Trabalho MBA Ciência de Dados Unifor - Análise das Notas do Enade 2017 - Educação Física Licenciatura

Diego Teixeira

25/06/2020

Carregando pacotes

library(tidyverse) library(readr) library(ggplot2) library(plotly) library(e1071) require(dplyr) require(Hmisc) library(DescTools) require(gridExtra) library(flexdashboard)

Carregando os dados

```
microdados_enade <- read.table("MICRODADOS_ENADE_2017.txt",
                               header = TRUE,
                               sep=";",
                               dec = ",",
                               colClasses=c(NT_OBJ_FG="numeric"))
```

Analisando base inicial

Quantidade de linhas e colunas

```
dim(microdados_enade)
```

```
## [1] 537436    150
```

Nomes das colunas

```
names(microdados_enade)
```

```
##   [1] "NU_ANO"             "CO_IES"             "CO_CATEGAD"
##   [4] "CO_ORGACAD"         "CO_GRUPO"           "CO_CURSO"
##   [7] "CO_MODALIDADE"      "CO_MUNIC_CURSO"     "CO_UF_CURSO"
##  [10] "CO_REGIAO_CURSO"    "NU_IDADE"           "TP_SEXO"
##  [13] "ANO_FIM_EM"         "ANO_IN_GRAD"        "CO_TURNO_GRADUACAO"
##  [16] "TP_INSCRICAO_ADM"   "TP_INSCRICAO"       "NU_ITEM_OFG"
##  [19] "NU_ITEM_OFG_Z"      "NU_ITEM_OFG_X"      "NU_ITEM_OFG_N"
##  [22] "NU_ITEM_OCE"        "NU_ITEM_OCE_Z"      "NU_ITEM_OCE_X"
##  [25] "NU_ITEM_OCE_N"      "DS_VT_GAB_OFG_ORIG" "DS_VT_GAB_OFG_FIN"
##  [28] "DS_VT_GAB_OCE_ORIG" "DS_VT_GAB_OCE_FIN"  "DS_VT_ESC_OFG"
##  [31] "DS_VT_ACE_OFG"      "DS_VT_ESC_OCE"      "DS_VT_ACE_OCE"
##  [34] "TP_PRES"            "TP_PR_GER"          "TP_PR_OB_FG"
##  [37] "TP_PR_DI_FG"        "TP_PR_OB_CE"        "TP_PR_DI_CE"
##  [40] "TP_SFG_D1"          "TP_SFG_D2"          "TP_SCE_D1"
##  [43] "TP_SCE_D2"          "TP_SCE_D3"          "NT_GER"
##  [46] "NT_FG"              "NT_OBJ_FG"          "NT_DIS_FG"
##  [49] "NT_FG_D1"           "NT_FG_D1_PT"        "NT_FG_D1_CT"
##  [52] "NT_FG_D2"           "NT_FG_D2_PT"        "NT_FG_D2_CT"
##  [55] "NT_CE"              "NT_OBJ_CE"          "NT_DIS_CE"
##  [58] "NT_CE_D1"           "NT_CE_D2"           "NT_CE_D3"
##  [61] "CO_RS_I1"           "CO_RS_I2"           "CO_RS_I3"
##  [64] "CO_RS_I4"           "CO_RS_I5"           "CO_RS_I6"
##  [67] "CO_RS_I7"           "CO_RS_I8"           "CO_RS_I9"
##  [70] "QE_I01"             "QE_I02"             "QE_I03"
##  [73] "QE_I04"             "QE_I05"             "QE_I06"
##  [76] "QE_I07"             "QE_I08"             "QE_I09"
##  [79] "QE_I10"             "QE_I11"             "QE_I12"
##  [82] "QE_I13"             "QE_I14"             "QE_I15"
##  [85] "QE_I16"             "QE_I17"             "QE_I18"
##  [88] "QE_I19"             "QE_I20"             "QE_I21"
##  [91] "QE_I22"             "QE_I23"             "QE_I24"
##  [94] "QE_I25"             "QE_I26"             "QE_I27"
##  [97] "QE_I28"             "QE_I29"             "QE_I30"
## [100] "QE_I31"             "QE_I32"             "QE_I33"
## [103] "QE_I34"             "QE_I35"             "QE_I36"
## [106] "QE_I37"             "QE_I38"             "QE_I39"
## [109] "QE_I40"             "QE_I41"             "QE_I42"
## [112] "QE_I43"             "QE_I44"             "QE_I45"
## [115] "QE_I46"             "QE_I47"             "QE_I48"
## [118] "QE_I49"             "QE_I50"             "QE_I51"
## [121] "QE_I52"             "QE_I53"             "QE_I54"
## [124] "QE_I55"             "QE_I56"             "QE_I57"
## [127] "QE_I58"             "QE_I59"             "QE_I60"
## [130] "QE_I61"             "QE_I62"             "QE_I63"
## [133] "QE_I64"             "QE_I65"             "QE_I66"
## [136] "QE_I67"             "QE_I68"             "QE_I69"
## [139] "QE_I70"             "QE_I71"             "QE_I72"
## [142] "QE_I73"             "QE_I74"             "QE_I75"
## [145] "QE_I76"             "QE_I77"             "QE_I78"
## [148] "QE_I79"             "QE_I80"             "QE_I81"
```

Filtrando somente as colunas necessárias para análise

CO_TURNO_GRADUACAO: Código do turno de graduação CO_REGIAO_CURSO: Código da região de funcionamento do curso QE_I02: Qual é a sua cor ou raça? CO_GRUPO: Código da área de enquadramento do curso no Enade NT_OBJ_FG: Nota bruta na parte objetiva da formação geral. (valor de 0 a 100)

```
microdados_enade_filtrados= microdados_enade %>% dplyr::select(NT_OBJ_FG, CO_GRUPO,
                                                               CO_REGIAO_CURSO, QE_I02,
                                                               CO_TURNO_GRADUACAO
)
```

Filtrando curso de Educação Física (licenciatura)

Quantidadde de linhas e colunas

```
microdados_ef= microdados_enade_filtrados %>% filter(CO_GRUPO==3502)
dim(microdados_ef)
```

```
## [1] 34763     5
```

Analisando o dataset filtrado

Resumo dos dados

```
summary(microdados_ef)
```

```
##    NT_OBJ_FG         CO_GRUPO    CO_REGIAO_CURSO    QE_I02
##  Min.   :  0.00   Min.   :3502   Min.   :1.000   Length:34763
##  1st Qu.: 25.00   1st Qu.:3502   1st Qu.:3.000   Class :character
##  Median : 37.50   Median :3502   Median :3.000   Mode  :character
##  Mean   : 40.08   Mean   :3502   Mean   :3.079
##  3rd Qu.: 50.00   3rd Qu.:3502   3rd Qu.:4.000
##  Max.   :100.00   Max.   :3502   Max.   :5.000
##  NA's   :6933
##  CO_TURNO_GRADUACAO
##  Min.   :1.000
##  1st Qu.:3.000
##  Median :4.000
##  Mean   :3.168
##  3rd Qu.:4.000
##  Max.   :4.000
##  NA's   :48
```

Nome das colunas

```
names(microdados_ef)
```

```
## [1] "NT_OBJ_FG"          "CO_GRUPO"           "CO_REGIAO_CURSO"
## [4] "QE_I02"             "CO_TURNO_GRADUACAO"
```

Transformando os dados numéricos em textos

```
# Regiao
microdados_ef = microdados_ef %>% mutate(regiao = case_when( CO_REGIAO_CURSO == 1 ~ "Norte",
                                                             CO_REGIAO_CURSO == 2 ~ "Nordest
e",
                                                             CO_REGIAO_CURSO == 3 ~ "Sudeste"
,
                                                             CO_REGIAO_CURSO == 4 ~ "Sul",
                                                             CO_REGIAO_CURSO == 5 ~ "Centro-O
este"
))

# Turno
microdados_ef = microdados_ef %>% mutate(turno = case_when( CO_TURNO_GRADUACAO == 1 ~ "Matuti
no",
                                                            CO_TURNO_GRADUACAO == 2 ~ "Verper
tino",
                                                            CO_TURNO_GRADUACAO == 3 ~ "Integr
al",
                                                            CO_TURNO_GRADUACAO == 4 ~ "Noturn
o",
))

# Raça
microdados_ef = microdados_ef %>% mutate(raca = case_when(  QE_I02 == "A" ~ "Branca",
                                                            QE_I02 == "B" ~ "Preta",
                                                            QE_I02 == "C" ~ "Amarela",
                                                            QE_I02 == "D" ~ "Parda",
                                                            QE_I02 == "E" ~ "Indígena",
                                                            QE_I02 == "F" ~ "Não quero declar
ar",
))
```

Verificando colunas criadas

```
names(microdados_ef)
```

```
## [1] "NT_OBJ_FG"          "CO_GRUPO"           "CO_REGIAO_CURSO"
## [4] "QE_I02"             "CO_TURNO_GRADUACAO" "regiao"
## [7] "turno"              "raca"
```

ANALISE DESCRITIVA

Analisando valores NA em porcentagem

```
totalLinhas = dim(microdados_ef[1])
# Coluna 01
na <- length(microdados_ef$NT_OBJ_FG[which(is.na(microdados_ef$NT_OBJ_FG))])
round(na/totalLinhas * 100,2)[1]
```

```
## [1] 19.94
```

```
# Coluna 02
na <- length(microdados_ef$CO_GRUPO[which(is.na(microdados_ef$CO_GRUPO))])
round(na/totalLinhas * 100,2)[1]
```

```
## [1] 0
```

```
# Coluna 03
na <- length(microdados_ef$CO_REGIAO_CURSO[which(is.na(microdados_ef$CO_REGIAO_CURSO))])
round(na/totalLinhas * 100,2)[1]
```

```
## [1] 0
```

```
# Coluna 04
na <- length(microdados_ef$QE_I02[which(is.na(microdados_ef$QE_I02))])
round(na/totalLinhas * 100,2)[1]
```

```
## [1] 0
```

```
# Coluna 05
na <- length(microdados_ef$CO_TURNO_GRADUACAO[which(is.na(microdados_ef$CO_TURNO_GRADUACA
O))])
round(na/totalLinhas * 100,2)[1]
```

```
## [1] 0.14
```

Nota-se que a coluna 01 tem 19,04% de valores nulos e a coluna 05 tem 0,14%. A seguir removeremos estes dados nulos.

Removendo as NAs

```
microdados_ef_final=microdados_ef %>% na.omit()
```

Verificando se a remoção ocorreu com sucesso e quantos porcento dos dados foram reduzidos

Porcentagem de dados removidos por NA

```
round((dim(microdados_ef_final)[1]/dim(microdados_ef)[1] -1) * 100, 2)
```

```
## [1] -21.65
```

```
summary(microdados_ef_final)
```

```
##    NT_OBJ_FG         CO_GRUPO    CO_REGIAO_CURSO    QE_I02
##  Min.   :  0.00   Min.   :3502   Min.   :1.000   Length:27238
##  1st Qu.: 25.00   1st Qu.:3502   1st Qu.:3.000   Class :character
##  Median : 37.50   Median :3502   Median :3.000   Mode  :character
##  Mean   : 40.18   Mean   :3502   Mean   :3.076
##  3rd Qu.: 50.00   3rd Qu.:3502   3rd Qu.:4.000
##  Max.   :100.00   Max.   :3502   Max.   :5.000
##  CO_TURNO_GRADUACAO    regiao             turno               raca
##  Min.   :1.000      Length:27238       Length:27238       Length:27238
##  1st Qu.:3.000      Class :character   Class :character   Class :character
##  Median :4.000      Mode  :character   Mode  :character   Mode  :character
##  Mean   :3.161
##  3rd Qu.:4.000
##  Max.   :4.000
```

Analisando as notas

```
quantidade <- length(microdados_ef_final$NT_OBJ_FG)
```

```
#Calculando a Média
media <- mean(microdados_ef_final$NT_OBJ_FG)
media
```

```
## [1] 40.17687
```

```
#Calculando a mediana
#De forma direta
mediana <- median(microdados_ef_final$NT_OBJ_FG)
mediana
```

```
## [1] 37.5
```

```
#Aplicando a Teoria
mediana2 <- (sort(microdados_ef_final$NT_OBJ_FG)[dim(microdados_ef_final)/2] +
    sort(microdados_ef_final$NT_OBJ_FG)[dim(microdados_ef_final)/2+1]) / 2
mediana2
```

```
## [1] 37.5  0.0
```

```
#Moda das notas
moda <- Mode(microdados_ef_final$NT_OBJ_FG)
moda
```

```
## [1] 37.5
## attr(,"freq")
## [1] 6796
```

```
consolidado_notas=data.frame("Quantidade"=quantidade,
                             "Media"=media,
                             "Mediana"=mediana,
                             "moda"=moda)
consolidado_notas
```

```
##   Quantidade    Media Mediana moda
## 1      27238 40.17687    37.5 37.5
```

Assimetria

```
assimetria=skewness(microdados_ef_final$NT_OBJ_FG)
assimetria
```

```
## [1] 0.172871
```

Curtose

```
curtose=kurtosis(microdados_ef_final$NT_OBJ_FG)
curtose
```

```
## [1] -0.2825687
```

```
consolidado_notas_completo=cbind(consolidado_notas,assimetria, curtose)
consolidado_notas_completo
```

```
##   Quantidade    Media Mediana moda assimetria    curtose
## 1      27238 40.17687    37.5 37.5   0.172871 -0.2825687
```

Conclusões nesta análise descritiva

Coefieciente de assimetria de pearson>0, logo terá assimetria positiva e concentração a esquerda dos maiores valores.

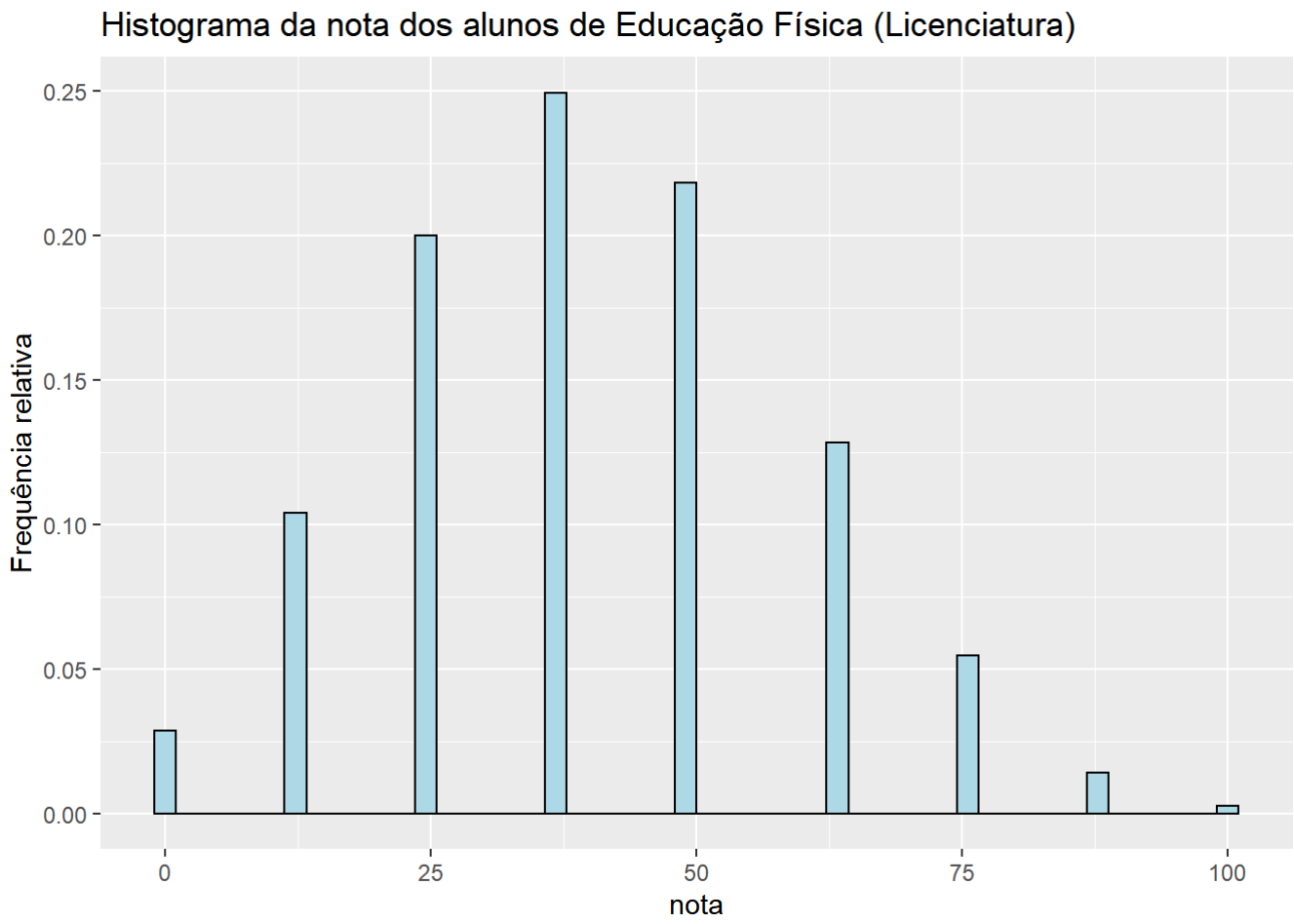
k>0, leptocúrtica k=0, Mesocúrtica k<0, Platicúrtica

Consideramos então platicúrtica.

ANALISE GRAFICA

Histograma das notas

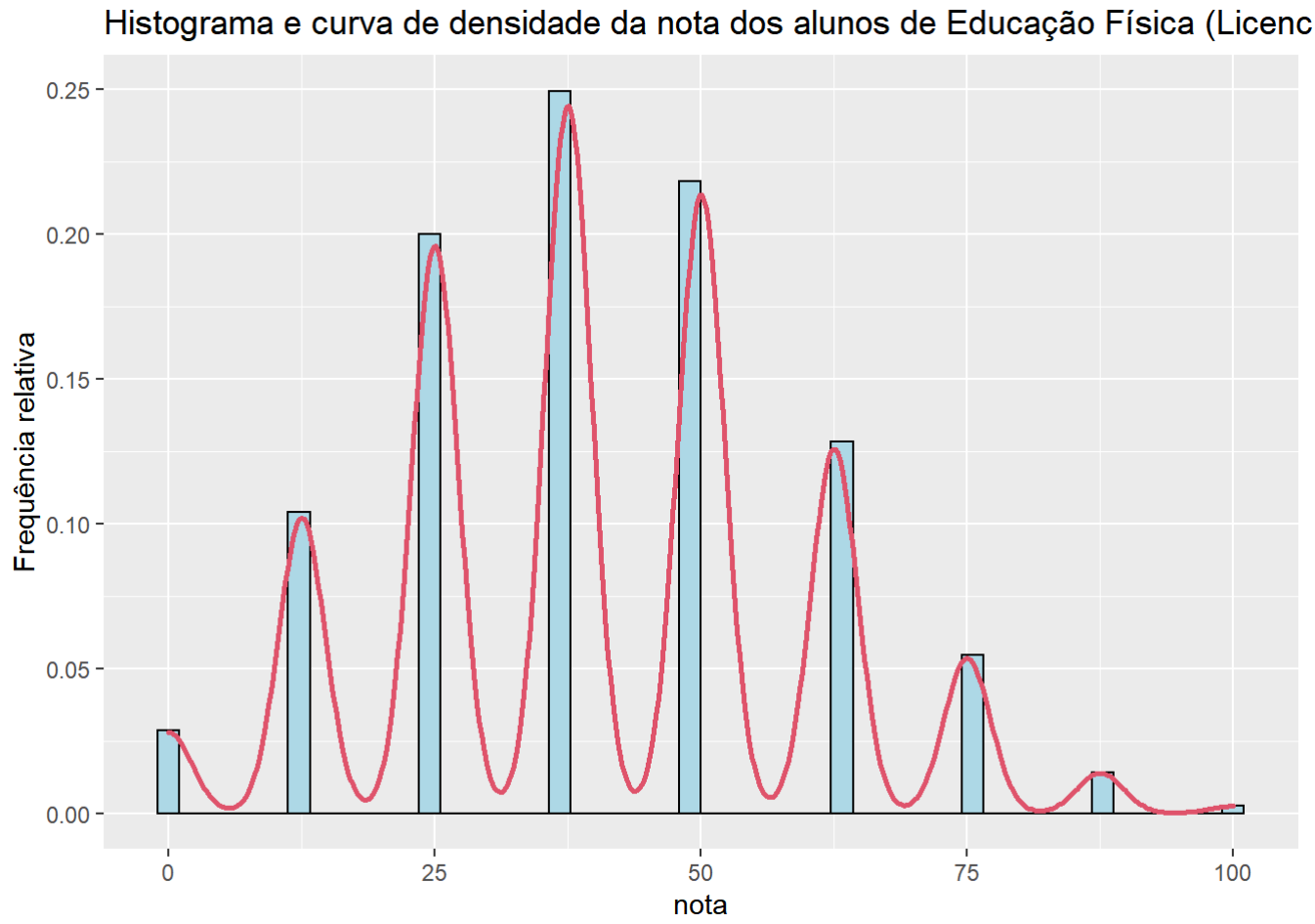
```
g_hist=ggplot(microdados_ef_final,aes(x=NT_OBJ_FG)) +
  geom_histogram(color = "black",fill="lightblue",bins =50,aes(y=(..count..)/sum(..count..)))
+
  ggtitle("Histograma da nota dos alunos de Educação Física (Licenciatura)")+
  xlab("nota") +
  ylab("Frequência relativa")
g_hist
```



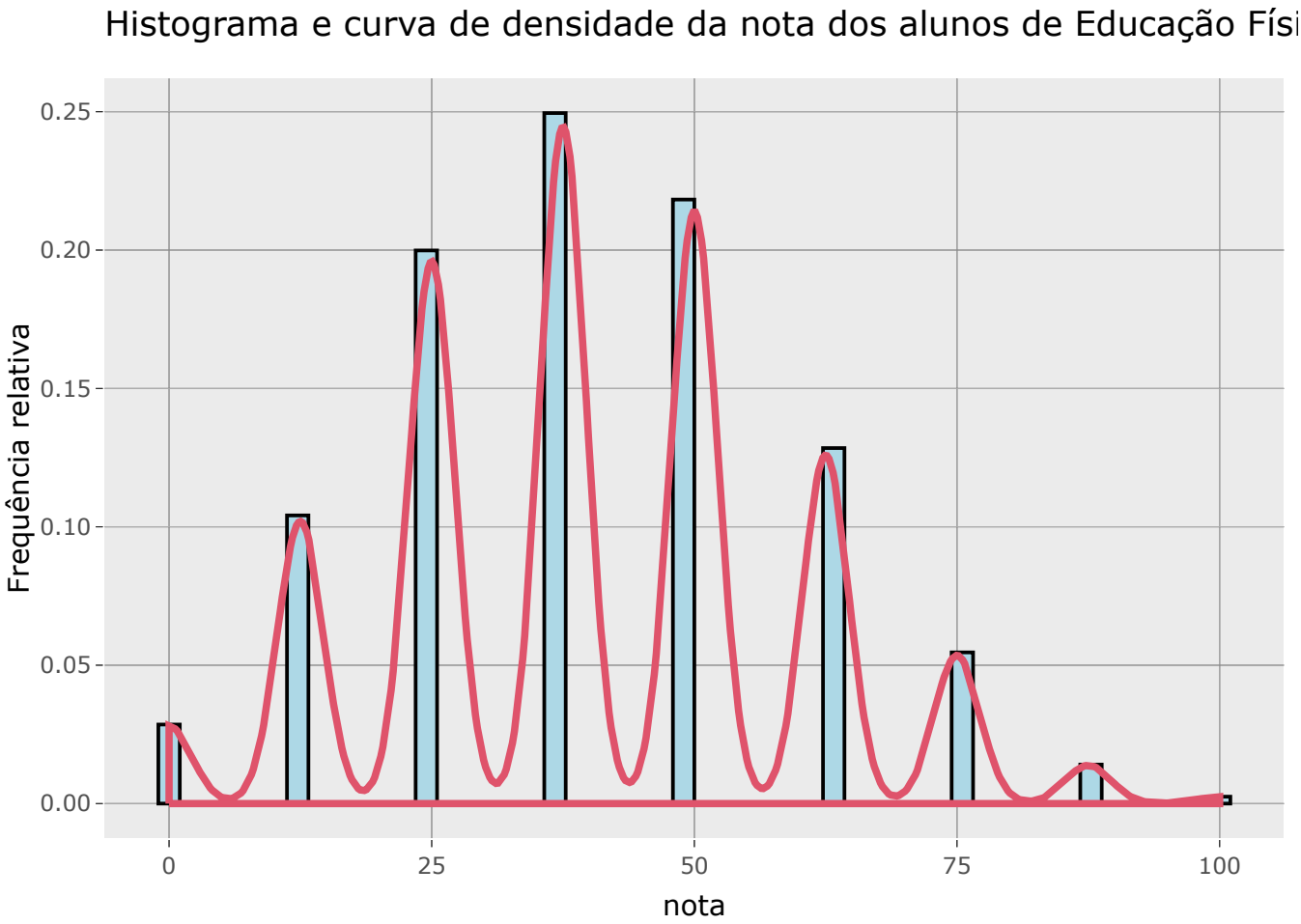
No histograma acima, oberva-se que a notas tem uma frequência maior nos valores entre 25 e 50, enquanto as notas entre 75 e 100 possuem uma menor aparição na base de dados.

Histograma e Densidade

```
g_hist_densidade = ggplot(microdados_ef_final,aes(x=NT_OBJ_FG)) +
  geom_histogram(color = "black",fill="lightblue",bins =50,aes(y=(..count..)/sum(..count..)))
+
  geom_density(col=2,size = 1, aes(y = 27 * (..count..)/sum(..count..))) +
  ggtitle("Histograma e curva de densidade da nota dos alunos de Educação Física (Licenciatur
a)")+
  xlab("nota") +
  ylab("Frequência relativa")
g_hist_densidade
```



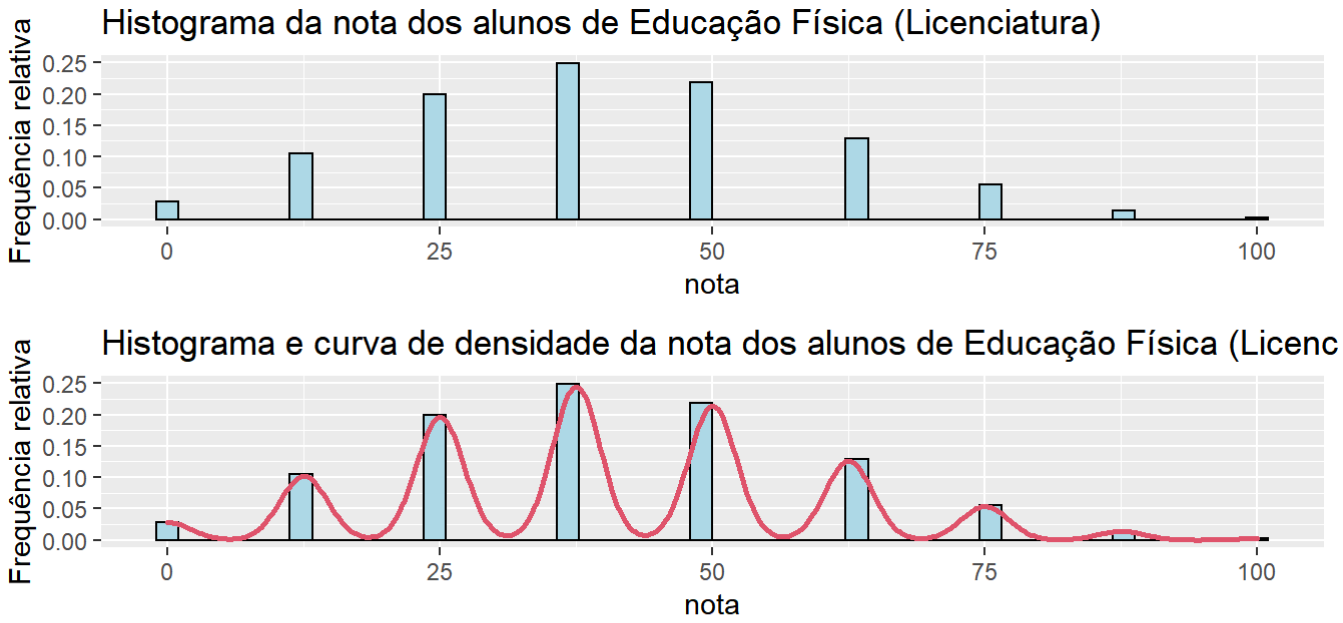

```
ggplotly(g_hist_densidade)
```



Acima é possível observar o histograma inicial com uma linha que representa a densidade das notas.

Agrupamento dos gráficos

```
grid.arrange( g_hist,
              g_hist_densidade,
              nrow=3,ncol=1)
```



ANALISE COMPARATIVA

Comparando notas por Região

```
microdados_ef_regiao= microdados_ef_final %>%
  select(regiao,NT_OBJ_FG) %>%
  group_by(regiao) %>%
  summarise(quantidade=n(),
              media = mean(NT_OBJ_FG,na.rm = T),
              mediana = median(NT_OBJ_FG,na.rm = T),
              cv=sd(NT_OBJ_FG,na.rm=T)/media*100,
              amplitude_interquartil=IQR(NT_OBJ_FG)) %>%
  arrange(desc(mediana))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
microdados_ef_regiao
```

```
## # A tibble: 5 x 6
##   regiao       quantidade media mediana    cv amplitude_interquartil
##   <chr>             <int> <dbl>   <dbl> <dbl>                  <dbl>
## 1 Centro-Oeste       1924  40.6    37.5  46.8                     25
## 2 Nordeste           4411  41.4    37.5  46.4                     25
## 3 Norte              1952  39.2    37.5  48.3                     25
## 4 Sudeste           12416  39.8    37.5  48.0                     25
## 5 Sul                6535  40.3    37.5  47.6                     25
```

A tabela acima mostra a quantidade de notas informadas por região, tendo a região sudeste em primeiro lugar com 12416 e centro-oeste em último com 1924 notas informadas. Também é possível identificar a região com maior média na prova, embora estejam muito próximas o centro-oeste possui uma média maior entre as região computando 40,64 de média, seguinda da região sul com 40,29. Em último lugar temos a região norte com 39,17.

Comparando notas por Raça

```
microdados_ef_raca= microdados_ef_final %>%
  select(raca,NT_OBJ_FG) %>%
  group_by(raca) %>%
  summarise(quantidade=n(),
              media = mean(NT_OBJ_FG,na.rm = T),
              mediana = median(NT_OBJ_FG,na.rm = T),
              cv=sd(NT_OBJ_FG,na.rm=T)/media*100,
              amplitude_interquartil=IQR(NT_OBJ_FG)) %>%
  arrange(desc(mediana))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
microdados_ef_raca
```

```
## # A tibble: 6 x 6
##   raca               quantidade media mediana    cv amplitude_interquartil
##   <chr>                   <int> <dbl>   <dbl> <dbl>                  <dbl>
## 1 Amarela                   571  40.6    37.5  47.2                     25
## 2 Branca                  12121  41.2    37.5  46.6                     25
## 3 Indígena                  166  37.3    37.5  54.2                     25
## 4 Não quero declarar        498  41.1    37.5  50.7                     34.4
## 5 Parda                   10400  39.1    37.5  48.6                     25
## 6 Preta                    3482  39.7    37.5  47.0                     25
```

Na tabela acima temos uma análise das notas informadas por Raça, identifica-se que a Raça branca possui uma média alta com 41,24 em comparação as demais que seguem com 41,11 referente a participantes que não declararam e 40,56 declarados amarela, em útlimo aparece a indígena com 37,34. Importante obervar que a média mais baixa só possui 166 notas informadas que são os indígenas e a Raça branca além de conter a média mais alta, possui também maior número de notas informadas.

Tabulação cruzada proporção entre Região e Raça(%)

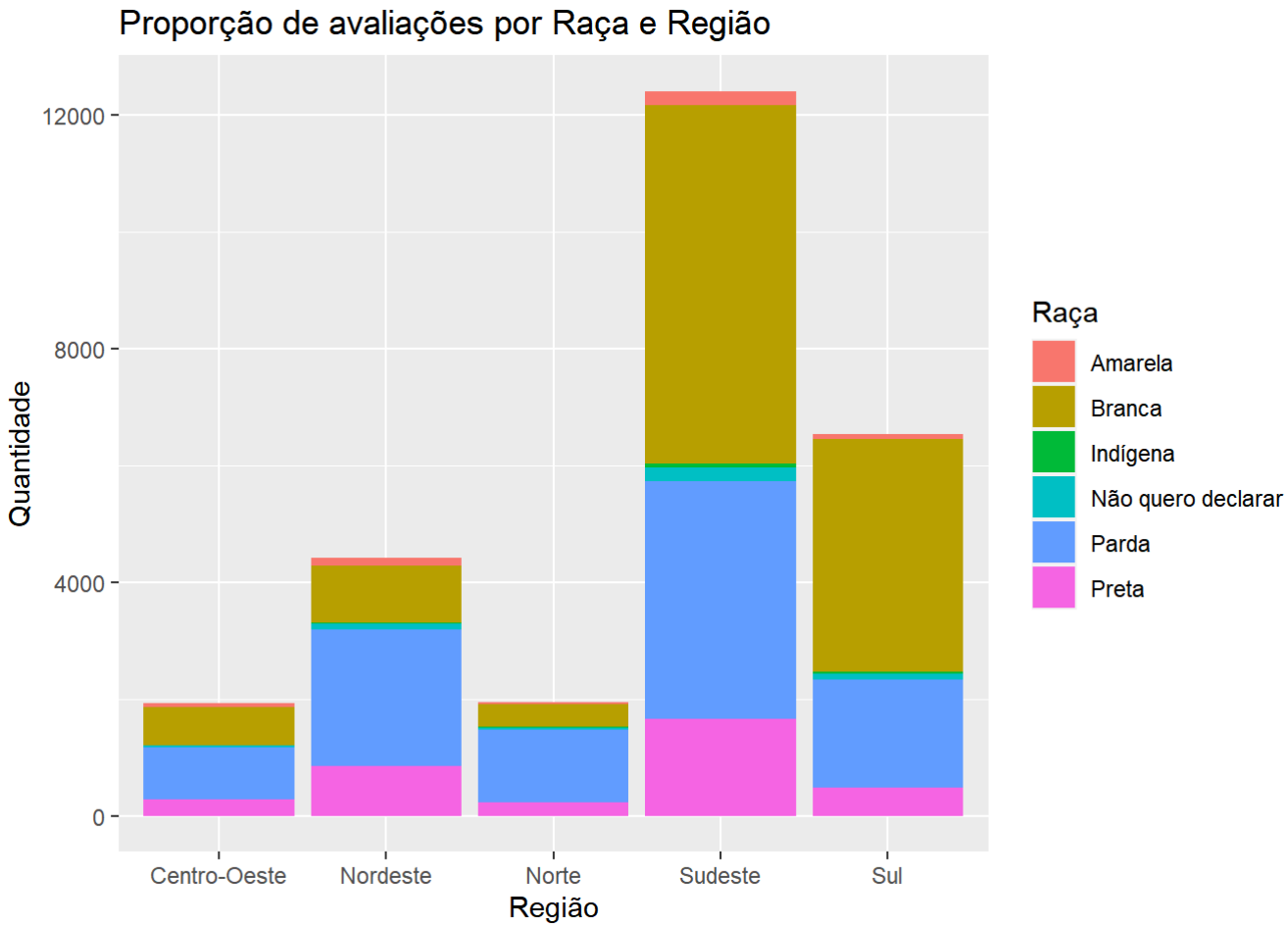
```
round(prop.table(table(microdados_ef_final$regiao,microdados_ef_final$raca)),4)*100
```

```
##
##                Amarela Branca Indígena Não quero declarar Parda Preta
##   Centro-Oeste    0.22   2.40     0.05                0.12  3.27  1.00
##   Nordeste        0.48   3.55     0.11                0.35  8.61  3.11
##   Norte           0.18   1.40     0.08                0.12  4.57  0.83
##   Sudeste         0.91  22.54     0.26                0.85 14.93  6.09
##   Sul             0.31  14.62     0.11                0.39  6.81  1.75
```

Acima podemos identificar as frequências de notas informadas entre Região e Raça. A raça preta possui uma maior participação na região Sudeste com 6,09% e menor participação na região Norte com 0,83%. A raça amarela possui seu pico na região Sudeste e a menor frequência na região Centro-Oeste. As raças Branca e Parda possuem maior participação geral com frequências de 22,54% e 14,93% respectivamente a qual estão concentradas na região Sudeste.

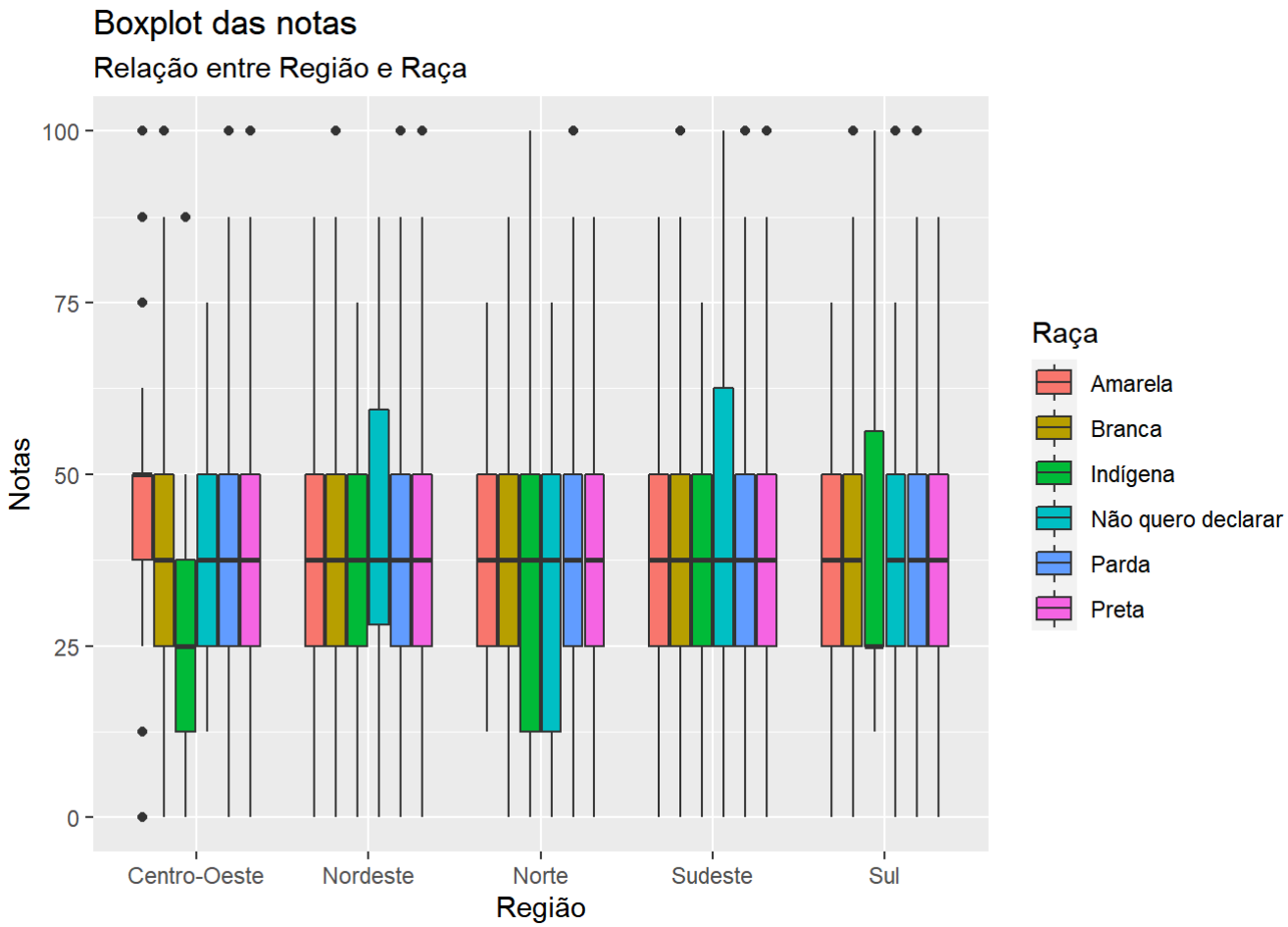
Outras comparações entre as variáveis

```
ggplot(microdados_ef_final) +
 aes(x = regiao, fill = raca, size = NT_OBJ_FG) +
 geom_bar() +
 scale_fill_hue() +
 labs(x = "Região", y = "Quantidade", title = "Proporção de avaliações por Raça e Região", fi
ll = "Raça") +
 theme_gray() +
 theme(legend.position = "right")
```



O gráfico de barras ilustrado acima reflete os números vistos na tabela anterior. A região sudeste possui mais notas informadas, seguida da região sul e nordeste. Pela análise visual, fica muito percepitível uma boa participação da raça branca em todas as regiões, apesar de não ser a maior proporção em todas. Em seguida obervamos a raça parda e depois os declarados de raça preta.

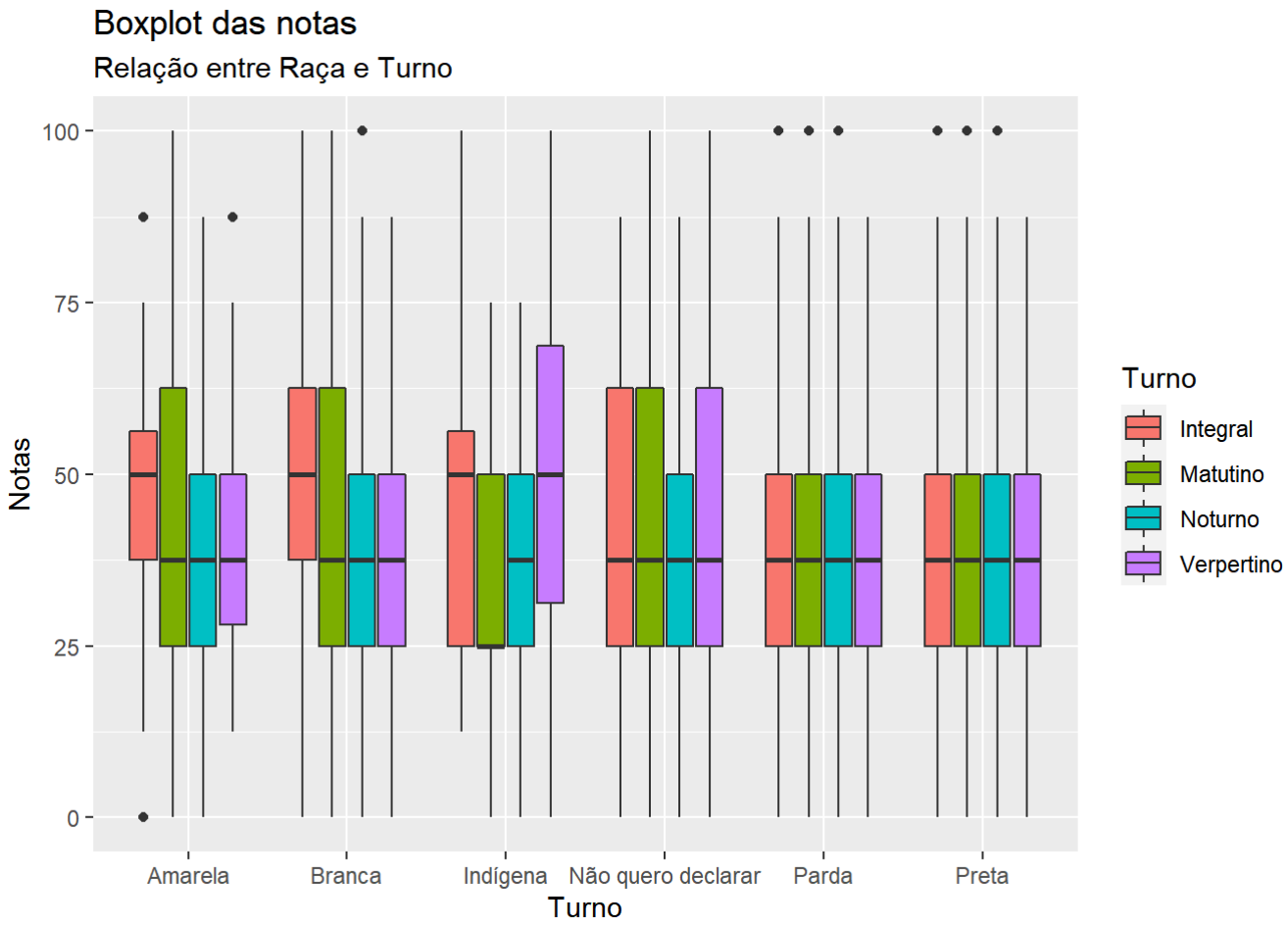
```
ggplot(microdados_ef_final) +
 aes(x = regiao, y = NT_OBJ_FG, fill = raca) +
 geom_boxplot() +
 scale_fill_hue() +
 labs(x = "Região", y = "Notas", title = "Boxplot das notas", subtitle = "Relação entre Regiã
o e Raça", fill = "Raça") +
 theme_gray() +
 theme(legend.position = "right")
```



Pelo boxplot acima podemos deduzir:

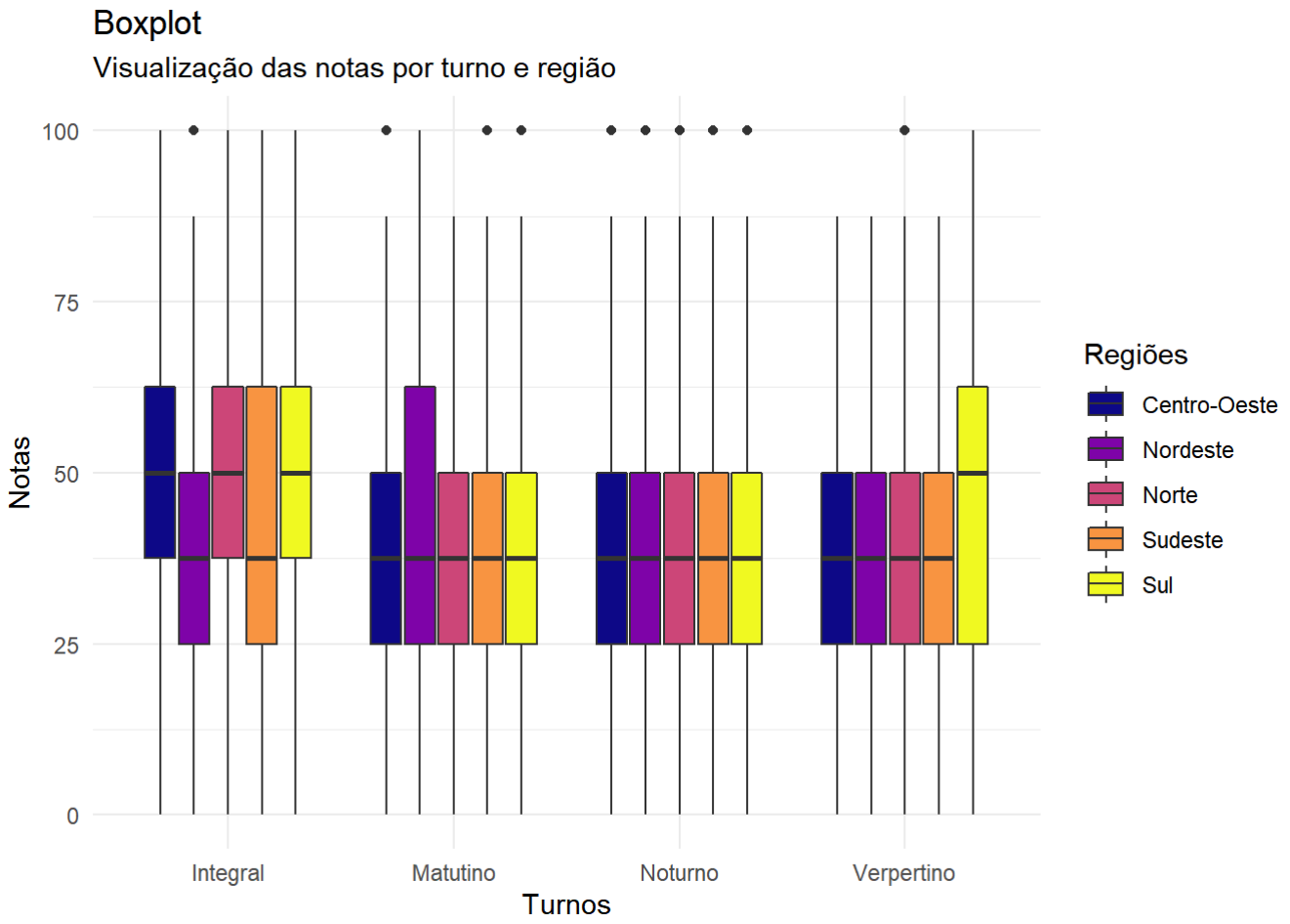
1. Região norte possui uma parcela de notas mais baixas, concentradas nas raças indígenas e nos que não declararam a que raça pertencem.Também temos uma concentração de notas baixas na região centro-oeste atribuídas a raça indígena.Em contrapartida, a raça indígena na região sul obtém um bom resultado se comparado as demais raças na mesma região.
2. Os que não quiseram declarar sua raça atingiram bons resultados nas regiões nordeste e sudeste se comparado a todas as outras regiões.

```
ggplot(microdados_ef_final) +
 aes(x = raca, y = NT_OBJ_FG, fill = turno) +
 geom_boxplot() +
 scale_fill_hue() +
 labs(x = "Turno", y = "Notas", title = "Boxplot das notas", subtitle = "Relação entre Raça e
Turno", fill = "Turno") +
 theme_gray() +
 theme(legend.position = "right")
```



No gráfico acima podemos ver uma porção de notas altas do turno vespertino, mas o turno integral apresenta uma tendência mais forte para com as notas altas, principalmente quando tratamos das raças branca e amarela.

```
ggplot(microdados_ef_final) +
 aes(x = turno, y = NT_OBJ_FG, fill = regiao) +
 geom_boxplot() +
 scale_fill_viridis_d(option = "plasma") +
 labs(x = "Turnos", y = "Notas", title = "Boxplot", subtitle = "Visualização das notas por tu
rno e região", fill = "Regiões") +
 theme_minimal()
```



Visualizando os dados através dos turnos e regiões, observamos uma maior distribuição quando tratamos o turno integral, onde também observamos uma proporção maior de notas altas. O Turno verpertino se destaca na região Sul com uma média acima das demais regiões.