Trabalho Probabilidade e Estatística 2.0
Professor: Petrúcio Barros
Alunos: Lívia Soares e Pedro Isidoro
Matrículas: 20112760 e 20212161

1) **Utilizando a coluna NOTA_ENEM, elaborar um intervalo de confiança com 5% e 1% de nível de significância:**
   - **5%**

```
#5%
alfa = 0.05
n = length(ENEM$NOTA_ENEN)
n
desvio = sd(ENEM$NOTA_ENEN)
desvio
media = mean(ENEM$NOTA_ENEN)
media

tc = qt(p = 1- alfa/2, df = n - 1)
tc = round(tc, 3)
tc

erro = tc * desvio/sqrt(n)
erro = round(erro, 3)
erro

cat('[',media - erro, ',',media + erro, ']')
```

```
[ 508.5722 , 509.7162 ]
```

   - **1%**

```
#1%

alfa = 0.01
n = length(ENEM$NOTA_ENEN)
n
desvio = sd(ENEM$NOTA_ENEN)
desvio
media = mean(ENEM$NOTA_ENEN)
media

tc = qt(p = 1- alfa/2, df = n - 1)
tc = round(tc, 3)
tc

erro = tc * desvio/sqrt(n)
erro = round(erro, 3)
erro

cat('[',media - erro, ',',media + erro, ']')
```

```
[ 508.5722 , 509.7162 ]
```

2) **Questão: Utilizando a base de dados, escolha um colégio, utilizando o atributo CO_ESCOLA.**

- **Faça um teste de normalidade para as notas das disciplinas do ENEM.**

```
#criando um data frame pra escola 27049140
ESCOLASdf <- data.frame(ENEM)
codigo <- c(27049140)
select.escola <- subset(ESCOLASdf, `CO_ESCOLA` %in% codigo)
select.escola

#shapiro teste
shapiro.test(select.escola$NU_NOTA_CN)
```

```
        Shapiro-Wilk normality test

data:  select.escola$NU_NOTA_CN
W = 0.95379, p-value = 0.1279
```

```
shapiro.test(select.escola$NU_NOTA_CH)

        Shapiro-Wilk normality test

data:  select.escola$NU_NOTA_CH
W = 0.96675, p-value = 0.3274
```

```
shapiro.test(select.escola$NU_NOTA_LC)

        Shapiro-Wilk normality test

data:  select.escola$NU_NOTA_LC
W = 0.95104, p-value = 0.1043
```

```
shapiro.test(select.escola$NU_NOTA_MT)

        Shapiro-Wilk normality test

data:  select.escola$NU_NOTA_MT
W = 0.9586, p-value = 0.1824
```

```
shapiro.test(select.escola$NU_NOTA_REDACAO)

        Shapiro-Wilk normality test

data:  select.escola$NU_NOTA_REDACAO
W = 0.95583, p-value = 0.1486
```

- **As notas que não passarem no teste de normalidade aplicar uma transformação para os dados (z-score ou raiz quadrada) e verificar se passam no teste de normalidade.**

  Todas as notas passaram no teste de normalidade, logo não será necessário.

- **Para as que atenderem ao critério de normalidade elabora testes de hipóteses para verificar se existem diferenças estatísticas entre as notas entre homens e mulheres.**

```
if(!require(RVAideMemoire)) install.packages("RVAideMemoire") # Instalação do pacote caso não esteja instala
library(RVAideMemoire)                                        # Carregamento do pacote
if(!require(car)) install.packages("car") # Instalação do pacote caso não esteja instalado
library(car)                              # Carregamento do pacote

#separando as meninas
SEXOdf <- data.frame(select.escola)
cod <- c('Feminino')
meninas <- subset(SEXOdf, `TP_SEXO` %in% cod)
meninas

#separando os meninos
cod2 <- c('Masculino')
meninos <- subset(SEXOdf, `TP_SEXO` %in% cod2)
meninos

#SHAPIRO PRA OS SEXOS
byf.shapiro(NU_NOTA_CN ~  TP_SEXO, select.escola)
```

```
        Shapiro-Wilk normality tests

data:  NU_NOTA_CN by TP_SEXO

               W p-value
Feminino  0.9638 0.73115
Masculino 0.8812 0.01535 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
byf.shapiro(NU_NOTA_CH ~  TP_SEXO, select.escola)
```

```
        Shapiro-Wilk normality tests

data:  NU_NOTA_CH by TP_SEXO

               W p-value
Feminino  0.9351  0.2930
Masculino 0.9352  0.1747
```

```
byf.shapiro(NU_NOTA_LC ~  TP_SEXO, select.escola)
```

```
        Shapiro-Wilk normality tests

data:  NU_NOTA_LC by TP_SEXO

               W p-value
Feminino  0.9545  0.5646
Masculino 0.9360  0.1818
```

```
byf.shapiro(NU_NOTA_MT ~  TP_SEXO, select.escola)
```

```
        Shapiro-Wilk normality tests

data:  NU_NOTA_MT by TP_SEXO

               W p-value
Feminino  0.9578  0.6228
Masculino 0.9295  0.1347
```

```
byf.shapiro(NU_NOTA_REDACAO ~  TP_SEXO, select.escola)

        Shapiro-Wilk normality tests

data:  NU_NOTA_REDACAO by TP_SEXO

               W p-value
Feminino  0.9458  0.4268
Masculino 0.9310  0.1442
```

p-value maior que 0.05 em todos, logo normal.

```
#verifica homogeneidade de variancias
leveneTest(NU_NOTA_CN ~ TP_SEXO, select.escola , center=mean)
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value Pr(>F)
group  1  0.2011 0.6566
      35

leveneTest(NU_NOTA_CH ~ TP_SEXO, select.escola , center=mean)
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value Pr(>F)
group  1  0.1627 0.6892
      35

leveneTest(NU_NOTA_LC ~ TP_SEXO, select.escola , center=mean)
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value Pr(>F)
group  1  1.2565 0.2699
      35

leveneTest(NU_NOTA_MT ~ TP_SEXO, select.escola , center=mean)
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value Pr(>F)
group  1  1.1872 0.2834
      35

leveneTest(NU_NOTA_REDACAO ~ TP_SEXO, select.escola , center=mean)
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value Pr(>F)
group  1  1.7576 0.1935
      35
```

p-value maior que 0.05 em todos os casos, então vamos fazer o teste-t

```r
#REALIZA O TESTE-T
t.test(NU_NOTA_CN ~ TP_SEXO, select.escola, var.equal=TRUE)
```

```
        Two Sample t-test

data:  NU_NOTA_CN by TP_SEXO
t = 1.6238, df = 35, p-value = 0.1134
alternative hypothesis: true difference in means between group Feminino and group Masculino is not equal to 0
95 percent confidence interval:
 -9.112579 81.942341
sample estimates:
 mean in group Feminino mean in group Masculino
               568.6625                532.2476
```

```r
t.test(NU_NOTA_CH ~ TP_SEXO, select.escola, var.equal=TRUE)
```

```
        Two Sample t-test

data:  NU_NOTA_CH by TP_SEXO
t = 1.272, df = 35, p-value = 0.2118
alternative hypothesis: true difference in means between group Feminino and group Masculino is not equal to 0
95 percent confidence interval:
 -17.96580  78.25746
sample estimates:
 mean in group Feminino mean in group Masculino
               606.0125                575.8667
```

```r
t.test(NU_NOTA_LC ~ TP_SEXO, select.escola, var.equal=TRUE)
```

```
        Two Sample t-test

data:  NU_NOTA_LC by TP_SEXO
t = 1.5823, df = 35, p-value = 0.1226
alternative hypothesis: true difference in means between group Feminino and group Masculino is not equal to 0
95 percent confidence interval:
 -6.580759 53.081950
sample estimates:
 mean in group Feminino mean in group Masculino
               588.3125                565.0619
```

```r
t.test(NU_NOTA_MT ~ TP_SEXO, select.escola, var.equal=TRUE)
```

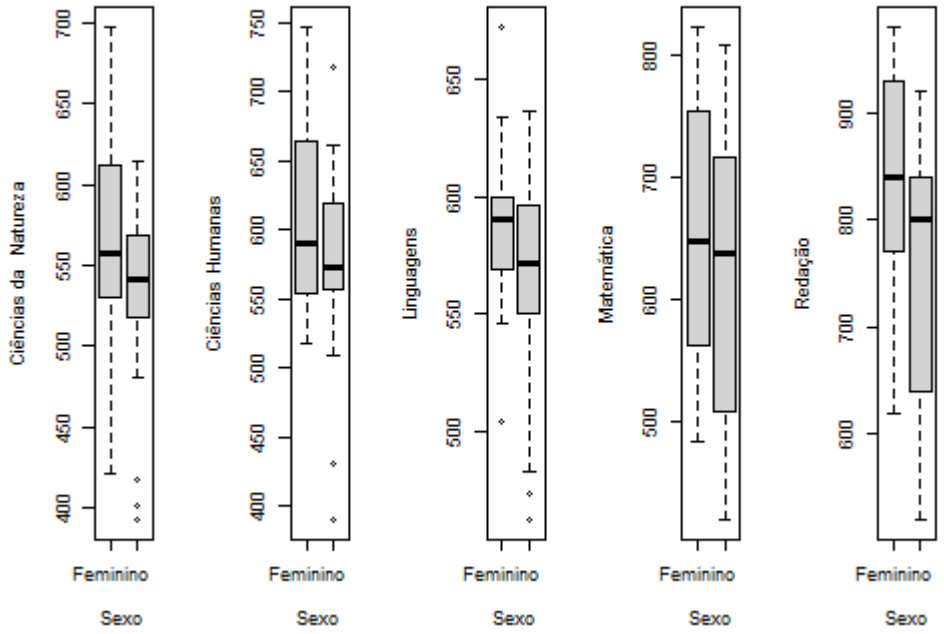
```
        Two Sample t-test

data:  NU_NOTA_MT by TP_SEXO
t = 1.1042, df = 35, p-value = 0.277
alternative hypothesis: true difference in means between group Feminino and group Masculino is not equal to 0
95 percent confidence interval:
 -35.37613 119.76065
sample estimates:
 mean in group Feminino mean in group Masculino
               655.8875                613.6952
```

```r
t.test(NU_NOTA_REDACAO ~ TP_SEXO, select.escola, var.equal=TRUE)
```

```
        Two Sample t-test

data:  NU_NOTA_REDACAO by TP_SEXO
t = 2.6379, df = 35, p-value = 0.01236
alternative hypothesis: true difference in means between group Feminino and group Masculino is not equal to 0
95 percent confidence interval:
  22.66992 174.11579
sample estimates:
 mean in group Feminino mean in group Masculino
               841.2500                742.8571
```

- **Elaborar gráficos box-plot para ajudar nas conclusões.**

```
### letra d
#Trabalho Probabilidade e Estatística Maria Calado e  Henrique Mesquita

par(mfrow=c(1,5)) # Estabeleci que quero que os gráficos saiam na mesma linha
boxplot(NU_NOTA_CN ~ TP_SEXO, data = select.escola, ylab="Ciências da  Natureza", xlab="Sexo")
boxplot(NU_NOTA_CH ~ TP_SEXO, data = select.escola, ylab="Ciências Humanas", xlab="Sexo")
boxplot(NU_NOTA_LC ~ TP_SEXO, data = select.escola, ylab="Linguagens", xlab="Sexo")
boxplot(NU_NOTA_MT ~ TP_SEXO, data = select.escola, ylab="Matemática", xlab="Sexo")
boxplot(NU_NOTA_REDACAO ~ TP_SEXO, data = select.escola, ylab="Redação", xlab="Sexo")
```



- **Elaborar uma conclusão para os dados apresentados.**

  As variâncias foram homogêneas para todas as notas(Linguagens, Redação, Matemática, Ciências humanas e Ciências da natureza).

  Para as disciplinas de linguagens, matemática, ciências humanas e ciências da natureza o valor de p no teste-t foi maior que 0.05, logo podemos concluir que prevalece a hipótese nula(H0), sendo assim não há diferença entre as médias para os dois grupos.

  Em contrapartida, os resultados em relação a nota de redação diferem das outras disciplinas. Como o p-value foi menor que 0.05 iremos considerar a hipótese alternativa(H1). Existe diferença na média de notas entre os dois grupos para redação. As meninas apresentaram, em média, notas de redação superiores às do grupo dos meninos.

3) **Em apenas 26 municípios alagoano houve provas do ENEM 2019. Calcule as médias das notas por municípios e utilizando a classificação em mesorregiões do estado (Leste, Agreste e Sertão), associe cada município a sua região e proceda uma Análise de Variância e verifique se existe diferenças entre médias de notas por região, informe as mesorregiões que diferem e faça gráficos que contribuam para fundamentar sua conclusão.**
Veja os municípios das mesorregiões no link:
https://www.brasilchannel.com.br/municipios/index.asp?nome=Alagoas&regiao=Lese

```
### leste de Alagoas
cidades1 <- c("Chã Preta", "Santana do Mundaú", "Ibateguara", "Viçosa", "São José da Laje", "Atalaia", "Camp
regiao1 <- subset(ESCOLASdf, `NO_MUNICIPIO_PROVA` %in% cidades1)

### Sertão de Alagoas
cidades2 <- c("Água Branca", "Mata Grande", "Canapi", "Pariconha", "Inhapi", "Delmiro Gouveia", "Olho d'Agua
regiao2 <- subset(ESCOLASdf, `NO_MUNICIPIO_PROVA` %in% cidades2)

### Agreste de Alagoas
cidades3 <- c("Belém", "Igaci", "Minerador do Negrão", "Quebrangulo", "Cacimbinhas", "Mar Vermelho", "Palmei
regiao3 <- subset(ESCOLASdf, `NO_MUNICIPIO_PROVA` %in% cidades3)
```

Região 1 é o leste, 2 é o sertão e 3 é o agreste.

```
media_regiao1 <- mean(regiao1$NOTA_ENEN)
media_regiao1
media_regiao2 <- mean(regiao2$NOTA_ENEN)
media_regiao2
media_regiao3 <- mean(regiao3$NOTA_ENEN)
media_regiao3
```

media_regiao1	512.060154919369
media_regiao2	489.316831587429
media_regiao3	508.745269695067

```
varia_regiao1 <- var(regiao1$NOTA_ENEN)
varia_regiao1
varia_regiao2 <- var(regiao2$NOTA_ENEN)
varia_regiao2
varia_regiao3 <- var(regiao3$NOTA_ENEN)
varia_regiao3
```

varia_regiao1	5680.99585648661
varia_regiao2	3843.62888312524
varia_regiao3	5469.50880287736

```
#TESTE DE VARIÂNCIA LESTE(1) X SERTÃO(2)
F.R1R2 <- varia_regiao1/varia_regiao2
var.test(regiao1$NOTA_ENEN, regiao2$NOTA_ENEN)
```

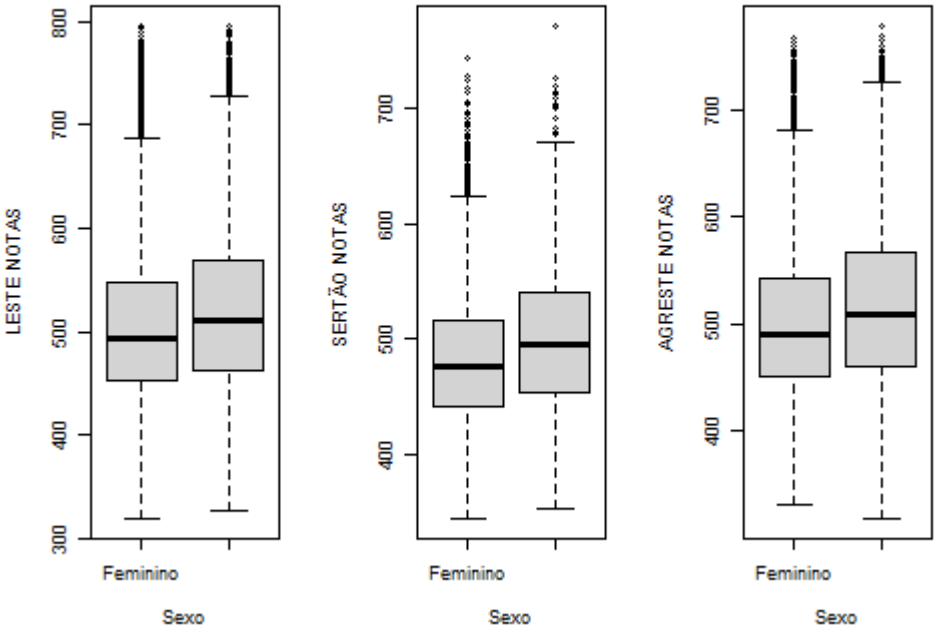
```
        F test to compare two variances

data:  regiao1$NOTA_ENEN and regiao2$NOTA_ENEN
F = 1.478, num df = 44151, denom df = 6204, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.423125 1.534224
sample estimates:
ratio of variances
          1.478029
```

```
#TESTE DE VARIÂNCIA LESTE(1) X AGRESTE(3)
F.R1R3 <- varia_regiao1/varia_regiao3
var.test(regiao1$NOTA_ENEN, regiao3$NOTA_ENEN)
```

```
        F test to compare two variances

data:  regiao1$NOTA_ENEN and regiao3$NOTA_ENEN
F = 1.0387, num df = 44151, denom df = 14330, p-value = 0.005455
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.011259 1.066621
sample estimates:
ratio of variances
          1.038667
```

```
#TESTE DE VARIÂNCIA SERTÃO(2) X AGRESTE(3)
F.R2R3 <- varia_regiao2/varia_regiao3
var.test(regiao2$NOTA_ENEN, regiao3$NOTA_ENEN)
```

```
        F test to compare two variances

data:  regiao2$NOTA_ENEN and regiao3$NOTA_ENEN
F = 0.70274, num df = 6204, denom df = 14330, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.6738673 0.7331055
sample estimates:
ratio of variances
          0.7027375
```



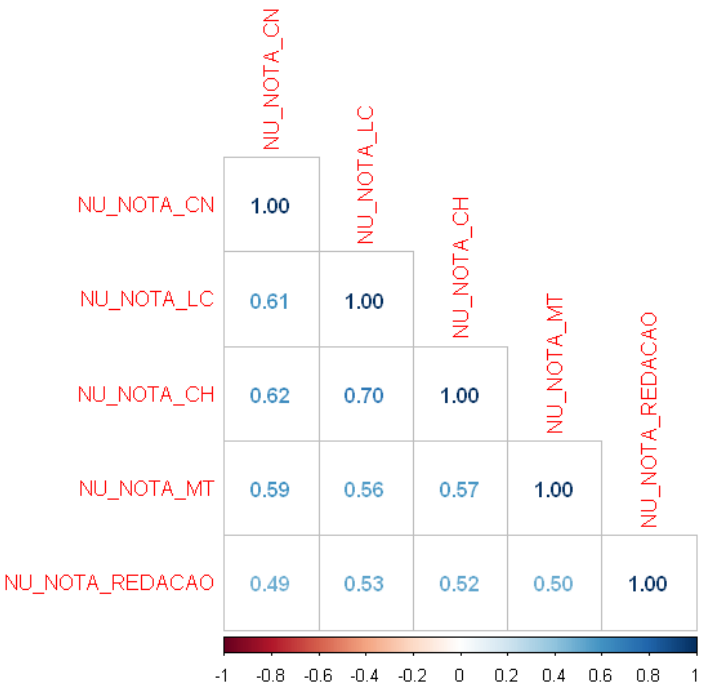
4) **Utilizando as médias dos 26 municípios de Alagoas elabore uma matriz de correlação entre as notas das 5 disciplinas do ENEM. Escolha as duas disciplinas de maior correlação e gerar a equação de regressão, o coeficiente de determinação, elabore um teste de hipótese para validar a correlação e comente os resultados.**

Inicialmente foi chamada as bibliotecas que serão usadas mais pra frente. Em seguida, feito uma separação das colunas desejadas e estas foram colocadas em uma tabela "Notas". Logo em seguida, realizado a correção de 2 em 2 das colunas e assim gerado a matriz:

```
library(corrplot)

Notas <- ENEM[,c('NU_NOTA_CN','NU_NOTA_LC', 'NU_NOTA_CH', 'NU_NOTA_MT','NU_NOTA_REDACAO')]

Matriz <- cor(Notas)
Matriz

corrplot(Matriz, method = "number", type = "lower")
```

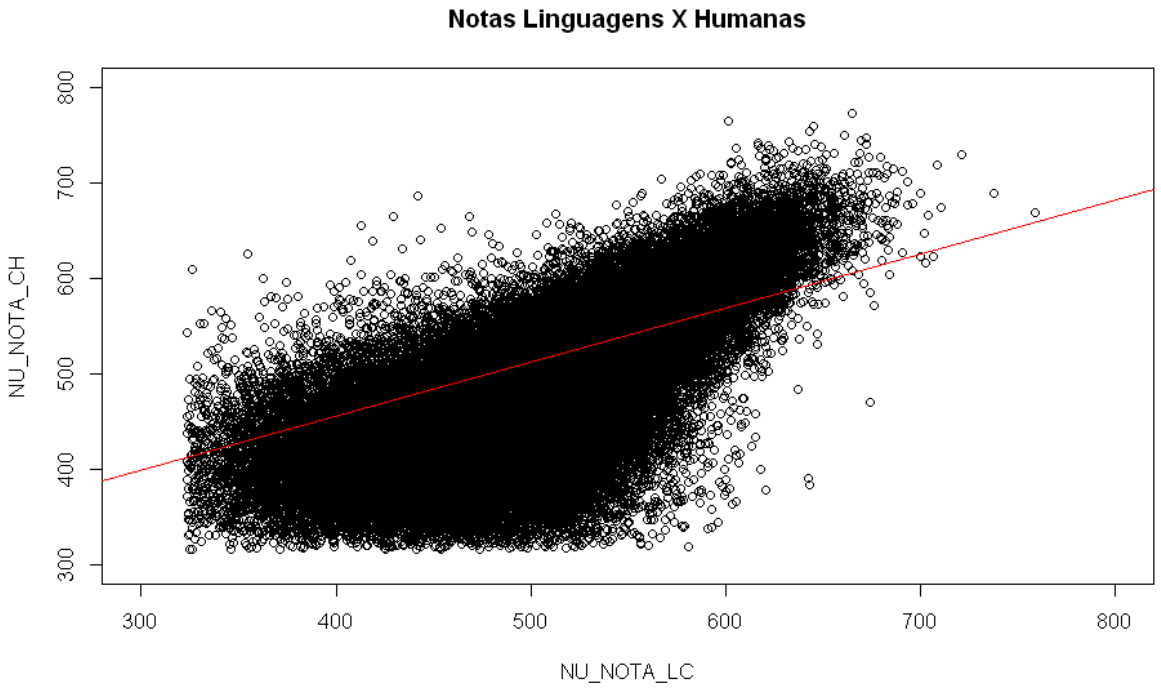
Matriz gerada :

```
> Matriz <- cor(Notas)
> Matriz
               NU_NOTA_CN NU_NOTA_LC NU_NOTA_CH NU_NOTA_MT NU_NOTA_REDACAO
NU_NOTA_CN      1.0000000  0.6052133  0.6244532  0.5933452       0.4856854
NU_NOTA_LC      0.6052133  1.0000000  0.7027909  0.5557151       0.5250297
NU_NOTA_CH      0.6244532  0.7027909  1.0000000  0.5714694       0.5153393
NU_NOTA_MT      0.5933452  0.5557151  0.5714694  1.0000000       0.4952407
NU_NOTA_REDACAO 0.4856854  0.5250297  0.5153393  0.4952407       1.0000000
```



Sendo assim, foi observado que as matérias de maior correlação foram Ciência Humanas e Linguagens. Portanto, foram as escolhidas para gerar a equação de regressão, o coeficiente de determinação e o teste de hipótese.

```r
#Maior correlação entre a NU_NOTA_LC e a NU_NOTA_CH

N_LC_CH <- Notas[,c('NU_NOTA_LC', 'NU_NOTA_CH')]

plot(N_LC_CH, main = "Notas Linguagens X Humanas", xlim=c(300,800),ylim=c(300,800))
regressao <- lm(N_LC_CH$NU_NOTA_LC ~N_LC_CH$NU_NOTA_CH, data = N_LC_CH)
abline(regressao, col = "red")
```



Em seguida, foi chamado o Summary para obter algumas informações dessa regressão Linear formada:

```
> summary(regressao)

Call:
lm(formula = N_LC_CH$NU_NOTA_LC ~ N_LC_CH$NU_NOTA_CH, data = N_LC_CH)

Residuals:
     Min       1Q   Median       3Q      Max
-247.721  -24.012    4.495   28.531  196.509

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.300e+02  1.115e+00   206.3   <2e-16 ***
N_LC_CH$NU_NOTA_CH 5.646e-01  2.247e-03   251.3   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43.93 on 64686 degrees of freedom
Multiple R-squared:  0.4939,    Adjusted R-squared:  0.4939
F-statistic: 6.313e+04 on 1 and 64686 DF,  p-value: < 2.2e-16
```

É possível observar que a equação de regressão é a seguinte:

$$y = 5,646e - 01\,x + 2,300e + 02$$

O coeficiente de determinação é: $0,4939$

E o teste de hipótese usado foi:

```
#Teste de hipótese:

t.test(ENEM$NU_NOTA_LC ~ENEM$TP_SEXO)
t.test(ENEM$NU_NOTA_CH ~ENEM$TP_SEXO)
```

```
> t.test(ENEM$NU_NOTA_LC ~ENEM$TP_SEXO )

        Welch Two Sample t-test

data:  ENEM$NU_NOTA_LC by ENEM$TP_SEXO
t = -10.61, df = 53182, p-value < 2.2e-16
alternative hypothesis: true difference in means between group Feminino and group Masculino is not equal to 0
95 percent confidence interval:
 -6.295527 -4.332203
sample estimates:
 mean in group Feminino mean in group Masculino
               504.6724                509.9863
```

```
> t.test(ENEM$NU_NOTA_CH ~ENEM$TP_SEXO)

        Welch Two Sample t-test

data:  ENEM$NU_NOTA_CH by ENEM$TP_SEXO
t = -26.01, df = 52526, p-value < 2.2e-16
alternative hypothesis: true difference in means between group Feminino and group Masculino is not equal to 0
95 percent confidence interval:
 -17.41933 -14.97800
sample estimates:
 mean in group Feminino mean in group Masculino
               483.7902                499.9889
```

H0: A correlação dos gêneros entre as matérias é maior ou igual à correlação entre as matérias

H1: A correlação dos gêneros entre as matérias é menor à correlação entre as matérias