Análise descritiva e estatística da base dos jogadores ativos da NBA

Marcos Perazo Viana

13 abril, 2025

Introdução

Este projeto utiliza como base de dados informações sobre jogadores ativos da NBA, obtidas no site oficial da NBA, na aba *Players*, no dia 22 de março de 2025. Os dados coletados incluem: nomes dos jogadores, equipes em que atuam, posições, altura, peso, data de nascimento, idade, média de pontos, rebotes e assistências por jogo da atual temporada, além de informações sobre o *Draft* e experiência profissional. Adicionalmente, foi criada uma variável indicando o mês de nascimento dos jogadores. Com esses dados, é possível investigar diversas correlações estatísticas relevantes que ajudam a compreender o desempenho e as características dos atletas. Por exemplo, pode-se analisar a relação entre altura dos jogadores e a média de rebotes por jogo. Partindo da hipótese de que jogadores mais altos possuem vantagem nesse quesito, essa análise busca confirmar ou refutar tal tendência. Outras correlações potencialmente interessantes incluem:
- Peso e média de pontos por jogo: Investiga se jogadores com maior massa corporal têm vantagens em marcar pontos, especialmente em posições como pivô.
- Assistências por jogo e pontos por jogo: Examina a relação entre jogadores que criam oportunidades para seus colegas de equipe e aqueles que executam as finalizações.
- Altura e assistências por jogo: Avalia se jogadores mais baixos, como armadores, têm maior tendência a contribuir com assistências. - Média de pontos por jogo e rebotes por jogo: Analisa a relação entre a capacidade de pontuar e pegar rebotes, considerando as diferentes posições em quadra.

Essas análises têm como objetivo explorar padrões e características que podem enriquecer a compreensão sobre o desempenho dos jogadores da NBA.

Carregamento da base.

```
dados_tratados <- read_csv("Dados_auxiliares/dados_apos_coluna_mes.csv", quote = "\"",
                           locale = locale(encoding = "UTF-8"), show_col_types = FALSE)
```

Observações iniciais:

1 - Fazendo uma análise preliminar, verificamos que a linha 507 não possui dados em nenhuma das colunas, e foi determinada a eliminação dessa linha.

```
dados <- dados %>% slice(-507)
```

2 - Após verificar as estatisticas individuais dos atletas que receberam NA nas colunas PPG, APG e RPG, verificamos que esses campos deveriam receber valor zero. O que foi realizado.

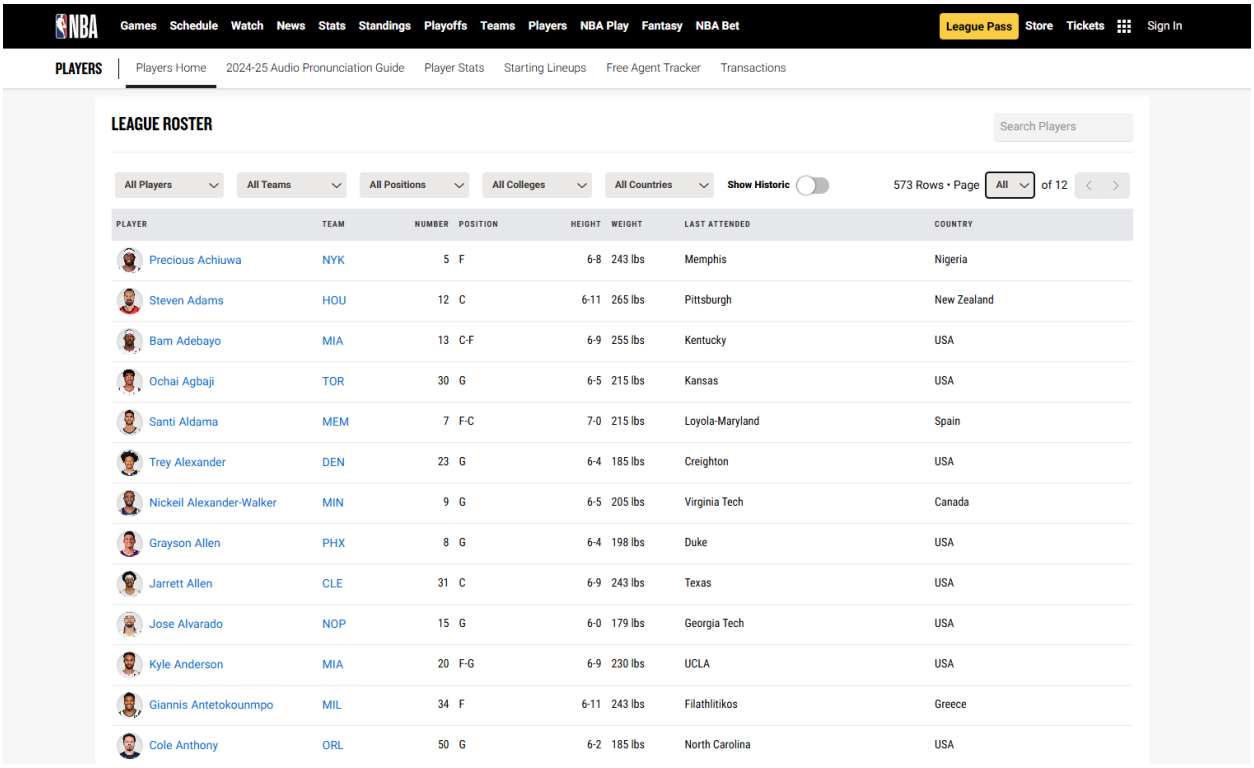


Figure 1: Site da NBA

```
dados <- dados %>%
  mutate(
    PPG = replace_na(PPG, 0),
    RPG = replace_na(RPG, 0),
    APG = replace_na(APG, 0)
  )
```

3 - Foram encontrados dois registros com o campo Peso com valor NA. Os dados faltantes serão omitidos.

```
dados <- dados %>% filter(!is.na(Peso))
```

Analisar a centralidade dos dados, dipersão, assimetria, bem como suas estatísticas de ordem, a fim de checar se há presença de outliers.

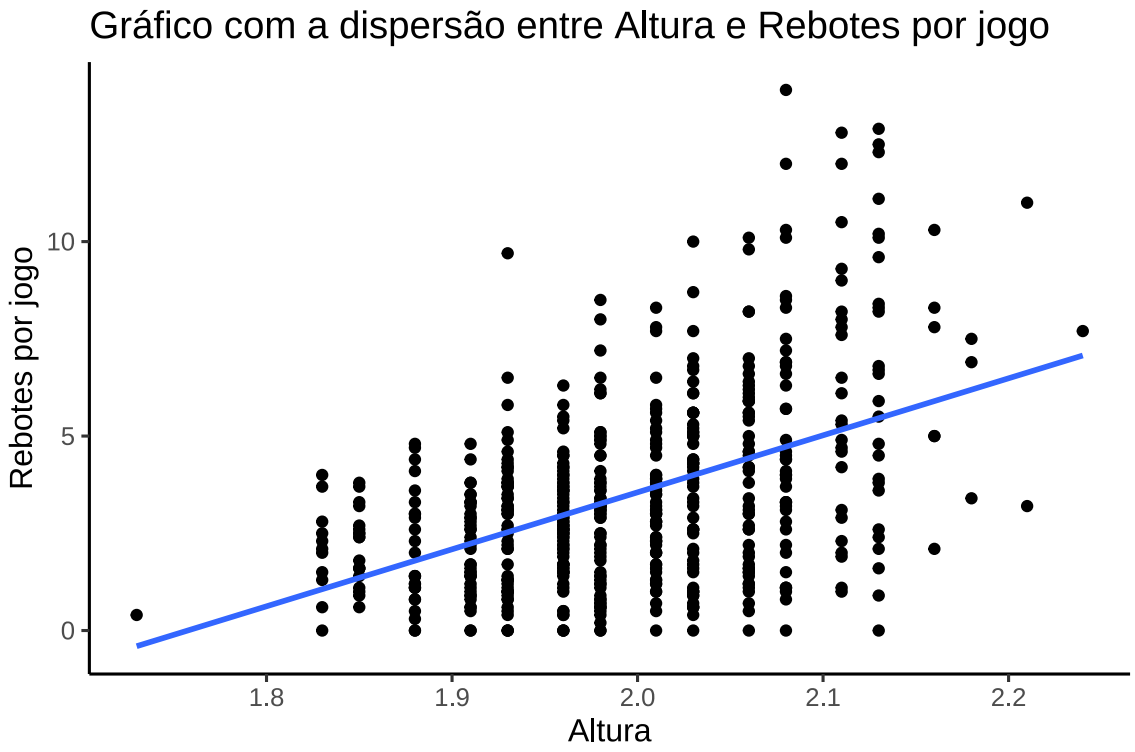
```
dados %>% dplyr::select(Altura, Peso, PPG, RPG, APG) %>%
  summarytools::descr() %>% kable()
```

```
## Error in table(names(candidates))[["tested"]]: índice fora dos limites
```

	Altura	APG	Peso	PPG	RPG
Mean	1.9946643	2.0153710	97.5265018	8.6017668	3.4757951
Std.Dev	0.0793925	1.8418518	10.5926258	6.7483014	2.4858746
Min	1.7300000	0.0000000	72.0000000	0.0000000	0.0000000

	Altura	APG	Peso	PPG	RPG
Q1	1.9300000	0.8000000	90.0000000	3.5000000	1.6000000
Median	1.9800000	1.4000000	97.0000000	7.0000000	3.1000000
Q3	2.0600000	2.7000000	104.0000000	11.9000000	4.6000000
Max	2.2400000	11.4000000	138.0000000	32.9000000	13.9000000
MAD	0.0741300	1.1860800	10.3782000	5.7821400	2.2239000
IQR	0.1300000	1.8750000	14.0000000	8.3750000	3.0000000
CV	0.0398024	0.9139021	0.1086128	0.7845250	0.7151960
Skewness	0.0650535	1.6143914	0.4680604	1.0153617	1.2225806
SE.Skewness	0.1026883	0.1026883	0.1026883	0.1026883	0.1026883
Kurtosis	-0.2812186	2.9171950	0.1500945	0.4372103	1.7806264
N.Valid	566.0000000	566.0000000	566.0000000	566.0000000	566.0000000
N	566.0000000	566.0000000	566.0000000	566.0000000	566.0000000
Pct.Valid	100.0000000	100.0000000	100.0000000	100.0000000	100.0000000

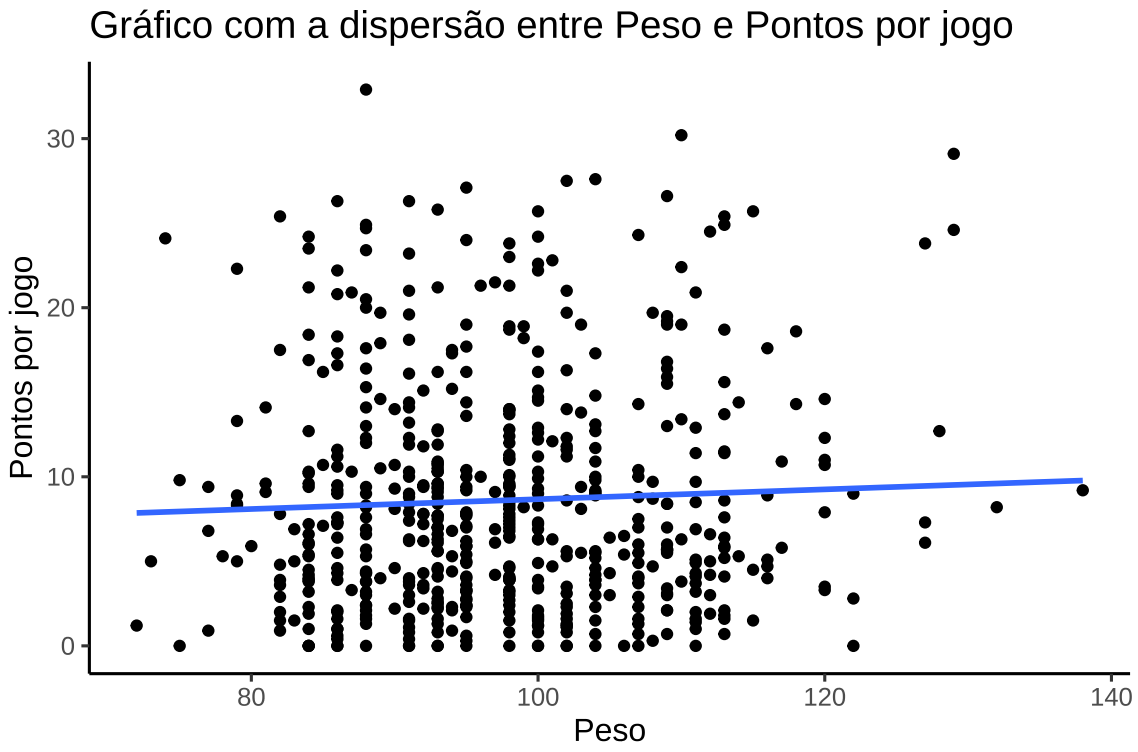
Dispersão Altura x Rebotes



Correlação entre as variaveis Altura e Rebotes por jogo

	Altura	RPG
Altura	1.0000000	0.4680633
RPG	0.4680633	1.0000000

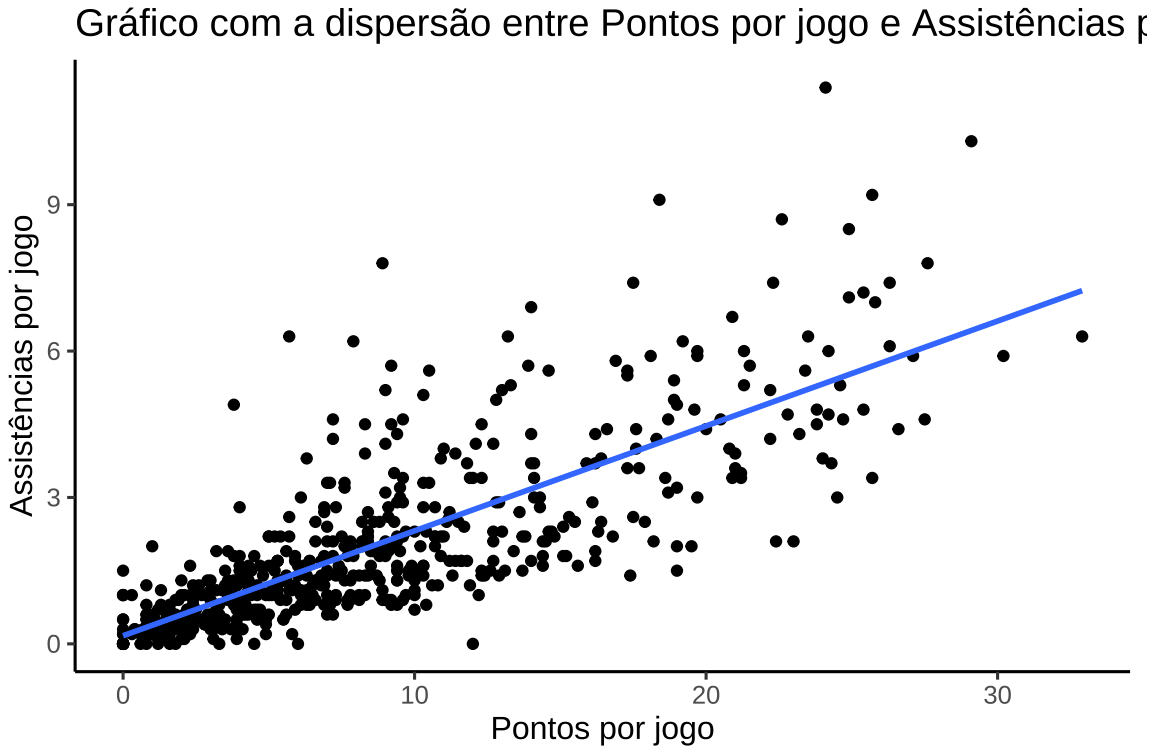
Dispersão Peso x Pontos



Correlação entre as variaveis Peso e Pontos por jogo

	Peso	PPG
Peso	1.0000000	0.0455853
PPG	0.0455853	1.0000000

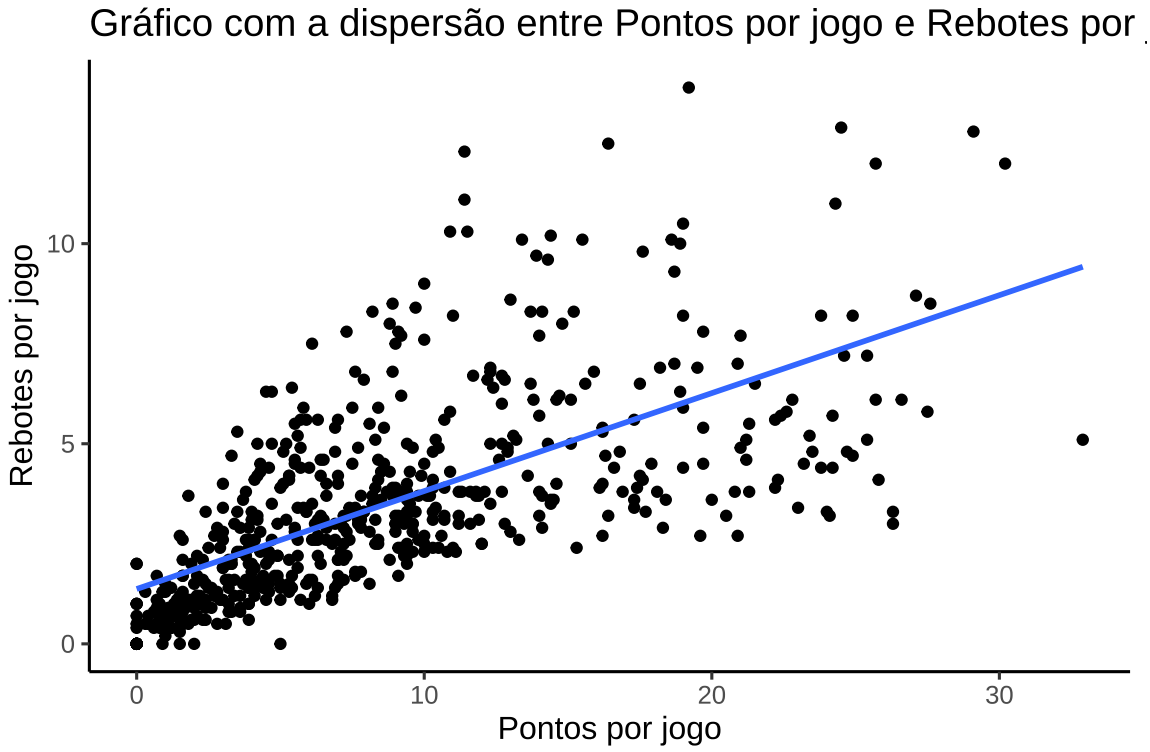
Dispersão Pontos x Assistências



Correlação entre as variaveis Pontos e AssistÊncias por jogo

	PPG	APG
PPG	1.0000000	0.7875714
APG	0.7875714	1.0000000

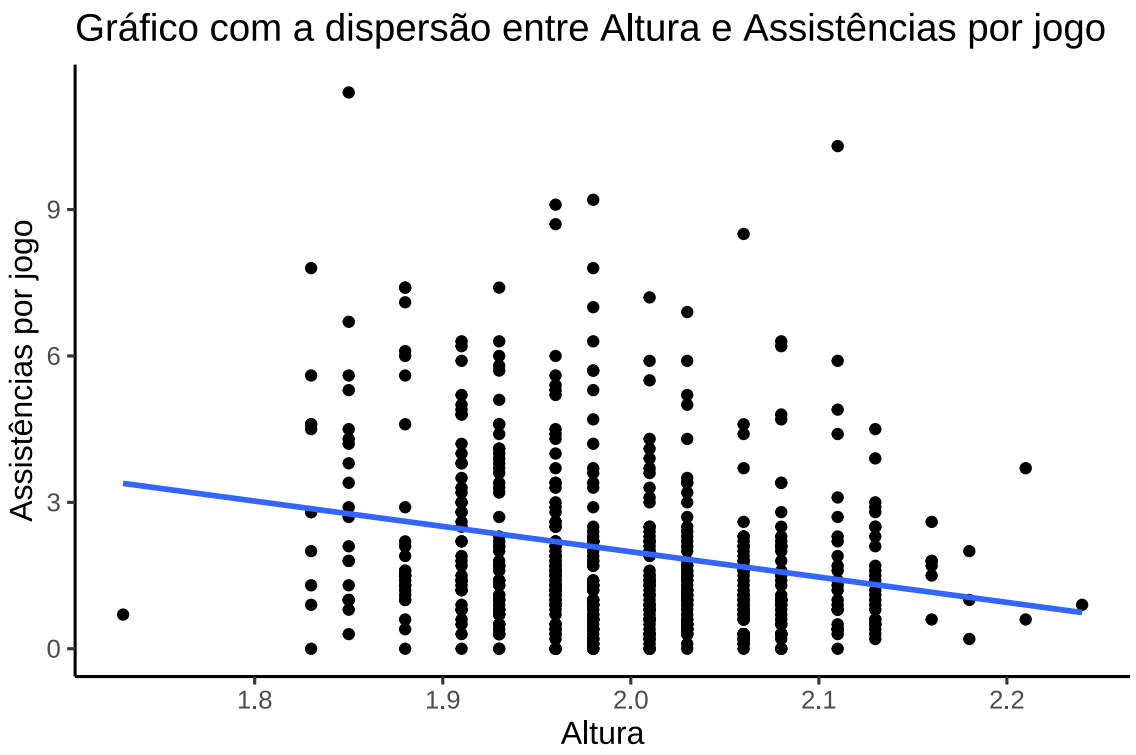
Dispersão Pontos x Rebotes



Correlação entre as variaveis Pontos e Rebotes por jogo

	PPG	RPG
PPG	1.0000000	0.6643496
RPG	0.6643496	1.0000000

Dispersão Altura x Assistências



Correlação entre as variaveis Altura e Assistências por jogo

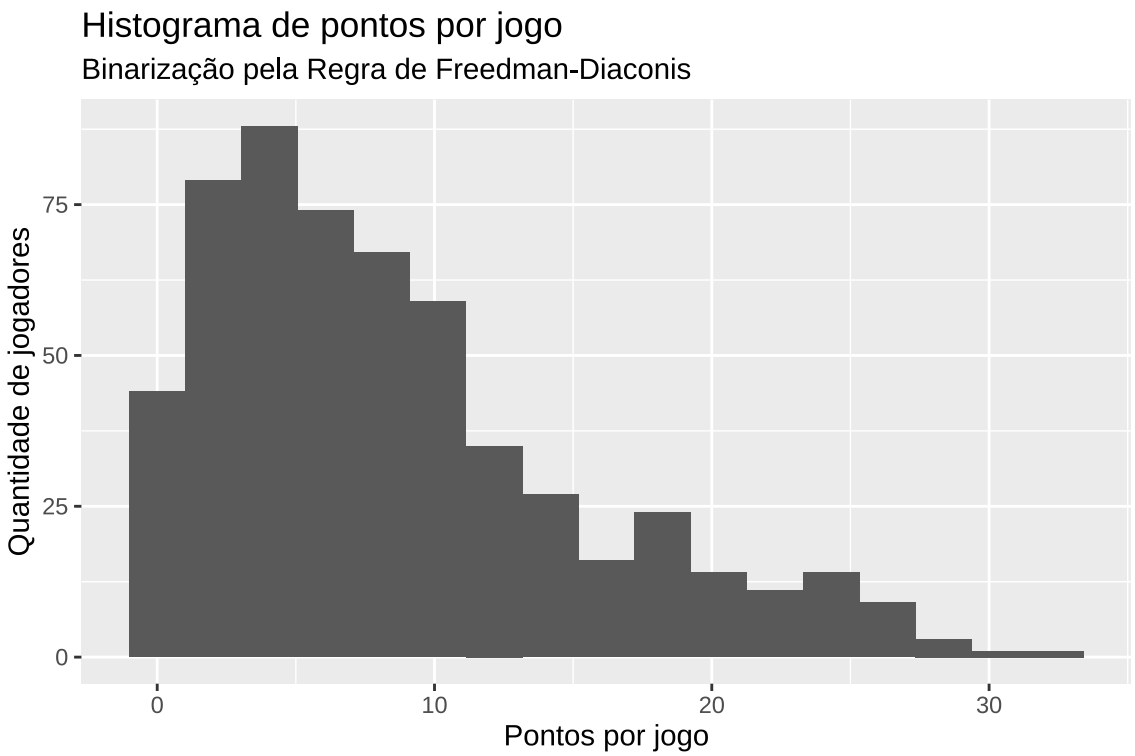
	Altura	APG
Altura	1.0000000	-0.2237194
APG	-0.2237194	1.0000000

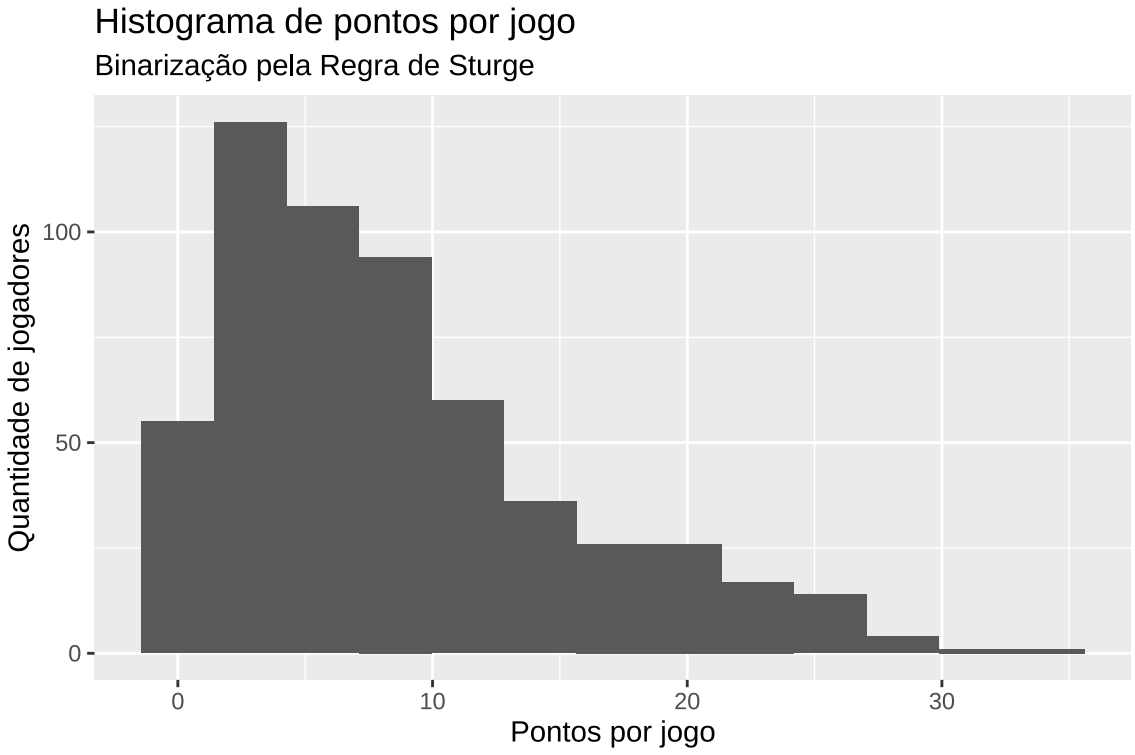
Normalidade das variáveis

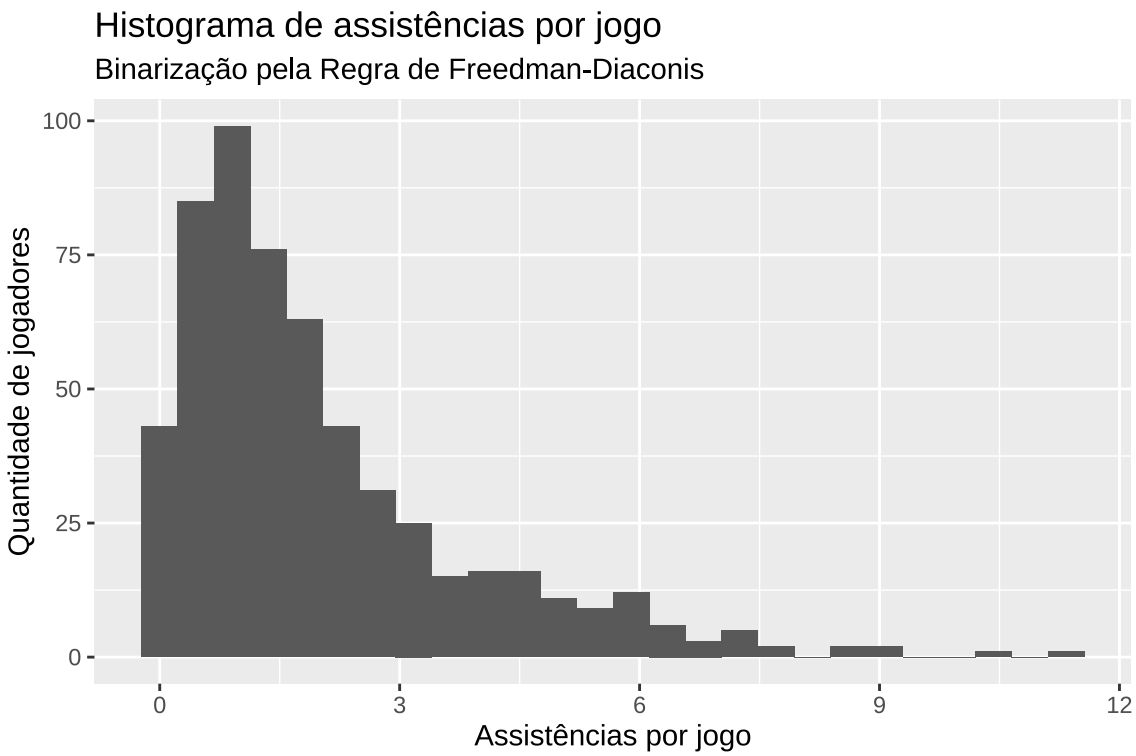
A normalidade das variáveis refere-se à ideia de que os dados seguem uma distribuição normal, também conhecida como distribuição gaussiana. Essa distribuição é simétrica em torno da média, com a maioria dos valores concentrados próximos a ela, e a frequência diminuindo à medida que os valores se afastam da média. É representada por uma curva em forma de sino.

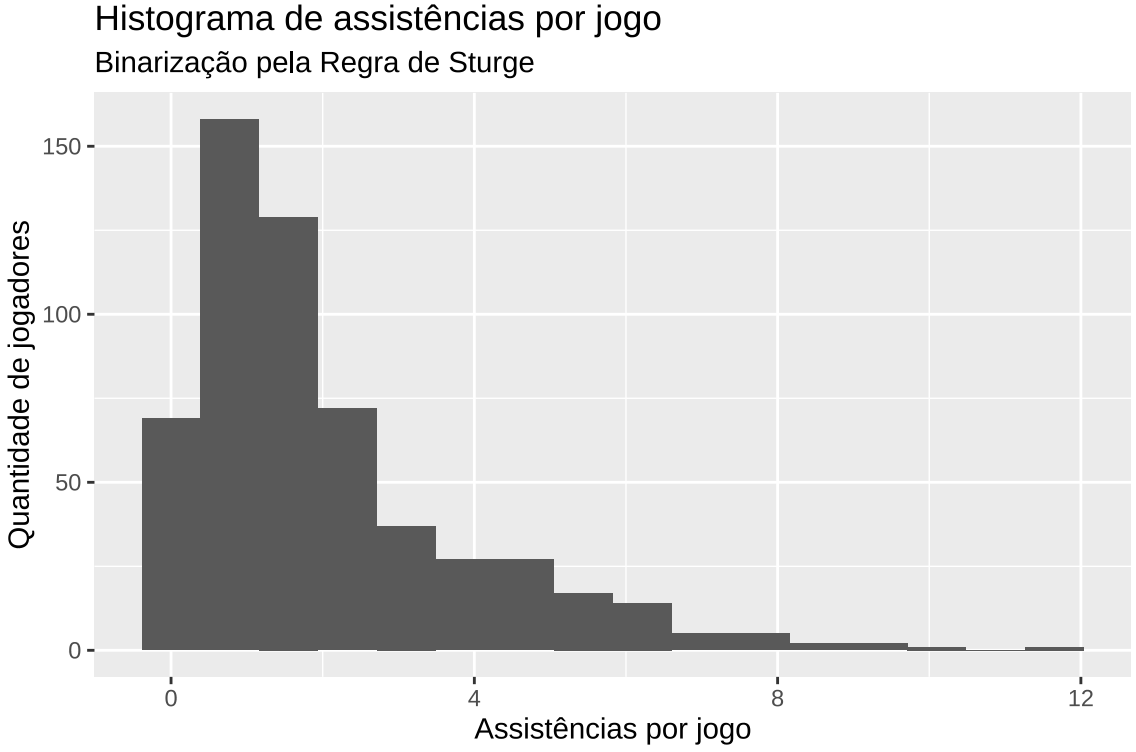
Definindo as funcoes geradoras de binwidths com as Regras de Freedman-Diaconis e Sturge

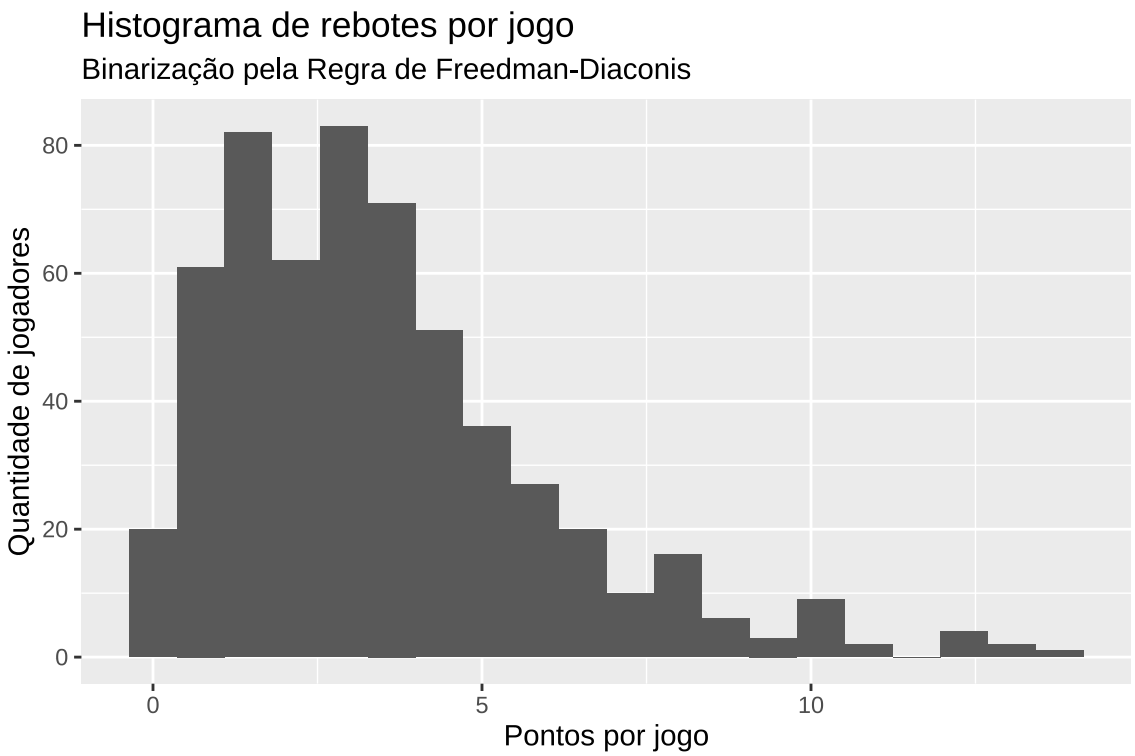
```
fd <- function(x) {
  n <-length(x)
  return((2*IQR(x))/n^(1/3))
}

sr <- function(x) {
  n <-length(x)
  return((3.49*sd(x))/n^(1/3))
}
```

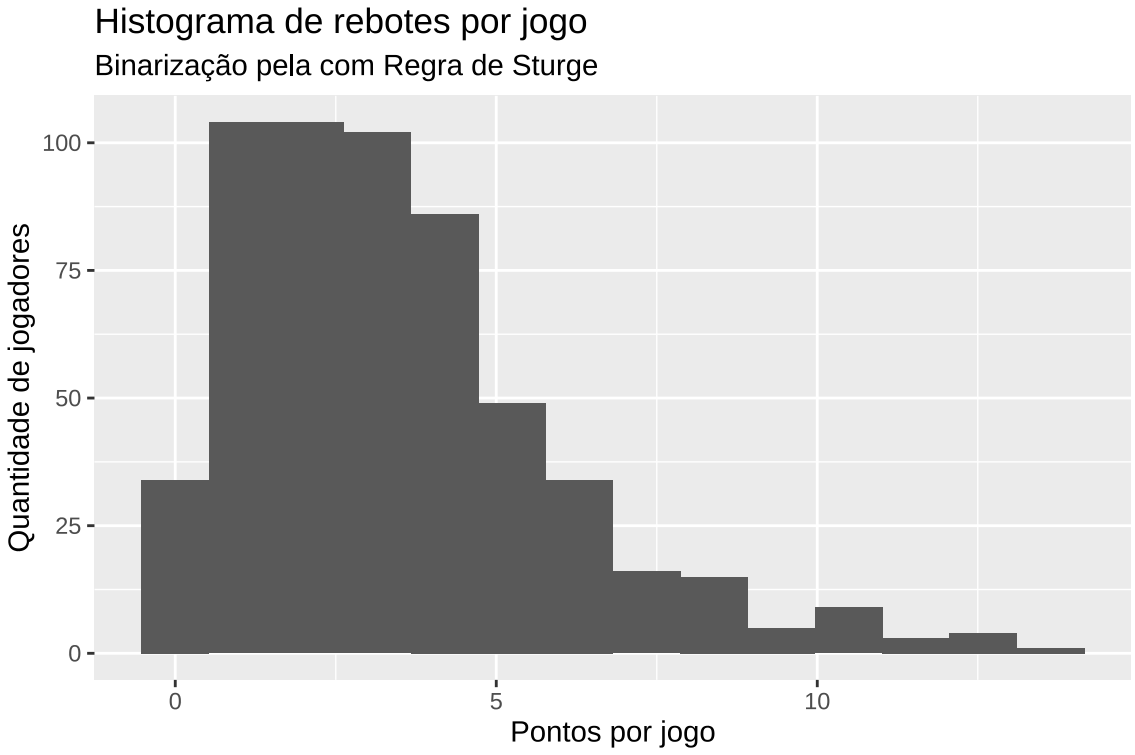

Histograma de pontos por jogo
Binarização pela Regra de Freedman-Diaconis
Quantidade de jogadores
75
50
25
0
0
10
20
30
Pontos por jogo

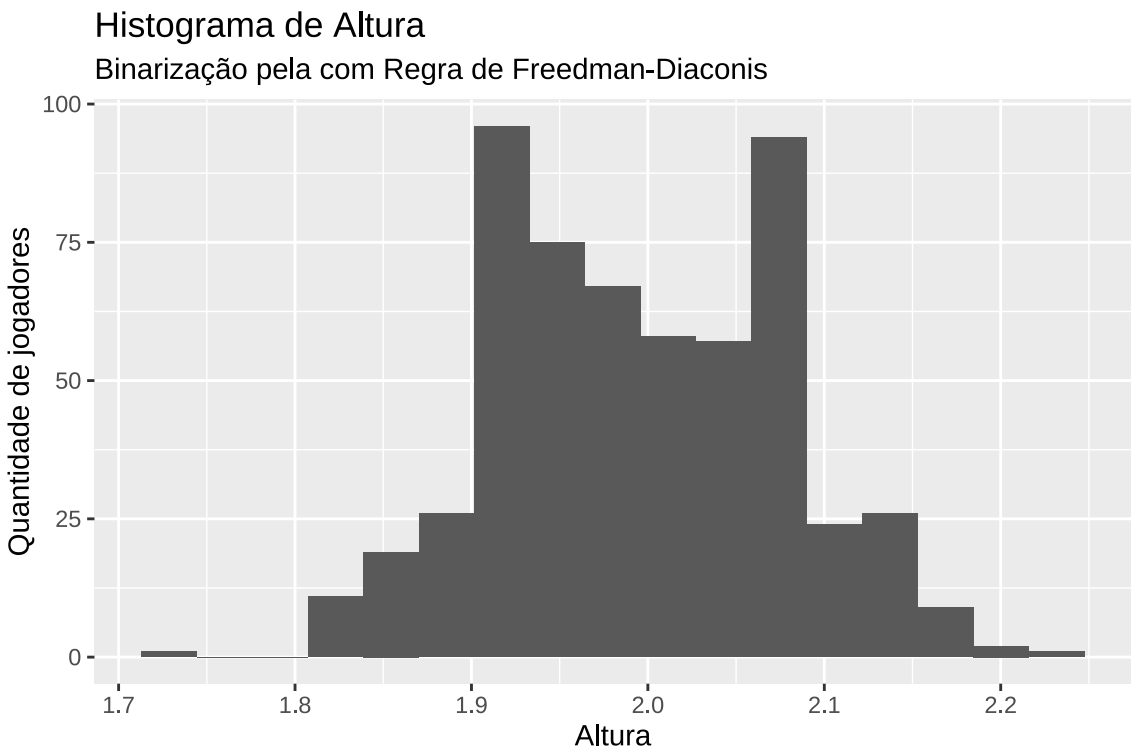

Histograma de pontos por jogo
Binarização pela Regra de Sturge
Quantidade de jogadores
100
50
0
0
10
20
30
Pontos por jogo

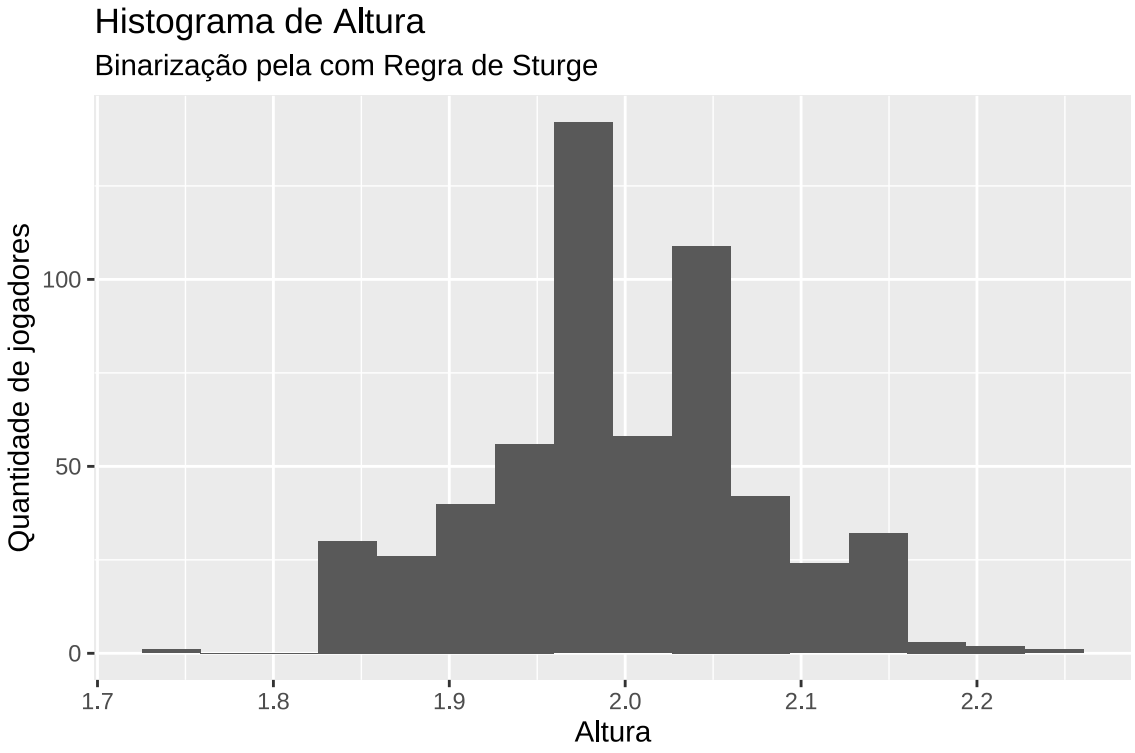

Histograma de assistências por jogo
Binarização pela Regra de Freedman-Diaconis
Quantidade de jogadores
100
75
50
25
0
0
3
6
9
12
Assistências por jogo

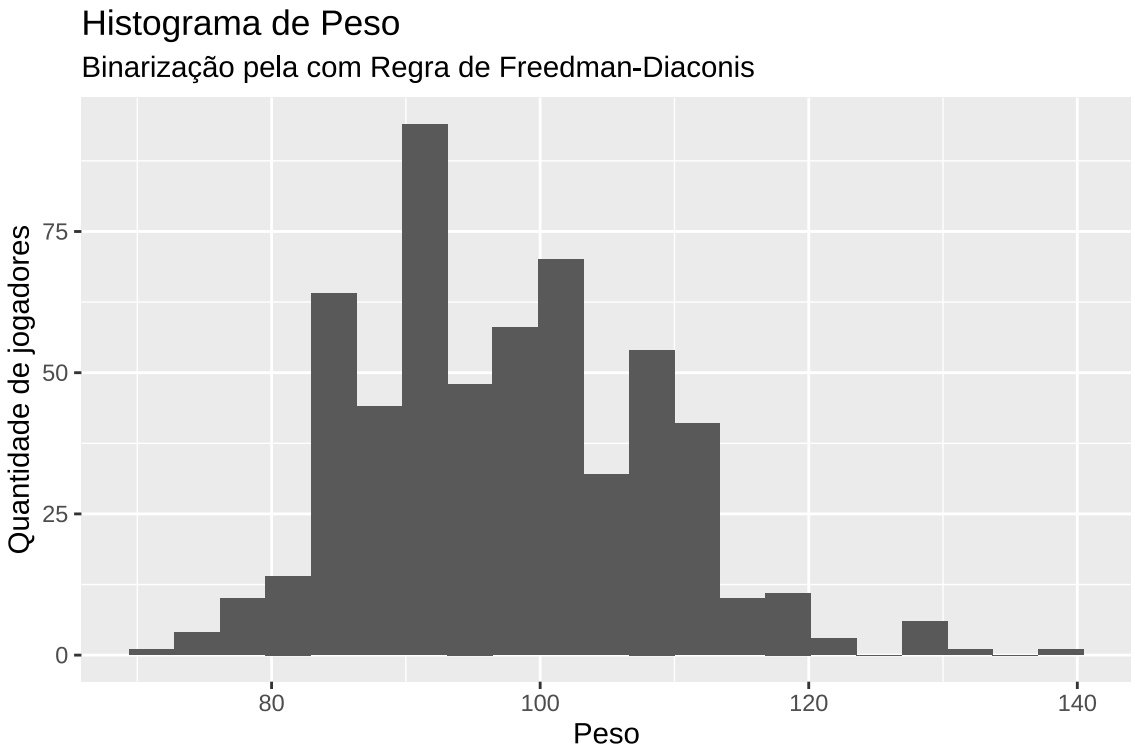

Histograma de assistências por jogo
Binarização pela Regra de Sturge
Quantidade de jogadores
150
100
50
0
0
4
8
12
Assistências por jogo

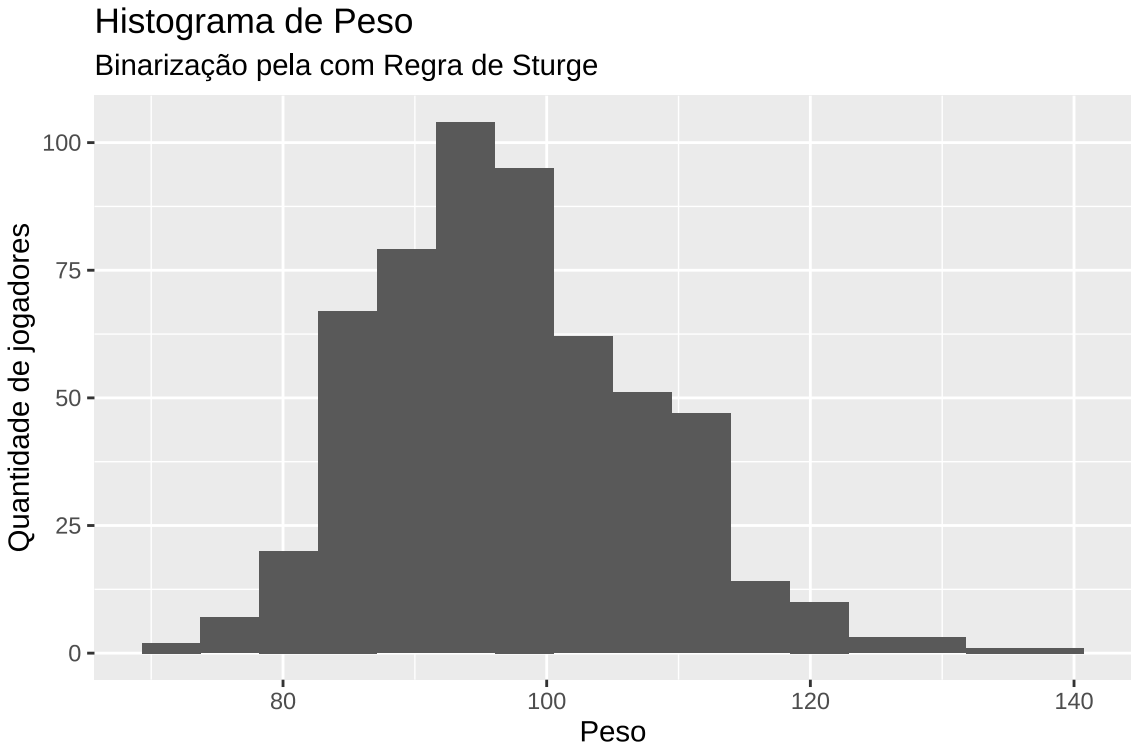

Histograma de rebotes por jogo
Binarização pela Regra de Freedman-Diaconis
Quantidade de jogadores
80
60
40
20
0
0
5
10
Pontos por jogo


Histograma de rebotes por jogo
Binarização pela com Regra de Sturge
Quantidade de jogadores
100
75
50
25
0
0
5
10
Pontos por jogo


Histograma de Altura
Binarização pela com Regra de Freedman-Diaconis
Quantidade de jogadores
100
75
50
25
0
1.7
1.8
1.9
2.0
2.1
2.2
Altura

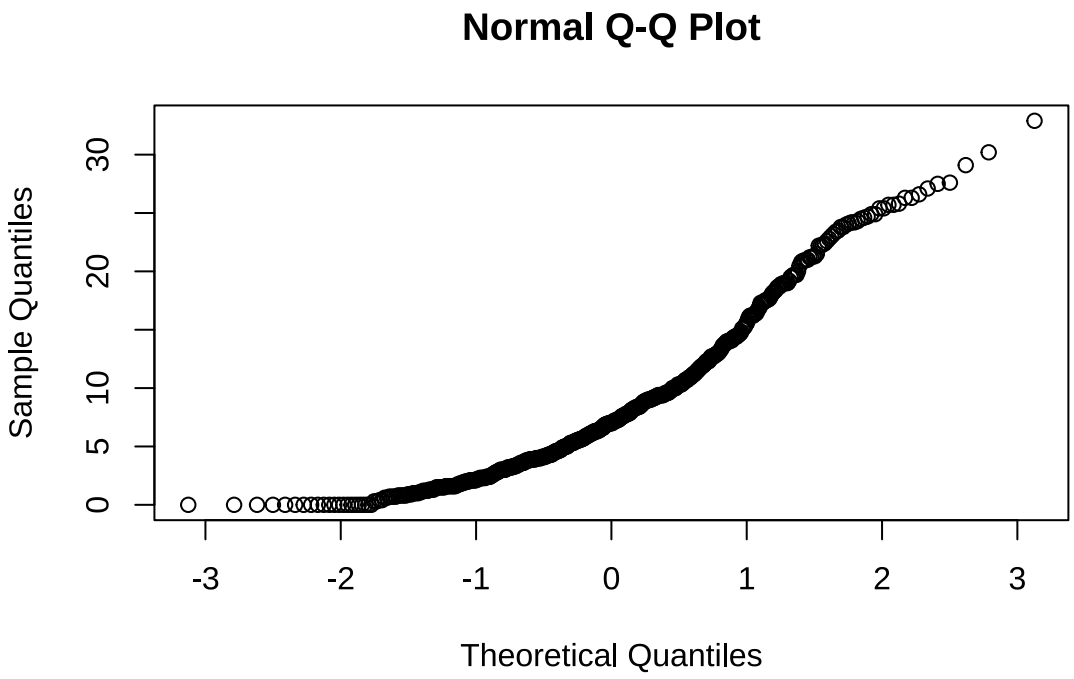

Histograma de Altura
Binarização pela com Regra de Sturge
Quantidade de jogadores
100
50
0
1.7
1.8
1.9
2.0
2.1
2.2
Altura



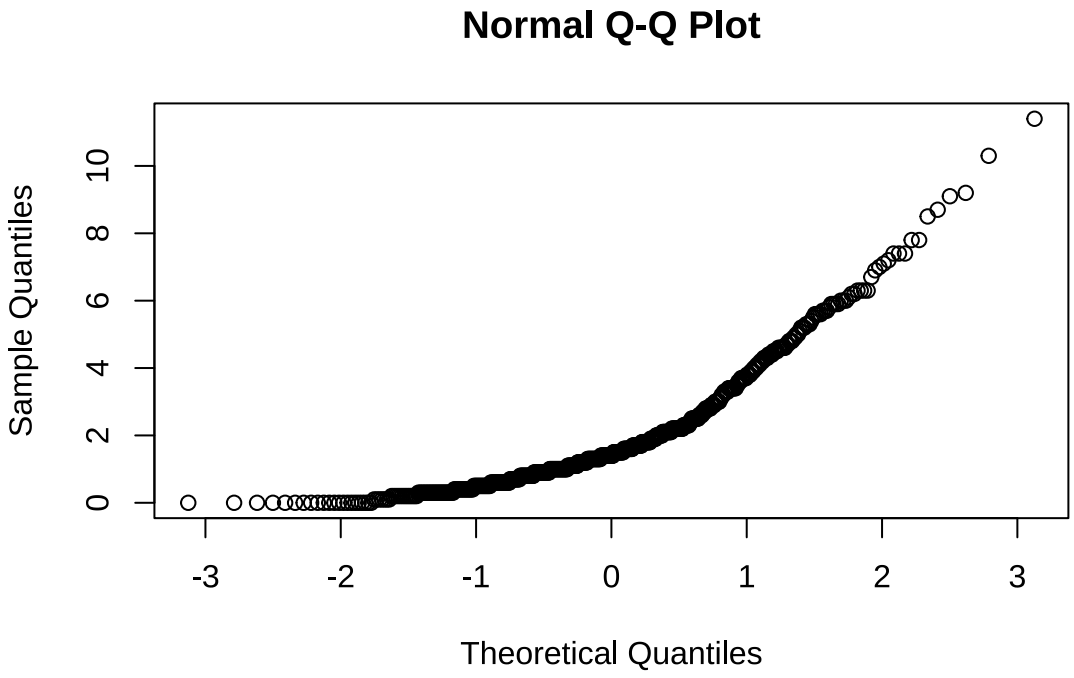


Q-Q Plot para checar visualmente a normalidade das distribuições

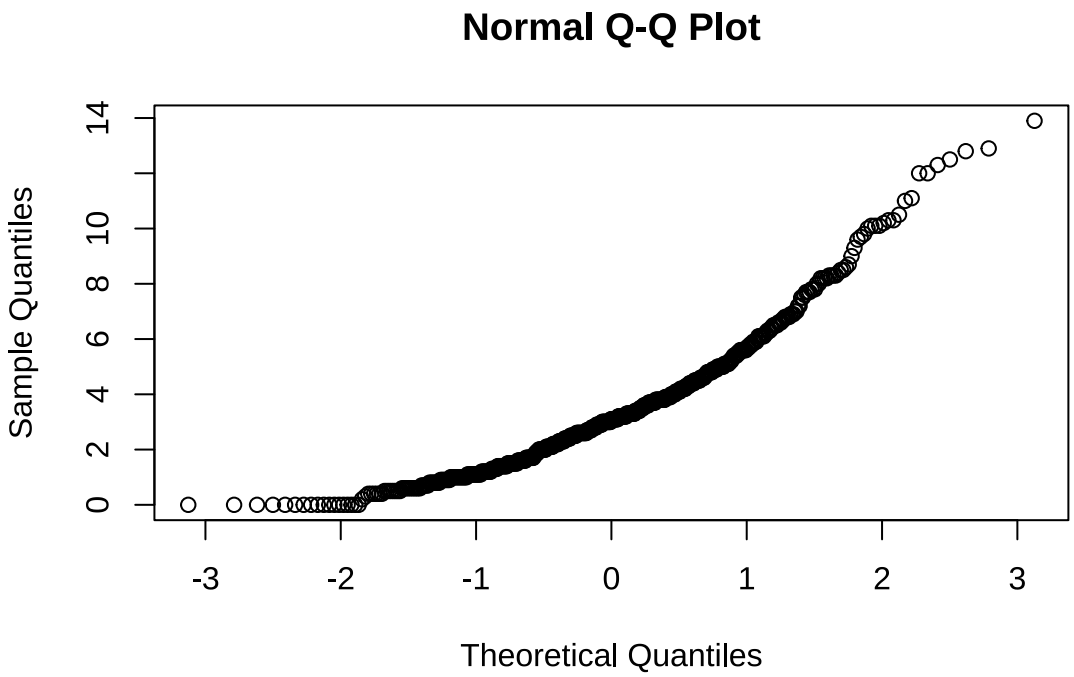
Q-Q Plot de Pontos por partida



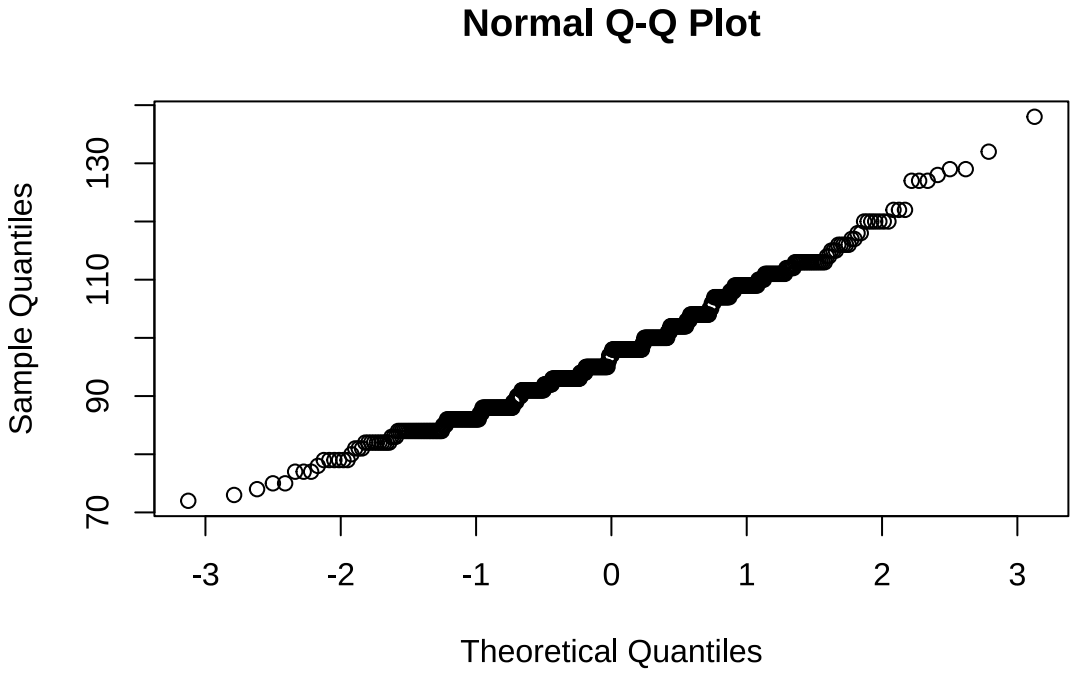
Q-Q Plot de Assistências por partida



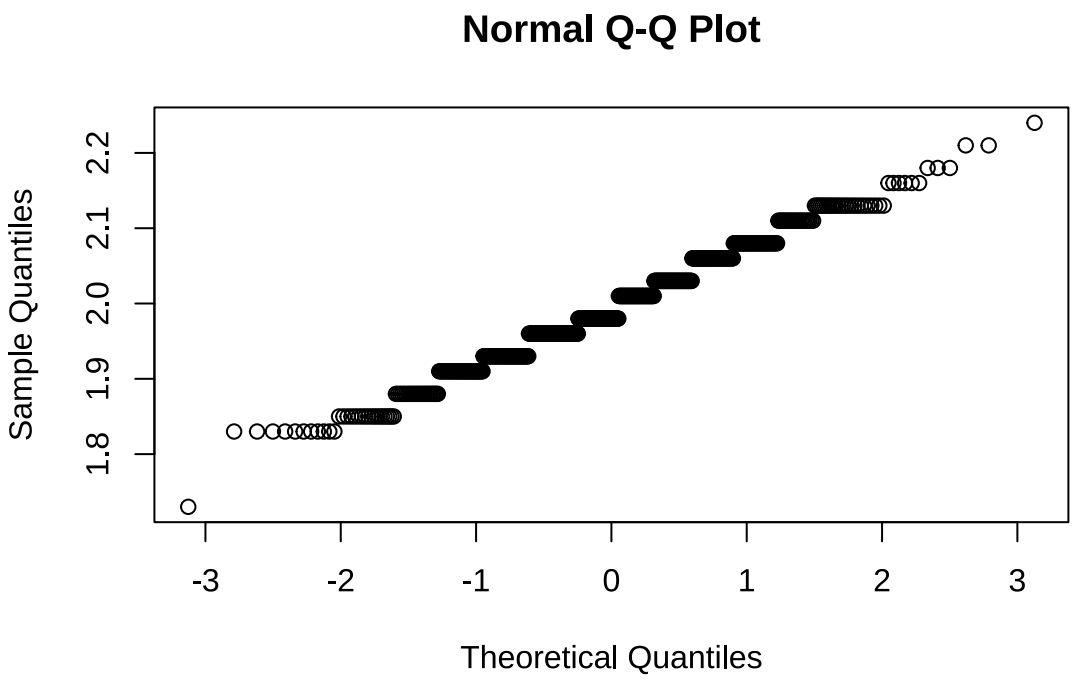
Q-Q Plot de Rebotes por partida



Q-Q Plot de Peso



Q-Q Plot de Altura



Teste de Shapiro-Wilk

A hipótese nula é: A distribuição de Pontos por partida segue distribuição normal.

A hipótese alternativa: A distribuição de Pontos por partida não segue distribuição normal.

```
##
##  Shapiro-Wilk normality test
##
## data:  dados$PPG
## W = 0.9116, p-value < 2.2e-16
```

O resultado indica que devemos rejeitar a hipótese nula.Portanto, a distribuição de Pontos por partida não segue distribuição normal.

Conclusão das normalidades:

Com base nos itens anteriores, é possível afirmar que as variáveis Altura e Peso, visualmente, são as que mais se aproximam de uma distribuição normal. Nos histogramas de Altura e Peso, verificamos um formato de sino e nos Q-Q Plots de Altura e Peso, o resultado foram retas muito próximas do que esperamos para as distribuições normais.

Qualidade dos dados

A Completude dos Dados se refere à presença ou ausência de dados essenciais em um conjunto de dados. Um conjunto de dados é considerado completo quando todos os campos relevantes possuem valores registrados para a abálise, sem dados faltantes. A falta de completude afeta significativamente a confiabilidade de

uma análise exploratória de dados, pelos motivos listados abaixo: Os Dados ausentes podem introduzir viés, levando a conclusões incorretas. Muitos funções e métodos assumem que os dados estão completos. Dados ausentes podem inviabilizar análises, como correlações ou regressões. Variáveis com valores faltantes podem ser excluídas no pré-processamento, reduzindo a quantidade de dados disponível para análise. A necessidade de lidar com valores ausentes, por meio de imputação ou outras técnicas, aumenta o esforço e o tempo necessários para realizar a análise. Ou seja, a completude dos dados é fundamental para garantir a consistência e a precisão de insights extraídos de uma análise.

A análise foi feita utilizando as estatísticas dos jogadores ativos da NBA, no projeto em questão, em relação aos dados que foram utilizados nas análises, verificamos, através das estatística por partida, que os dados faltantes relacionados com as variáveis Pontos, Assistências e Rebotes por partida deveriam ser preenchidos com o valor 0 (zero), o que foi feito. Na coluna referente ao Peso dos atletas, apenas dois registros possuiam dados faltantes. Embora a quantidade de dados faltantes tenha sido pequena, a quantidade de registros utilizados após o tratamento (466 registros) também é pequena. O ideal para a análise em questão teria sido a utilização de dados dos jogadores ativos e inativos, de forma a tornar a base de dados mais completa.

Realize uma operação de imputação de dados usando o pacote MICE.

```
dados_brutos <- read_csv("Dados_auxiliares/dados_apos_coluna_mes.csv", quote = "\"",
                         locale = locale(encoding = "UTF-8"), show_col_types = FALSE)
```

```
dados_para_imputacao <- dados_para_imputacao %>% slice(-507)
```

```
dados_para_imputacao <- dados_para_imputacao %>%
  mutate(
    PPG = replace_na(PPG, 0),
    RPG = replace_na(RPG, 0),
    APG = replace_na(APG, 0)
  )
```

```
dados_para_imputacao <- dados_para_imputacao %>% select(Altura, Peso)
```

```
dados_para_imputacao <- as.data.frame(lapply(dados_para_imputacao, function(x) {
  if (is.character(x)) {
    Encoding(x) <- "UTF-8"
  }
  return(x)
}))
```

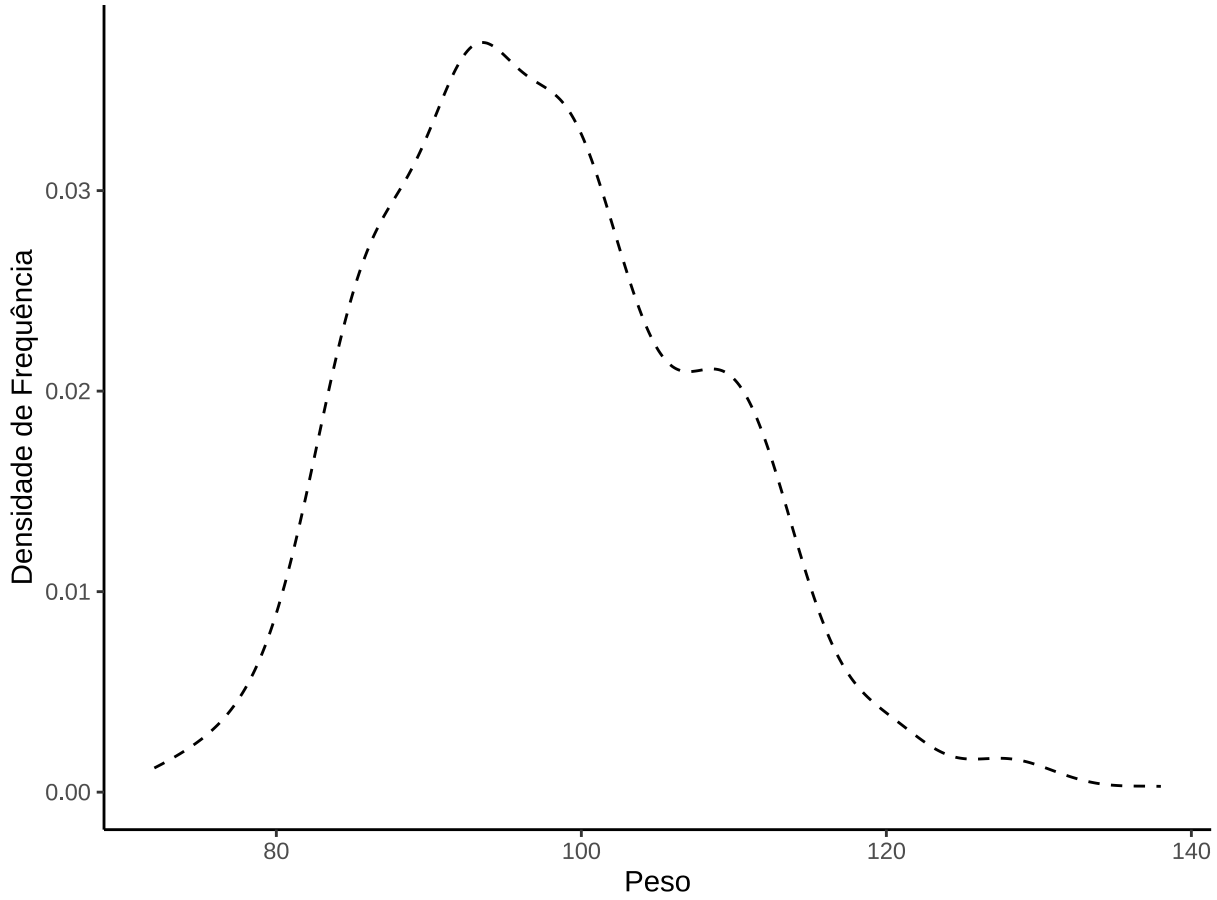
```
imputed_data <- mice(dados,
                     method = "pmm",
                     m = 5)
```

```
##
##  iter imp variable
##   1   1
##   1   2
##   1   3
##   1   4
##   1   5
##   2   1
```

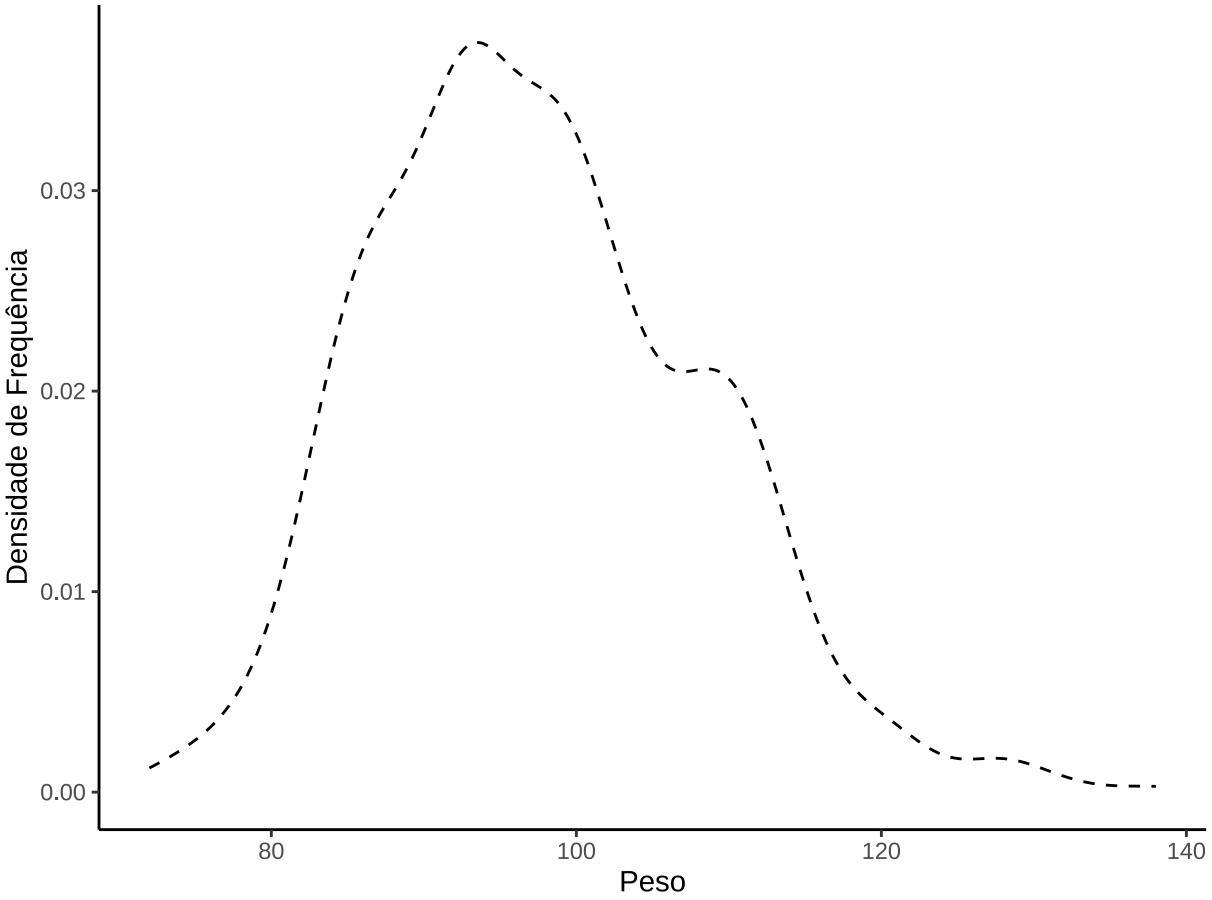
```
##   2   2
##   2   3
##   2   4
##   2   5
##   3   1
##   3   2
##   3   3
##   3   4
##   3   5
##   4   1
##   4   2
##   4   3
##   4   4
##   4   5
##   5   1
##   5   2
##   5   3
##   5   4
##   5   5

## Warning: Number of logged events: 7
```

```
dados_imputados <- complete(imputed_data)
```

```
dados %>% dplyr::select(Peso) %>% ggplot(aes(x=Peso, y = after_stat(density))) + geom_density(linetype =
```




```
dados_imputados %>% dplyr::select(Peso) %>% ggplot(aes(x=Peso, y = after_stat(density))) + geom_density
```



Abaixo segue o link do repositório no Github, onde pode ser encontrado os arquivos RMarkdown e Shiny: Repositório no Github