Assignment 1 - Modern Regression Analysis

Balaji Padmanathan - 22202290

2022-10-30

Section 1: Exploratory Data Analysis

Data Loading

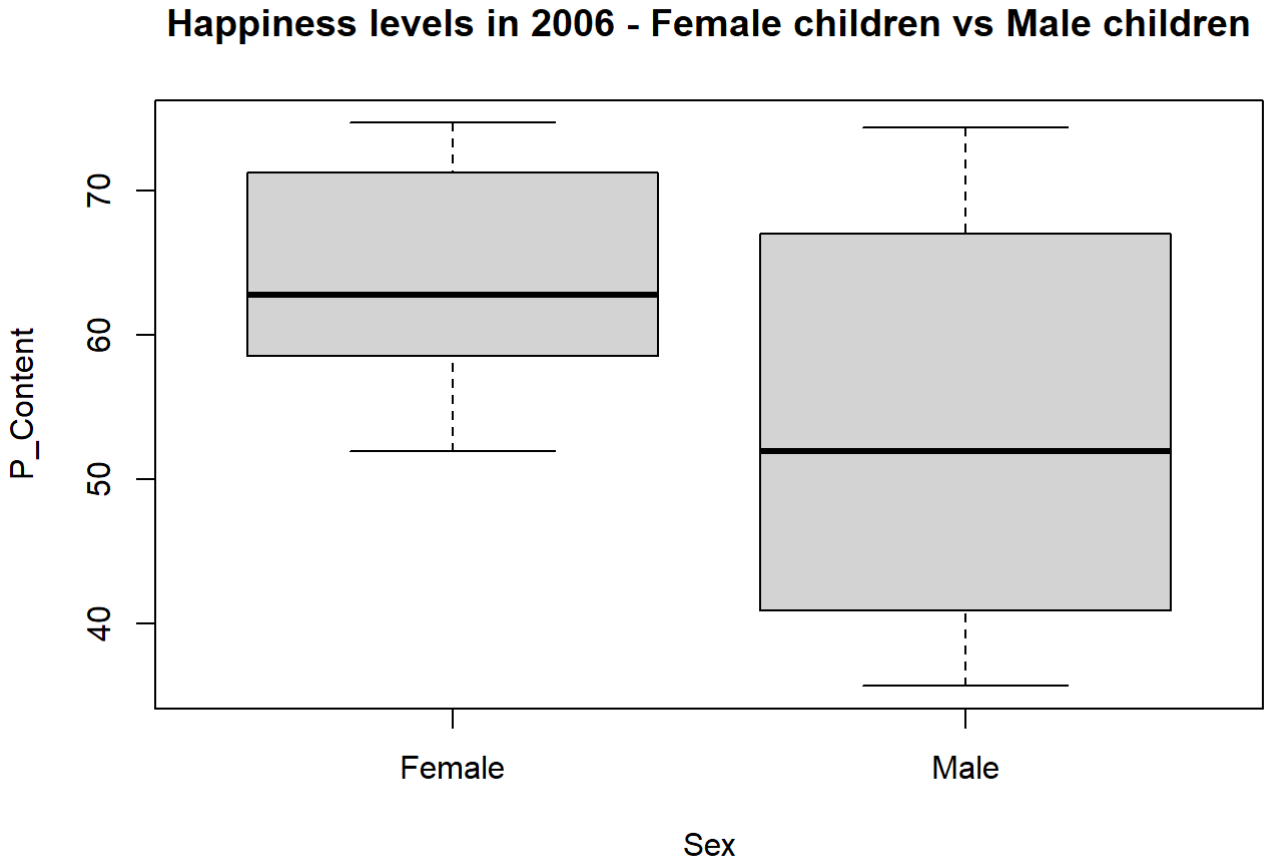
Loading the Irish children's life satisfaction data which contains information on the percentage of children that reported being happy along with their age, gender and the year of survey. After loading, using head() to check if the data is loaded correctly.

```
life_satisfaction <- read.csv(file.choose())
head(life_satisfaction)
```

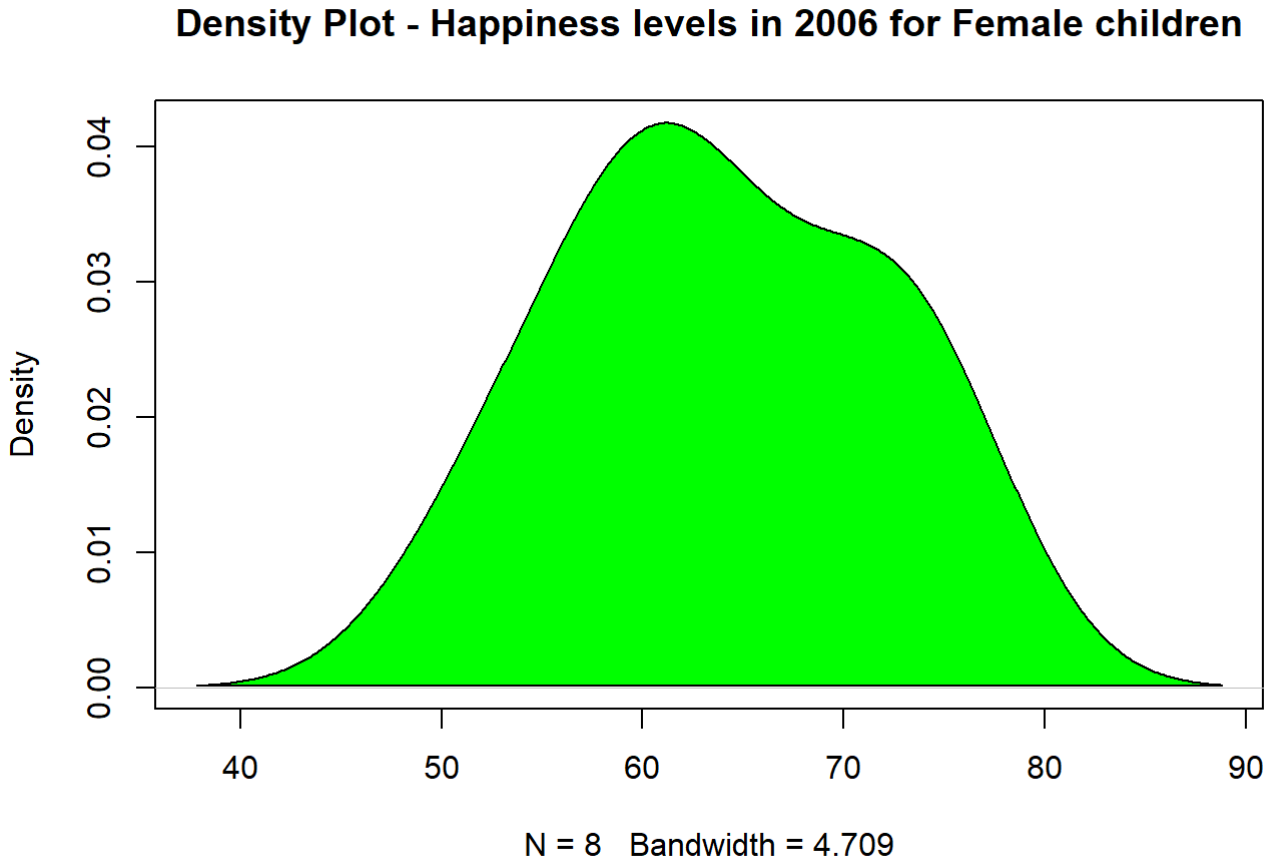
```
##   Year                Age    Sex P_Content
## 1 2006 10 years and under   Male     67.72
## 2 2006 10 years and under Female     69.39
## 3 2010 10 years and under   Male     78.41
## 4 2010 10 years and under Female     71.77
## 5 2014 10 years and under   Male     76.70
## 6 2014 10 years and under Female     77.64
```

Q1) Using a boxplot, the density and the descriptive statistics (mean, min, max, median, and quantiles), describe the distributions and the difference in the distributions for the percentage of school aged children that reported being happy with the way they are in 2006 with respect to sex (i.e. female vs male).

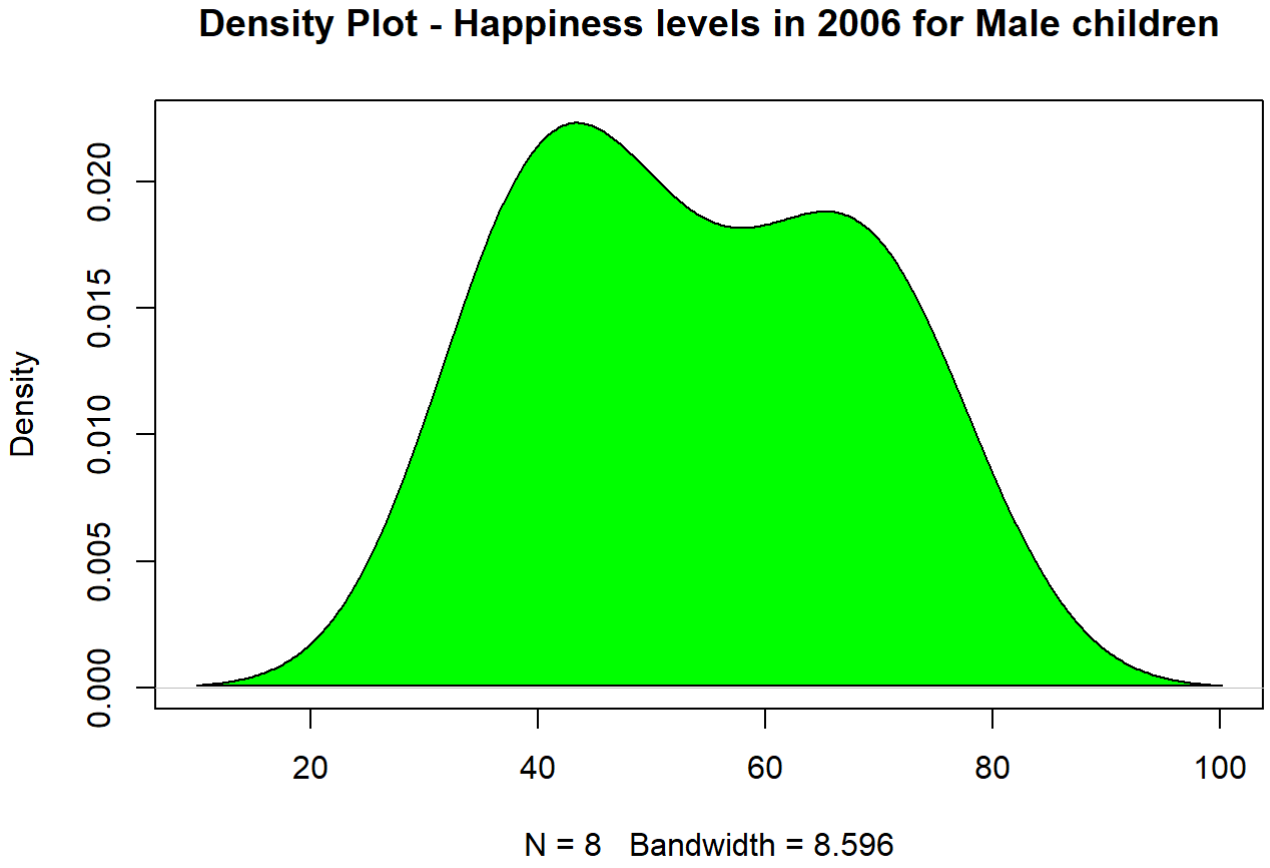
```
library(moments)

boxplot(P_Content[Year == 2006] ~ Sex[Year == 2006], data = life_satisfaction, main = 'Happin
ess levels in 2006 - Female children vs Male children', xlab = 'Sex', ylab = 'P_Content')
```




```
plot(density(life_satisfaction[life_satisfaction$Year == 2006 & life_satisfaction$Sex == 'Fem
ale',]$P_Content), main = 'Density Plot - Happiness levels in 2006 for Female children')
polygon(density(life_satisfaction[life_satisfaction$Year == 2006 & life_satisfaction$Sex ==
'Female',]$P_Content), col="green")
```




```
plot(density(life_satisfaction[life_satisfaction$Year == 2006 & life_satisfaction$Sex == 'Mal
e',]$P_Content), main = 'Density Plot - Happiness levels in 2006 for Male children')
polygon(density(life_satisfaction[life_satisfaction$Year == 2006 & life_satisfaction$Sex ==
'Male',]$P_Content), col="green")
```




```
summary(life_satisfaction[life_satisfaction$Year == 2006 & life_satisfaction$Sex == 'Female',
"P_Content"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   51.93   59.19   62.80   63.98   70.33   74.69
```

```
summary(life_satisfaction[life_satisfaction$Year == 2006 & life_satisfaction$Sex == 'Male',
"P_Content"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   35.71   41.44   51.93   53.72   66.62   74.37
```

```
quantile(life_satisfaction[life_satisfaction$Year == 2006 & life_satisfaction$Sex == 'Female'
, "P_Content"])
```

```
##      0%     25%     50%     75%    100%
## 51.9300 59.1925 62.7950 70.3325 74.6900
```

```
quantile(life_satisfaction[life_satisfaction$Year == 2006 & life_satisfaction$Sex == 'Male',
"P_Content"])
```

```
##      0%     25%     50%     75%    100%
## 35.7100 41.4425 51.9300 66.6175 74.3700
```

```
skewness(life_satisfaction[life_satisfaction$Year == 2006 & life_satisfaction$Sex == 'Female'
, "P_Content"])
```

```
## [1] 0.007478656
```

```
skewness(life_satisfaction[life_satisfaction$Year == 2006 & life_satisfaction$Sex == 'Male',
"P_Content"])
```

```
## [1] 0.1633307
```

```
IQR(life_satisfaction[life_satisfaction$Year == 2006 & life_satisfaction$Sex == 'Female', "P_
Content"])
```

```
## [1] 11.14
```

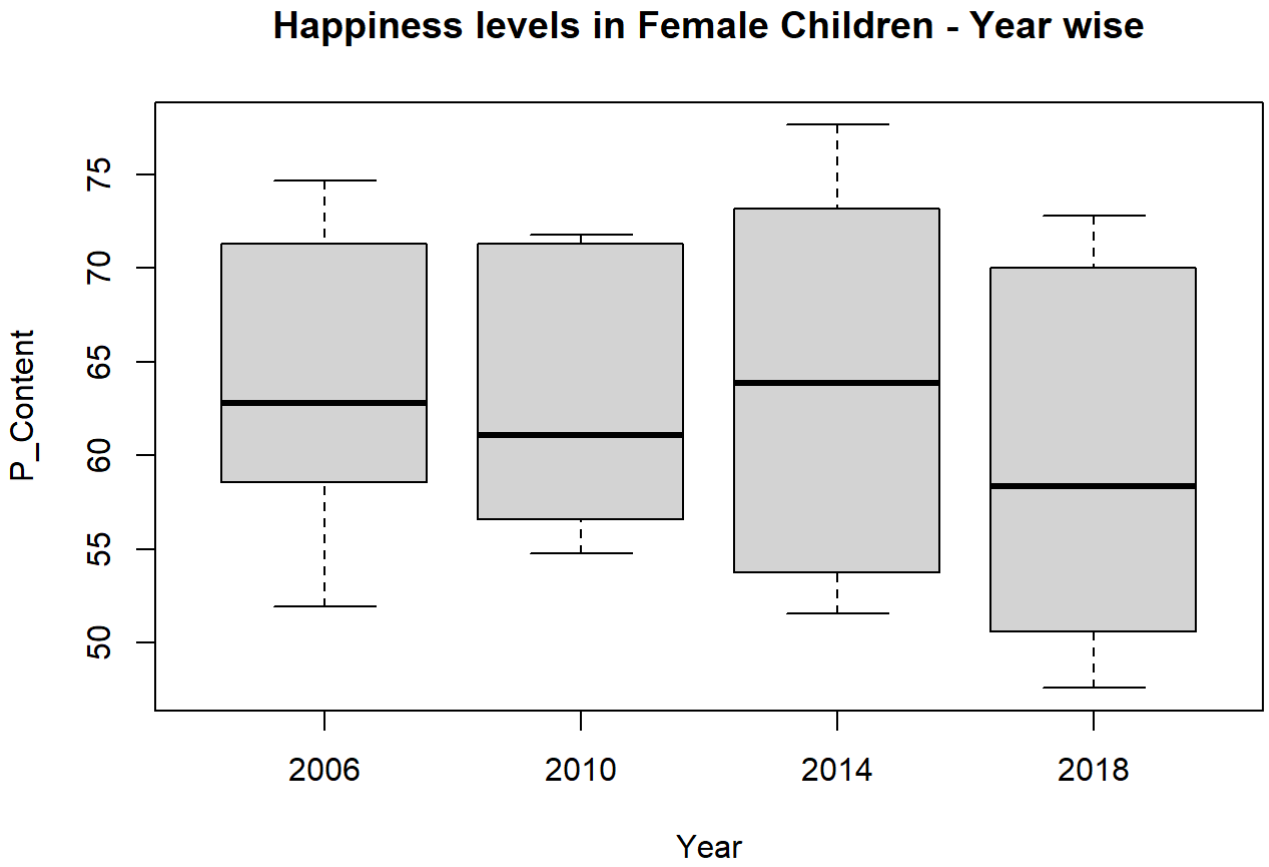
```
IQR(life_satisfaction[life_satisfaction$Year == 2006 & life_satisfaction$Sex == 'Male', "P_Co
ntent"])
```

```
## [1] 25.175
```

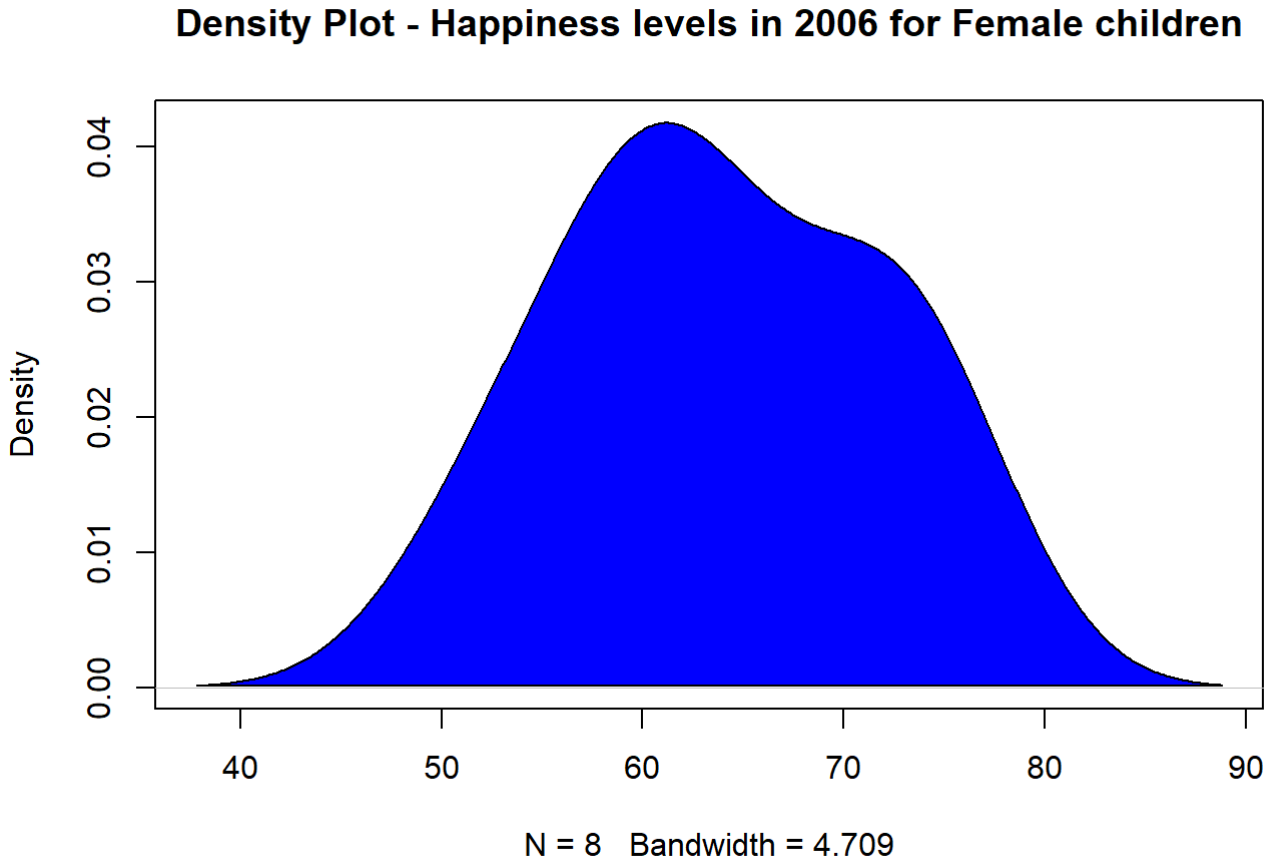
Based on the above plots and descriptive statistics, we can infer the following about the happiness levels of school children in 2006 with respect to sex:
i) On average, Female children are happier than Male children (Both Mean and Median is higher for Female children)
ii) The highest happiness percentage (among all children) was recorded for Female children (74.69%) while the lowest (among all children) was recorded for Male children (35.71%)
iii) The density plots for both Female and Male children are positive/right skewed as we can see from the plots and also by the fact that the Mean is greater than the Median for both. This is further confirmed by measuring their skewness
iv) The density plot of Male children is more skewed than Females
v) The Happiness levels of Male children is more spread out than the Female children as can be seen from the boxplot. We can further confirm this by looking at the IQR (25.175 for Male children vs 11.14 for Female children)

Q2) Using a boxplot, the density and the descriptive statistics (mean, min, max, median, and quantiles), describe the difference in the distributions for the percentage of female school aged children that reported being happy with the way they are with respect to year (2006; 2010; 2014; 2018)

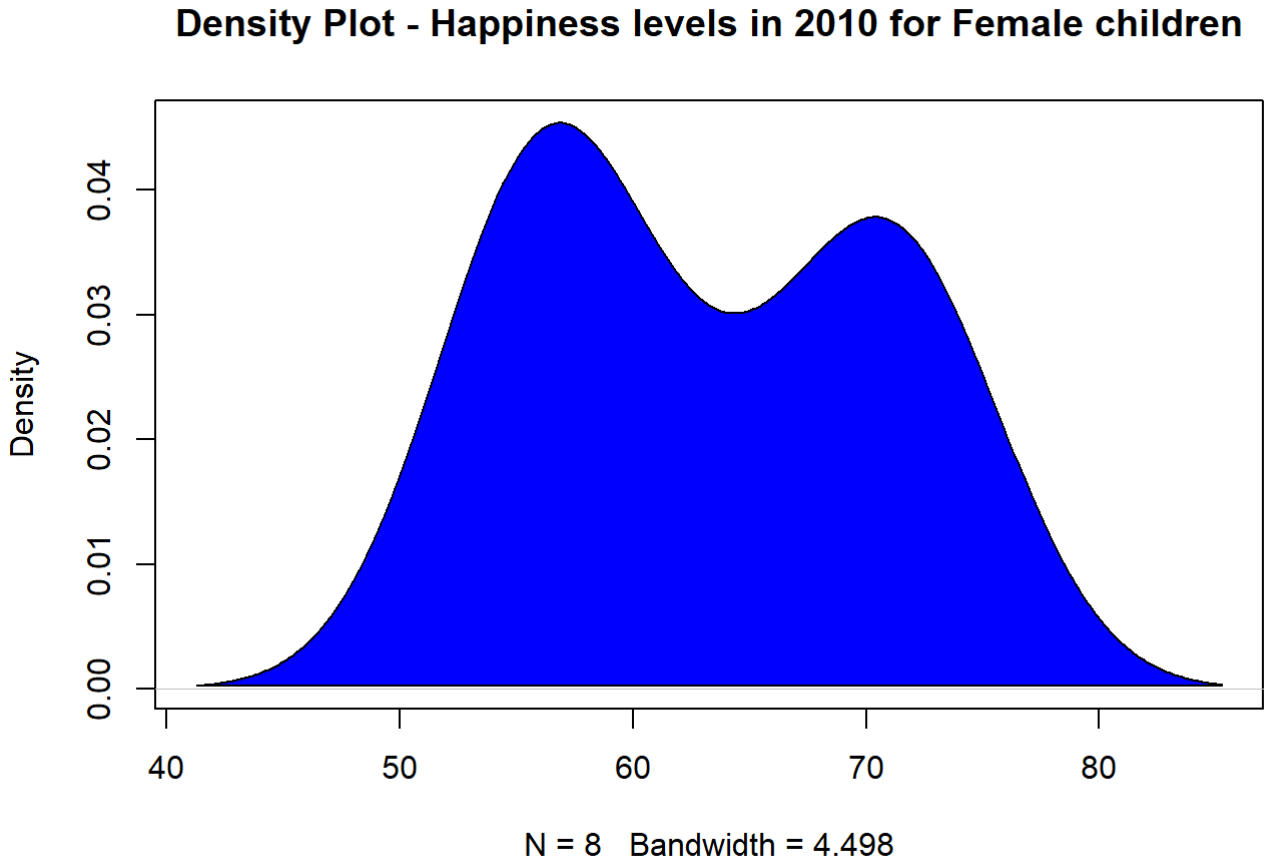
```
boxplot(P_Content[Sex == 'Female'] ~ Year[Sex == 'Female'], data = life_satisfaction, main =
'Happiness levels in Female Children - Year wise', xlab = 'Year', ylab = 'P_Content')
```



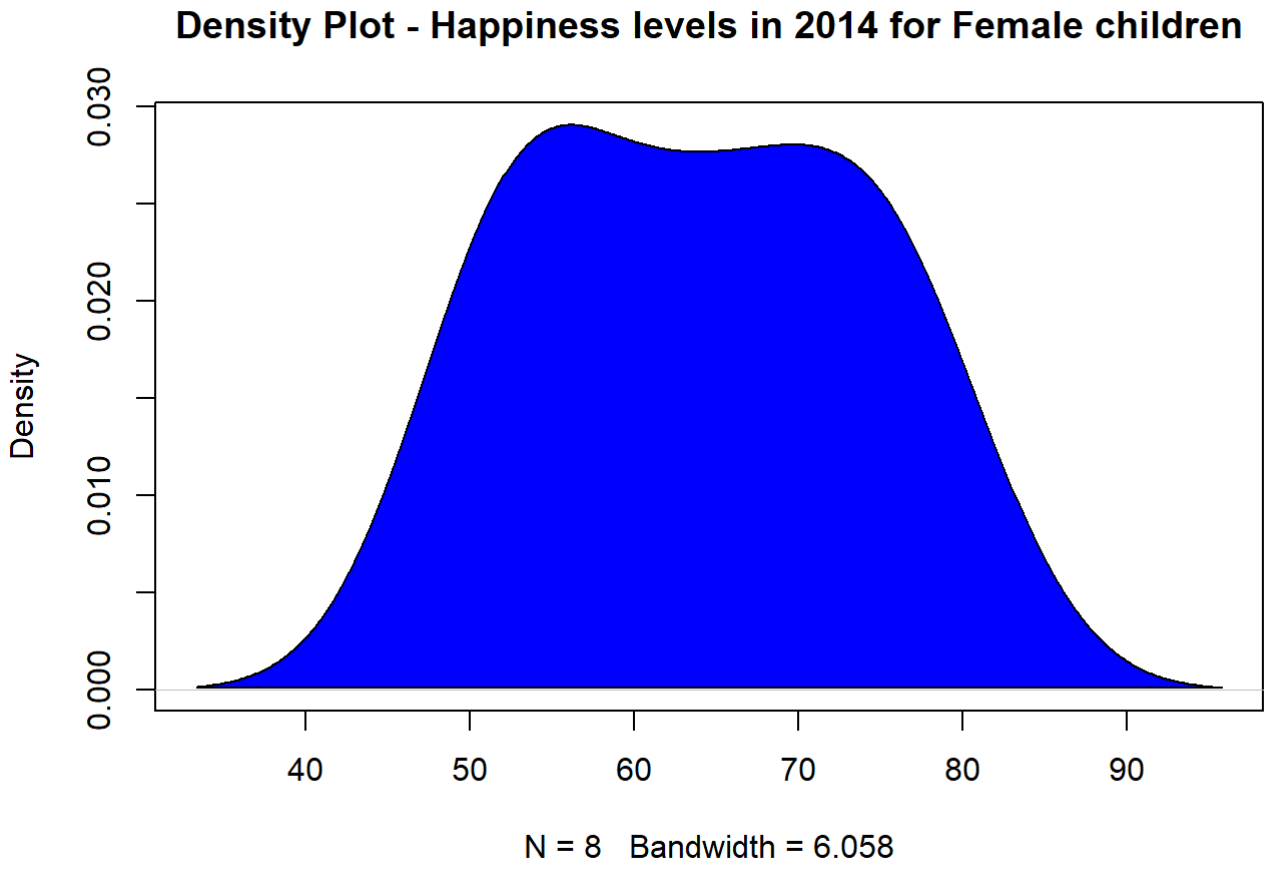

```
plot(density(life_satisfaction[life_satisfaction$Year == 2006 & life_satisfaction$Sex == 'Fem
ale',]$P_Content), main = 'Density Plot - Happiness levels in 2006 for Female children')
polygon(density(life_satisfaction[life_satisfaction$Year == 2006 & life_satisfaction$Sex ==
'Female',]$P_Content), col="blue")
```



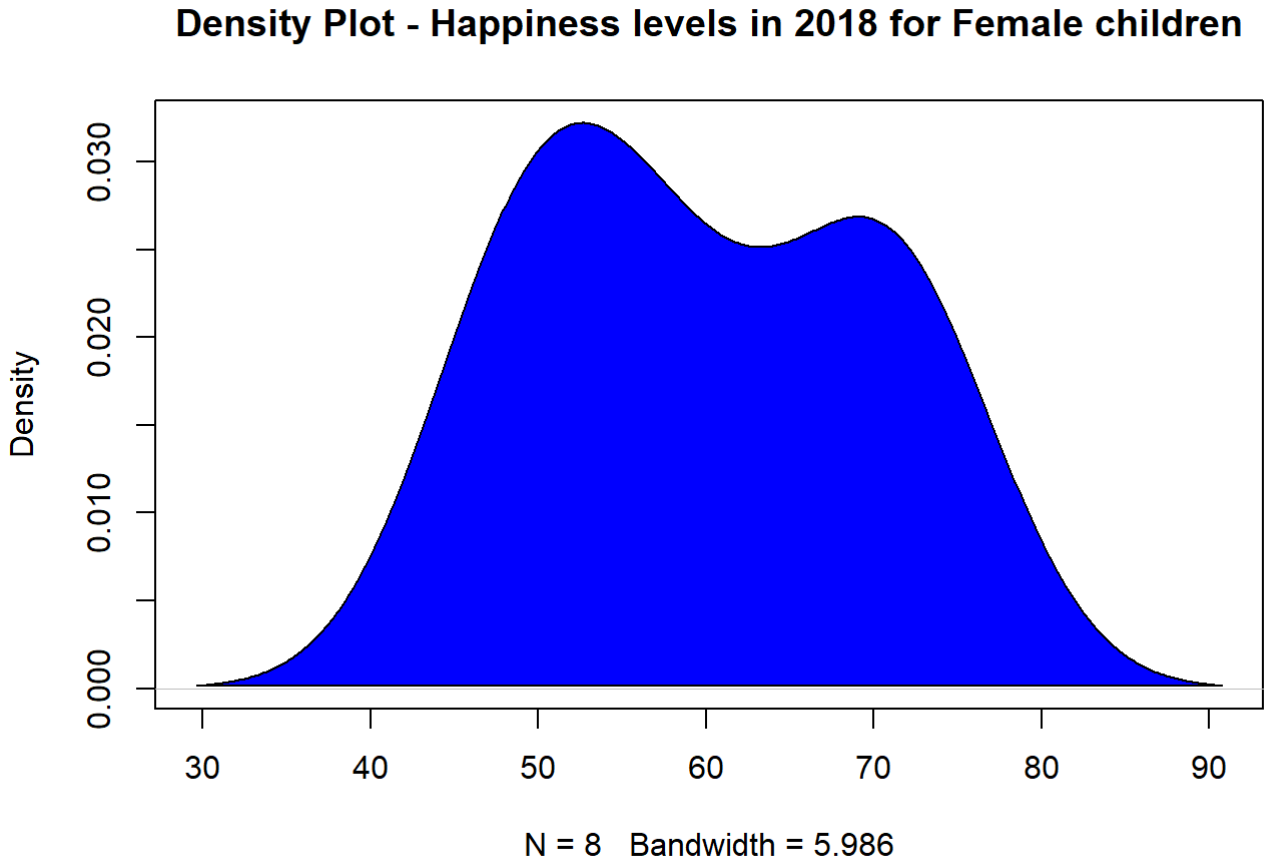

```
plot(density(life_satisfaction[life_satisfaction$Year == 2010 & life_satisfaction$Sex == 'Fem
ale',]$P_Content), main = 'Density Plot - Happiness levels in 2010 for Female children')
polygon(density(life_satisfaction[life_satisfaction$Year == 2010 & life_satisfaction$Sex ==
'Female',]$P_Content), col="blue")
```




```
plot(density(life_satisfaction[life_satisfaction$Year == 2014 & life_satisfaction$Sex == 'Fem
ale',]$P_Content), main = 'Density Plot - Happiness levels in 2014 for Female children')
polygon(density(life_satisfaction[life_satisfaction$Year == 2014 & life_satisfaction$Sex ==
'Female',]$P_Content), col="blue")
```




```
plot(density(life_satisfaction[life_satisfaction$Year == 2018 & life_satisfaction$Sex == 'Fem
ale',]$P_Content), main = 'Density Plot - Happiness levels in 2018 for Female children')
polygon(density(life_satisfaction[life_satisfaction$Year == 2018 & life_satisfaction$Sex ==
'Female',]$P_Content), col="blue")
```




```
years <- seq(2006,2018,4)
print(years)
```

```
## [1] 2006 2010 2014 2018
```

```
for (year in years)
{
    print(summary(life_satisfaction[life_satisfaction$Year == year & life_satisfaction$Sex ==
'Female', "P_Content"]))
}
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   51.93   59.19   62.80   63.98   70.33   74.69
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   54.77   56.90   61.07   63.06   71.24   71.77
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   51.54   54.26   63.85   63.85   72.26   77.64
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   47.58   51.07   58.37   59.79   69.38   72.81
```

```
for (year in years)
{
    print(quantile(life_satisfaction[life_satisfaction$Year == year & life_satisfaction$Sex =
= 'Female', "P_Content"]))
}
```

```
##      0%     25%     50%     75%    100%
## 51.9300 59.1925 62.7950 70.3325 74.6900
##      0%     25%     50%     75%    100%
## 54.7700 56.8950 61.0700 71.2375 71.7700
##      0%     25%     50%     75%    100%
## 51.5400 54.2575 63.8550 72.2625 77.6400
##      0%     25%     50%     75%    100%
## 47.5800 51.0725 58.3700 69.3750 72.8100
```

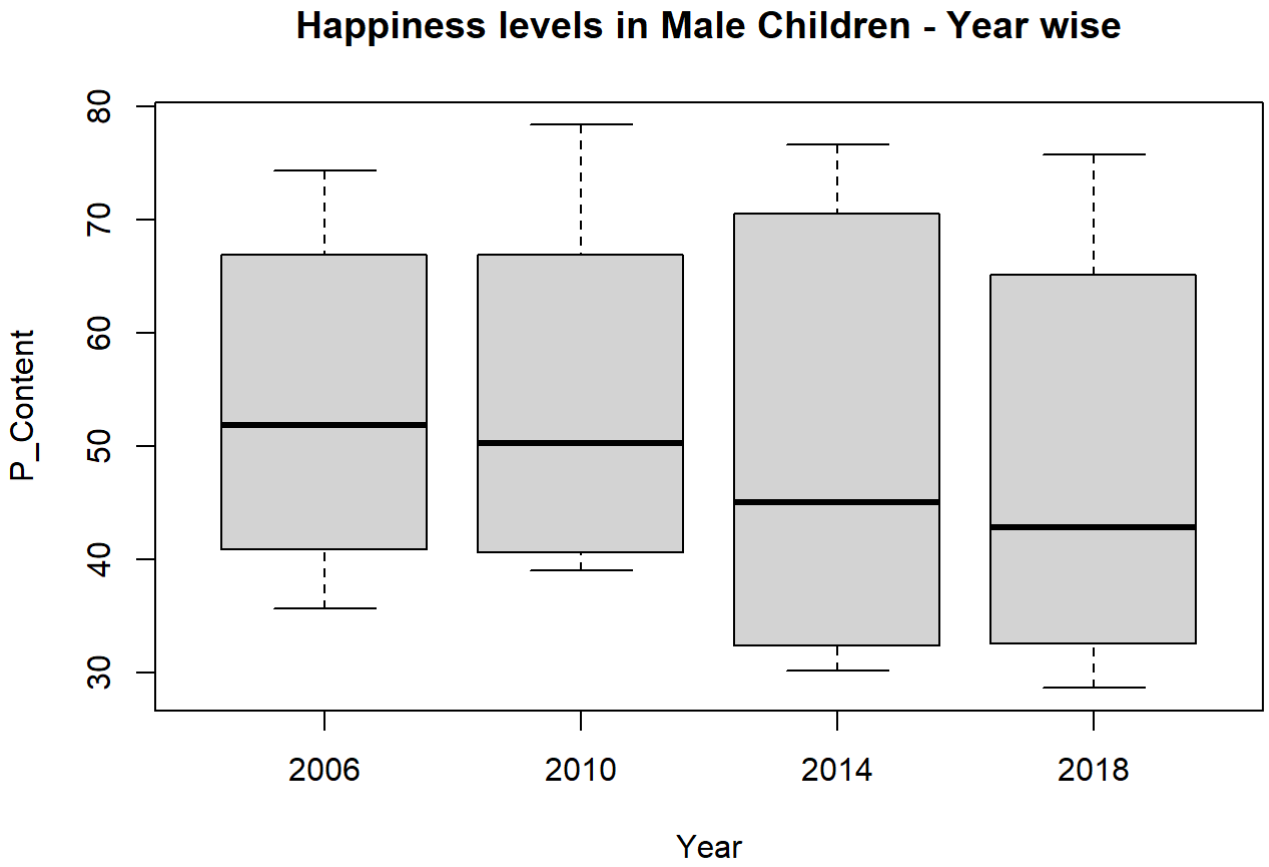
```
for (year in years)
{
    print(IQR(life_satisfaction[life_satisfaction$Year == year & life_satisfaction$Sex == 'Fe
male', "P_Content"]))
}
```

```
## [1] 11.14
## [1] 14.3425
## [1] 18.005
## [1] 18.3025
```

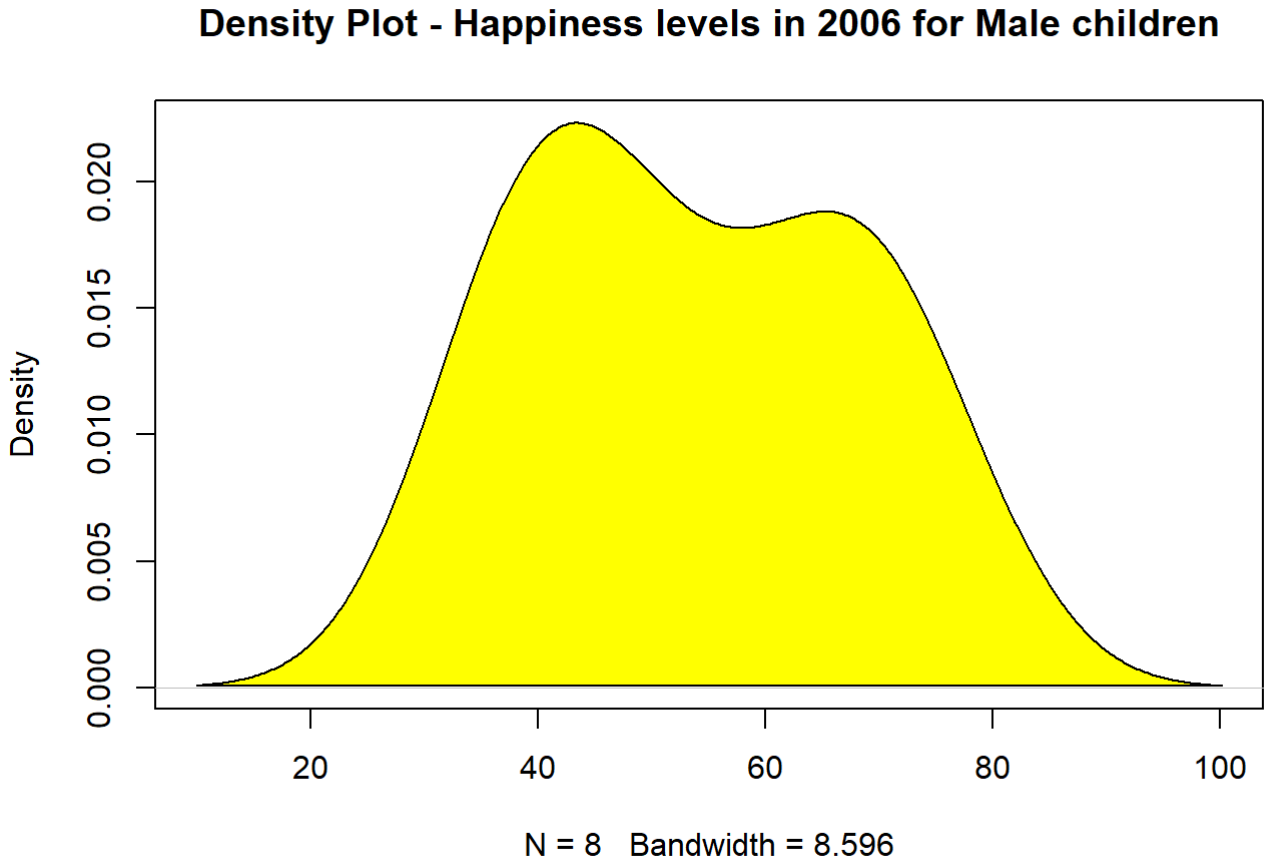
Based on the above plots and descriptive statistics, we can infer the following about the happiness levels of Female school children over time:
i) The Median happiness was highest in 2014 (63.85%) while the Mean happiness was highest in 2006 (63.98%)
ii) The Median and Mean happiness was lowest in 2018 (58.37% & 59.79% respectively)
iii) The happiness levels in 2018 is the most spread out while it's the least in 2006. (IQR 18.3025 vs IQR 11.14)
iv) The highest happiness percentage was recorded in 2014 (77.64%) while the lowest was recorded in 2018 (47.58%)

Q3) Using a boxplot, the density and the descriptive statistics (mean, min, max, median, and quantiles), describe the difference in the distributions for the percentage of male school aged children that reported being happy with the way they are with respect to year (2006; 2010; 2014; 2018)

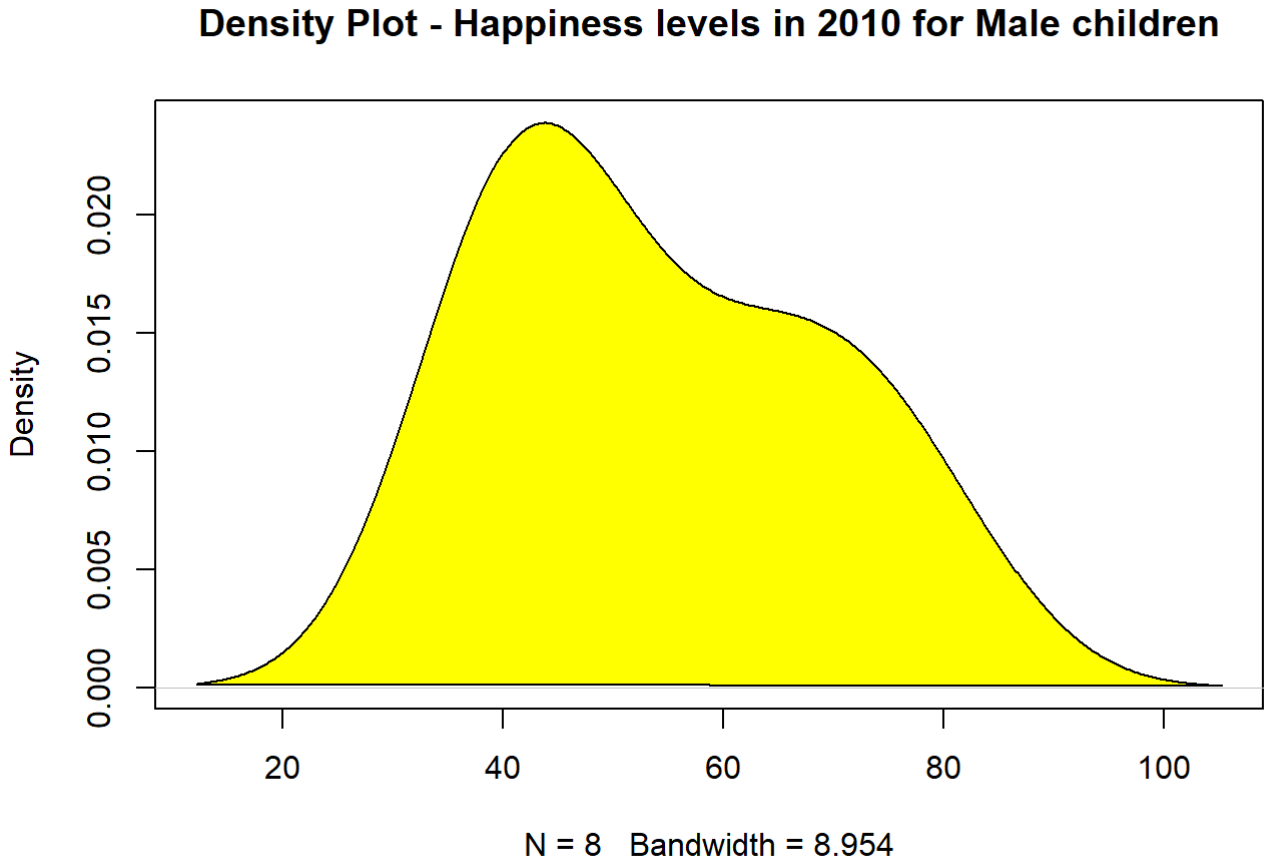
```
boxplot(P_Content[Sex == 'Male'] ~ Year[Sex == 'Male'], data = life_satisfaction, main = 'Hap
piness levels in Male Children - Year wise', xlab = 'Year', ylab = 'P_Content')
```



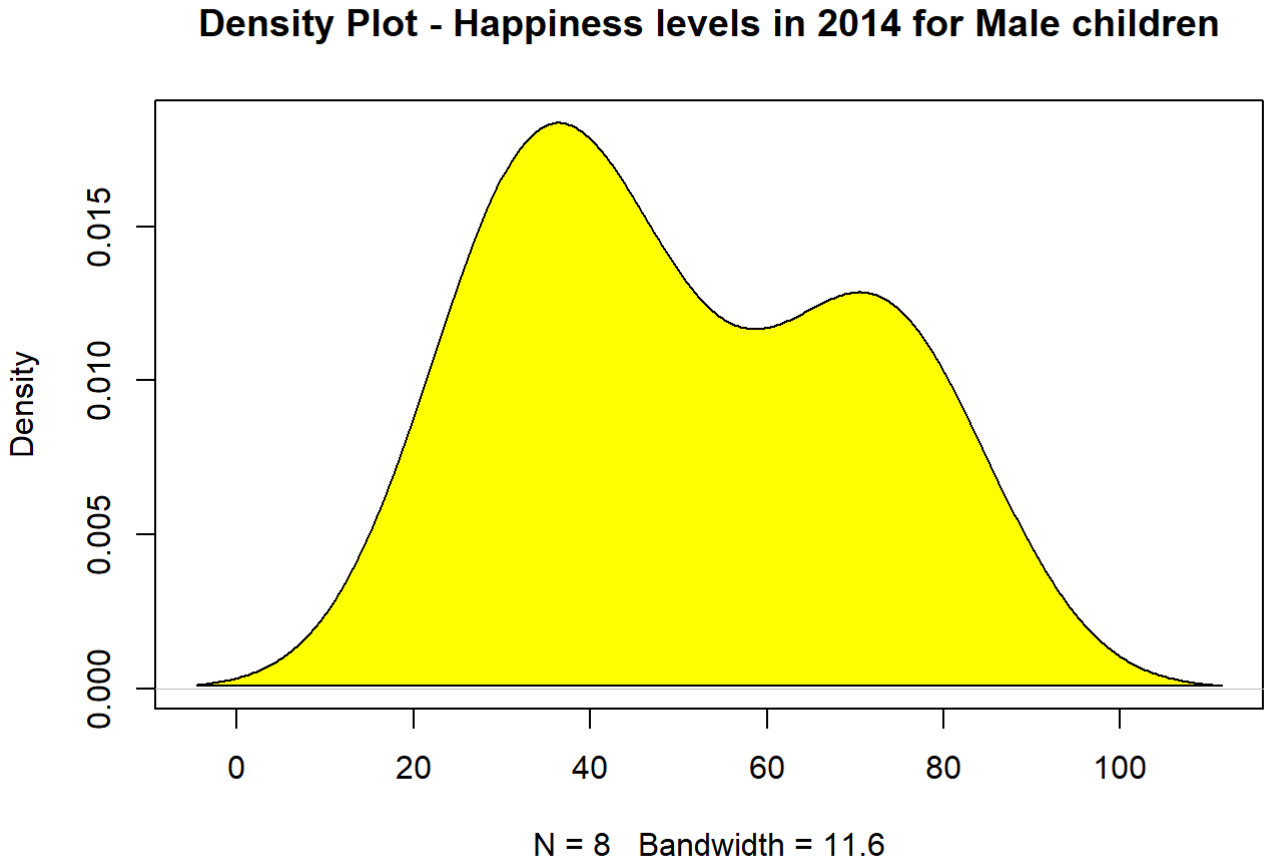

```
plot(density(life_satisfaction[life_satisfaction$Year == 2006 & life_satisfaction$Sex == 'Mal
e',]$P_Content), main = 'Density Plot - Happiness levels in 2006 for Male children')
polygon(density(life_satisfaction[life_satisfaction$Year == 2006 & life_satisfaction$Sex ==
'Male',]$P_Content), col="yellow")
```



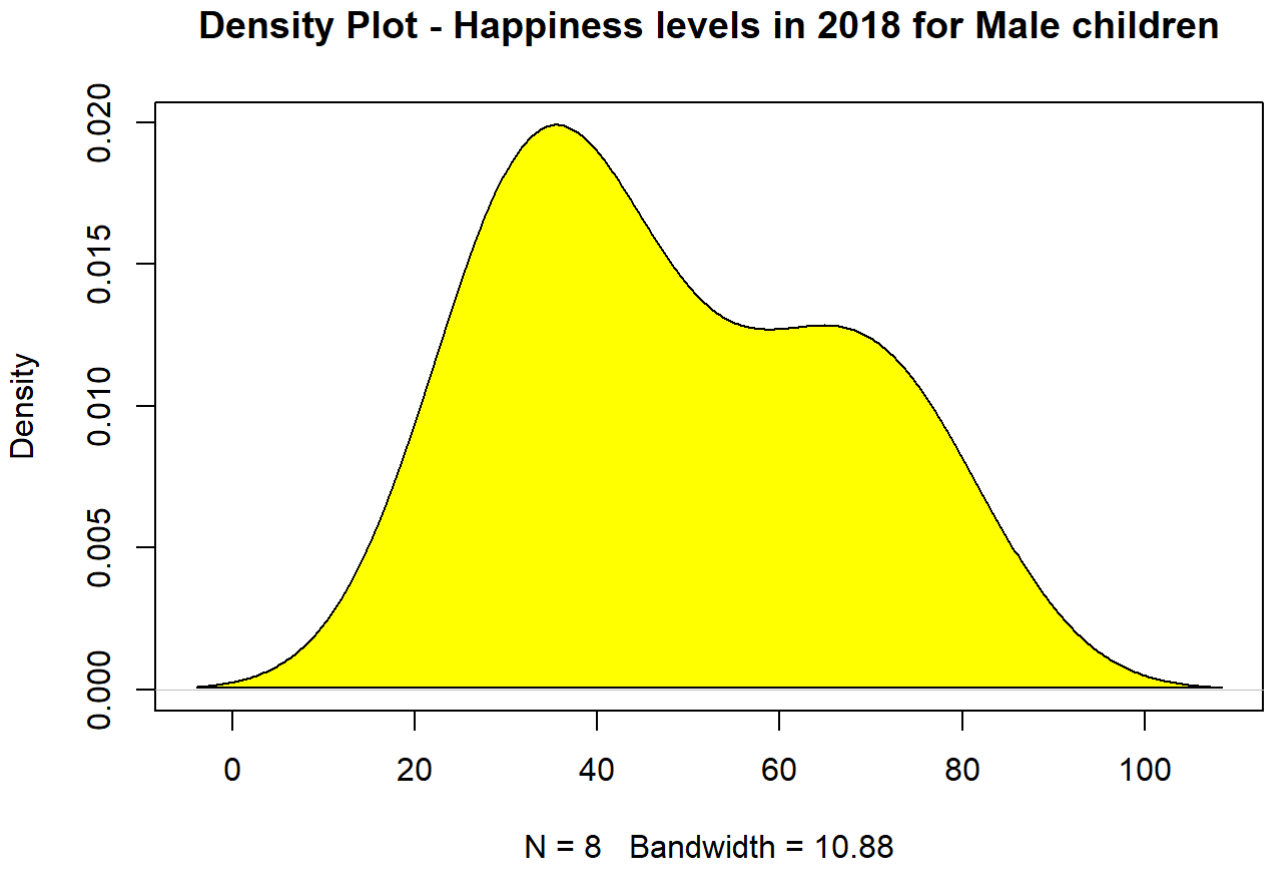

```
plot(density(life_satisfaction[life_satisfaction$Year == 2010 & life_satisfaction$Sex == 'Mal
e',]$P_Content), main = 'Density Plot - Happiness levels in 2010 for Male children')
polygon(density(life_satisfaction[life_satisfaction$Year == 2010 & life_satisfaction$Sex ==
'Male',]$P_Content), col="yellow")
```




```
plot(density(life_satisfaction[life_satisfaction$Year == 2014 & life_satisfaction$Sex == 'Mal
e',]$P_Content), main = 'Density Plot - Happiness levels in 2014 for Male children')
polygon(density(life_satisfaction[life_satisfaction$Year == 2014 & life_satisfaction$Sex ==
'Male',]$P_Content), col="yellow")
```




```
plot(density(life_satisfaction[life_satisfaction$Year == 2018 & life_satisfaction$Sex == 'Mal
e',]$P_Content), main = 'Density Plot - Happiness levels in 2018 for Male children')
polygon(density(life_satisfaction[life_satisfaction$Year == 2018 & life_satisfaction$Sex ==
'Male',]$P_Content), col="yellow")
```




```
for (year in years)
{
    print(summary(life_satisfaction[life_satisfaction$Year == year & life_satisfaction$Sex ==
'Male', "P_Content"]))
}
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   35.71   41.44   51.93   53.72   66.62   74.37
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   39.07   41.37   50.34   54.19   65.48   78.41
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   30.23   32.60   45.05   50.40   68.68   76.70
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   28.70   32.62   42.84   48.23   62.99   75.82
```

```
for (year in years)
{
    print(quantile(life_satisfaction[life_satisfaction$Year == year & life_satisfaction$Sex =
= 'Male', "P_Content"]))
}
```

```
##      0%     25%     50%     75%    100%
## 35.7100 41.4425 51.9300 66.6175 74.3700
##      0%     25%     50%     75%    100%
## 39.0700 41.3650 50.3400 65.4775 78.4100
##      0%     25%     50%     75%    100%
## 30.2300 32.5975 45.0550 68.6800 76.7000
##      0%     25%     50%     75%    100%
## 28.7000 32.6250 42.8400 62.9875 75.8200
```

```
for (year in years)
{
    print(IQR(life_satisfaction[life_satisfaction$Year == year & life_satisfaction$Sex == 'Ma
le', "P_Content"]))
}
```

```
## [1] 25.175
## [1] 24.1125
## [1] 36.0825
## [1] 30.3625
```

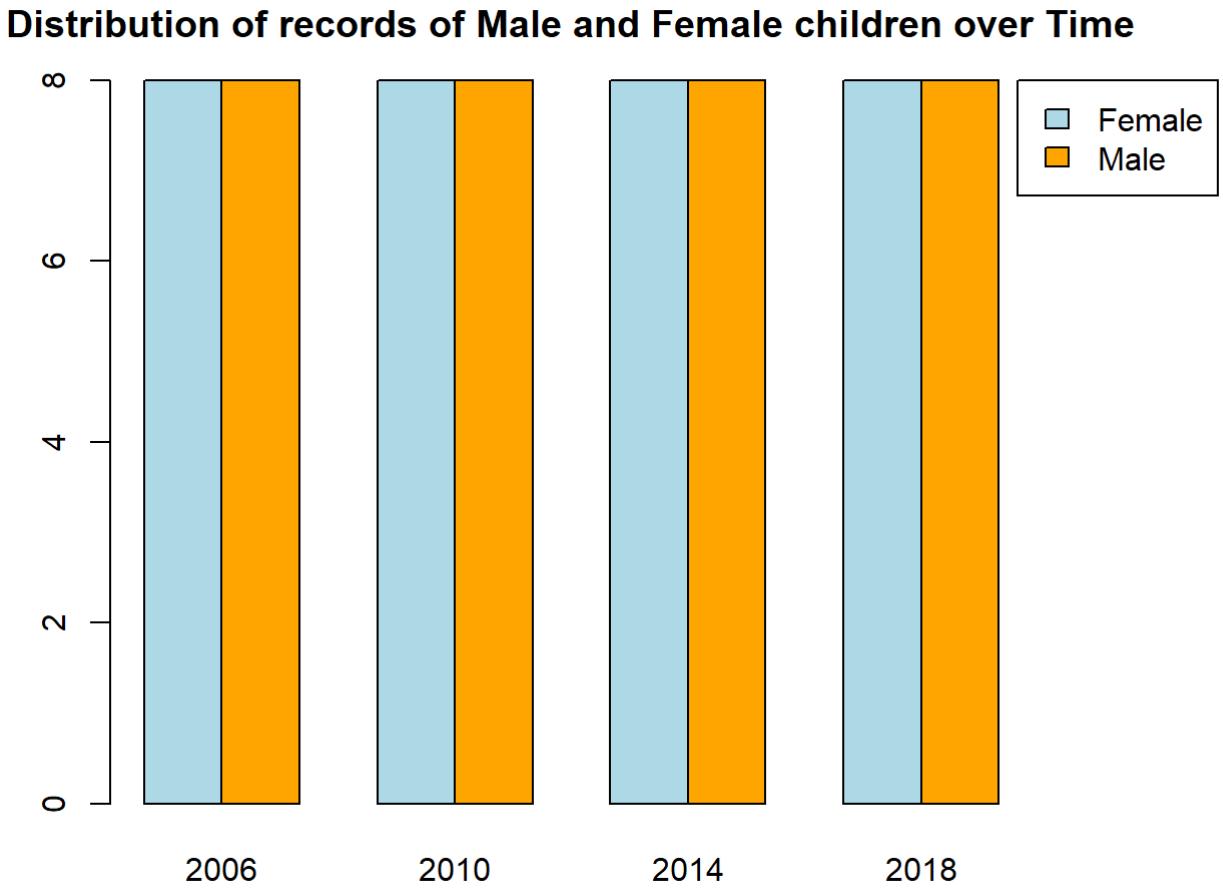
Based on the above plots and descriptive statistics, we can infer the following about the happiness levels of Male school children over time:
i) The Median happiness was highest in 2006 (51.93%) while the Mean happiness was highest in 2010 (54.19%)
ii) The Median and Mean happiness was lowest in 2018 (42.84% & 48.23% respectively)
iii) The Median happiness has been steadily decreasing over time
iv) The highest happiness percentage was recorded in 2010 (78.41%) while the lowest was recorded in 2018 (28.70%)
v) The happiness levels in 2014 is the most spread out while it's the least in 2010. (IQR 36.0825 vs IQR 24.1125)

Q4) Convert the categorical variable Sex to a factor. Describe and illustrate the frequency of the categorical variable Sex with respect to year (2006; 2010; 2014; 2018)

```
life_satisfaction$Sex <- as.factor(life_satisfaction$Sex)
is.factor(life_satisfaction$Sex)
```

```
## [1] TRUE
```

```
table1 <- table(life_satisfaction$Sex, life_satisfaction$Year)

par(mar=c(3, 3, 3, 8))
barplot(table1, main="Distribution of records of Male and Female children over Time", xlab="Y
ear", col=c("lightblue","orange"), legend = rownames(table1), beside=TRUE, args.legend = list
(x = "topright", inset = c(-0.20, 0)))
```



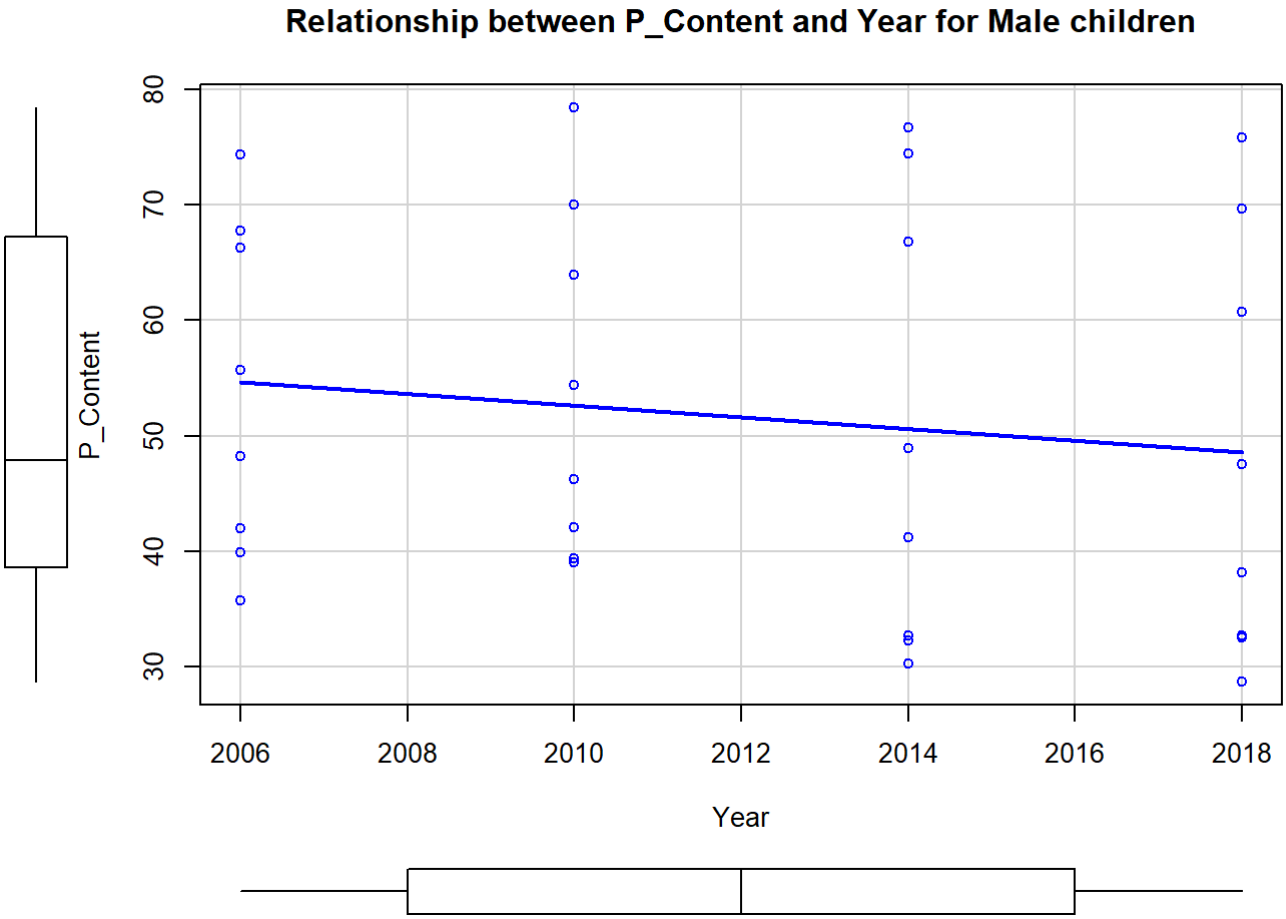
We can clearly see that there are 8 records each for both Male and Female children for every year (2006, 2010, 2014, 2018)

Q5) Using the correlation and scatter plots discuss the relationship between P_Content and Year for Males and Females separately

```
library(car)
```

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

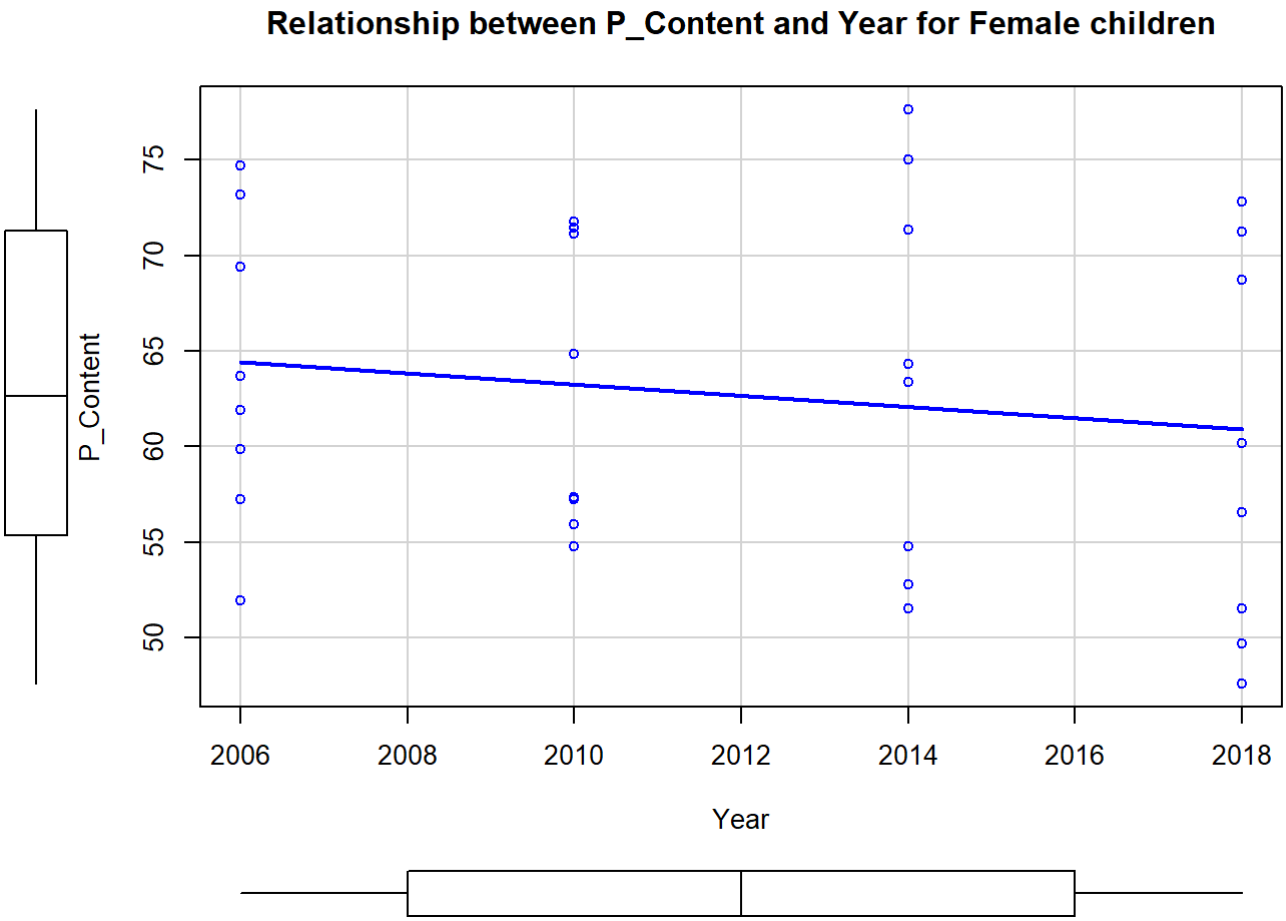
```
male <- life_satisfaction[life_satisfaction$Sex == 'Male',c('P_Content','Year')]
female <- life_satisfaction[life_satisfaction$Sex == 'Female',c('P_Content','Year')]

scatterplot(P_Content ~ Year, data = male, smooth = FALSE, main = 'Relationship between P_Con
tent and Year for Male children')
```




```
cor(male)
```

```
##           P_Content      Year
## P_Content  1.000000 -0.141013
## Year      -0.141013  1.000000
```

```
scatterplot(P_Content ~ Year, data = female, smooth = FALSE, main = 'Relationship between P_C
ontent and Year for Female children')
```




```
cor(female)
```

```
##             P_Content       Year
## P_Content  1.0000000 -0.1528523
## Year      -0.1528523  1.0000000
```

Based on the scatter plot and the correlation, we can deduce the following about the relationship between P_Content and Year for Male and Female children:
i) There's a low negative correlation (weak linear relationship) between Year and P_Content for both Male (-0.1410) and Female children (-0.1528)
ii) The negative correlation is almost identical for both Male and Female children with a slightly higher correlation for Female children (diff of 0.0118393)

Section 2: Regression Model

Q1) Using R fit a simple linear regression model to the data with P_Content as the response variable and Year as a numeric predictor variable for females. Define and describe the terms in your mathematical equation for the model. (Also provide you R code)

```
linearModel <- lm(P_Content ~ Year, data = female)  # female data set created above

linearModel
```

```
##
## Call:
## lm(formula = P_Content ~ Year, data = female)
##
## Coefficients:
## (Intercept)         Year
##    655.0157      -0.2944
```

```
# Also cross checked using the codes below:

SXX = sum((female$Year - mean(female$Year))^2)
SXY = sum((female$Year - mean(female$Year))*(female$P_Content - mean(female$P_Content)))

beta1 <- SXY / SXX
beta0 <- mean(female$P_Content) - beta1 * mean(female$Year)

print(beta0)
```

```
## [1] 655.0157
```

```
print(beta1)
```

```
## [1] -0.2944062
```

The linear regression model is described as, P_Content = 655.0157 - 0.2944 * Year
Here 655.0157 is the intercept while -0.2944 is the slope.

The model estimates that increasing the year by 1 unit would reduce the P_Content by 0.2944.

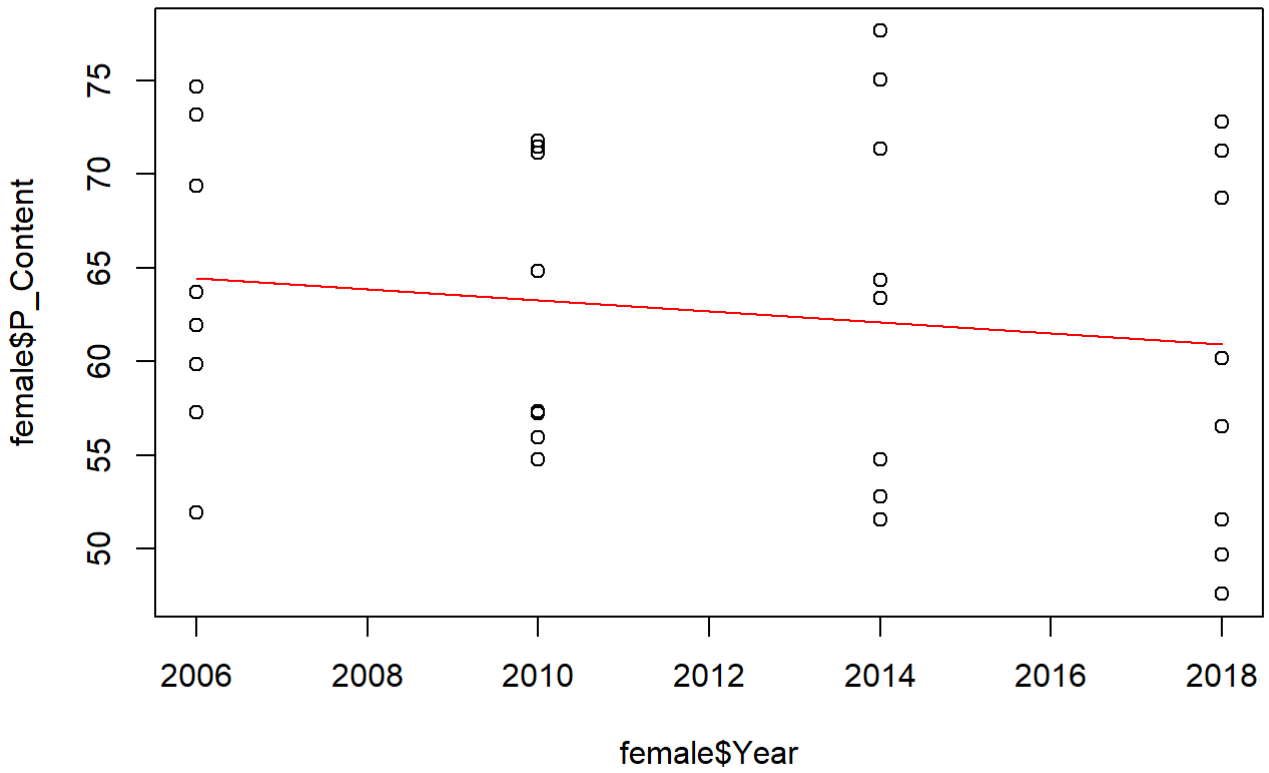
If we assume the year to be 0 (we know year won't be 0. But let's just assume it for argument's sake), the P_Content comes out to be 655.0157. But it would make more sense if the P_Content lies between 0 - 100.

So, in order to aid in the interpretation of the intercept, we can subtract the Average Year from the predictor variable Year. This will be 0 when the Year is at the average Year (which is 2012).

```
female$Cen_Year <- female$Year - mean(female$Year)
linearModel <- lm(P_Content ~ Cen_Year, data = female)

linearModel
```

```
##
## Call:
## lm(formula = P_Content ~ Cen_Year, data = female)
##
## Coefficients:
## (Intercept)     Cen_Year
##     62.6703      -0.2944
```

```
plot(female$Year, female$P_Content)
lines(female$Year, fitted(linearModel), col = 'red')
```



The linear regression model is now described as, P_Content = 62.6703 - 0.2944 * (Year(i) - average(Year))
Here 62.6703 is the intercept while the slope continues to be -0.2944. P_Content is the response variable while Year is the predictor variable.

Q2) Interpret the estimate of the intercept term

From the previous question, we know that the linear regression model can be described as
P_Content = 62.6703 - 0.2944 * (Year(i) - average(Year))

So, the intercept is 62.6703. This indicates that when the Year is 2012 (equals the Average Year (2012 in this case)), the P_Content for Female children comes out to be 62.67%

Q3) Interpret the estimate of the slope

From question 1, we know that the linear regression model can be described as
P_Content = 62.6703 - 0.2944 * (Year(i) - average(Year))

So, the slope is -0.2944. This indicates that increasing the year by 1 (when compared to the average Year) would reduce the P_Content by 0.2944%

Q4) What is the standard error of a parameter? Calculate and comment on the standard error of the estimate of the intercept and slope term.

```
summary(linearModel)
```

```
##
## Call:
## lm(formula = P_Content ~ Cen_Year, data = female)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3239  -7.3147  -0.7353   8.2859  15.5585
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  62.6703     1.5542  40.324   <2e-16 ***
## Cen_Year     -0.2944     0.3475  -0.847    0.404
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.792 on 30 degrees of freedom
## Multiple R-squared:  0.02336,    Adjusted R-squared:  -0.009191
## F-statistic: 0.7177 on 1 and 30 DF,  p-value: 0.4036
```

The standard deviation of the sampling distribution of a parameter is known as the standard error of the parameter. As we can see from the summary above, the standard error for the intercept and slope is 1.5542 and 0.3475 respectively. This means that the the standard deviation (degree of spread) of the sampling distribution of the intercept is 1.5542 while the standard deviation (degree of spread) for the slope is 0.3475.

Q5) Calculate and interpret the confidence intervals for β0 (Provide you R code)

```
confint(linearModel)
```

```
##                 2.5 %     97.5 %
## (Intercept) 59.496293 65.8443324
## Cen_Year    -1.004139  0.4153262
```

By using the confint function, we were able to get the confidence intervals for β0
The confidence intervals for β0 is [59.49, 65.84]. This indicates that we're 95% confident that β0 lies between 59.49 and 65.84

Q6) Calculate and interpret the confidence intervals for β1 (Provide you R code)

```
confint(linearModel)
```

```
##                 2.5 %     97.5 %
## (Intercept) 59.496293 65.8443324
## Cen_Year    -1.004139  0.4153262
```

By using the confint function, we were able to get the confidence intervals for β1
The confidence intervals for β1 is [-1.004, 0.415]. This indicates that we're 95% confident that β1 lies between -1.004 and 0.415

Q7) What does the confidence interval of a parameter measure?

Confidence interval is a range of estimates for an unknown parameter. Confidence level refers to the percentage of probability, or certainty, that the confidence interval would contain the true population parameter when you draw a random sample many times. They can take any number of probability limits, with the most common being a 95% or 99% confidence level.

To give an example, in our model, in the case of parameter β0, we're 95% confident (Confidence level) that β0 would lie between [59.49, 65.84] (Confidence interval)

Q8) Does a 95% confidence interval always contain the population parameter?

No. It may or may not contain the population parameter. The 95% Confidence Interval tells you that that there is a 95% probability that the population parameter falls within the interval. So there's always a 5% chance that it could fall outside the interval.

Q9) Compute and interpret the hypothesis test H0 : β0 = 0 vs Ha : β0 ≠ 0. State the test statistic. Compare the test statistic to the correct distribution value and state your conclusion. Also, report the p-value and the conclusion in the context of the problem.

```
# Computing the test statistic where H0 : β0 = 0 vs Ha : β0 != 0

summary(linearModel)
```

```
##
## Call:
## lm(formula = P_Content ~ Cen_Year, data = female)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3239  -7.3147  -0.7353   8.2859  15.5585
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  62.6703     1.5542  40.324   <2e-16 ***
## Cen_Year     -0.2944     0.3475  -0.847    0.404
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.792 on 30 degrees of freedom
## Multiple R-squared:  0.02336,    Adjusted R-squared:  -0.009191
## F-statistic: 0.7177 on 1 and 30 DF,  p-value: 0.4036
```

```
# Also cross checked using the codes below:

N = length(female$Cen_Year)
MSE = sum(linearModel$residuals^2/(N-2))
SXX = sum((female$Cen_Year- mean(female$Cen_Year))^2)
VARB0 = MSE*(1/ N + (mean(female$Cen_Year)^{2}/SXX))
T = (linearModel$coefficients[1]-0)/sqrt(VARB0)

print(T)
```

```
## (Intercept)
##    40.32421
```

```
# Comparing the test statistic to the correct distribution value and finding the p value

alpha = 0.05
TDIST = qt(1-alpha/2, N-2)
PVALUE = 2 *( 1- pt(abs(T), df = N- 2))

print(TDIST)
```

```
## [1] 2.042272
```

```
print(PVALUE)
```

```
## (Intercept)
##           0
```

As we can see from the results from our R codes, the t value for β0 = 40.324. The distribution value is found to be 2.042. Thus we can clearly see that |40.324| > 2.04 and hence we can reject the null hypothesis. At the 5% level of significance, the evidence is not strong enough to indicate that β0 = 0. We also found the p value to be 0 which is less than 0.05. So we can safely reject the null hypothesis in favor of the alternative hypothesis. This indicates that when the year is 2012 (the mean) the P_Content is non-zero.

Q10) Compute and interpret the hypothesis test H0 : β1 = 0 vs Ha : β1 ≠ 0. State the test statistic. Compare the test statistic to the correct distribution value and state your conclusion. Also, report the p-value and the conclusion in the context of the problem.

```
# Computing the test statistic where H0 : β1 = 0 vs Ha : β1 != 0

summary(linearModel)
```

```
##
## Call:
## lm(formula = P_Content ~ Cen_Year, data = female)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3239  -7.3147  -0.7353   8.2859  15.5585
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  62.6703     1.5542  40.324   <2e-16 ***
## Cen_Year     -0.2944     0.3475  -0.847    0.404
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.792 on 30 degrees of freedom
## Multiple R-squared:  0.02336,    Adjusted R-squared:  -0.009191
## F-statistic: 0.7177 on 1 and 30 DF,  p-value: 0.4036
```

```
# Also cross checked using the codes below:

N = length(female$Cen_Year)
MSE = sum(linearModel$residuals^2/(N-2))
SXX = sum((female$Cen_Year- mean(female$Cen_Year))^2)
VARB1 = MSE/SXX
T = (linearModel$coefficients[2]-0)/sqrt(VARB1)

print(T)
```

```
##   Cen_Year
## -0.8471612
```

```
# Comparing the test statistic to the correct distribution value and finding the p value

alpha = 0.05
TDIST = qt(1-alpha/2, N-2)
PVALUE = 2 *( 1- pt(abs(T), df = N- 2))

print(TDIST)
```

```
## [1] 2.042272
```

```
print(PVALUE)
```

```
##  Cen_Year
## 0.4036132
```

As we can see from the results from our R codes, the t value for β1 = -0.847. The distribution value is found to be 2.042. Thus we can clearly see that |-0.847| < 2.04 and hence we cannot reject the null hypothesis. At the 5% level of significance, the evidence is strong enough to indicate that β1 = 0. We also found the p value to be

0.403 which is greater than 0.05. So we cannot reject the null hypothesis in favor of the alternative hypothesis. This indicates that there's no relationship between Year and P_Content.

Q11) Interpret the F-statistic in the output in the summary of the regression model. Hint: State the hypothesis being tested, the test statistic and p-value and the conclusion in the context of the problem.

```
# Computing the F test statistic H0 : Yi = β0 + e vs. HA : Yi = β0 + β1Xi + e

summary(linearModel)
```

```
##
## Call:
## lm(formula = P_Content ~ Cen_Year, data = female)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3239  -7.3147  -0.7353   8.2859  15.5585
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  62.6703     1.5542  40.324   <2e-16 ***
## Cen_Year     -0.2944     0.3475  -0.847    0.404
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.792 on 30 degrees of freedom
## Multiple R-squared:  0.02336,    Adjusted R-squared:  -0.009191
## F-statistic: 0.7177 on 1 and 30 DF,  p-value: 0.4036
```

```
# Using anova function to cross-check the F-statistic

anova(linearModel)
```

```
## Analysis of Variance Table
##
## Response: P_Content
##           Df  Sum Sq Mean Sq F value Pr(>F)
## Cen_Year   1   55.47  55.472  0.7177 0.4036
## Residuals 30 2318.80  77.293
```

```
# Comparing the F test statistic to the correct distribution value

alpha = 0.05
FDIST = qf(1-alpha,1, N-2)

print(FDIST)
```

```
## [1] 4.170877
```

As we can see from the results from our R codes, the F test statistic = 0.7177. The distribution value is found to be 4.170. Thus we can clearly see that |0.7177| < 4.170 and hence we cannot reject the null hypothesis. At the 5% level of significance, the evidence is strong enough to indicate that Yi = β0 provides a better fit to the data. We also found the p value to be 0.4036 which is greater than 0.05. So we cannot reject the null hypothesis in favor of the alternative hypothesis. This indicates that there's no relationship between Year and P_Content.

Q12) Interpret the R-squared value.

```
# Getting the R-Squared value

summary(linearModel)
```

```
##
## Call:
## lm(formula = P_Content ~ Cen_Year, data = female)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3239  -7.3147  -0.7353   8.2859  15.5585
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  62.6703     1.5542  40.324   <2e-16 ***
## Cen_Year     -0.2944     0.3475  -0.847    0.404
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.792 on 30 degrees of freedom
## Multiple R-squared:  0.02336,    Adjusted R-squared:  -0.009191
## F-statistic: 0.7177 on 1 and 30 DF,  p-value: 0.4036
```

```
# Cross-checking it through the formula

SST = sum((female$P_Content - mean(female$P_Content))^2)
SSE = sum(linearModel$residuals^2)
R2 <- (SST - SSE) /SST

print(R2)
```

```
## [1] 0.02336381
```

R squared value gives a numerical measure of the degree of association between the response variable (P_Content) and the predictor variable (Year). We can see that the R squared value is 0.023 which is very close to 0. This implies that there is no discernible relationship between P_Content and Year. (Similar to what was predicted through the T-test and F-test)

Q13) Interpret the residual standard error of the simple linear regression model.

```
# Getting the residual standard error

summary(linearModel)
```

```
##
## Call:
## lm(formula = P_Content ~ Cen_Year, data = female)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3239  -7.3147  -0.7353   8.2859  15.5585
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  62.6703     1.5542  40.324   <2e-16 ***
## Cen_Year     -0.2944     0.3475  -0.847    0.404
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.792 on 30 degrees of freedom
## Multiple R-squared:  0.02336,    Adjusted R-squared:  -0.009191
## F-statistic: 0.7177 on 1 and 30 DF,  p-value: 0.4036
```

```
# Cross-checking it through the formula

N = length(female$Cen_Year)
SSE = sum(linearModel$residuals^2)
RMSE = sqrt(SSE/(N-2))

print(RMSE)
```

```
## [1] 8.791661
```

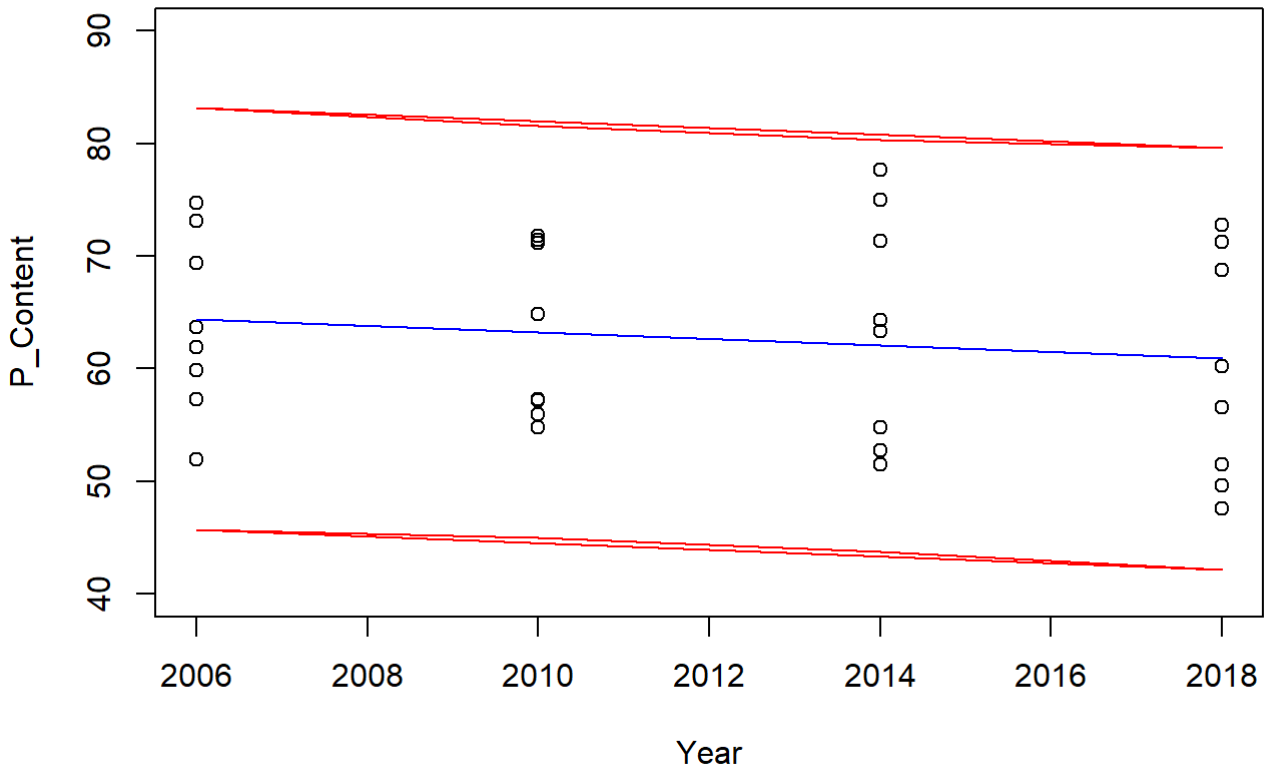
Residual Standard Error is a measure of the quality of a linear regression fit. Due to the presence of error in the model, we are not capable of perfectly predicting our response variable (P_Content) from the predictor variable (Year). Based on our finding, we can say that the Year accurately predicts P_Content with about 8.79% error on average.

Q14) Calculate, plot and comment on the shape of the prediction intervals for the estimated values of Y.

```
N = length(female$Cen_Year)
SXX = sum((female$Cen_Year - mean(female$Cen_Year))^2)
MSE = SSE/(N-2)
Var_E = MSE*(1 + 1/N + (female$Cen_Year-mean(female$Cen_Year))^2/SXX)
Yhat = fitted(linearModel)
cbind(Yhat- qt(1-alpha/2,N-2)*sqrt(Var_E), Yhat + qt(1-alpha/2,N-2)*sqrt(Var_E))
```

```
##        [,1]     [,2]
## 2  45.71272 83.16078
## 4  44.97060 81.54765
## 6  43.79297 80.37003
## 8  42.17985 79.62790
## 10 45.71272 83.16078
## 12 44.97060 81.54765
## 14 43.79297 80.37003
## 16 42.17985 79.62790
## 18 45.71272 83.16078
## 20 44.97060 81.54765
## 22 43.79297 80.37003
## 24 42.17985 79.62790
## 26 45.71272 83.16078
## 28 44.97060 81.54765
## 30 43.79297 80.37003
## 32 42.17985 79.62790
## 34 45.71272 83.16078
## 36 44.97060 81.54765
## 38 43.79297 80.37003
## 40 42.17985 79.62790
## 42 45.71272 83.16078
## 44 44.97060 81.54765
## 46 43.79297 80.37003
## 48 42.17985 79.62790
## 50 45.71272 83.16078
## 52 44.97060 81.54765
## 54 43.79297 80.37003
## 56 42.17985 79.62790
## 58 45.71272 83.16078
## 60 44.97060 81.54765
## 62 43.79297 80.37003
## 64 42.17985 79.62790
```

```
plot(female$Year,female$P_Content, xlab="Year", ylab="P_Content", ylim = c(40,90))
lines(female$Year,Yhat, col="blue")
lines(female$Year,Yhat + qt(1-alpha/2,N-2)*sqrt(Var_E),col="red")
lines(female$Year,Yhat - qt(1-alpha/2,N-2)*sqrt(Var_E),col="red")
```

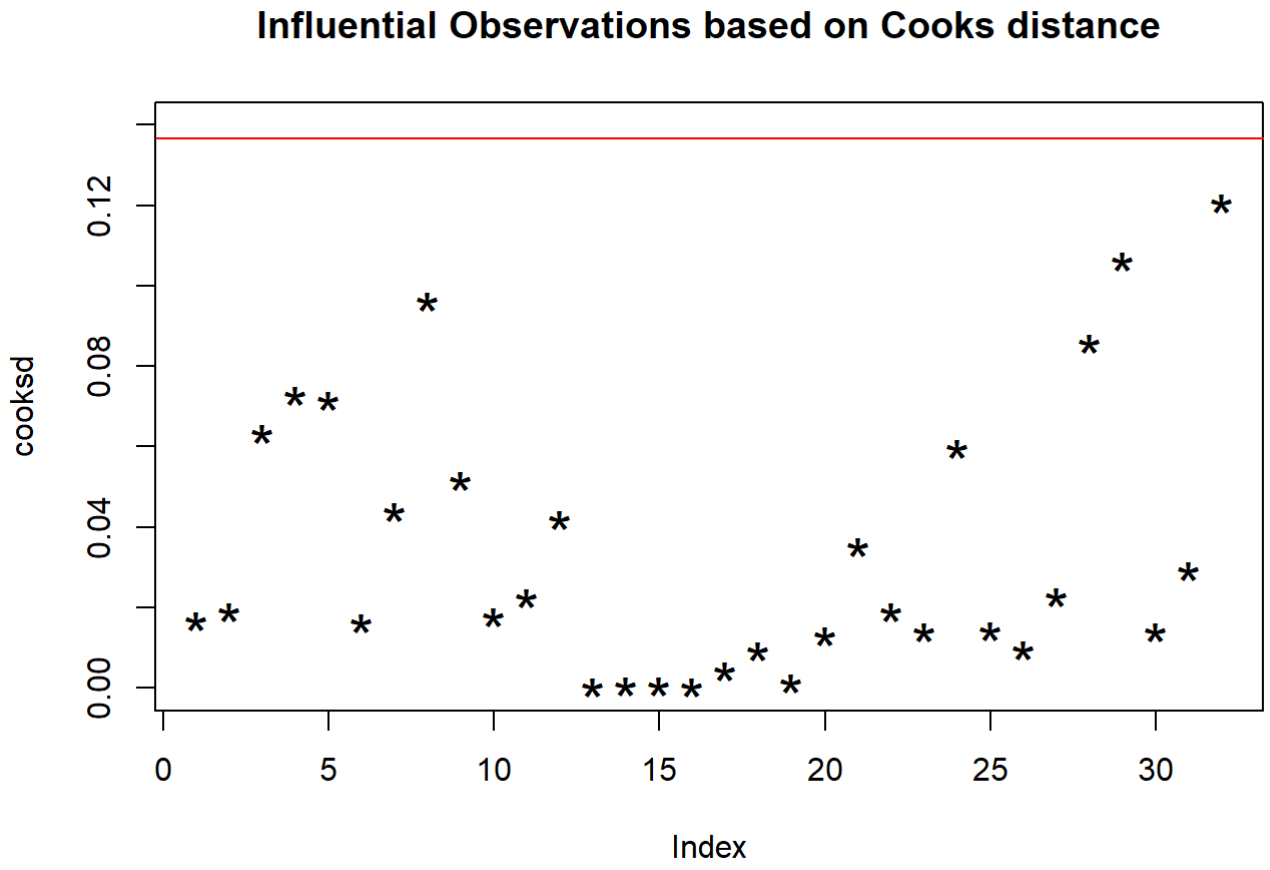


A prediction interval is a range of values that predicts the value of a new observation, based on our existing model. In other words, a prediction interval is where you expect a future value to fall. A 95% prediction interval tells you that future values will fall into that range 95% of the time. The predicted interval seems to have similar negative linear shape as the model on both sides of the model. This predicted interval can be used to confirm with a 95% chance that any new value for Female children P_Content will fall within this interval.

Section 3: Regression Model Diagnostics

Q1) Are there any influential observations?

```
cooksd <- cooks.distance(linearModel)
plot(cooksd, pch="*", cex=2, main="Influential Observations based on Cooks distance", ylim =
 c(0,0.14))
abline(h = 4*mean(cooksd, na.rm=T), col="red")
```



Cook's distance computes the influence exerted by each data point on the predicted outcome. In general use, those observations that have a Cook's distance greater than 4 times the mean may be classified as influential. As we can see from the chart above, there are no observations that cross the red line (The red line denotes the distance which equates to 4 times the mean). Hence we can say that there are no influential observations in the female children data.

Q2) Examine the residuals of the regression model and comment on whether you think the residuals satisfy the assumptions required for small sample inference. Provide the rationale for your answer

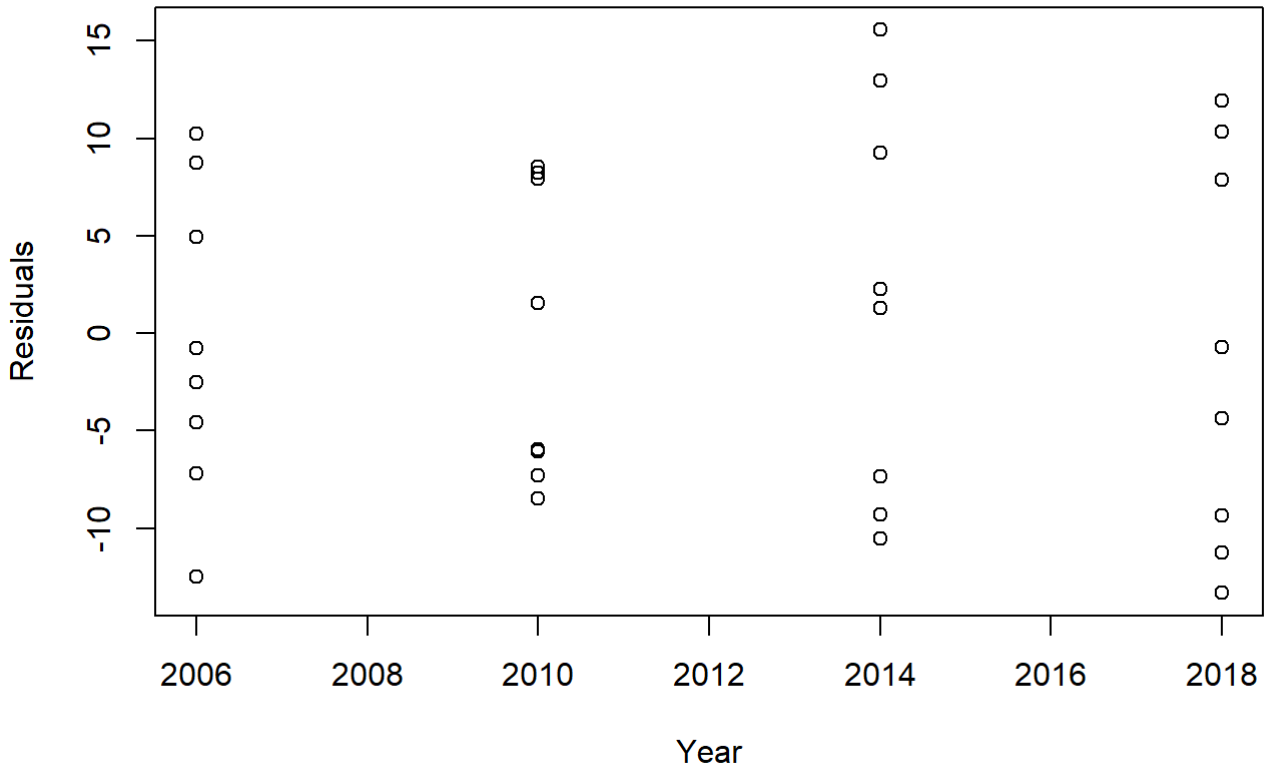
The assumptions that need to be satisfied to perform the small sample inference (t and F tests) are
i) Variation in the data input X
ii) Random Sampling
iii) Linearity in Parameters
iv) Zero Conditional Mean
v) Homoskedasticity (Constant variance)
vi) Normality of Errors

Using the residuals, we will be able to identify if they satisfy zero conditional mean, constant variance and normality of errors.

```
summary(residuals(linearModel))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -13.3239  -7.3147  -0.7353   0.0000   8.2859  15.5585
```

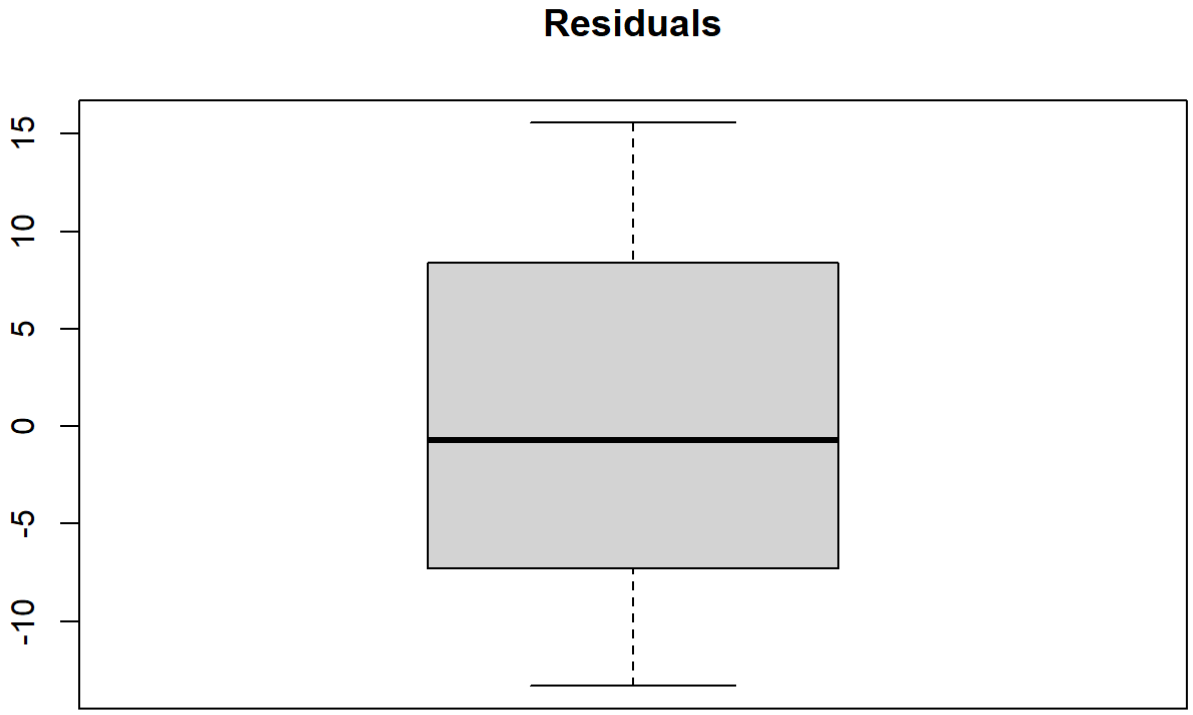
```
plot(female$Year,residuals(linearModel), xlab = 'Year', ylab = 'Residuals')
```



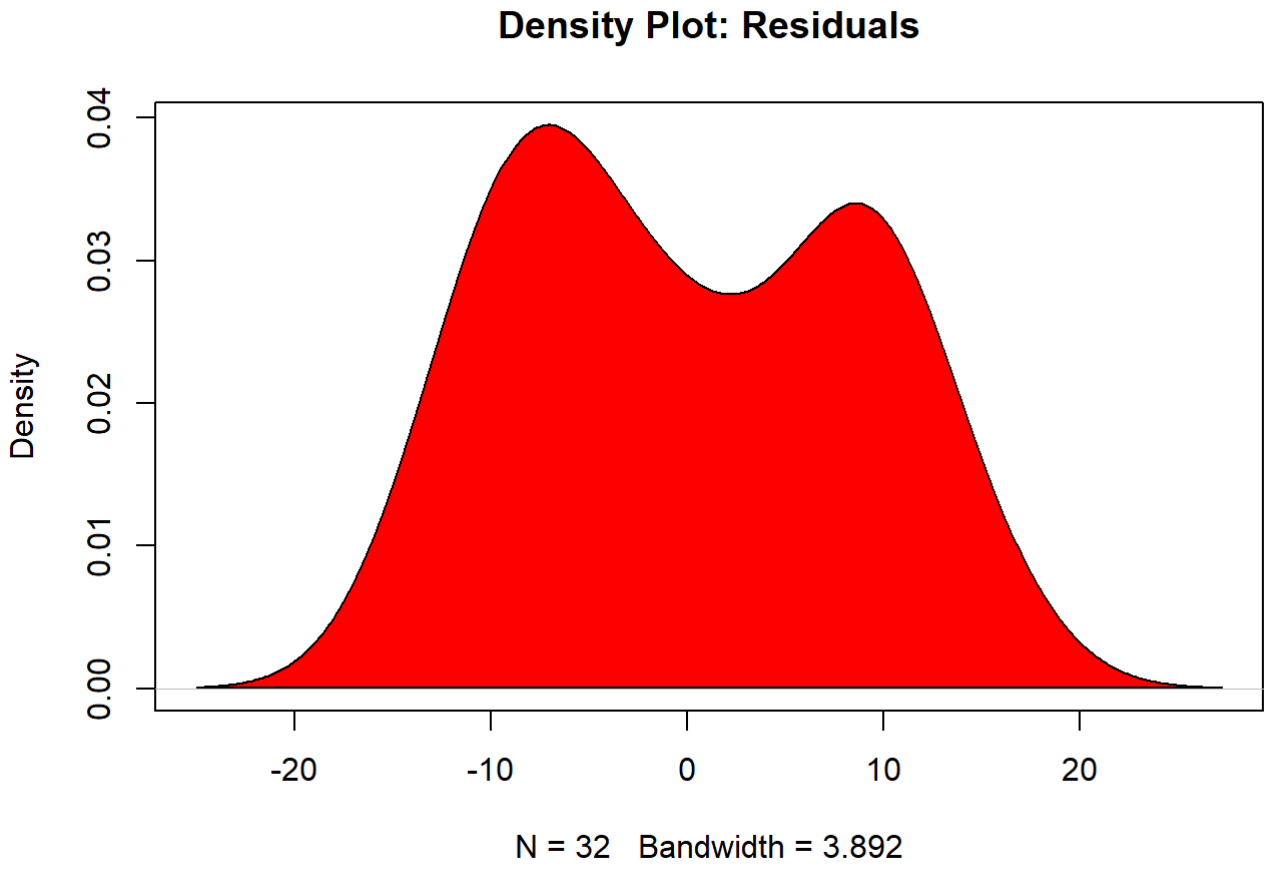
Based on the plot of the residuals, we can see that the residuals don't seem to have a constant variance about the mean and hence dont satisfy the condition of Homoskedasticity.

To perform model inference (T-test, F-test, CI, PI) the errors must be normally distributed or at least approximately normally distributed. We can check that by looking at the box plot and density plot of the residuals.

```
par(mfrow=c(1, 1))
boxplot(residuals(linearModel), main="Residuals")
```

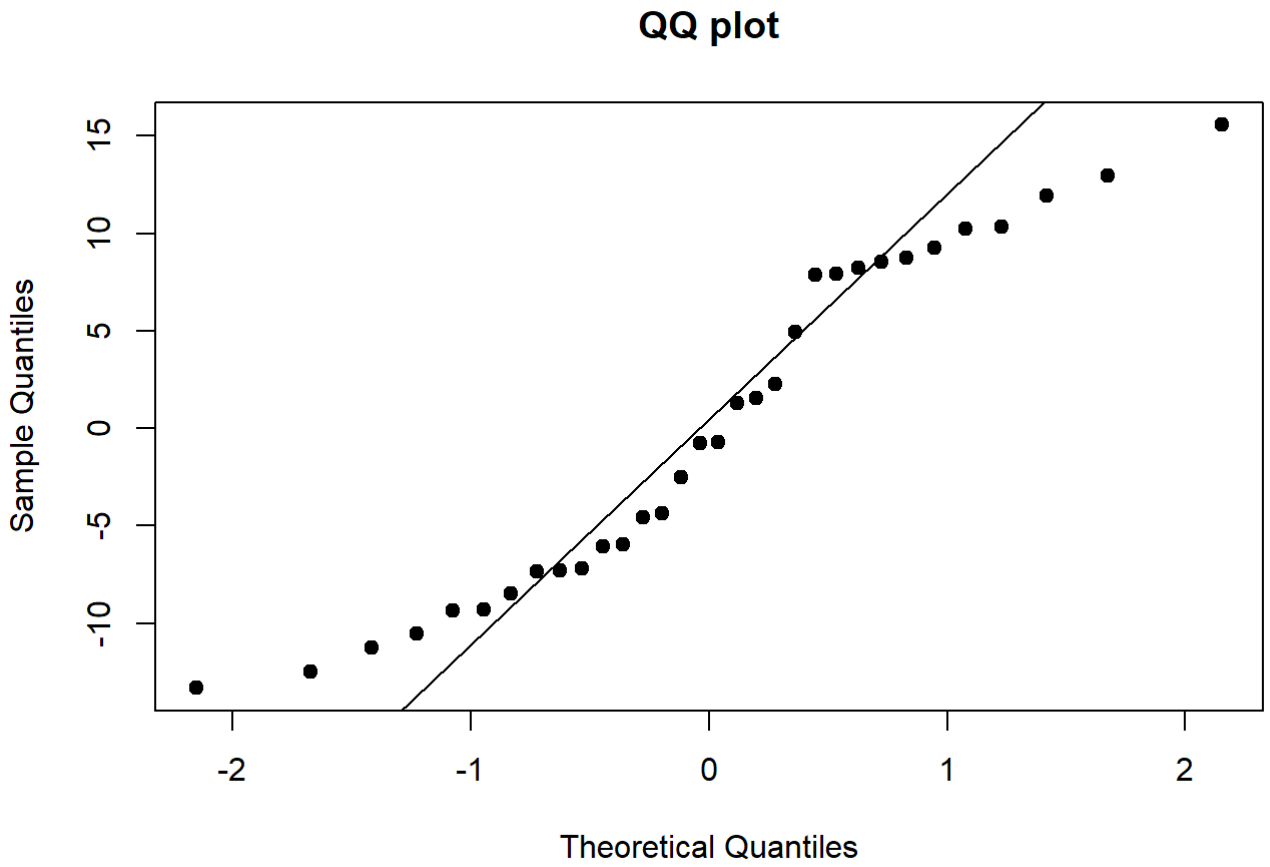



```
plot(density(residuals(linearModel)),
main = "Density Plot: Residuals")
polygon(density(residuals(linearModel)), col="red")
```



We can clearly see that the errors don't seem to follow a normal distribution indicating that there's no Normality of errors. We can further confirm this by plotting the qqnorm and perform Shapiro Wilk normality test.

```
qqnorm(residuals(linearModel),main="QQ plot",pch=19)
qqline(residuals(linearModel))
```




```
shapiro.test(residuals(linearModel))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(linearModel)
## W = 0.93299, p-value = 0.04748
```

Based on the qqnorm plot, we can see that the residuals are not lined well on the straight dashed line indicating that we cannot assume normality.

Similarly, based on Shapiro-Wilk test, The p-value 0.047 < 0.05 implying that the distribution of the data is different from a normal distribution. So, we cannot assume normality.

Thus, based on all of our findings, we can clearly see that the residuals don't satisfy the assumptions required for small sample inference.

Q3) Based on the information in Q2. How could you use the other information in the dataset to potentially improve your simple linear regression model?

We can look at the correlation between other metrics in the data to understand if there are other metrics that play a stronger role in P_Content. Since we are looking only at female children, we can only consider 3 columns (Year, P_Content, Age) as the Sex column is irrelevant (it will only have 1 value - Female)

We have already seen that there's no clear relationship between P_Content and Year. Let's look at the correlation with Age and see if there's any impact.

```
life_satisfaction_new <- life_satisfaction

life_satisfaction_new$Age_Updated <- with(life_satisfaction_new, substr(Age, start = 1, stop
 = 2))

life_satisfaction_new = subset(life_satisfaction_new, select = -c(Age))

life_satisfaction_new <- transform(life_satisfaction_new, Age_Updated = as.numeric(Age_Update
d))

head(life_satisfaction_new)
```

```
##   Year    Sex P_Content Age_Updated
## 1 2006   Male     67.72          10
## 2 2006 Female     69.39          10
## 3 2010   Male     78.41          10
## 4 2010 Female     71.77          10
## 5 2014   Male     76.70          10
## 6 2014 Female     77.64          10
```

```
unique(life_satisfaction_new$Age_Updated)
```

```
## [1] 10 11 12 13 14 15 16 17
```

```
female_new <- life_satisfaction_new[life_satisfaction_new$Sex == 'Female', c('Year','P_Conten
t','Age_Updated')]
cor(female_new)
```

```
##                   Year  P_Content Age_Updated
## Year         1.0000000 -0.1528523   0.0000000
## P_Content   -0.1528523  1.0000000  -0.9211129
## Age_Updated  0.0000000 -0.9211129   1.0000000
```

As we can see, there's a high negative correlation (-0.92) indicating a strong negative linear relationship between Age and P_Content. This implies that as the Age increases, the P_Content decreases. So, in order to improve our linear regression model, we can create a model based on P_Content and Age instead of using Year. This would help us to predict future patterns in a better way.