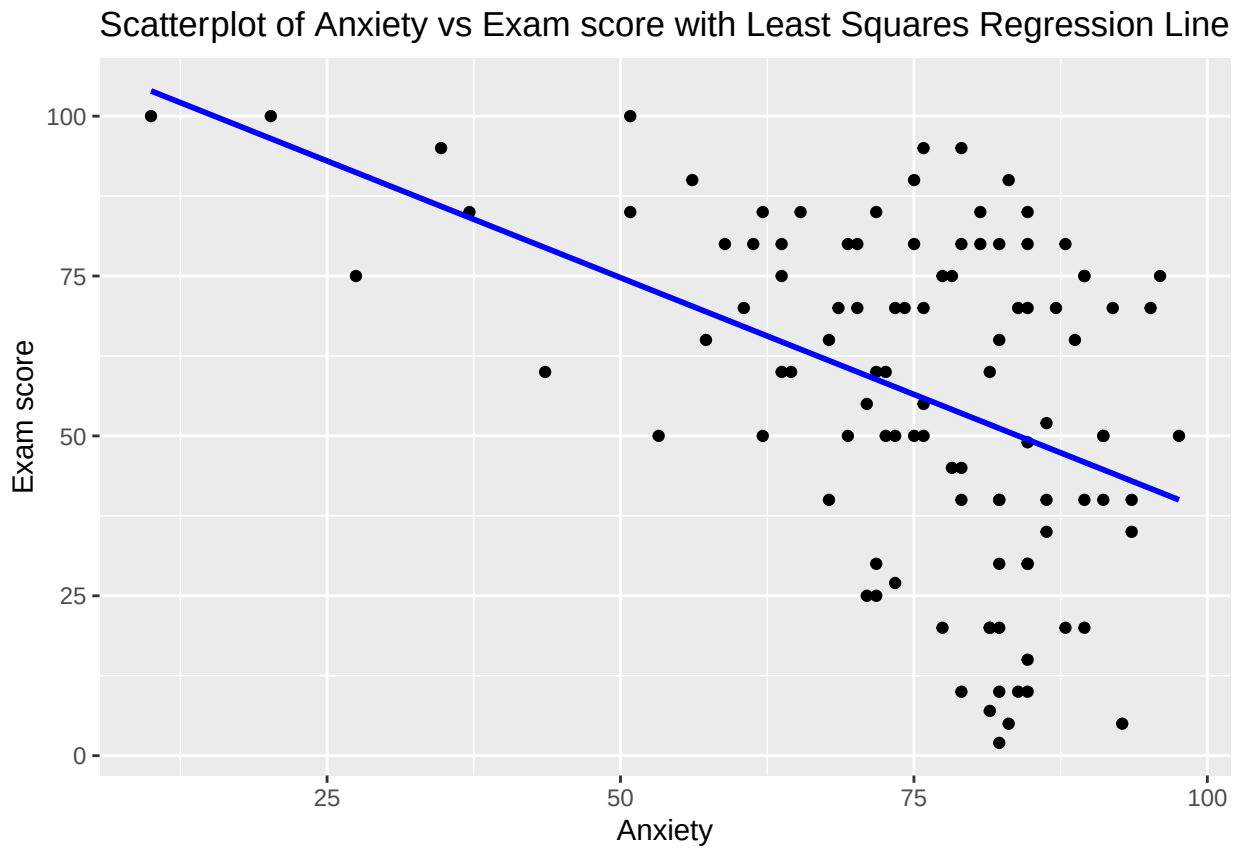
Question 2b: Lets draw the scatter plot with the regression line:

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```r
scatterplot <- ggplot(exam_anxiety, aes(x = Anxiety, y = Exam)) +
  geom_point() +  # Scatterplot
  geom_smooth(method = "lm", se = FALSE, color = "blue") +  # Add regression line
  labs(x = "Anxiety", y = "Exam score", title = "Scatterplot of Anxiety vs Exam score with Least Squares

# Print the plot
print(scatterplot)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```r
model <- lm(Exam ~ Anxiety, data = exam_anxiety)
summary(model)
```

```
##
## Call:
## lm(formula = Exam ~ Anxiety, data = exam_anxiety)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.185 -16.046   1.166  19.856  41.461
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 111.2444    11.3498   9.801 2.46e-16 ***
## Anxiety      -0.7300     0.1484  -4.920 3.37e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.41 on 101 degrees of freedom
## Multiple R-squared:  0.1933, Adjusted R-squared:  0.1853
## F-statistic:  24.2 on 1 and 101 DF,  p-value: 3.374e-06
```

From the results and the plot,it looks like there is a downward trend in the data, When the anxiety levels are higher, the marks tend to be all over the place.The regression model does a horrible job in in capturing the variance in the data set.Linear Regression may not be our best model in predicting the exam score based on anxiety levels.

```
library(broom)
```

```
## Warning: package 'broom' was built under R version 4.2.3
```

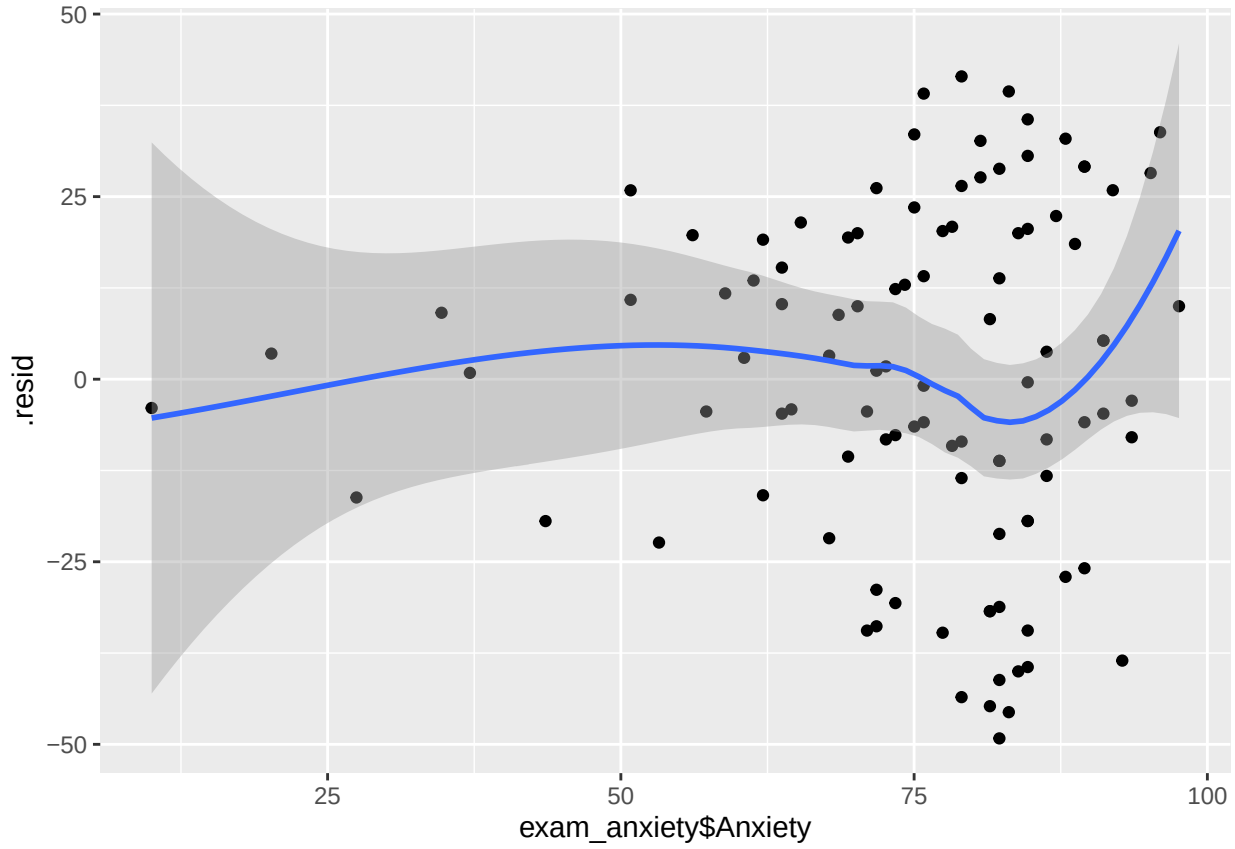
Question 2c:

i. Linearity:

Scatter plot:

```
my.lm.df <- augment(model)
ggplot(my.lm.df, aes(exam_anxiety$Anxiety, .resid)) + geom_point() + geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

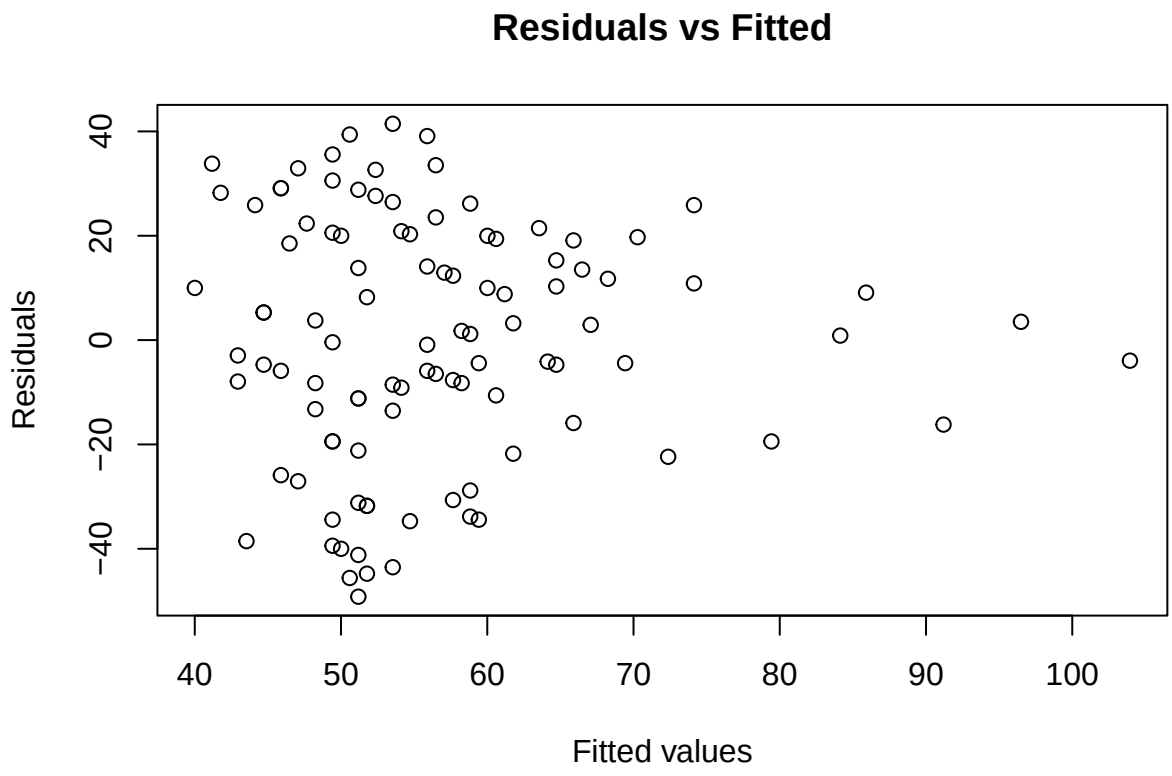


It is clear that there is no strong linear relationship between the two variables. There does seem to be a lot of outliers, and most of the data seems to be scattered on the right side of the plot with varying exam scores for a small range of anxiety levels.

ii.Independence.

Let's plot a residuals vs. fitted values plot.

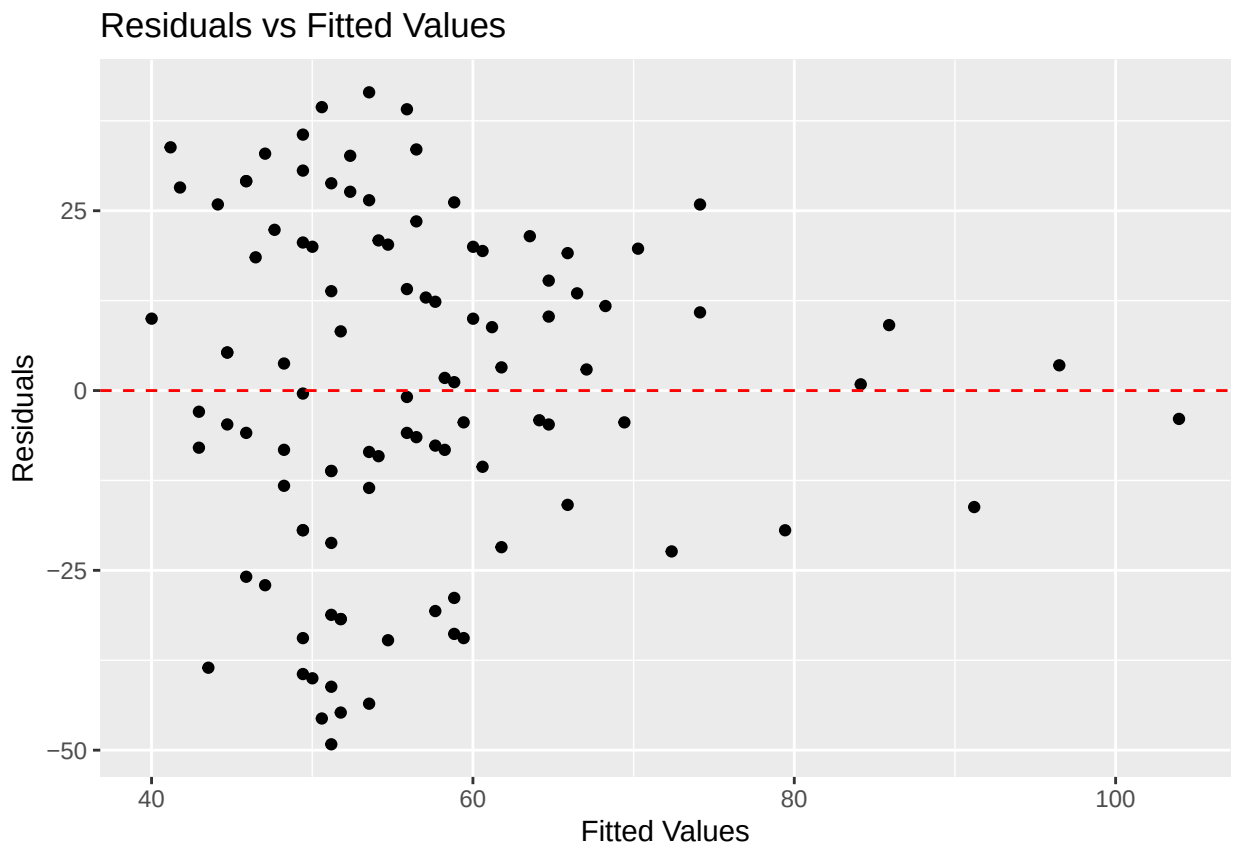
```
model <- lm(Exam ~ Anxiety, data = exam_anxiety)
fit <- fitted.values(model)
res <- residuals(model)
plot(fit, res, main = "Residuals vs Fitted", xlab = "Fitted values", ylab = "Residuals")
```



Because the exam scores recorded are for different students, ideally the data set should be independent as long as there were no multiple measurements taken for the same students.

iii. Homoskedasticity:

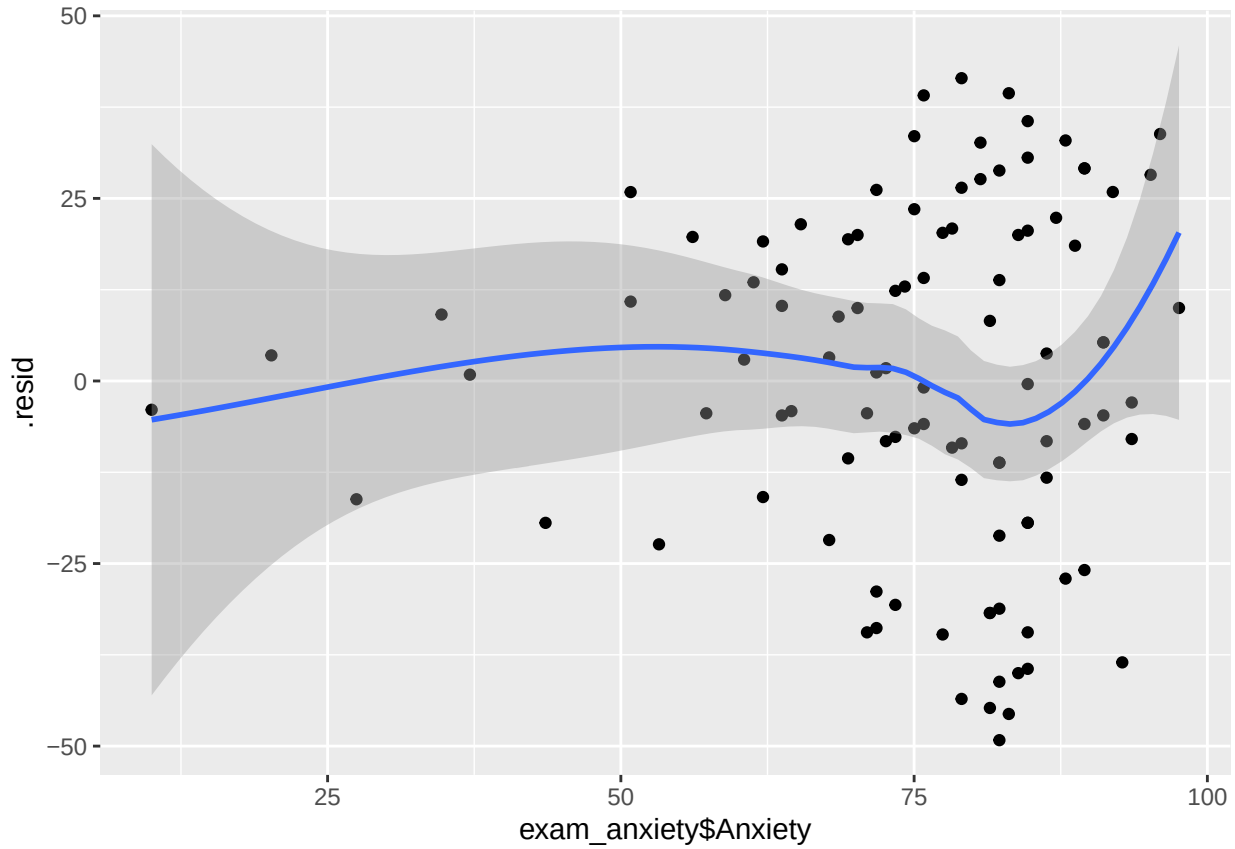
```
ggplot(model, aes(x = fitted(model), y = resid(model))) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Residuals vs Fitted Values",
       x = "Fitted Values",
       y = "Residuals")
```



Let's plot absolute residual plot:

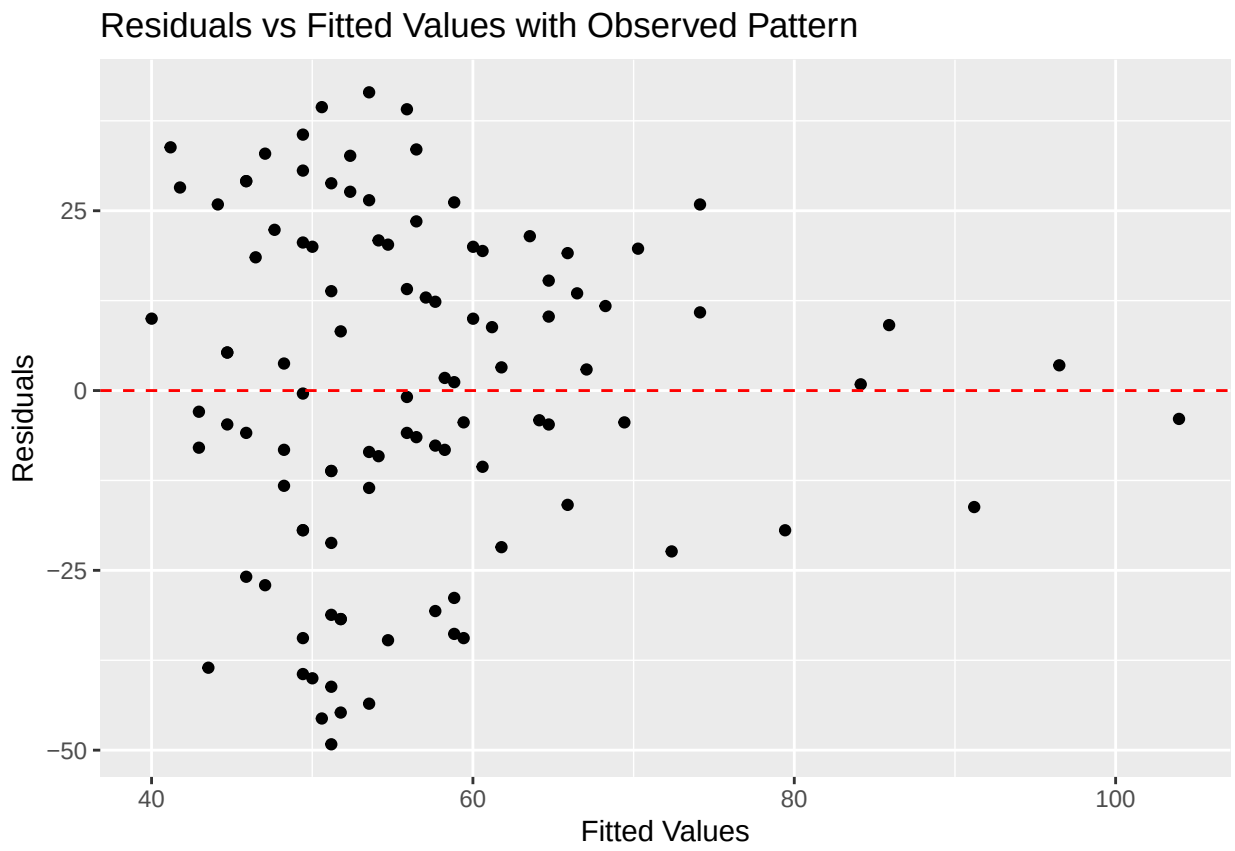
```
aug_model_df <- augment(model)
ggplot(aug_model_df, aes(exam_anxiety$Anxiety, .resid)) + geom_point() + geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



The residuals get smaller as we go right, and hence there is no constant spread. Homoskedasticity check fails.

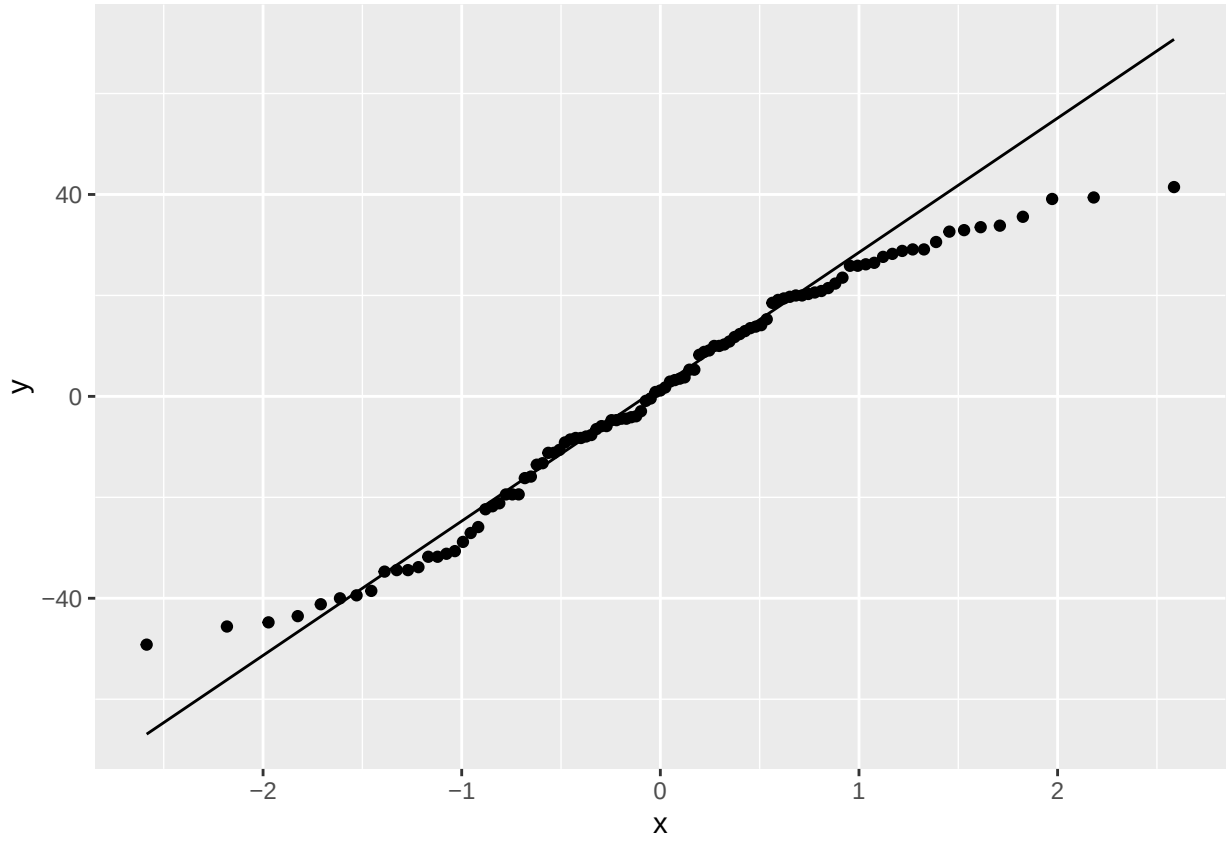
```
residuals_df <- data.frame(
  Fitted_Values = fitted(model),
  Residuals = residuals(model)
)

library(ggplot2)
ggplot(residuals_df, aes(x = Fitted_Values, y = Residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Residuals vs Fitted Values with Observed Pattern",
       x = "Fitted Values",
       y = "Residuals")
```



Normality of errors :

lets feed the residual to a normal QQ plot:

```
ggplot(aug_model_df, aes(sample = .resid)) + stat_qq() + stat_qq_line()
```



It definitely is not straight.At the tails it deviates significantly from the qqline. Hence, one could conclude that the residuals are barely normal.

Question 3:

Question 3a: Assumption of ANOVA

1. Observations are independent: Yes, Because the rats were randomly divided into four groups and were put in a cage, to a certain extent one could say this was a randomized control experiment. So the observations are independent.

2.All the populations are normal : From the given QQ plots, except at the tails, the observations does look like they fall on the straight line. But perhaps, if we had more sample we could have been sure of the normality. From the 35 observations for each group we have, we can assume that the population is normal.

3. Homoscedasticity:

Apart from the fruit diet sample that has a standard deviation of 16.9, other sample's standard deviation are very close to each other. To a certain extent, one can conclude that the stds of all four samples (16.9, 14.6,14.2,14.1) are kinda close to each other. hence Homoskedasticity checks out.

```
N <- 140
n<- 35

fruitMean <- 83.5
fruitSD <- 16.9
carbsMean <- 92.3
```

```
carbsSD <- 14.6
meatMean <- 88.6
meatSD <- 14.2
mixedMean <- 99.4
mixedSD <- 14.1


means <- c(fruitMean,carbsMean,meatMean,mixedMean)
SD <- c(fruitSD,carbsSD,meatSD,mixedSD)
grandMean <- mean(means)
SSB <- n * (fruitMean-grandMean)^2 + n * (carbsMean-grandMean)^2 + n * (meatMean-grandMean)^2 + n * (mi
betweenDF <- 4-1
between.meansquare <- SSB/betweenDF
SSB
```

Question 3b:

```
## [1] 4698.75
```

```
betweenDF
```

```
## [1] 3
```

```
between.meansquare
```

```
## [1] 1566.25
```

```
ssw <- (n-1) * fruitSD^2 + (n-1) * carbsSD^2 + (n-1) * meatSD^2 + (n-1) * mixedSD^2
dfw <- N-4 # 4 groups
within.meansqaure <- ssw/dfw
ssw
```

```
## [1] 30573.48
```

```
within.meansqaure
```

```
## [1] 224.805
```

```
dfw
```

```
## [1] 136
```

```
F <- between.meansquare/within.meansqaure
F
```

```
## [1] 6.967149
```

```
1 - pf(F, df1 = betweenDF, df2 = dfw)
```

```
## [1] 0.0002140835
```

Hence, SSB = 4698.75, BDF = 3, between mean square = 1566.25

SSW = 30573.48, within mean square = 224.805, WDF = 136

F = 6.967149

Total = 35272.23

Total DF = 136 + 3 = 139

p-value = 0.00021

Question 3c: There is infact strong evidence against the null hypothesis with a p-value of 0.00021.Yes, our sample size is small. But to a certain extent we can conclude that there is significant difference in the means of the group which means one of the four groups has a significant difference in the amount of weight gained. Do we know which group? we don't. But with further analysis, one could find out the answer.

Question 4: Empathy vs Game

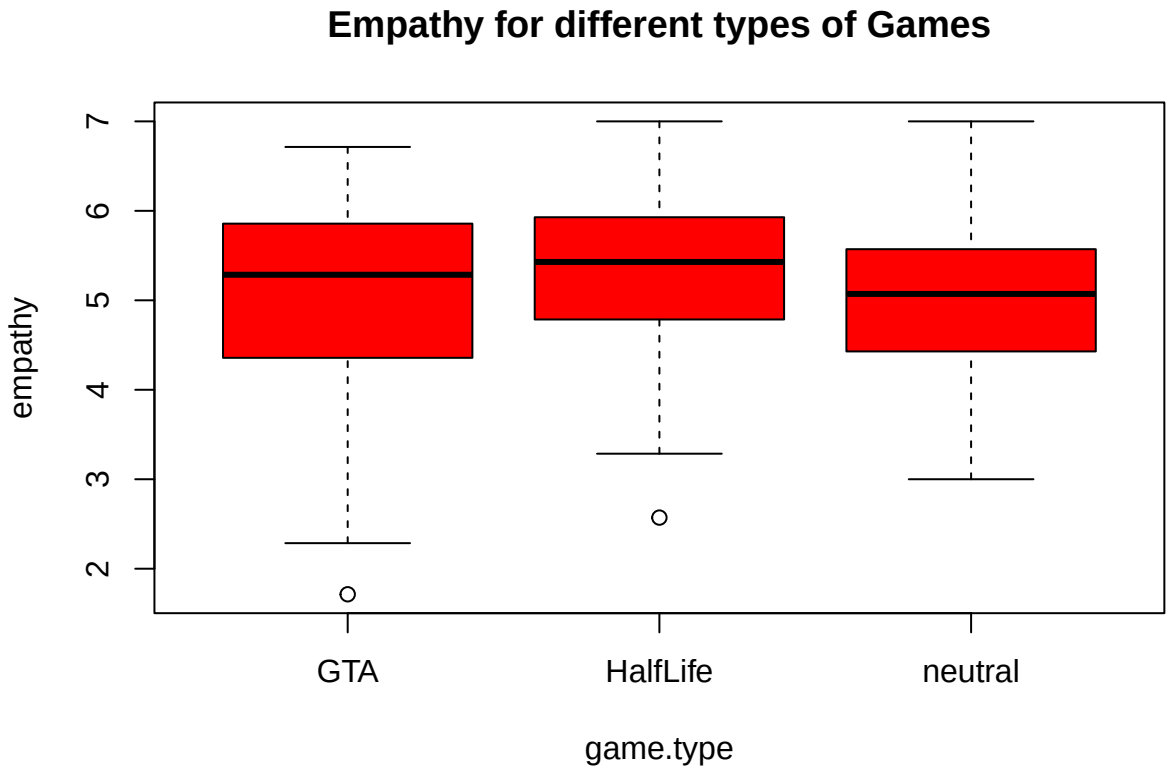
```
game_data <- read.table("GameEmpathy.txt", header = TRUE)
head(game_data)
```

Question 4a

```
##      sex game.type identify  empathy
## 1 female   neutral 3.333333 5.285714
## 2 female   neutral 1.833333 5.571429
## 3   male   neutral 1.000000 4.714286
## 4 female   neutral 1.000000 5.571429
## 5 female   neutral 3.333333 3.142857
## 6 female   neutral 1.000000 5.571429
```

Visualization:

```
boxplot(empathy ~ game.type, data = game_data, col = "red", main = "Empathy for different types of Games
```




```
GTA <- subset(game_data, game.type == "GTA")
HalfLife <- subset(game_data, game.type == "HalfLife")
neutral <- subset(game_data, game.type == "neutral")
```

```r
#sample empthy observations for each games
mean(GTA$empathy)
```

```
## [1] 5.029762
```

```r
mean(HalfLife$empathy)
```

```
## [1] 5.293939
```

```r
mean(neutral$empathy)
```

```
## [1] 5.05381
```

```r
NROW(GTA)
```

```
## [1] 48
```

```r
NROW(HalfLife)
```

```
## [1] 55
```

```r
NROW(neutral)
```

```
## [1] 50
```

ANOVA test:

```r
anova_model <- aov(empathy ~ game.type, data = game_data)
summary(anova_model)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## game.type     2   2.25   1.125   1.092  0.338
## Residuals   150 154.47   1.030
```

```r
F <- 1.092
bdf <- 2
wdf <- 150
1 - pf(F, df1 = bdf, df2 = wdf)
```

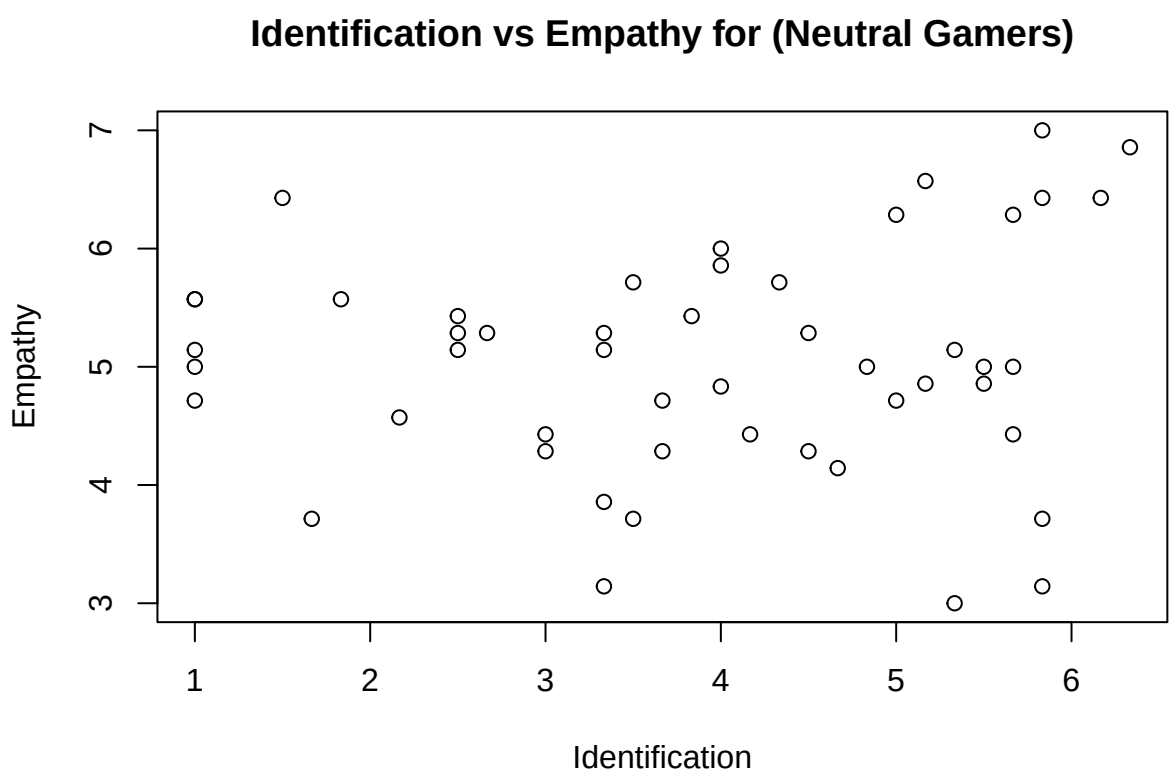
```
## [1] 0.338197
```

Because we dont have a tiny p-value (0.05), we can conclude that indeed the null hypotheses is true, there is no significant difference in the means between the three samples.

Question 4b: To understand the relationship between identification and empathy among gamers who played different games, lets find the correlation and visualize scatterplots.

i. Student who played neutral games:

```r
plot(neutral$identify, neutral$empathy,
     xlab = "Identification", ylab = "Empathy",
     main = "Identification vs Empathy for (Neutral Gamers)")
```



Correlation :

```
# Calculate correlation coefficient
corr_neutral <- cor(neutral$identify, neutral$empathy)
corr_neutral
```

```
## [1] 0.08991878
```

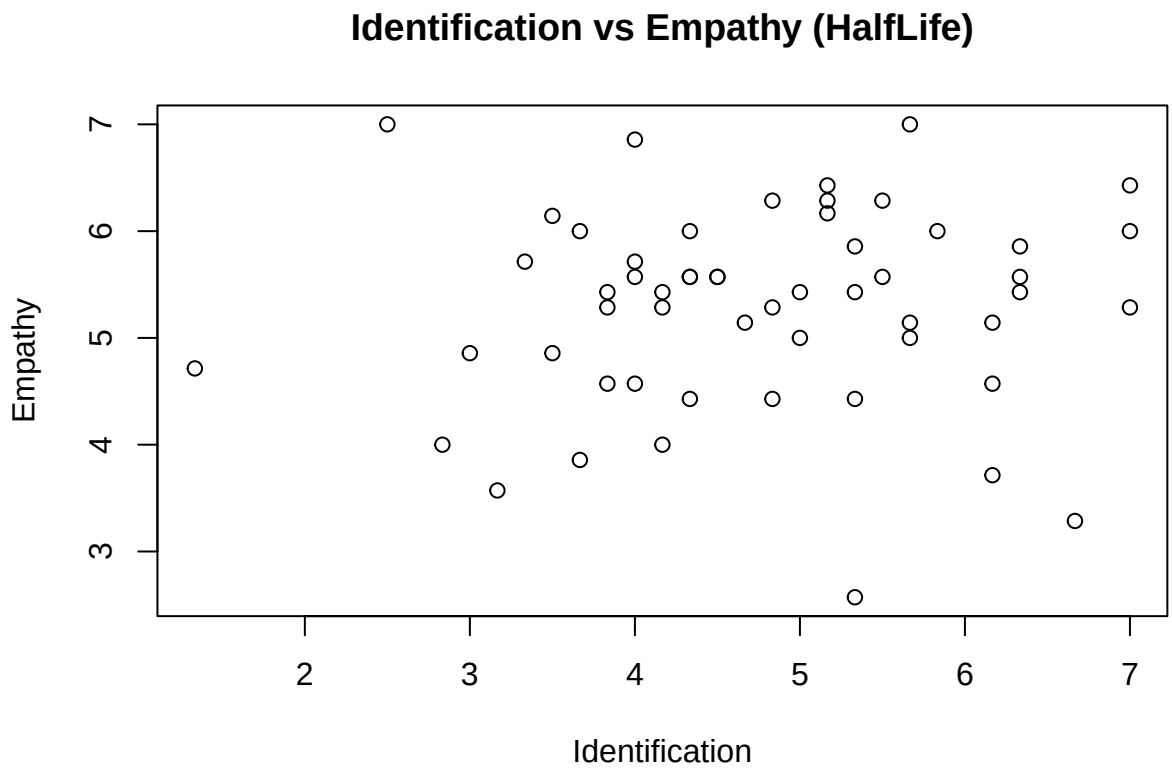
```
#calculate the p-value
cor_test_neutral <- cor.test(neutral$identify, neutral$empathy, method = "pearson")
p_value_neutral <- cor_test_neutral$p.value
p_value_neutral
```

```
## [1] 0.5345996
```

A small positive correlation between identification and empathy of 0.089 is observed in neutral players

ii. Half Life:

```
plot(HalfLife$identify, HalfLife$empathy,
     xlab = "Identification", ylab = "Empathy",
     main = "Identification vs Empathy (HalfLife)")
```




```
corr_HalfLife<- cor(HalfLife$identify, HalfLife$empathy, method = "pearson")
corr_HalfLife
```

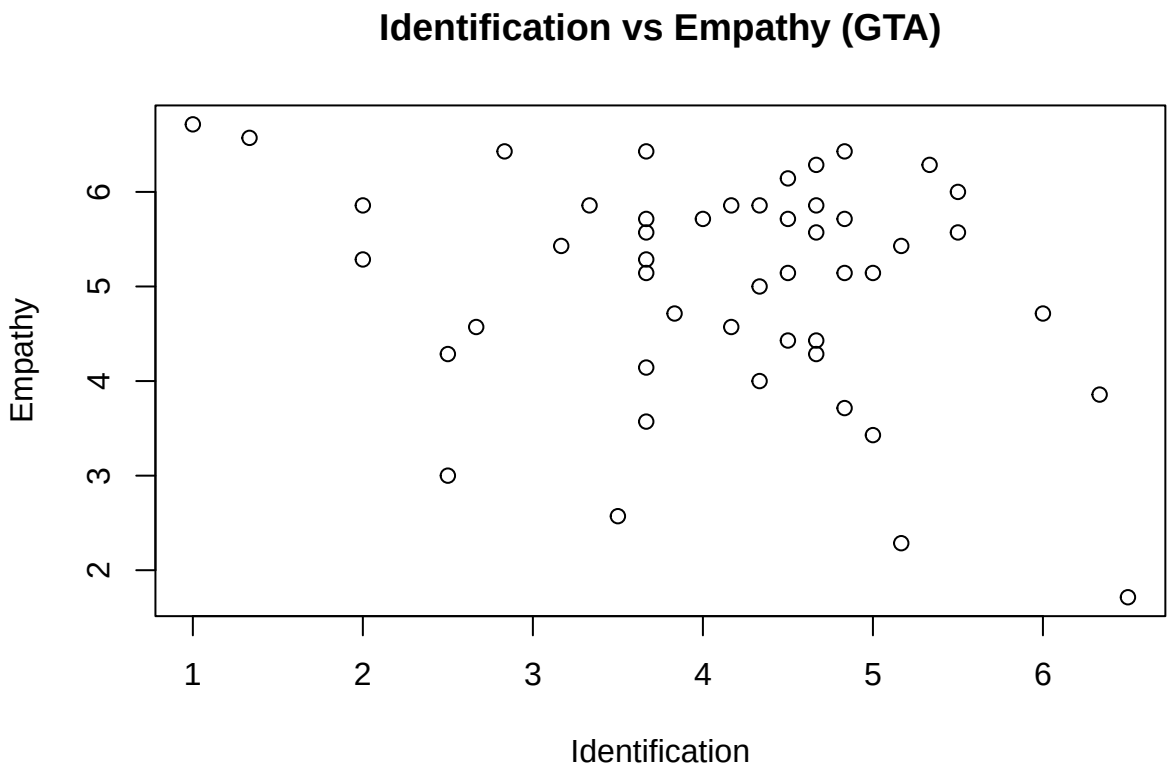
```
## [1] 0.07164441
```

```
#calculate the p-value
cor_test_HalfLife <- cor.test(HalfLife$identify, HalfLife$empathy, method = "pearson")
p_value_HalfLife <- cor_test_HalfLife$p.value
p_value_HalfLife
```

```
## [1] 0.6032072
```

iii. GTA:

```
plot(GTA$identify, GTA$empathy,
     xlab = "Identification", ylab = "Empathy",
     main = "Identification vs Empathy (GTA)")
```



Correlation:

```
# Calculate correlation coefficient
corr_GTA <- cor(GTA$identify, GTA$empathy)
corr_GTA
```

```
## [1] -0.2722745
```

```
#calculate the p-value
cor_test_GTA <- cor.test(GTA$identify, GTA$empathy, method = "pearson")
p_value_GTA <- cor_test_GTA$p.value
p_value_GTA
```

```
## [1] 0.06118009
```

Adjusting the p-values:

```
#  adjust for multiple testing using bonferroni method
p_values <- c(p_value_neutral, p_value_HalfLife, p_value_GTA)
p_value_adjust <- p.adjust(p_values, method = "bonferroni")
p_value_adjust
```

```
## [1] 1.0000000 1.0000000 0.1835403
```

Well, with the adjusted p-values using bonferroni method, we can clearly see that these are not tiny. With the assumed significance level of 0.05/3, we can say the data is consistent with the null hypotheses that there is no significant difference between identity and empathy among students that play gta, halflife or neutral games. Are we 100% sure? The sample size we have is about 50 for each class which in my opinion is too less to draw conclusion on the population. But perhaps, with more data we could be 100% sure.