**UNIVERSIDADE PERNAMBUCO**

**CURSO DE EXTENSÃO EM CIÊNCIA DOS DADOS E ANALYTICS**

# ESTATÍSTICA COMPUTACIONAL

**PROF. DRA. ROBERTA FAGUNDES**

**Recife/PE**

**2021**

**Projeto Final de Disciplina**

**Aluno: Breno Luiz Santos Soares**

## 1. O Projeto:

Trata-se da análise estatística de dados da plataforma UCI, a base é denominada "Combined Cycle Power Plant Data Set", encontrada no seguinte endereço da internet https://archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant.

O projeto tem início com a elaboração de um problema vinculado à base de dados com que se trabalha.

Objetiva-se aplicar os conhecimentos adquiridos em sala de aula, inclusive com a utilização da linguagem R, de tal forma que se entenda melhor o problema e que se tenha maior criticidade na análise dos resultados.

## 2. Relatório:

### 2.1 Justificativa:

Uma usina de ciclo combinado (UCB) é composta por turbinas de gás, turbinas de vapor e geradores à vapor de recuperação de calor. Em uma UCB, a eletricidade é gerada por turbinas de gás e a vapor, que são combinadas em um ciclo, e é transferida de uma turbina para outra. O número de variáveis que possuem influência na produção de energia de uma UCB são várias, tais como temperatura, pressão, humidade, dentre outras.

O objetivo desse trabalho é estimar quais são as melhores condições para a produção de energia elétrica em uma UCB. Para tanto, será utilizado o modelo de regressão linear, preferencialmente com variáveis cuja correlação com o valor da produção de energia seja maior.

Por fim, serão comparados alguns modelos utilizando-se testes de hipótese, para ao final tentar concluir qual condição seria a melhor e teria a menor redução na produção de energia elétrica de uma UCB.

**2.2 Base de Dados:**

Cada registro da base de dados é composto de cinco variáveis, que são:

- Temperatura (AT)
- Pressão ambiente (AP)
- Umidade relativa (RH)
- Vácuo de exaustão (V)
- Produção energia (PE)

**2.3 Análise Estatística:**

A programação em R será utilizada para realizar a análise exploratória, formulação de hipóteses e análise dos resultados. O código em R será mostrado em itálico, para diferenciar-se das demais colocações. As saídas do programa, quando relevantes, também serão mostradas.

*#Abrir a base de dados:*

*getwd()*

*setwd("C:\Users\BSOARES\Documents\CCPP")*

*base = read.csv2("base.csv", header=TRUE, sep=",")*

*#Ver como estão os dados, quais os seus tipos:*

*head(dados)*

```
     AT     V      AP    RH     PE
1 14.96 41.76 1024.07 73.17 463.26
2 25.18 62.96 1020.04 59.08 444.37
3  5.11 39.40 1012.16 92.14 488.56
4 20.86 57.32 1010.24 76.64 446.48
5 10.82 37.50 1009.23 96.62 473.90
6 26.27 59.44 1012.23 58.77 443.67
```

*str(dados)*

'data.frame': 9568 obs. of 5 variables:

$ AT: num 14.96 25.18 5.11 20.86 10.82 ...

$ V : num 41.8 63 39.4 57.3 37.5 ...

$ AP: num 1024 1020 1012 1010 1009 ...

$ RH: num 73.2 59.1 92.1 76.6 96.6 ...

$ PE: num 463 444 489 446 474 ...

*#Identificar a amplitude, média e mediana para cada uma das variáveis:*

*> summary(base$AT)*

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.81 13.51 20.34 19.65 25.72 37.11

*> summary(base$V)*

Min. 1st Qu. Median Mean 3rd Qu. Max.

25.36 41.74 52.08 54.31 66.54 81.56

*> summary(base$AP)*

Min. 1st Qu. Median Mean 3rd Qu. Max.

992.9 1009.1 1012.9 1013.3 1017.3 1033.3

*> summary(base$RH)*

Min. 1st Qu. Median Mean 3rd Qu. Max.

25.56 63.33 74.97 73.31 84.83 100.16

*> summary(base$PE)*

Min. 1st Qu. Median Mean 3rd Qu. Max.

420.3 439.8 451.6 454.4 468.4 495.8

A partir dos resultados obtidos acima, é possível observar que todas as variáveis possuem média muito próxima da mediana, sugerindo uma distribuição simétrica. Pode-se observar que na variável umidade relativa (RH) que a maior parte dos dados está concentrada em valores maiores

*#Calcular a variância e desvio padrão para cada uma das variáveis:*

*> var(base$AT)*

[1] 55.53936

*> sd(base$AT)*

[1] 7.452473

*> var(base$V)*

[1] 161.4905

*> sd(base$V)*

[1] 12.70789

*> var(base$AP)*

[1] 35.26915

*> sd(base$AP)*

[1] 5.938784

*> var(base$RH)*

[1] 213.1678

*> sd(base$RH)*

[1] 14.60027

*> sd(base$PE)*

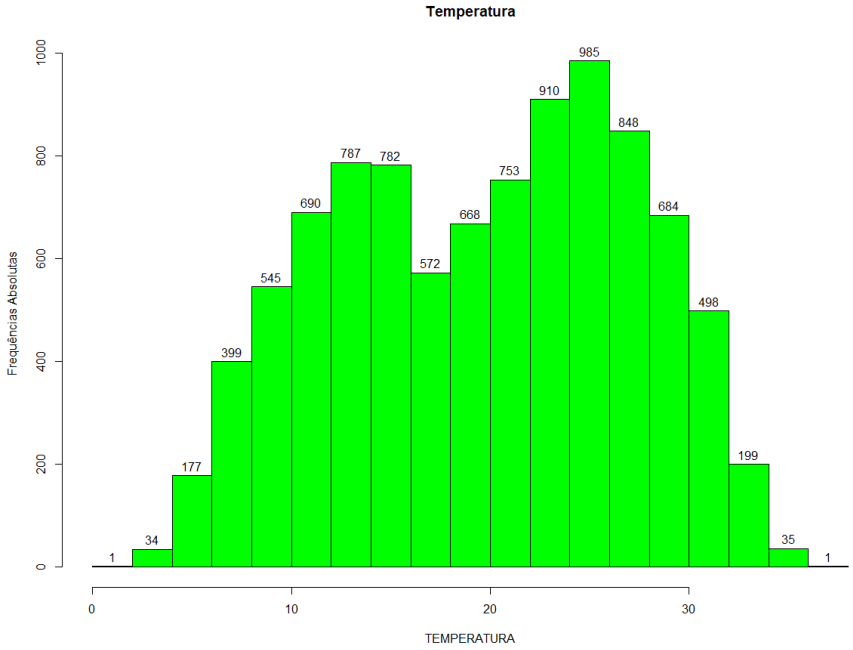
[1] 17.06699

*> var(base$PE)*

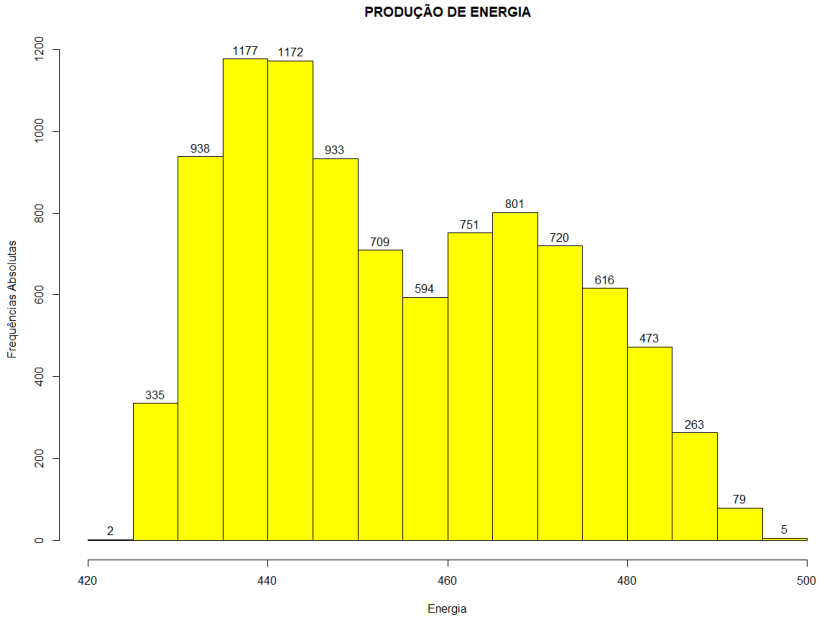
[1] 291.2823

*#Plotar o melhor gráfico para cada uma das variáveis:*

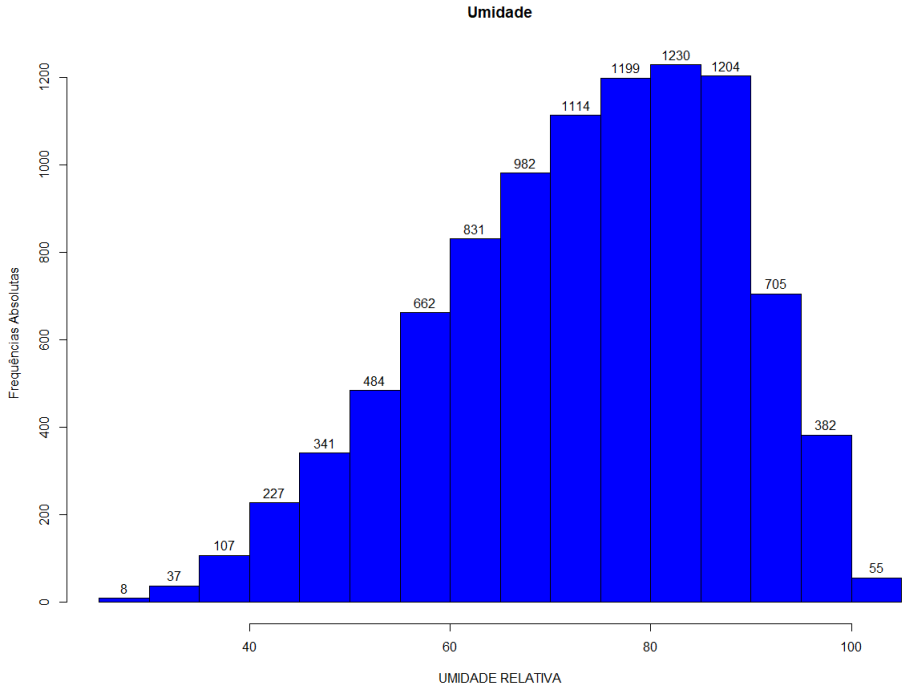
*> hist(base$AT, labels = T, col = "green", main = "Temperatura", xlab = "TEMPERATURA", ylab = "Frequências Absolutas")*



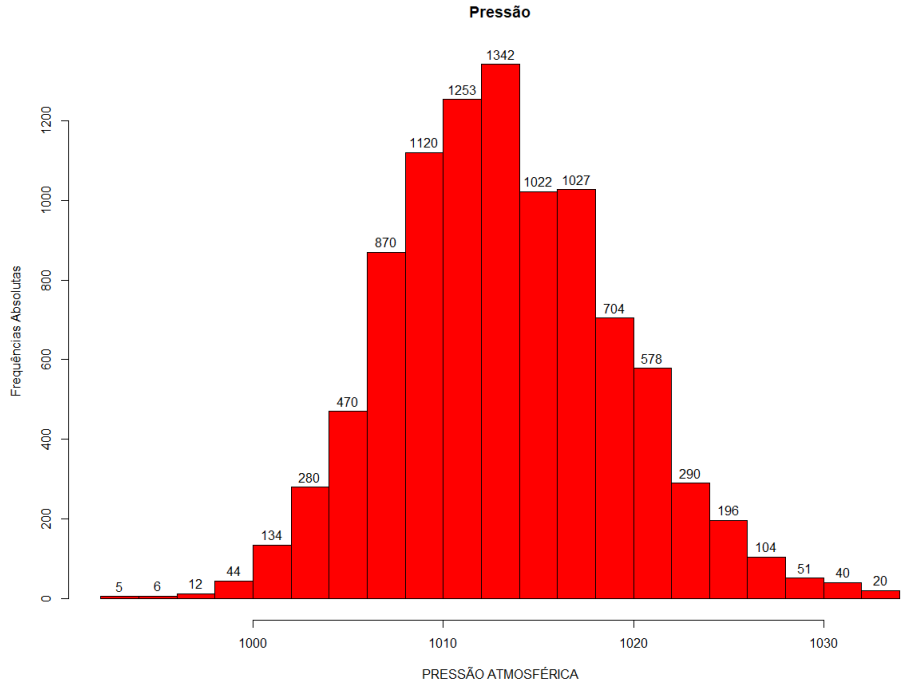

*> hist(base$PE, labels = T, col = "yellow", main = "PRODUÇÃO DE ENERGIA", xlab = "Energia", ylab = "Frequências Absolutas")*



*> hist(base$RH, labels = T, col = "blue", main = "Umidade", xlab = "UMIDADE RELATIVA", ylab = "Frequências Absolutas")*



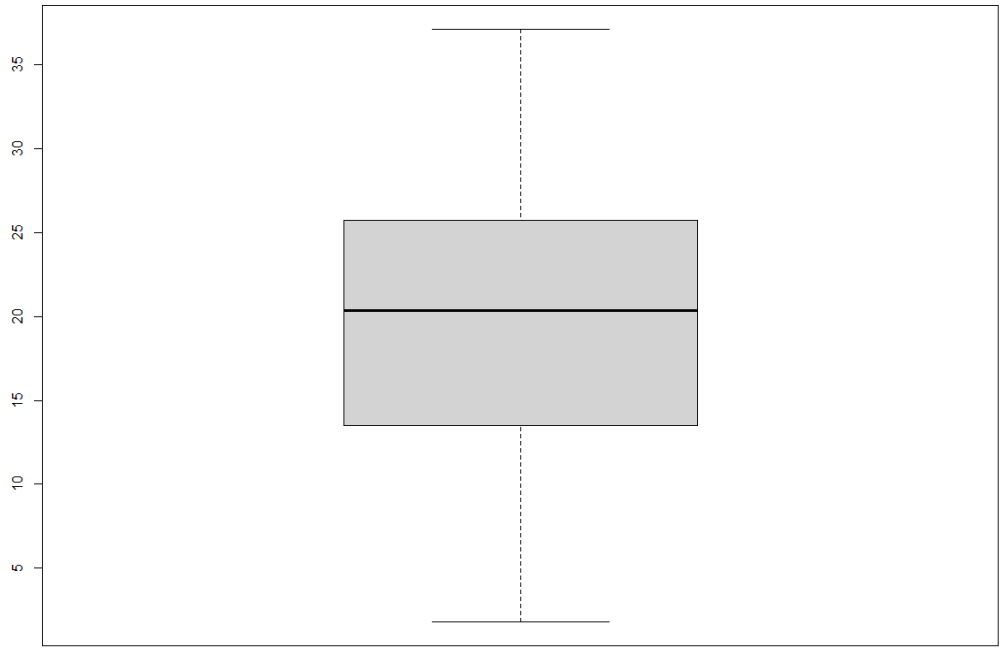

*> hist(base$AP, labels = T, col = "red", main = "Pressão", xlab = "PRESSÃO ATMOSFÉRICA", ylab = "Frequências Absolutas")*



A partir da análise dos gráficos acima, é possível observar que a variável que mais se aproxima de uma distribuição normalizada é a pressão atmosférica (AP), além disso, pode-se observar que a variável umidade relativa (RH) possui maior parte de seus registros concentrados em valores acima da média.

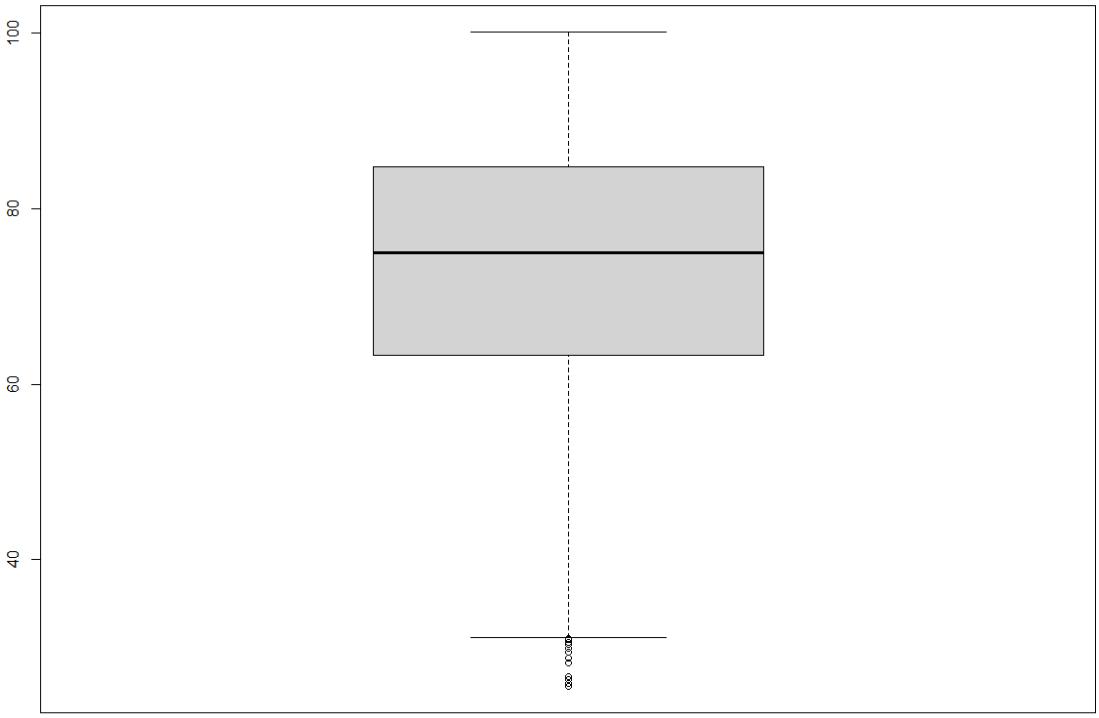
*#Plotar o bloxplot para todas as variáveis quantitativas:*

*> boxplot(base$AT)*



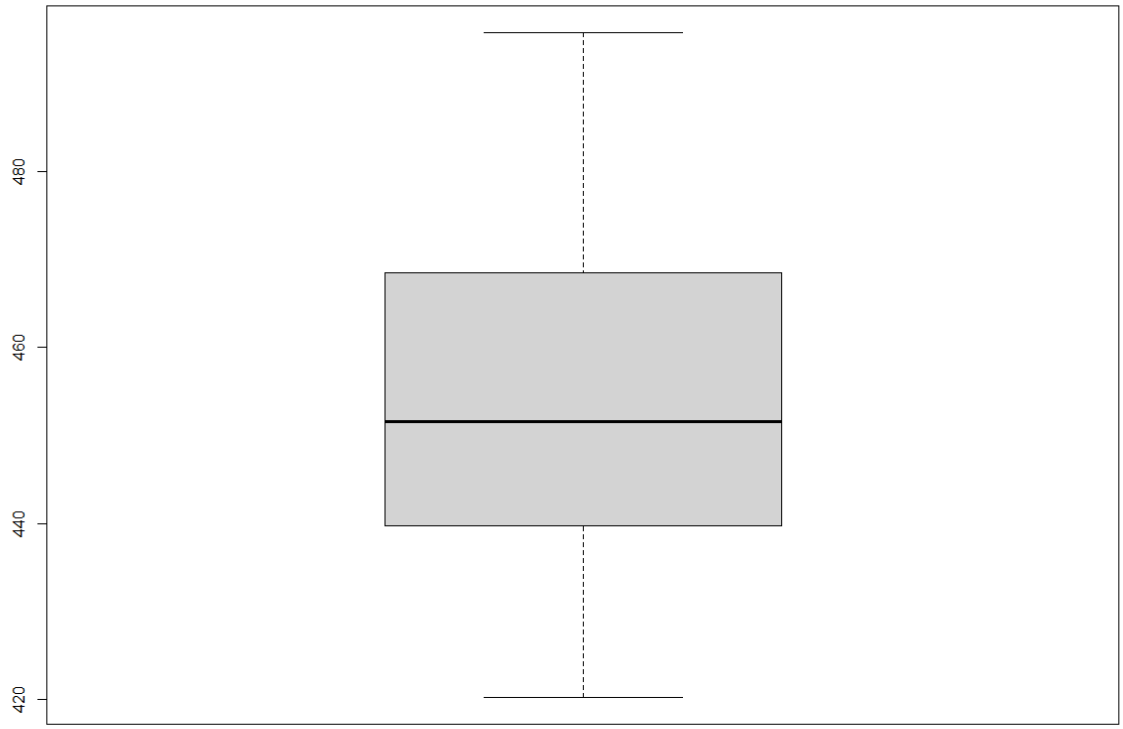

Como foi visto anteriormente, pode-se observar que a mediana da variável temperatura (AT) se aproxima de 20.

*> boxplot(base$RH)*



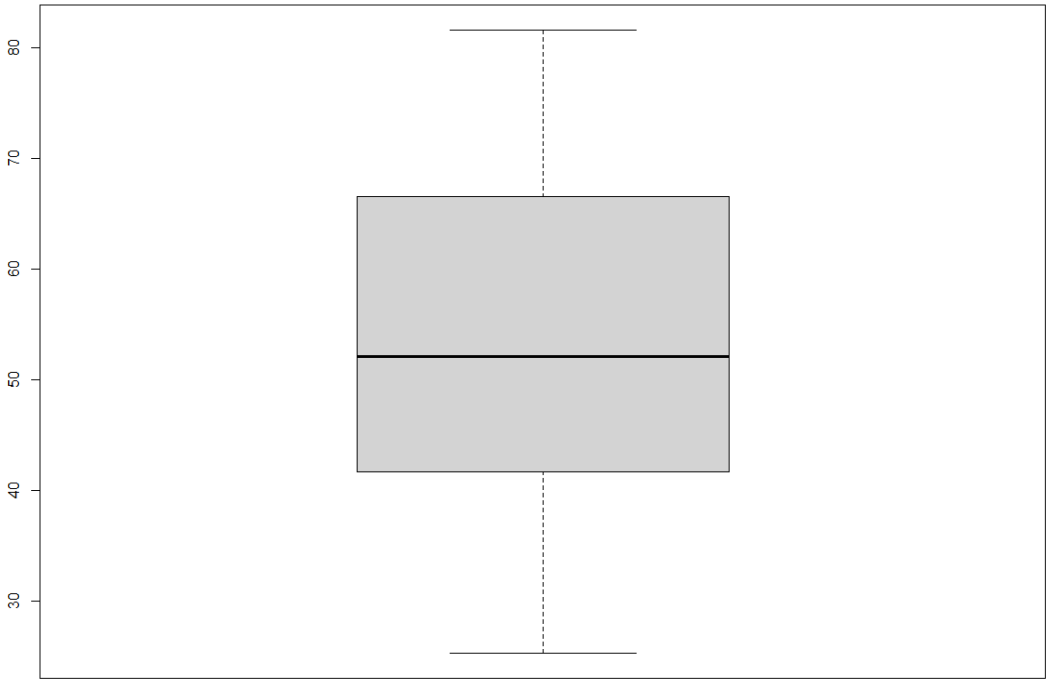

Podemos observar que a mediana está próxima de 80, como foi calculado anteriormente e a sugestão da existência de outliers perto de 50.

*> boxplot(base$PE)*



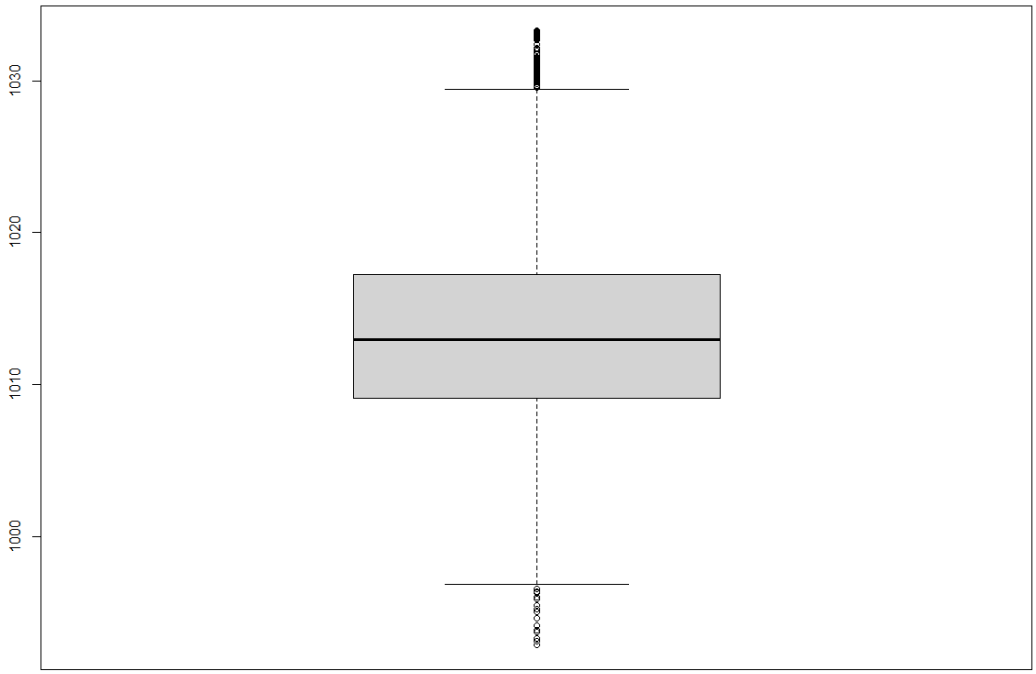

Pode-se observar que a mediana está próxima de 450, como foi visto anteriormente e os dados possuem uma distribuição quase simétrica.

*> boxplot(base$V)*




Como calculado anteriormente, pode-se observar que a mediana está próxima de 50 e esse gráfico sugere que os dados possuem uma destruição quase simétrica.

*> boxplot(base$AP)*



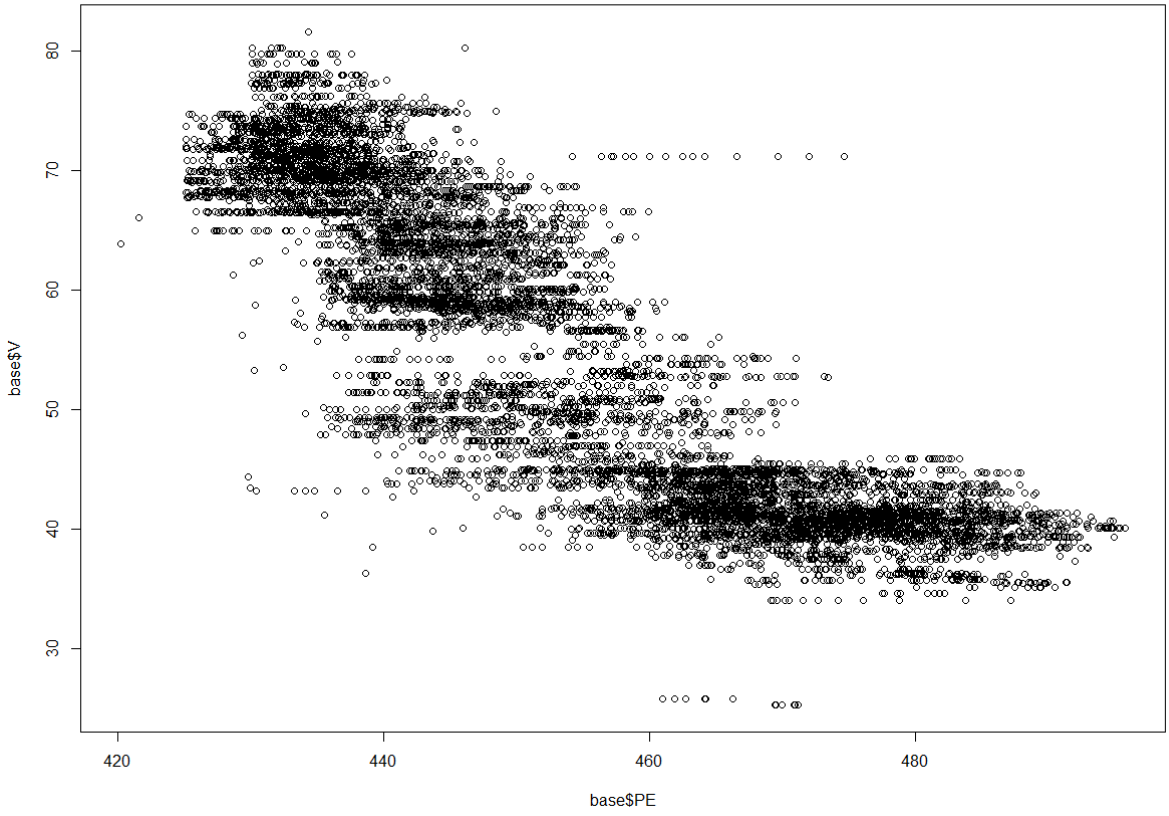

Pode-se observar que nesse gráfico é clara a presença de outliers acima de 1030 e abaixo de 1000 e como viso anteriormente a mediana encontra-se próxima de 1010.

*#Calcular a correlacão e plotar o gráfico de dispersão:*

*> cor(base$PE, base$V)*

[1] -0.8697803

*> plot(base$PE, base$V)*

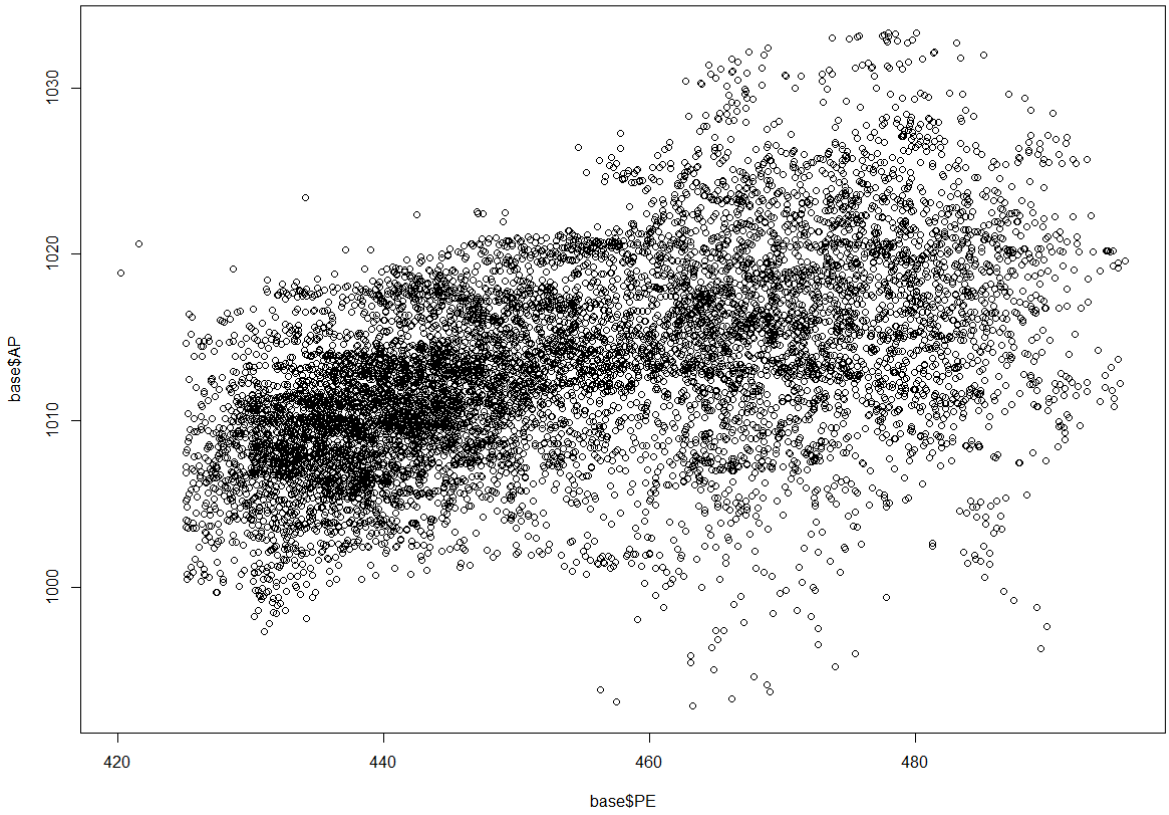



Pode-se observar que a correlação da produção de energia com o vácuo é negativa e indica que as duas variáveis se movem em direções opostas, porém como a correlação está próxima de -1 é possível afirmar que as duas variáveis possuem uma relação forte. Ainda é possível observar no gráfico a presença de outliers.

*> cor(base$PE, base$AP)*

[1] 0.518429

*> plot(base$PE, base$AP)*

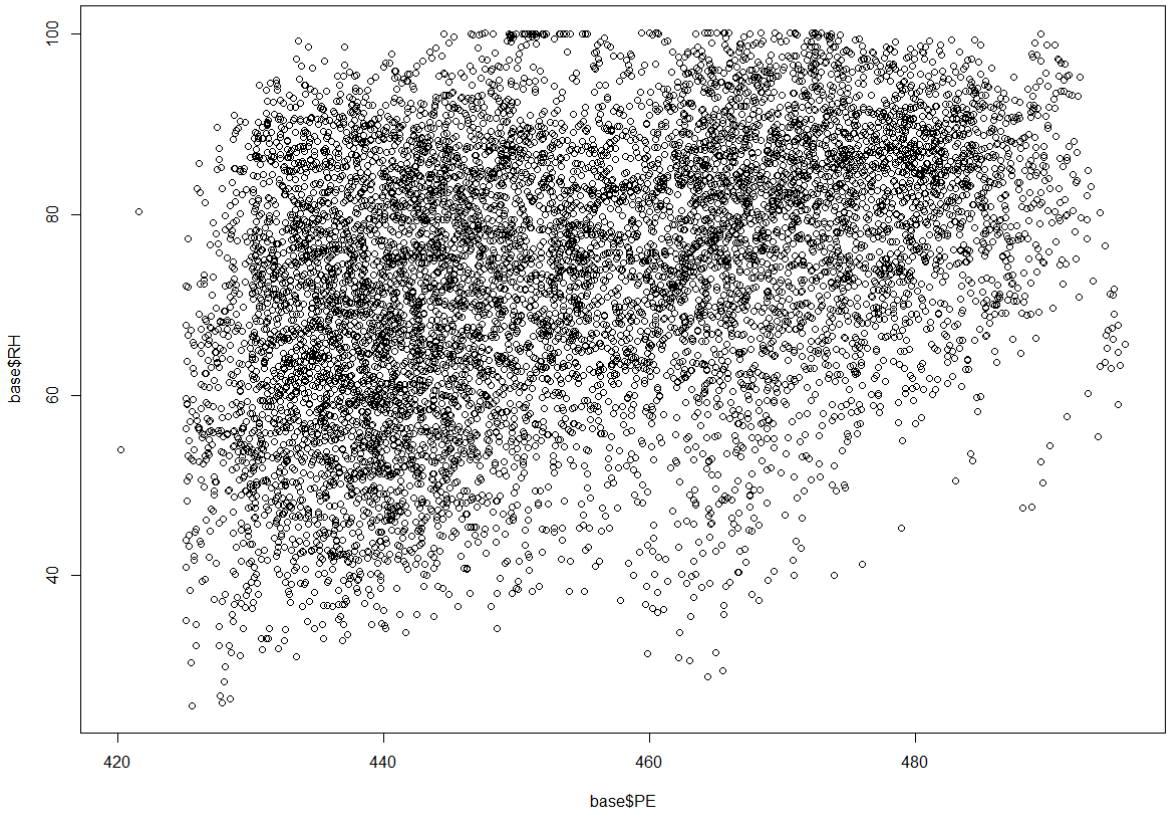



Pode-se observar que a correlação entre a produção de energia e a pressão atmosférica não é tão forte e ainda podemos observar no gráfico uma grande presença de outliers depois dos 1030 e antes de 1000, como foi visto anteriormente no boxplot da variável.

> *cor(base$PE, base$RH)*

[1] 0.3897941

> *plot(base$PE, base$RH)*

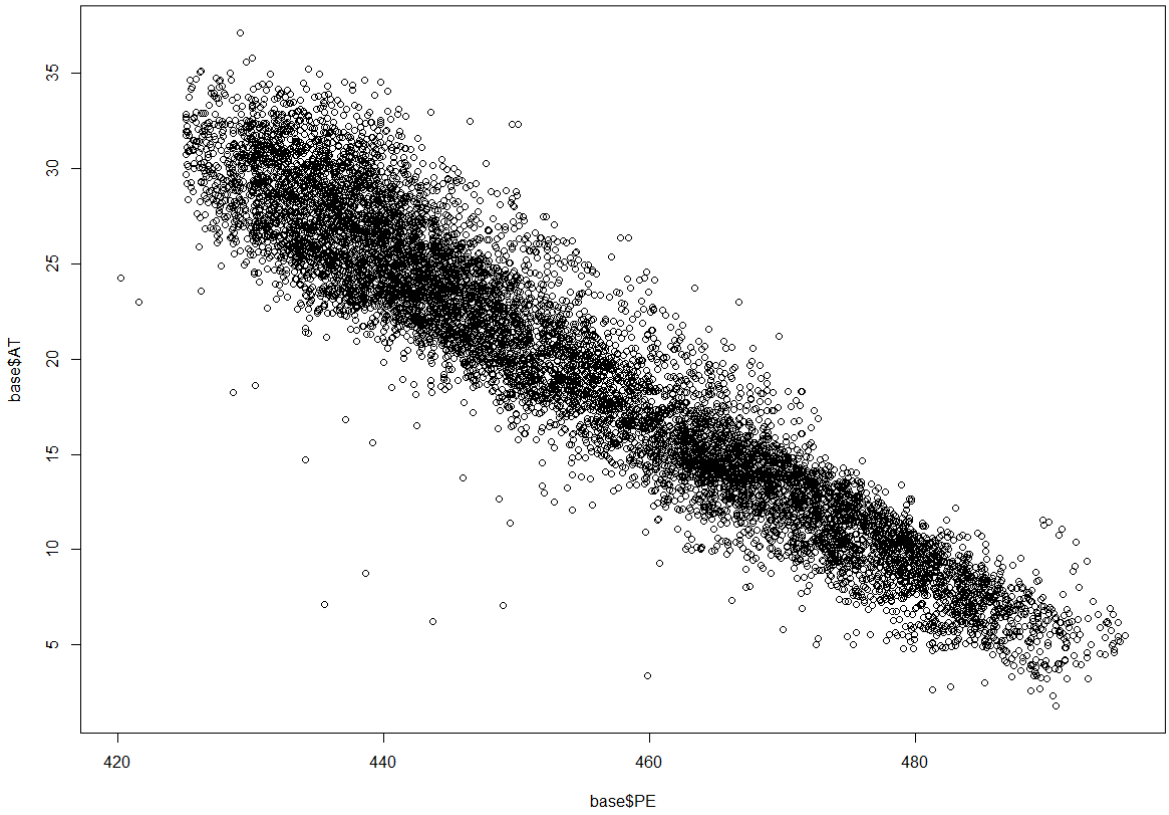



A correlação da produção de energia com a umidade relativa foi ainda mais fraca que a da pressão atmosférica.

*> cor(base$PE, base$AT)*

[1] -0.9481285

*> plot(base$PE, base$AT)*




Observa-se que a correlação entre a produção de energia e a temperatura é fortíssima e negativa, se aproximando bastante de -1. Ou seja, isso indica que a temperatura possui uma grande influência na produção de energia em UCB.

*#Formular um problema de regressão e analisá-lo como também discuti-lo:*

Como foi possível observar no gráfico anterior, pelo cálculo das correlações e dos gráficos de dispersão, percebe-se que a correlação da variável independente com as variáveis dependentes metade são fracas e a outra metade é forte. A correlação mais forte foi com a variável temperatura, como se era de esperar, e a mais fraca foi com a pressão. Porém, será utilizado o método de regressão linear para estimar os valores de produção de energia: primeiramente com a variável temperatura, cuja correlação foi mais forte e depois, com todas as variáveis quantitativas dependentes, para ao final comparar os resultados.

*#Regressão Linear Simples:*

*> #MAIOR CORRELAÇÃO*

*> reg1 = lm(base$AT~base$PE)*

*> a1 = reg1$coefficients[1]*

*> b1 = reg1$coefficients[2]*

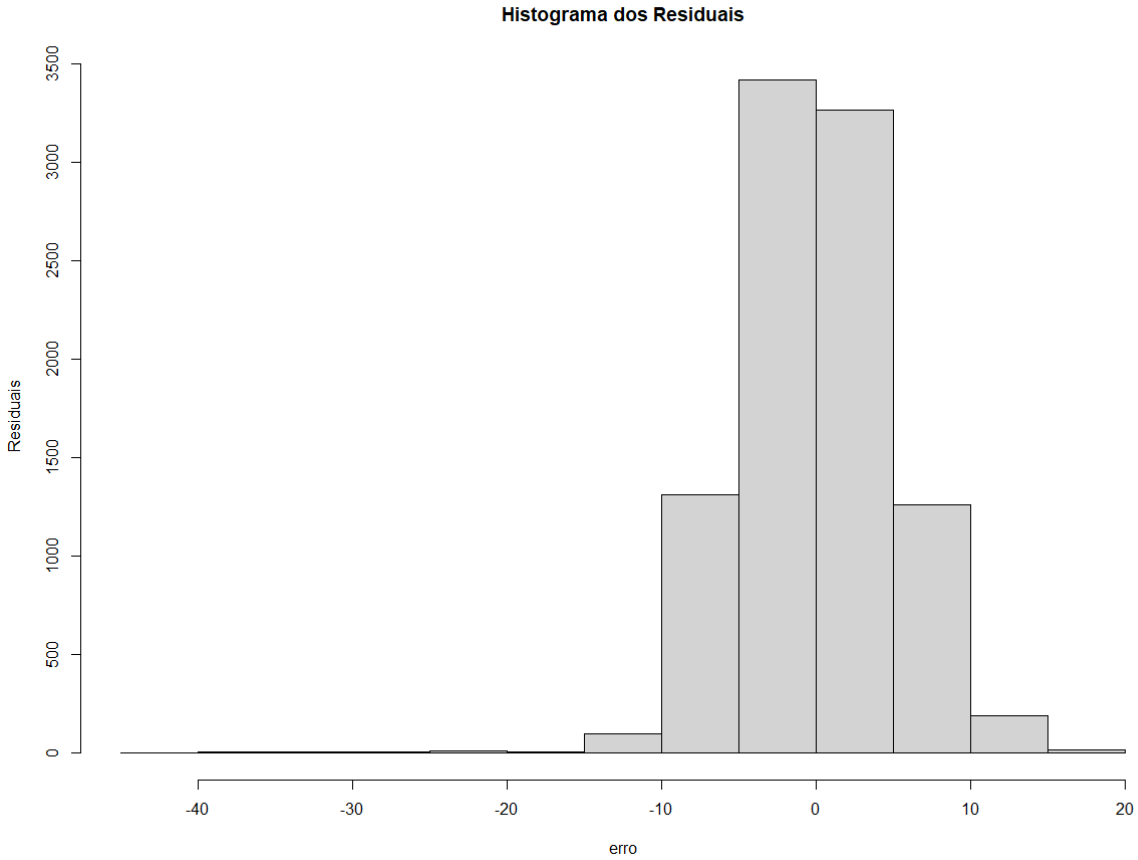
*> ajustados1 = reg1$fitted.values*

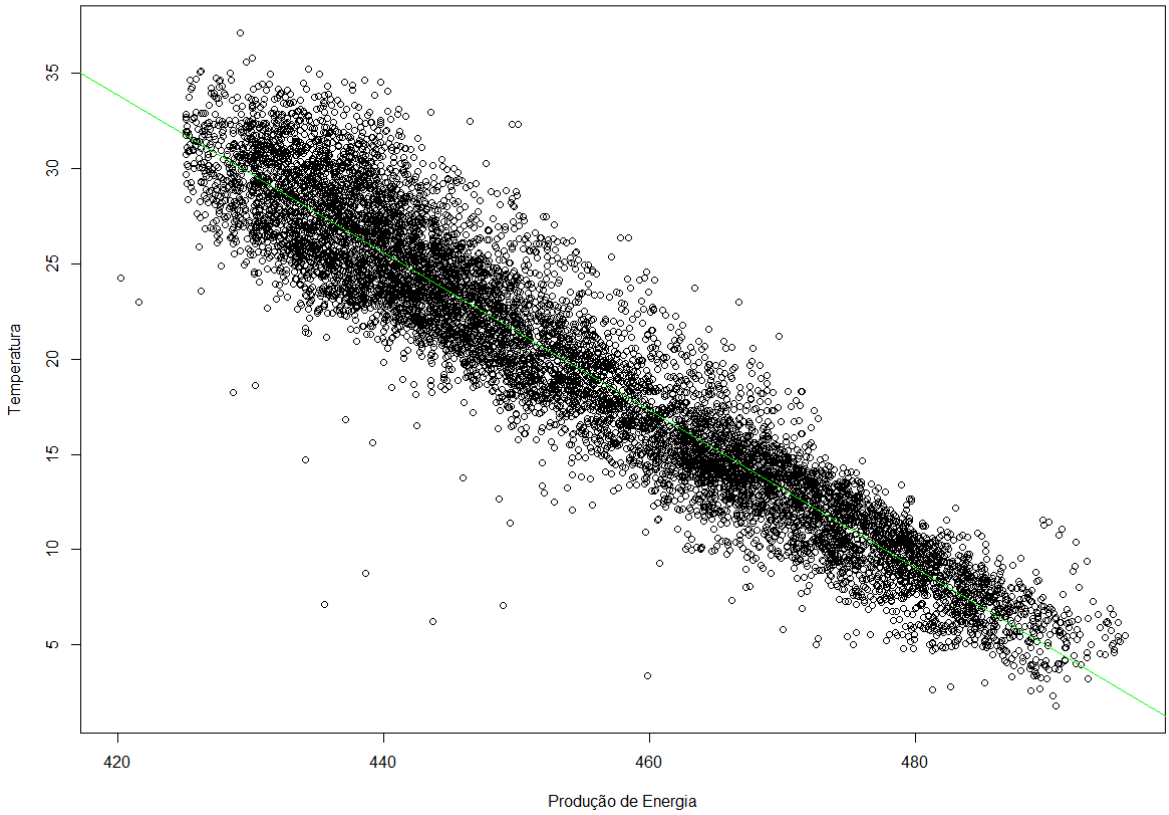
*> erro1 = base$AT - ajustados*

*> hist(erro, main = "Histograma dos Residuais", ylab = "Residuais")*

*> plot(base$PE, base$AT, xlab = "Produção de Energia", ylab = "Temperatura")*

*> abline(reg1, col = "green")*






Como podemos observar a correlação é altíssima entre a produção de energia e a temperatura, isso apenas confirma o que era esperado, pois em turbinas de gás e a vapor a temperatura é uma variável que possui uma grande influência no processo de funcionamento normal delas. Ainda é possível observar que quanto menor a temperatura maiores são os níveis de produção de energia elétrica, indicando um melhor funcionamento das turbinas da UCB.

*> #CORRELAÇÃO MEDIANA*

*> reg2 = lm(base$AP~base$PE)*

*> a2 = reg2$coefficients[1]*

*> b2 = reg2$coefficients[2]*

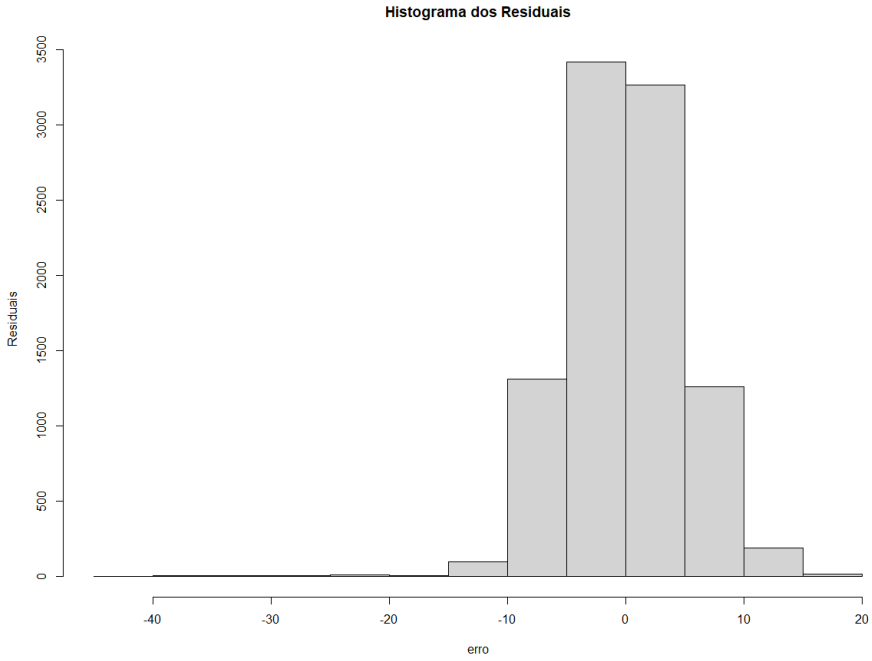
*> ajustados2 = reg2$fitted.values*

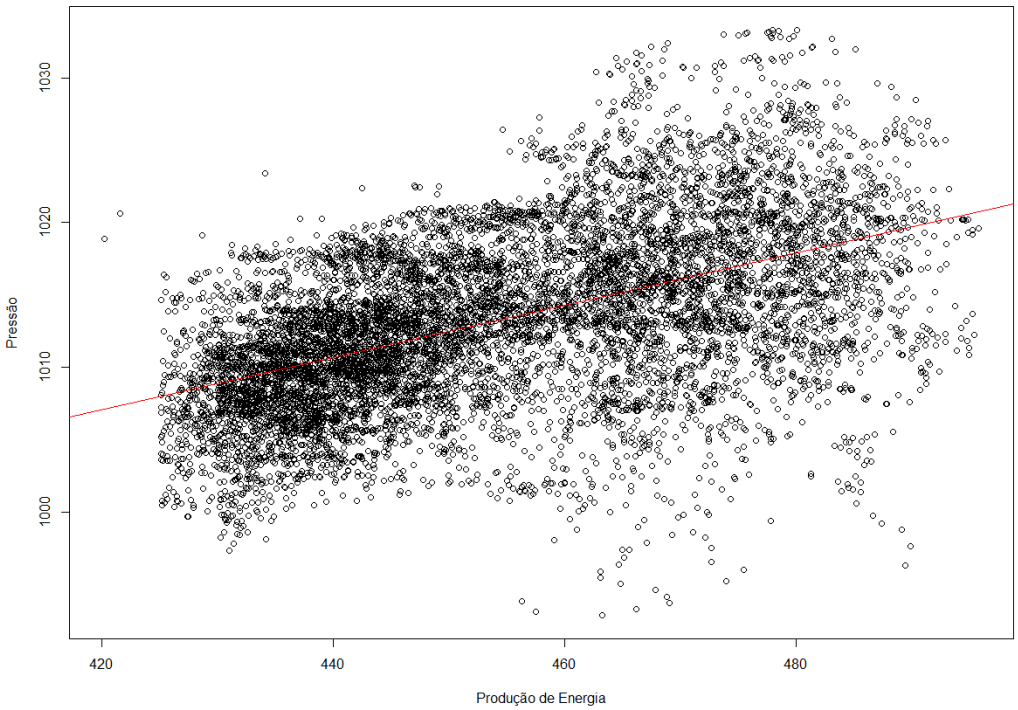
*> erro2 = base$AP - ajustados*

*> hist(erro, main = "Histograma dos Residuais", ylab = "Residuais")*

*> plot(base$PE, base$AP, xlab = "Produção de Energia", ylab = "Pressão")*

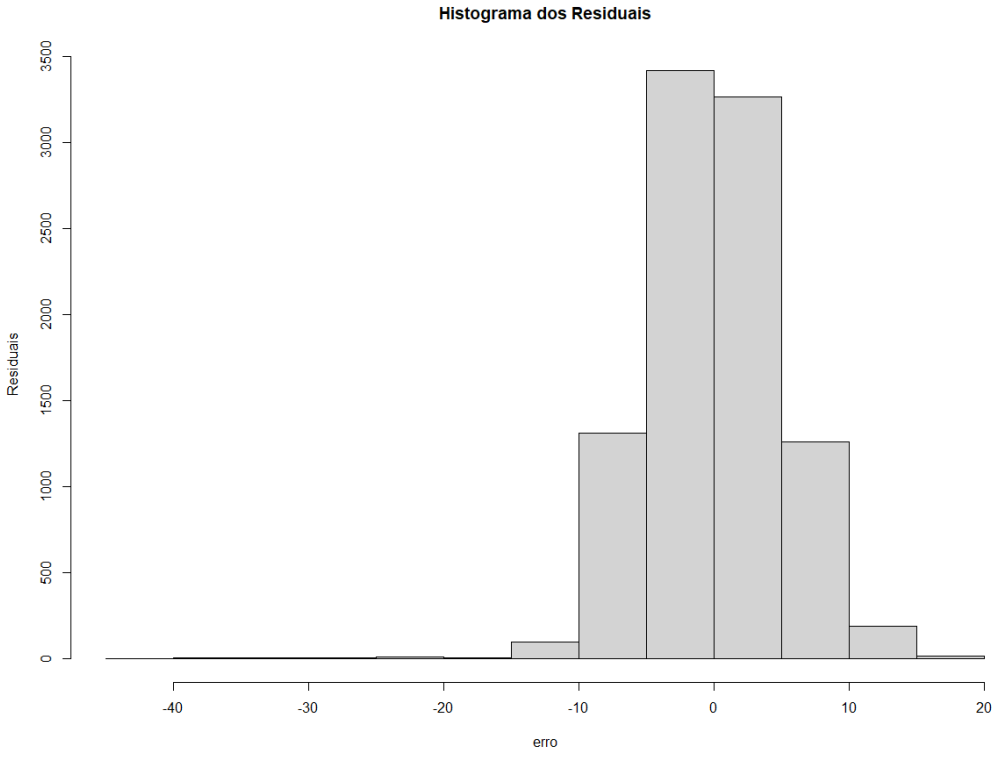
*> abline(reg2, col = "red")*

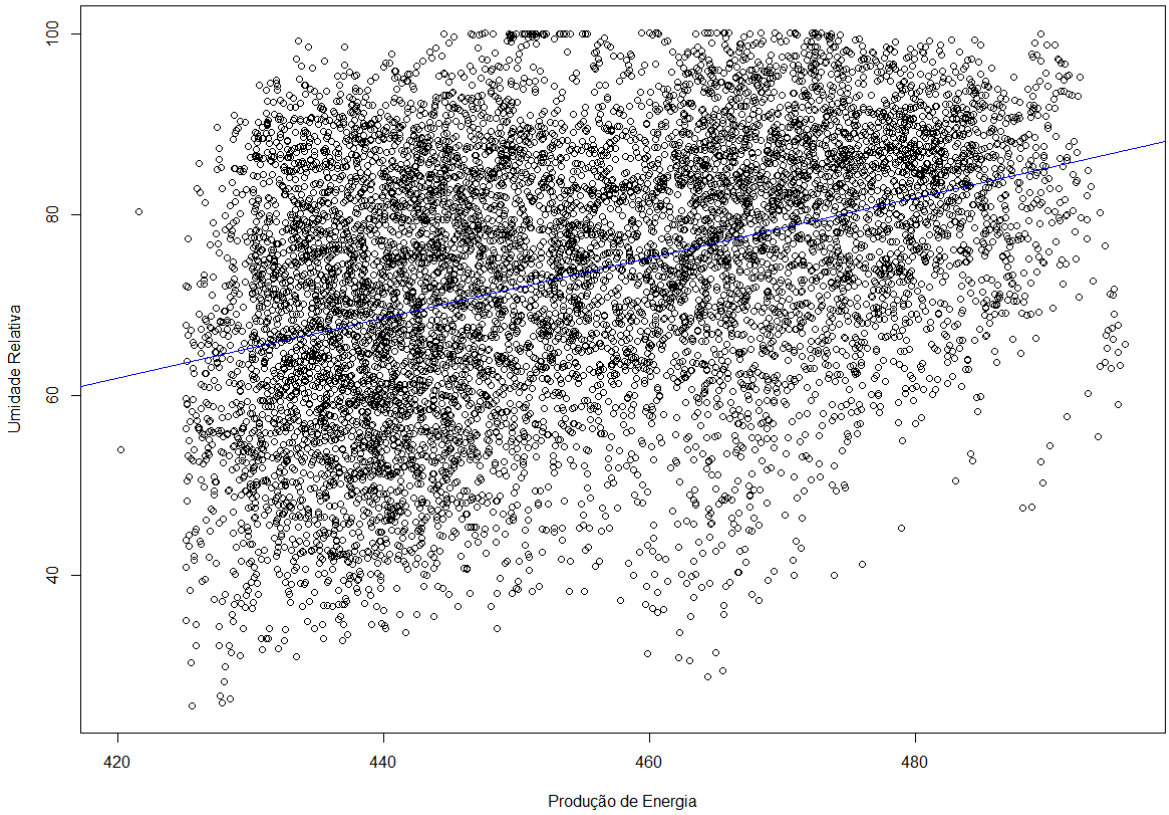


Como pode-se observar uma grande presença de outliers entre as variáveis produção de energia e pressão atmosférica, além disso a correlação entre as duas está em patamar médio, indicando que a variável não possui tanta influência no processo de produção de energia, esse fato está de acordo com o que se diz na literatura pois a pressão atmosférica não tem ligação direta no processo principalmente de turbinas.

```
> #MENOR CORRELAÇÃO
> reg3 = lm(base$RH~base$PE)
> a3 = reg3$coefficients[1]
> b3 = reg3$coefficients[2]
> ajustados3 = reg3$fitted.values
> erro3 = base$RH - ajustados
> hist(erro, main = "Histograma dos Residuais", ylab = "Residuais")
> plot(base$PE, base$RH, xlab = "Produção de Energia", ylab = "Umidade Relativa")
> abline(reg3, col = "blue")
```






Entre as variáveis a que possui menor correlação com a variável produção de energia é a variável umidade relativa, que também é possível observar que existe um grande número de outliers e a umidade não tem praticamente nenhuma influência na produção de energia elétrica na UCB.

*#Regressão linear múltipla, com todas as variáveis quantitativas independentes:*

*> #REGRESSÃO MULTIPLA*

*> regressao_multipla = lm(base$PE~base$AT+base$RH+base$AP)*

*> regressao_multipla$coefficients[1]*

*(Intercept)*

490.3237

*> regressao_multipla$coefficients[2]*

*base$AT*

-2.377708

*> regressao_multipla$coefficients[3]*

*base$RH*

-0.2038319

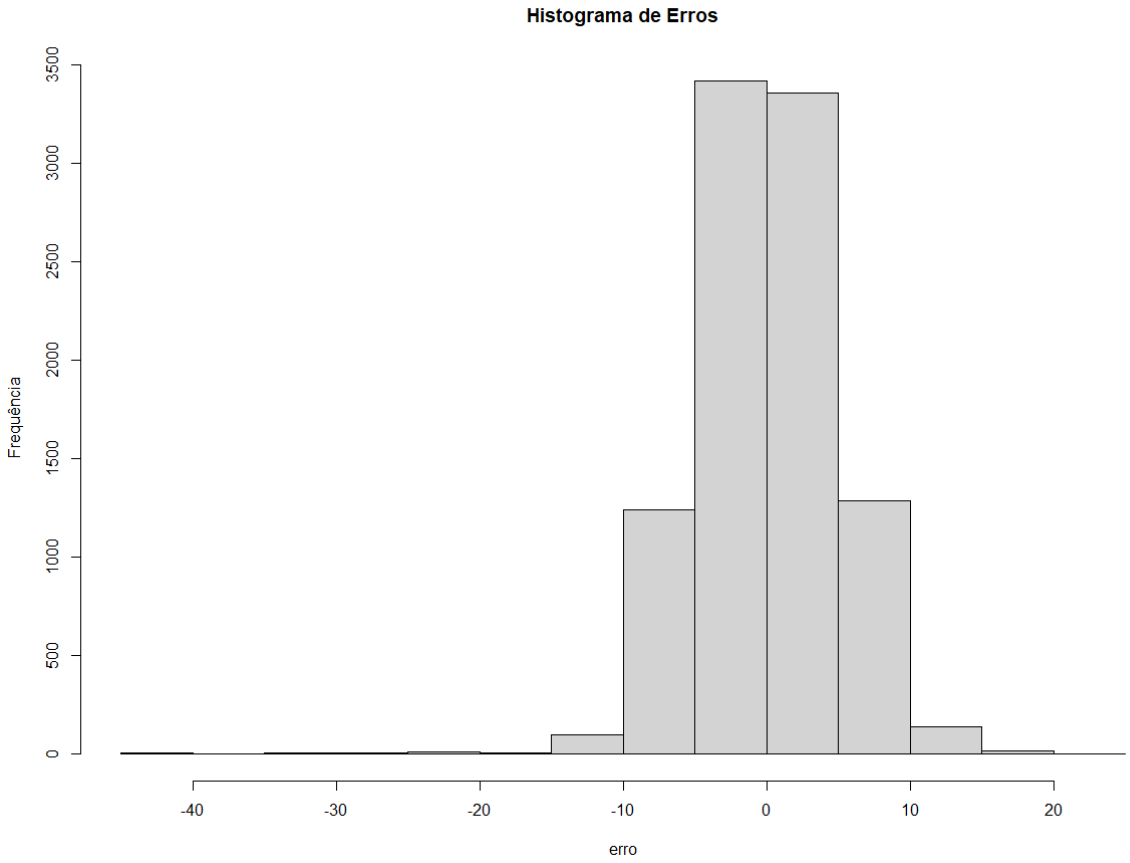
*> regressao_multipla$fitted.values*

1 2 3 4 5 6 7 8 9 10 11 12 13

465.8220 444.2916 485.0736 450.7353 470.5093 441.5649 462.9338 480.1001 473.0311 473.7511 457.8449 455.0910 440.9573 ....

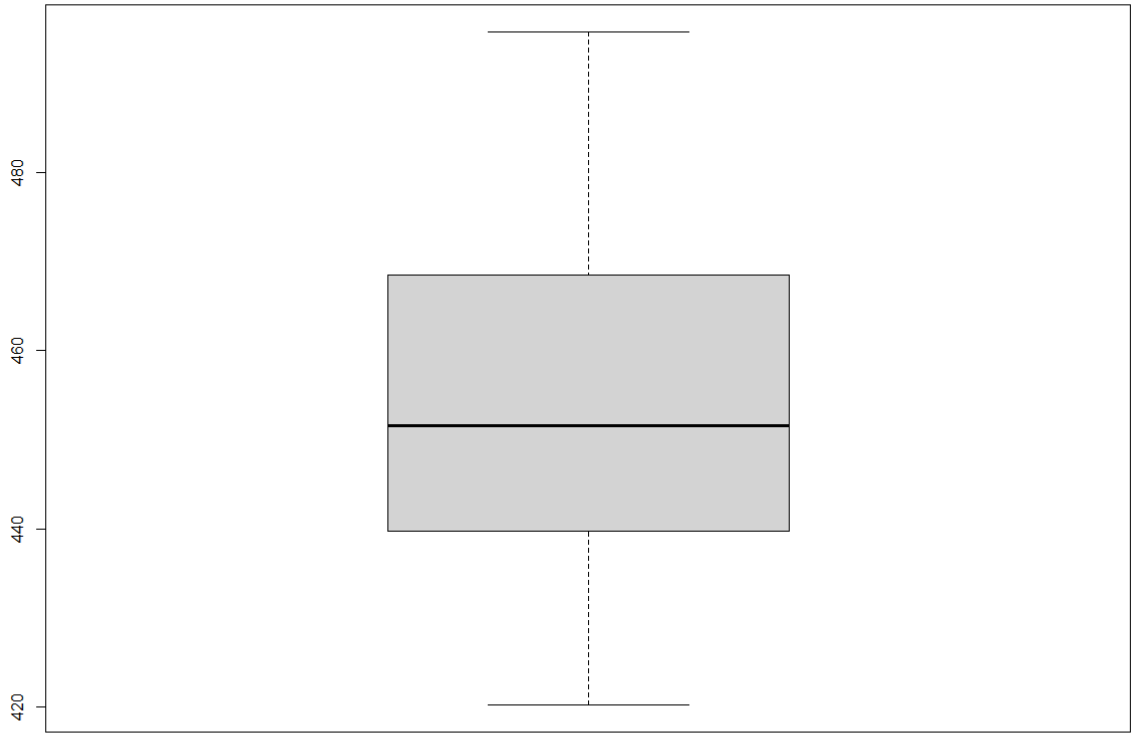
*> erro = base$PE - regressao_multipla$fitted.values*

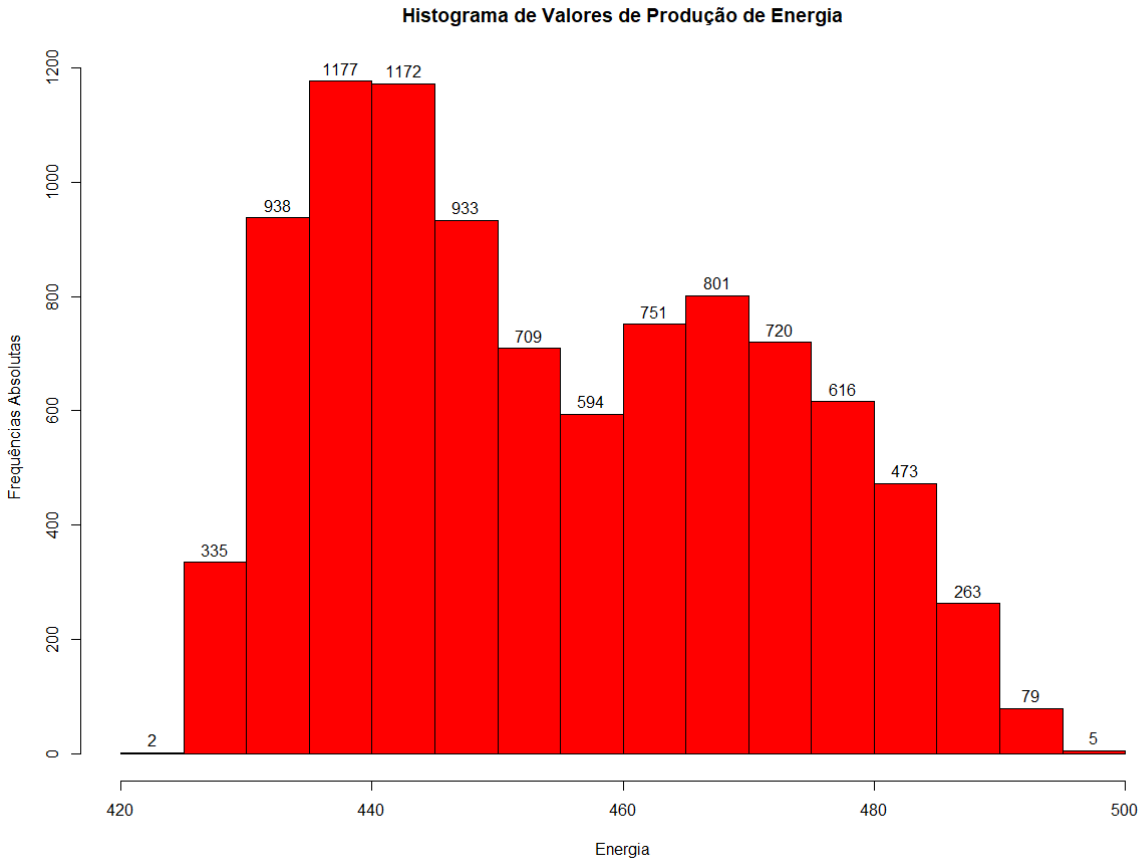
*> hist(erro, main = "Histograma de Erros", ylab = "Frequência")*




Podemos observar no histograma que o erro possui uma distribuição normal, indicando que os erros estão muito próximos.

*#Realizar teste de normalidade usando boxplot, histograma e shapiro.test, discutir os resultados:*






Pelo histograma, é possível observar que os dados dos valores de produção de energia não seguem uma distribuição normal. Vamos aplicar o teste de aderência de Shapiro para chegarmos a uma conclusão a esse respeito.

*> library(caret)*

***#Modelo 1***

*> ModeloAjustado2 = c(0)*

*> ErroAbsoluto2 = c(0)*

*> ErroMedioQuadratico2 = c(0)*

*> for (i in 1:50){*

*+ #Criando os índices para Holdout*

*+ ind = createDataPartition(base$PE,p = 2/3, list = FALSE)*

*+ #Dividindo em Treino e Teste*

```
+   train.data <- base[ind, ]
+   test.data <- base[-ind, ]
+   # Construção do Modelo de Regressão
+   modelo2 =lm(base$PE~base$AT+base$V+base$AP+base$RH, data = train.data)
+   # Calculo do Valor predito
+    ValoresPreditos2 = modelo2$coefficients[1] + modelo2$coefficients[2]*test.data$AT + modelo2$coefficients[3]*test.data$V + modelo2$coefficients[4]*test.data$AP + modelo2$coefficients[5]*test.data$RH
+   ModeloAjustado2 [i] = c(R2(ValoresPreditos2, test.data$PE))
+   ErroAbsoluto2 [i] = c(MAE(ValoresPreditos2, test.data$PE))
+   ErroMedioQuadratico2 [i] = c(RMSE(ValoresPreditos2, test.data$PE))
+ }
> mean(ModeloAjustado2)
[1] 0.9288119
> mean(ErroAbsoluto2)
[1] 3.623074
> mean(ErroMedioQuadratico2)
[1] 4.556838
```

***#Modelo 2***

*> ModeloAjustado3 = c(0)*

*> ErroAbsoluto3 = c(0)*

*> ErroMedioQuadratico3 = c(0)*

*> for (i in 1:50){*

*+ #Criando os índices para Holdout*

*+ ind = createDataPartition(base$AT,p = 2/3, list = FALSE)*

*+ #Dividindo em Treino e Teste*

*+ train.data <- base[ind, ]*

*+ test.data <- base[-ind, ]*

*+ # Construção do Modelo de Regressão*

*+ modelo3 =lm(base$PE~base$AT+base$V, data = train.data)*

*+ # Calculo do Valor predito*

*+ValoresPreditos3=modelo3$coefficients[1]+modelo3$coefficients[2]*test.data$AT+model o3$coefficients[3]*test.data$V*

*+ #Métricas para Avaliar o modelo*

*+ ModeloAjustado3 [i] = c(R2(ValoresPreditos3, test.data$PE))*

*+ ErroAbsoluto3 [i] = c(MAE(ValoresPreditos3, test.data$PE))*

*+ ErroMedioQuadratico3 [i] = c(RMSE(ValoresPreditos3, test.data$PE))*

*+ }*

*> mean(ModeloAjustado3)*

[1] 0.9150037

*> mean(ErroAbsoluto3)*

[1] 3.926851

*> mean(ErroMedioQuadratico3)*

[1] 4.971302

***#Modelo 3***

```
> ModeloAjustado4 = c(0)
> ErroAbsoluto4 = c(0)
> ErroMedioQuadratico4 = c(0)
> for (i in 1:50){
+   #Criando os índices para Holdout
+   ind = createDataPartition(base$AP,p = 2/3, list = FALSE)
+   #Dividindo em Treino e Teste
+   train.data <- base[ind, ]
+   test.data <- base[-ind, ]
+   # Construção do Modelo de Regressão
+   modelo4 =lm(base$PE~base$AT, data = train.data)
+   # Calculo do Valor predito
+   ValoresPreditos4 = modelo4$coefficients[1] + modelo4$coefficients[2]*test.data$AT
+   #Métricas para Avaliar o modelo
+   ModeloAjustado4 [i] = c(R2(ValoresPreditos4, test.data$PE))
+   ErroAbsoluto4 [i] = c(MAE(ValoresPreditos4, test.data$PE))
+   ErroMedioQuadratico4 [i] = c(RMSE(ValoresPreditos4, test.data$PE))
+ }
> mean(ModeloAjustado4)
[1] 0.8979915
> mean(ErroAbsoluto4)
```

[1] 4.299107

*> mean(ErroMedioQuadratico4)*

[1] 5.445082

*> media_erro2 = mean(ErroAbsoluto2)*

*> media_erro3 = mean(ErroAbsoluto3)*

*> media_erro4 = mean(ErroAbsoluto4)*

*> media_erro2*

[1] 3.623074

*> media_erro3*

[1] 3.926851

*> media_erro4*

[1] 4.299107

Foram criados 3 modelos:

- O modelo2 leva em consideração todas as variáveis independentes quantitativas;
- O modelo3 leva em consideração as 2 variáveis independentes quantitativas com maior correlação com a variável dependente;
- O modelo4 leva em consideração apenas a variável independente quantitativa com maior correlação com a variável dependente.

Como se percebe após o cálculo dos vetores de erro num loop de 50 repetições, a média do erro do modelo 1 é a menor, o que torna esse modelo com mais variáveis teoricamente o "menos ruim". Mas isso é apenas uma hipótese que será testada mais adiante.

> *shapiro.test(ErroAbsoluto2)*

Shapiro-Wilk normality test

data: ErroAbsoluto2

W = 0.95949, p-value = 0.0846

> *shapiro.test(ErroAbsoluto3)*

Shapiro-Wilk normality test

data: ErroAbsoluto3

W = 0.98063, p-value = 0.5787

> *shapiro.test(ErroAbsoluto4)*

Shapiro-Wilk normality test

data: ErroAbsoluto4

W = 0.97661, p-value = 0.4194

Pelo teste de normalidade de Shapiro, em todos os modelos os dados seguem distribuição normal, pois o p-value foi superior a 0.05.

*#Formular teste de hipótese e ANOVA, discutir os resultados:*

*#Utilizando o Teste t de Student:*

*> t.test(ErroAbsoluto2,ErroAbsoluto3, alternative = "two.sided")*

Welch Two Sample t-test

data: ErroAbsoluto2 and ErroAbsoluto3

t = -35.764, df = 96.707, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.3206360 -0.2869183

sample estimates:

mean of x mean of y

3.623074 3.926851

Como p-value deu menor que 0.05, então rejeita-se H0 que os erros 2 e 3 são iguais, acatando-se H1 que são diferentes. Em sendo diferentes, vejamos qual dos dois é menor:

*> t.test(ErroAbsoluto2,ErroAbsoluto3, alternative = "less")*

Welch Two Sample t-test

data: ErroAbsoluto2 and ErroAbsoluto3

t = -35.764, df = 96.707, p-value < 2.2e-16

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf -0.2896707

sample estimates:

mean of x mean of y

3.623074 3.926851

Como p-value deu menor que 0.05, então rejeita-se H0 e acata-se a hipótese alternativa, ou seja, erro2 realmente é significativamente menor que erro3 e, portanto, o modelo1 é melhor.

*#Comparando erro 2 e erro 4:*

*> t.test(ErroAbsoluto2,ErroAbsoluto4, alternative = "two.sided")*

Welch Two Sample t-test

data: ErroAbsoluto2 and ErroAbsoluto4

t = -80.369, df = 97.063, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.6927276 -0.6593384

sample estimates:

mean of x mean of y

3.623074 4.299107

Como p-value deu menor que 0.05, então rejeita-se H0 que os erros 2 e 4 são iguais, acatando-se H1 que são diferentes. Em sendo diferentes, vejamos qual dos dois é menor:

*> t.test(ErroAbsoluto2,ErroAbsoluto4, alternative = "less")*

Welch Two Sample t-test

data: ErroAbsoluto2 and ErroAbsoluto4

t = -80.369, df = 97.063, p-value < 2.2e-16

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf -0.6620637

sample estimates:

mean of x mean of y

3.623074 4.299107

Como p-value deu menor que 0.05, então rejeita-se H0 e acata-se a hipótese alternativa, ou seja, erro2 realmente é significativamente menor que erro4 e, portanto, o modelo 1 é melhor.

*> t.test(ErroAbsoluto3,ErroAbsoluto4, alternative = "two.sided")*

Welch Two Sample t-test

data: ErroAbsoluto3 and ErroAbsoluto4

t = -41.856, df = 97.97, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.3899052 -0.3546064

sample estimates:

mean of x mean of y

3.926851 4.299107

Como p-value deu menor que 0.05, então rejeita-se H0 que os erros 3 e 4 são iguais, acatando-se H1 que são diferentes. Em sendo diferentes, vejamos qual dos dois é menor:

*> t.test(ErroAbsoluto3,ErroAbsoluto4, alternative = "less")*

Welch Two Sample t-test

data: ErroAbsoluto3 and ErroAbsoluto4

t = -41.856, df = 97.97, p-value < 2.2e-16

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf -0.3574872

sample estimates:

mean of x mean of y

3.926851 4.299107

Como p-value deu menor que 0.05, então rejeita-se H0 e acata-se a hipótese alternativa, ou seja, erro3 realmente é significativamente menor que erro4 e, portanto, o modelo 2 é melhor.

*#Análise de Variância (ANOVA):*

*> dados_anova = aov(base$PE~base$AT+base$AP)*

*> summary(dados_anova)*

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) | |
|---|---|---|---|---|---|---|
| base$AT | 1 | 2505095 | 2505095 | 86688.5 | <2e-16 | *** |
| base$AP | 1 | 5196 | 5196 | 179.8 | <2e-16 | *** |
| Residuals | 9565 | 276406 | 29 | | | |

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Utilizando a ANOVA é possível obter as seguintes conclusões:

- Tem-se que existe uma diferença de produção de energia entre temperatura com nível de confiança de 99%;
- Tem-se que existe uma diferença de produção de energia entre pressão com nível de confiança de 99%;