Análise de dados com R

Média de alunos por região

Professores Manuel Martins e Sérgio Assunção

Ismael Weslley Neves de Brito - 2018101710
Leonardo Monteiro Assumpção - 2018200873

Recursos utilizados

R Studio - Disponível em:
https://www.rstudio.com/

Base de dados utilizada “Taxas – Escolas 2010” Disponível em:
https://dados.gov.br/dataset/media-de-alunos-por-turma-na-educacao-basica

Bibliotecas R utilizadas

```
#Importa a biblioteca necessária para ler os dados da PLanilha
library(readxl)
```

Detalhes iniciais

A base de dados estava inicialmente dividida em 6 partes, então para evitar alterações foi mantida a divisão com 6 regiões, sendo:

- NordesteA: NORDESTE - EXT MA E BA
- NordesteB: NORDESTE - SOMENTE MA E BA

Todas as informações mostradas, foram calculadas sem a remoção dos outliers, ou seja utilizando integralmente a base de dados com todos os valores disponíveis para cada coluna.

Objetivo

Utilizar uma base de dados, para obter as seguintes informações individuais separadas por regiões:

Calcular as Médias de cada uma dos 9 Anos/série e demonstrar visualmente os resultados em um gráfico de barras.

1º Ano	1ª série/ 2° ano	2ª série/ 3° ano	3ª série/ 4° ano	4ª série/ 5° ano	5ª série/ 6° ano	6ª série/ 7° ano	7ª série/ 8° ano	8ª série/ 9° ano

Fazer uma análise exploratória dos dados das colunas que engloba pelo menos as seguintes informações:

- Valor Mínimo e Máximo;
- Desvio padrão;
- Mediana;
- Quartis;
- Outliers.

Etapas realizadas

Importar dados → Tratar dados → Manipular dados

Carregar base de dados

Os dados devem ser corretamente carregados e armazenados para serem utilizados.

Remoção de nulos

Dados que não estão corretamente descritos devem ser eliminados dos cálculos que serão realizados.

Transformar em informação

Realizar os devidos cálculos e operações com uma solução escolhida, objetivando a obtenção das informações através dos dados.

Importação dos dados

```
#Define o Endereço em que a planilha de dados se localiza
#Alterando essa linha, as outras importações também serão alteradas
url<-c("C:\\Users\\leloe\\Desktop\\TrabalhoR\\escolas_media_alunos_turma_2010.xls")

#Define o tipo dos dados de cada coluna da planilha
colunas<-c("numeric","text","text","numeric","text","numeric","text","text","text","numeric","numeric","numeric","n
umeric","numeric","numeric","numeric","numeric","numeric","numeric","numeric","numeric","numeric","numeric"
,"numeric","numeric","numeric","numeric","numeric","numeric","numeric","numeric")

#Importa as regiões para cada variável individual, transformando – em NA e ignorando linhas iniciais
N     <- read_excel(path=url,sheet = 1,skip=8,na ="--",col_types=colunas) #NORTE
NDa <- read_excel(path=url,sheet = 2,skip=8,na ="--",col_types=colunas) #NORDESTE A  EXT MA E BA
NDb <- read_excel(path=url,sheet = 3,skip=8,na ="--",col_types=colunas) #NORDESTE B SOMENTE MA E BA
SD    <- read_excel(path=url,sheet = 4,skip=8,na ="--",col_types=colunas) #SUDESTE
S      <- read_excel(path=url,sheet = 5,skip=8,na ="--",col_types=colunas) #SUL
CO    <- read_excel(path=url,sheet = 6,skip=8,na ="--",col_types=colunas) #CENTRO-OESTE
```

Tratamento dos dados

Os dados foram tratados da seguinte forma:

Primeiro foi definido o tipo dos dados de cada coluna, para evitar eventuais erros de tipo incorreto durante a utilização de funções que recebem números como parâmetros:

```
#Define o tipo dos dados de cada coluna da planilha
colunas<-c("numeric","text","text","numeric","text","numeric","text","text","text","numeric","numeric","numeric","n
umeric","numeric","numeric","numeric","numeric","numeric","numeric","numeric","numeric","numeric","numeric"
,"numeric","numeric","numeric","numeric","numeric","numeric","numeric","numeric")
```

Também foi utilizado o argumento **na.rm = TRUE** dentro das funções de cálculo utilizadas, que tem como objetivo, remover todos os eventuais valores não numéricos que possam alterar os resultados finais nos cálculos e resultar em uma análise incorreta e errônea dos dados.

Solução escolhida

Criação de uma função geral:

dados<-function(tb,visual,nome)

Retorna um data frame com os dados exploratórios (1° até 9°ano) de uma região, exibindo também o seu gráfico de médias no final de sua execução.

A função dados() foi criada para receber como parâmetro respectivamente:

- Uma das 6 regiões delimitadas na base de dados;
- Uma cor para o gráfico resultante;
- O nome da região para serem utilizadas na criação do gráfico.

Resumo das funções R utilizadas

```
round()              #Arredonda valores
c()                  #Concatena valores em um vetor
read_excel()         #Importa uma planilha em uma variável
rbind()              #Combina vetores,matrizes e dataframes
mean()               #Média aritmética dos valores
barplot()            #Exibe um gráfico de barras
text()               #Utilizada para exibir um texto adicional no gráfico
print()              #Exibir valor na tela
colorRampPalette()   #Gerar paleta de cores
min()                #Retorna o valor mínimo do conjunto de dados
max()                #Retorna o valor máximo do conjunto de dados
sd()                 #Retorna o desvio padrão do conjunto de dados
median()             #Retorna a mediana do conjunto de dados
quantile()           #Retorna um determinado quartil do conjunto de dados
paste()              #Concatena um vetor depois de converter em caractere
sort()               #Ordena um conjunto de dados
boxplot.stats()$out  #Exibe outliers de um boxplot
unique()             #Exibe apenas valores que não sejam repetidos
```

```
View()       #Exibe valores de variável
data.frame() #cria um data frame
```

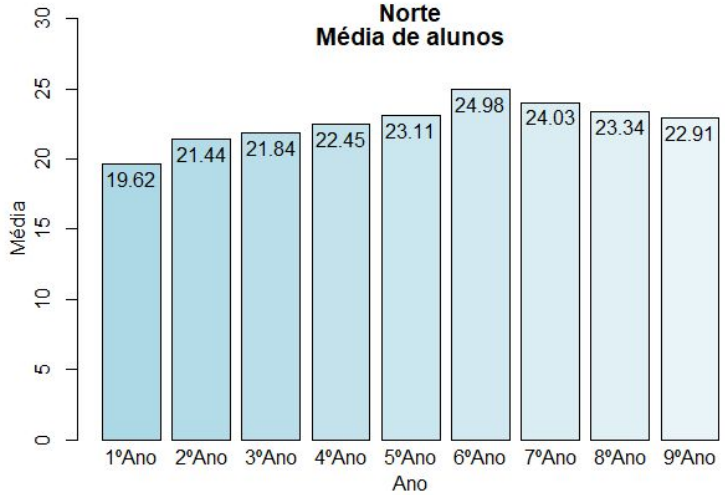
Exemplo de utilização da função geral

Temos por exemplo, a seguinte execução:

```
#Variável onde os dados da tabela norte foram armazenados, cor lightblue para o visual do gráfico
#A string “Norte” para título do gráfico
N_Dados<-dados(N,"lightblue","Norte")
```

Será então obtido o seguinte resultado:

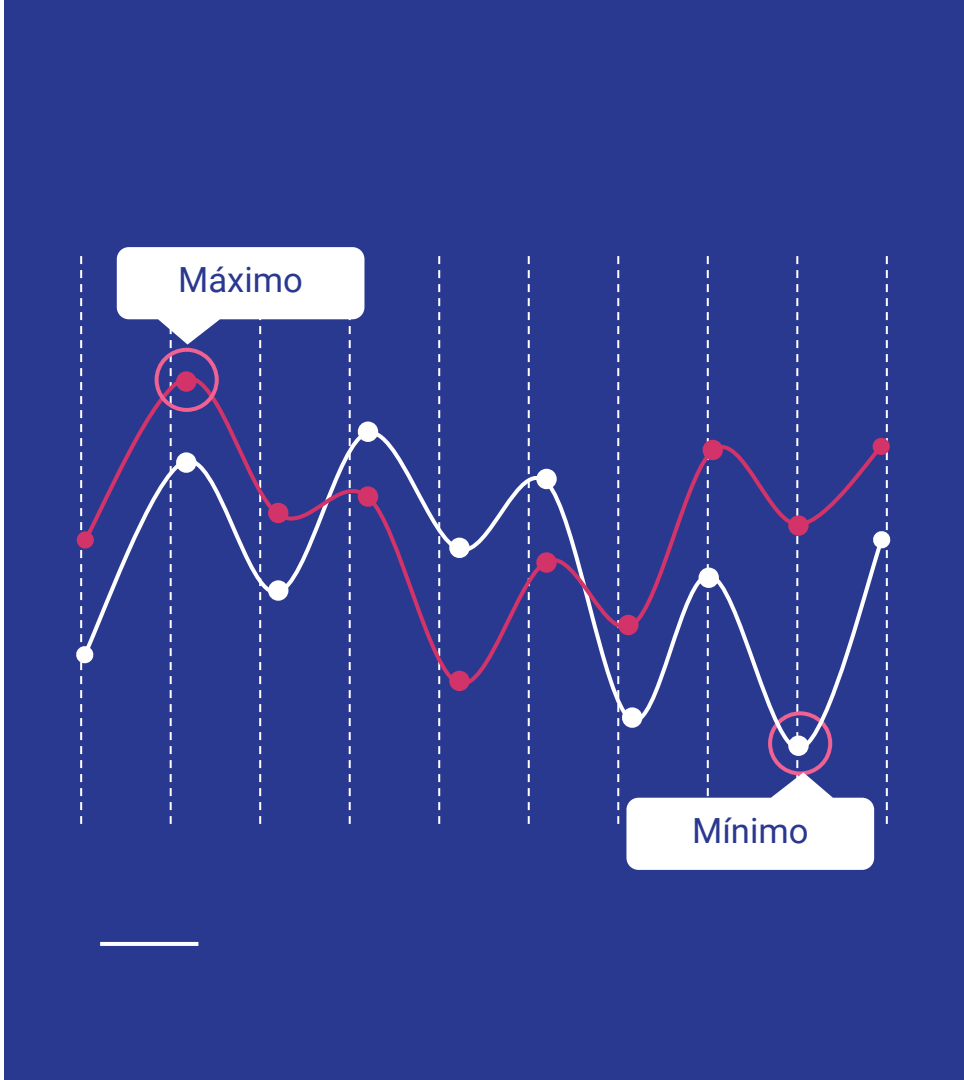
	ano	Medias	Min	Max	Desvio	Mediana	Quartis	Outliers
1	1ºAno	19.62	1	51	7.403420	20.0	15 20 25 51	40.5 41 42 43 45 46 50 51
2	2ºAno	21.44	1	79	7.838771	22.0	16 22 27 79	44 45 46.5 47 47.5 50 53 79
3	3ºAno	21.84	1	80	8.367538	22.0	16 22 28 80	48 50 52 53 57 80
4	4ºAno	22.45	1	80	8.775906	23.0	16.5 23 29 80	48 51 53 57 60 80
5	5ºAno	23.11	1	80	9.177292	23.5	17 23.5 30 80	50 50.3 51 73 80
6	6ºAno	24.98	1	110	10.738367	26.0	17 26 33 110	59 61 62 70 90 110
7	7ºAno	24.03	1	90	11.145596	25.0	15 25 32.725 90	62 66 80 90
8	8ºAno	23.34	1	96	11.303170	24.0	14 24 32 96	60 61 63 73 75 91 96
9	9ºAno	22.91	1	89	11.797027	23.0	13 23 32 89	63 64 67 73 81 89



Gráficos das regiões

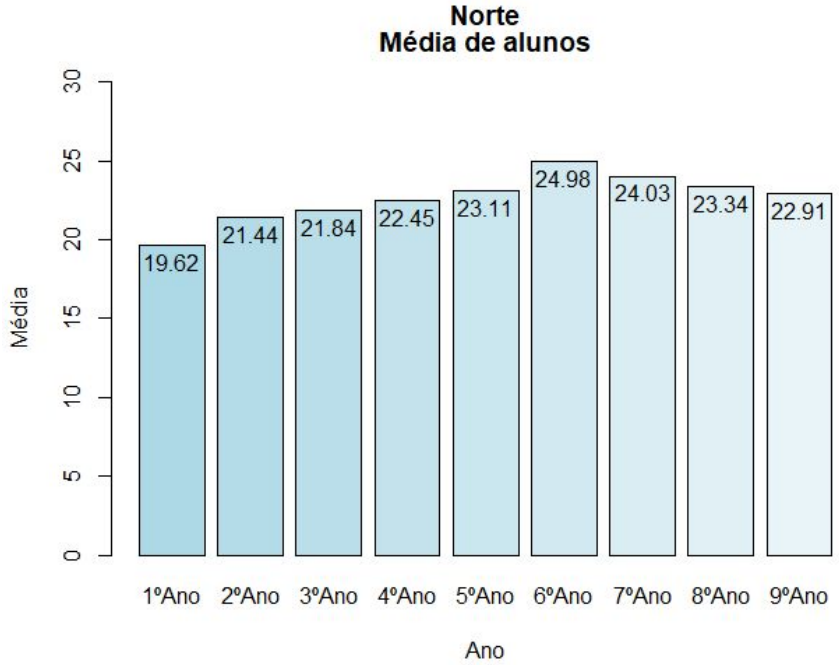
Médias de alunos de cada ano separado por região.

Dados exploratórios.



Norte

Valores avaliados na menor coluna: 5059
Valores avaliados na maior coluna: 7219
Quantidade de linhas (incluindo NA): 24032

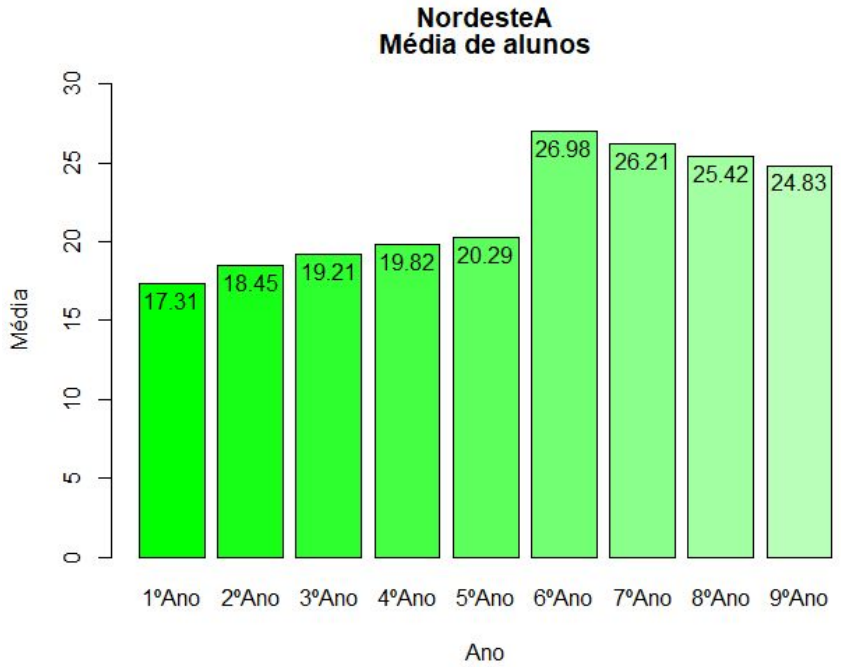


#Região com baixa quantidade de outliers distintos
#Região com menor diferença do 5° para 6° ano (1.87)
#Possui as maiores médias de 1° até o 5° ano

Norte							
Ano	Media	Menor	Maior	Desvio	Mediana	Quartis	Outliers
1ºAno	19.62	1	51	7.403	20.0	15-20-25-51	40.5 41 42 43 45 46 50 51
2ºAno	21.44	1	79	7.838	22.0	16-22-27-79	44 45 46.5 47 47.5 50 53 79
3ºAno	21.84	1	80	8.367	22.0	16-22-28-80	48 50 52 53 57 80
4ºAno	22.45	1	80	8.775	23.0	16.5-23-29-80	48 51 53 57 60 80
5ºAno	23.11	1	80	9.177	23.5	17-23.5-30-80	50 50.3 51 73 80
6ºAno	24.98	1	110	10.738	26.0	17-26-33-110	59 61 62 70 90 110
7ºAno	24.03	1	90	11.145	25.0	15-25-32.7-25-90	62 66 80 90
8ºAno	23.34	1	96	11.303	24.0	14-24-32-96	60 61 63 73 75 91 96
9ºAno	22.91	1	89	11.797	23.0	13-23-32-89	63 64 67 73 81 89

NordesteA

Valores avaliados na menor coluna: 10598
Valores avaliados na maior coluna: 17644
Quantidade de linhas (incluindo NA): 42014

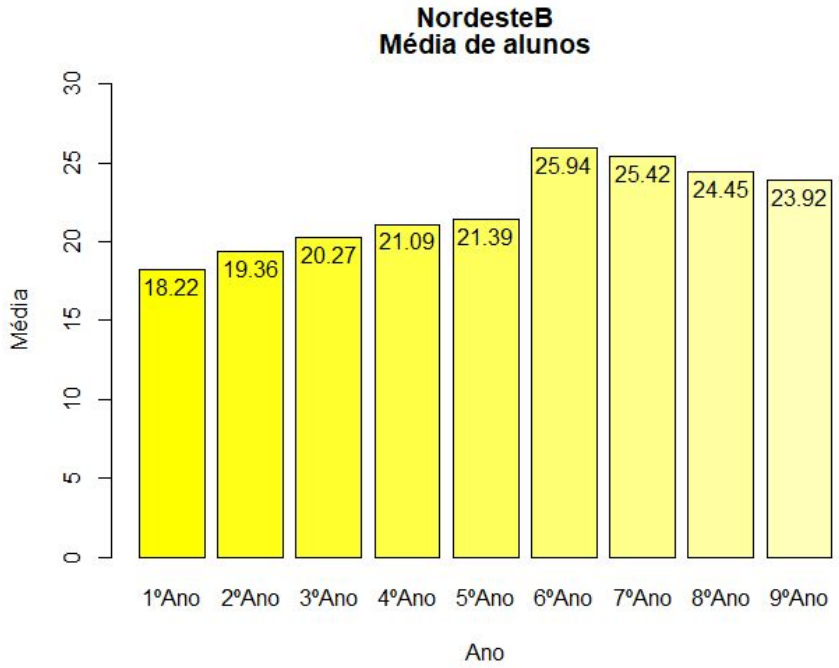


#Região com alta quantidade de outliers distintos
#Possui a menor média em 1°, 2°,3 e 5° ano
#Região com maior diferença do 5° para 6° ano (6.69)

Nordeste A							
Ano	Media	Menor	Maior	Desvio	Mediana	Quartis	Outliers
1ºAno	17.31	1	56	7.528	17.5	12-17.5-23-56	40 40.5 41 41.5 42 42.5 42.8 43 43.5 44 45 46 46.5 47 48 49 50 51 52 53 56
2ºAno	18.45	1	77	7.929	19.0	13-19-24-77	41 41.5 42 42.5 43 43.5 43.7 44 45 45.5 45.7 46 47 48 49 50 51 52 53 54 55 57 61 71 77
3ºAno	19.21	1	65	8.439	19.0	13-19-25-65	44 44.5 45 45.5 46 46.5 46.7 47 48 48.3 49 50 51 53 54 55 57 61 65
4ºAno	19.82	1	59	8.838	20.0	13-20-26-59	46 47 48 48.5 49 50 51 52 53 56 57 59
5ºAno	20.29	1	60	9.391	20.5	13-20.5-27-60	48.5 49 50 50.5 51 52 53 54 55 56 57 60
6ºAno	26.98	1	90	10.056	27.3	20-27.3-34-90	55.5 55.7 56 57 58 59 60 60.8 61 61.5 62 63 64 65 66 67 68 80 82 90
7ºAno	26.21	1	100	10.136	26.5	19-26.5-33.5-100	55.4 55.5 56 56.5 56.7 57 58 58.5 59 60 61 62.5 63 64 64.8 87 100
8ºAno	25.42	1	98	10.411	25.0	18-25-33-98	56 57 57.5 58 59 60 61 62 66 69 72 73 98
9ºAno	24.83	1	91	10.790	24.7	16.5-24.7-32.5-91	57 58 59 60 61 62 63.5 64 65 66 66.5 75 91

NordesteB

Valores avaliados na menor coluna: 6656
Valores avaliados na maior coluna: 10356
Quantidade de linhas (incluindo NA): 34411



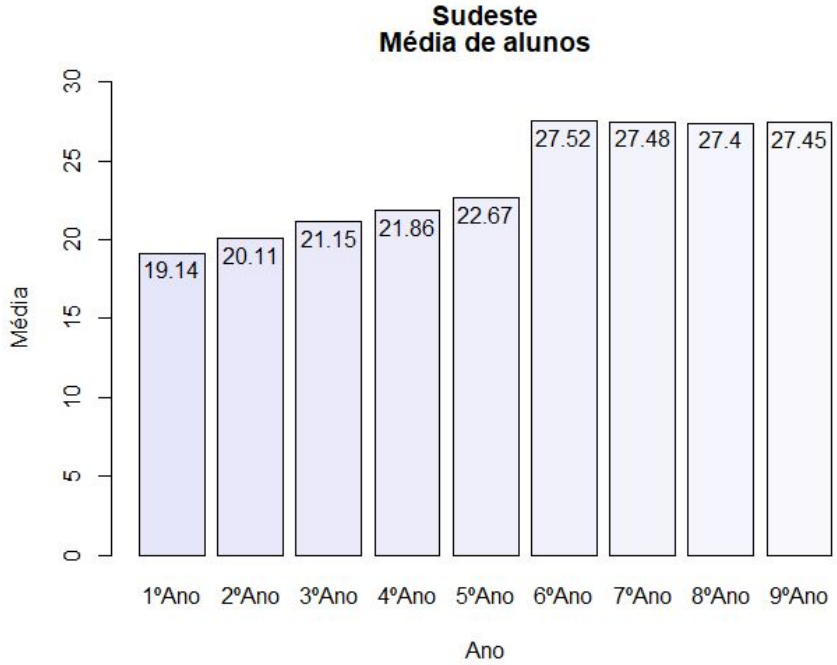
#Região com alta quantidade de outliers distintos
#Região com menor quantidade de valores NA totais
#Região com menor média 1° ano

Nordeste B							
Ano	Media	Menor	Maior	Desvio	Mediana	Quartis	Outliers
1ºAno	18.22	1	72	7.492	18.50	13-18.5-23-72	38.5 39 39.5 40 41 42 42.3 43 43.5 44 45 45.5 46 47 49 50 51 52 57 61 66 72
2ºAno	19.36	1	101	7.650	20.00	14-20-24.5-101	40.5 41 41.5 42 42.5 42.8 43 44 45 45.7 46 47 48 48.7 54 55 57 58 58.5 69 101
3ºAno	20.27	1	97	8.039	20.50	15-20.5-26-97	42.8 43 44 44.5 45 45.5 46 47.5 48 49 50 50.5 53 53.5 54 55 65 83 97
4ºAno	21.09	1	66	8.484	21.00	15-21-27-66	45.3 46 47 48 49 50 51 54 57 59 60 61 66
5ºAno	21.39	1	81	8.873	21.30	15-21.3-27.3-81	46 46.5 47 47.5 48 48.5 49 49.5 50 51 52 53 54 55 55.3 57 58 59 62 63.5 66 69 71 81
6ºAno	25.94	1	116	9.535	26.35	.5-26.35-32.7-116	53.5 54 55 56 57 58 59 61 62 64 70 71 79 89 102 116
7ºAno	25.42	1	106	9.729	26.00	18.5-26-32.5-106	54.5 56 58 59 63.5 64 65 65.5 75 77 85 98 106
8ºAno	24.45	1	104	9.915	24.50	17-24.5-31.6-104	54 56 60 62 64 66 67 70.5 73.7 79 104
9ºAno	23.92	1	79	10.188	24.00	16-24-31.5-79	55 56 58 60.5 61 62 62.5 64 70 74.5 78.5 79

Sudeste

Valores avaliados na menor coluna: 17547
Valores avaliados na maior coluna: 23971
Quantidade de linhas (incluindo NA): 56345

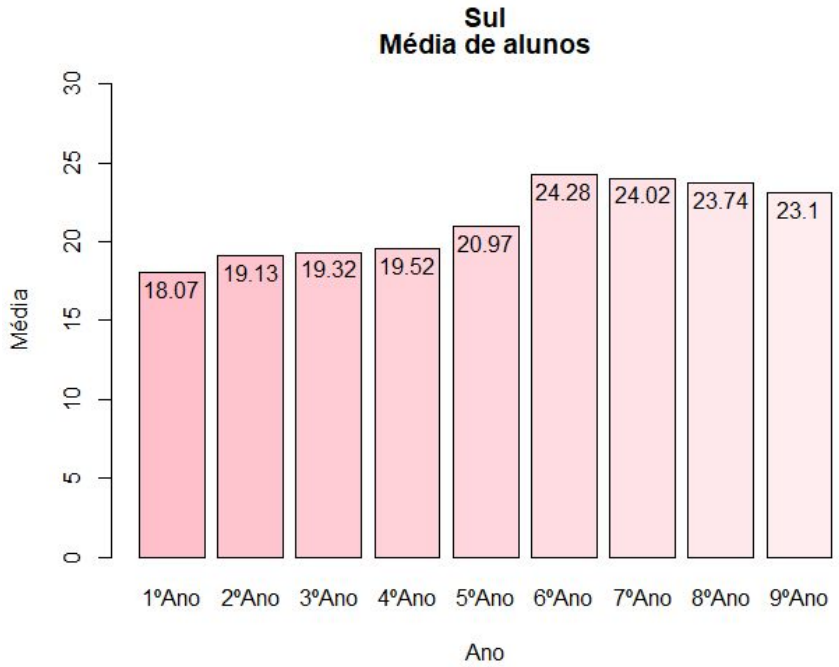
#Região com maior quantidade de dados
#Região com alta quantidade de outliers distintos
#Região com maiores médias de alunos por ano
#Possui a maior média de 6° até o 9° ano



Sudeste							
Ano	Media	Menor	Maior	Desvio	Mediana	Quartis	Outliers
1ºAno	19.14	1	73	7.470	20	14-20-24.3-73	40 40.3 41 41.5 41.7 42 43 43.3 44 45 46 48 49 50 52.5 55 57 59 73
2ºAno	20.11	1	60	7.704	21	15-21-25.6-60	41.7 42 43 45 46 47 49 50 52 55 60
3ºAno	21.15	1	60	8.174	22	15-22-27-3 60	46.5 48 50 52 53 56 57 60
4ºAno	21.86	1	122	8.488	23	16-23-28-122	46.5 47 47.5 48 50 52 57 58 59 61 64 87 122
5ºAno	22.67	1	109	8.785	24	17-24-29-109	48 48.3 49 51 51.5 52 53 55 56 58 66 72 73 88 109
6ºAno	27.52	1	165	8.635	29	22.5-29-33.6-165	1 2 3 3.5 4 4.5 4.7 5 50.5 51 52 52.3 52.5 53 54 54.5 55 56 56.8 57 58.5 59 60 61 62 65 68 69 75 77.5 79 84 91 100 165
7ºAno	27.48	1	147	8.950	29	22-29-34-147	1 2 3 3.5 53 53.3 54 54.3 55 56 57 58 60 61.5 62 62.7 65 67 68 73 74 77 78 84 88 147
8ºAno	27.40	1	133	9.090	29	21.7-29-34-133	1 1.5 2 2.5 2.8 3 53 54 56 57 58 58.5 59 60 61 64 66 74 76 84 102 103 133
9ºAno	27.45	1	100	9.462	29	21-29-34.5-100	55 56 57 58 59 60 63 64 65 67 73 96 99 100

Sul

Valores avaliados na menor coluna: 8544
Valores avaliados na maior coluna: 10952
Quantidade de linhas (incluindo NA): 24065

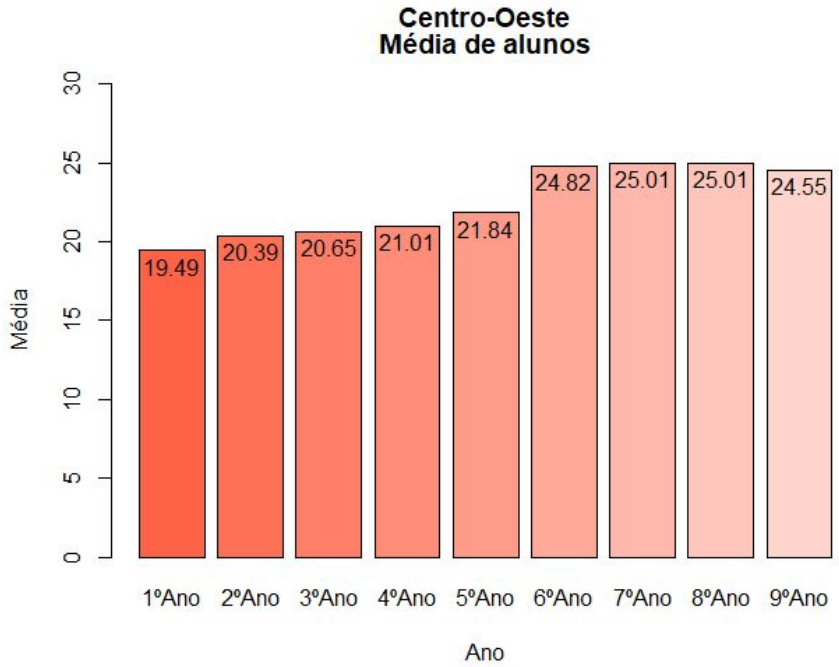


#Possui a menor média em 4°, 6° e 7° ano
#Região com baixa quantidade de outliers distintos
#Região com menores médias de alunos por ano

Sul							
Ano	Media	Menor	Maior	Desvio	Mediana	Quartis	Outliers
1ºAno	18.07	1	54.0	6.588	19.0	14-19-23-54	37 38 40 41 44 54
2ºAno	19.13	1	41.0	6.809	20.0	15-20-24-41	1 38 39 39.5 40 41
3ºAno	19.32	1	42.5	7.128	20.0	14.5-20-24.5-42.5	40 40.5 42 42.5
4ºAno	19.52	1	45.0	7.529	20.0	14-20-25-45	43 45
5ºAno	20.97	1	56.0	7.648	22.0	16-22-26.7-56	43 44 56
6ºAno	24.28	1	72.0	7.845	25.0	19-25-30-72	1 2 47 51 56 64 72
7ºAno	24.02	1	61.0	7.806	25.0	19-25-30-61	1 2 48 50 54 61
8ºAno	23.74	1	62.0	8.034	24.5	8-24.5-29.8-75	62 48 51 62
9ºAno	23.10	1	88.5	8.543	24.0	17-24-29.5-88.5	48.3 53 61 79 88.5

Centro-Oeste

Valores avaliados na menor coluna: 4042
Valores avaliados na maior coluna: 5304
Quantidade de linhas (incluindo NA): 9653



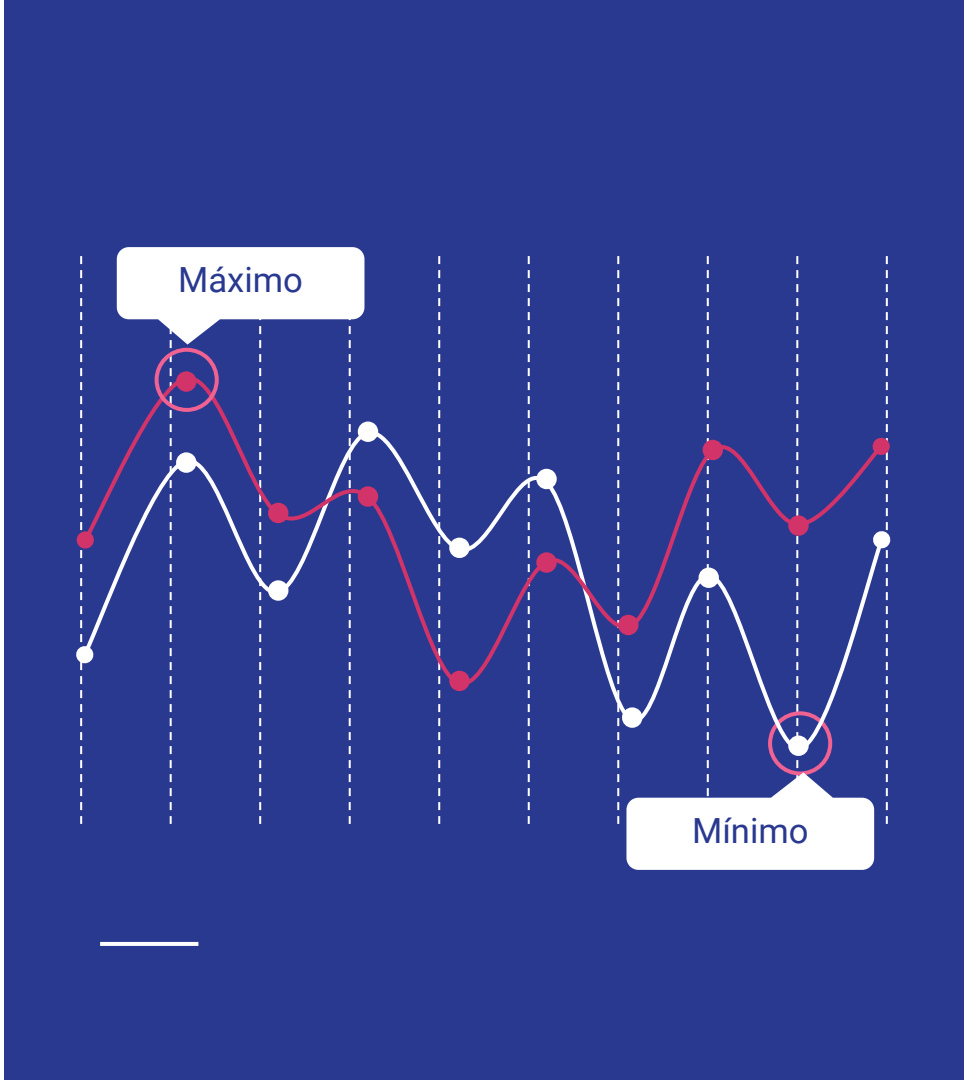
#Região com menor quantidade de valores NA totais
#Região com menor quantidade de dados
#Região com baixa quantidade de outliers distintos

Centro-Oeste							
Ano	Media	Menor	Maior	Desvio	Mediana	Quartis	Outliers
1ºAno	19.49	1	44	7.159	20.50	15-20.5-25-44	42 44
2ºAno	20.39	1	53	7.531	21.50	15.5-21.5-26-53	43 46 53
3ºAno	20.65	1	76	8.126	21.75	15-21.75-26.5-76	44 45 76
4ºAno	21.01	1	116	8.658	22.00	15-22 27.5-116	116
5ºAno	21.84	1	126	8.952	23.00	16-23-28.5-126	126
6ºAno	24.82	1	289	9.967	26.00	19-26-31-289	49.4 49.6 50 50.8 52.5 56 61 69 86 137 289
7ºAno	25.01	1	292	10.507	26.00	19-26-31.3-292	52 52.5 53 54 57.7 60 64 70 207 292
8ºAno	25.01	1	372	11.392	26.00	18-26-32-372	65 98 209 372
9ºAno	24.55	1	102	9.810	25.50	18-25.5-31.5-102	52.2 53 56 67 102

Gráficos dos Anos

Médias de alunos de cada região.

Análise do crescimento e decrescimento.



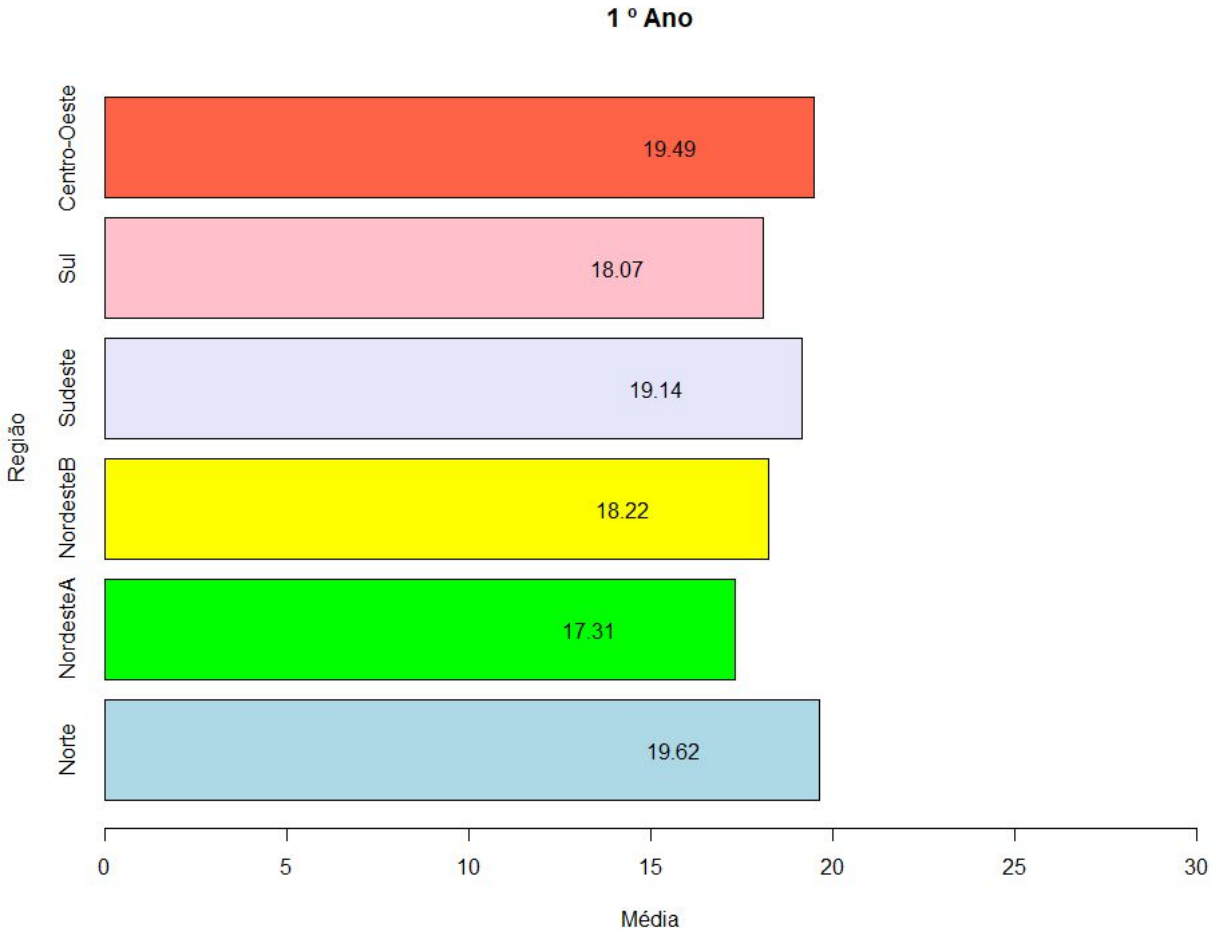


Gráfico de 1° ano

#Menor média de alunos:
#NordesteA
#17.31

#Maior média de alunos:
#Norte
#19.62

#Menores médias de alunos

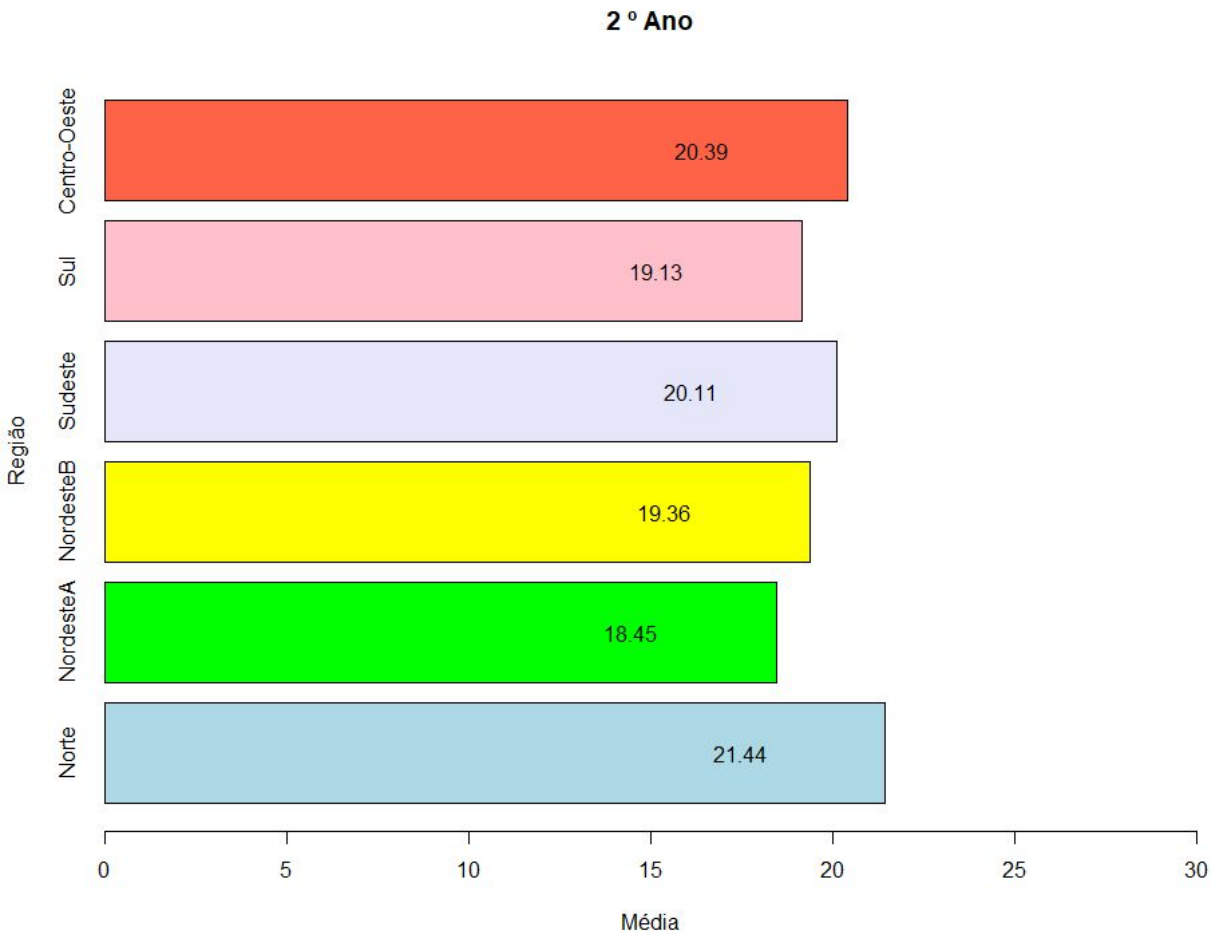


Gráfico de 2° ano

#Menor média de alunos:
#NordesteA
#18.45

#Maior média de alunos:
#Norte
#21.44

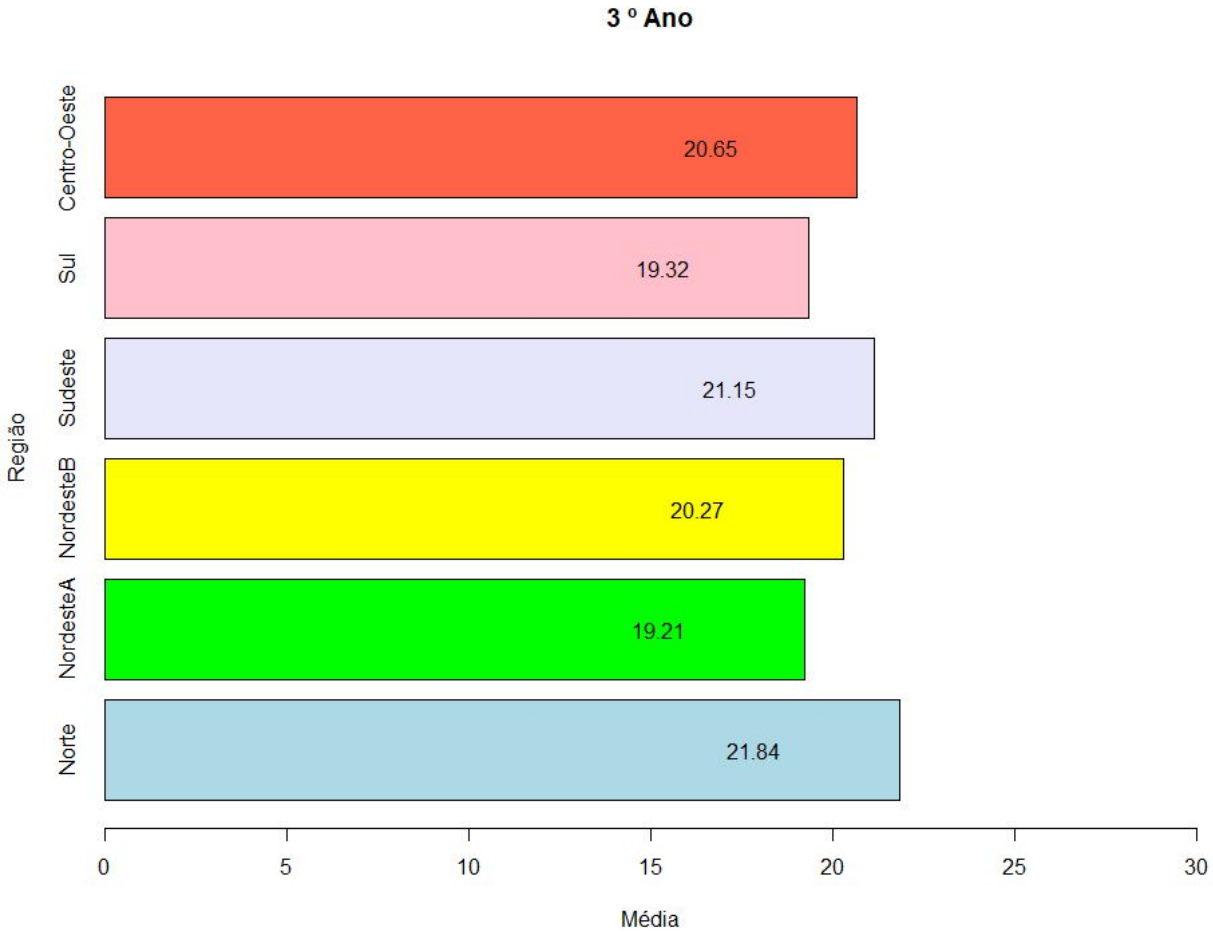


Gráfico de 3° ano

#Menor média de alunos:
#NordesteA
#19.21

#Maior média de alunos:
#Norte
#21.84

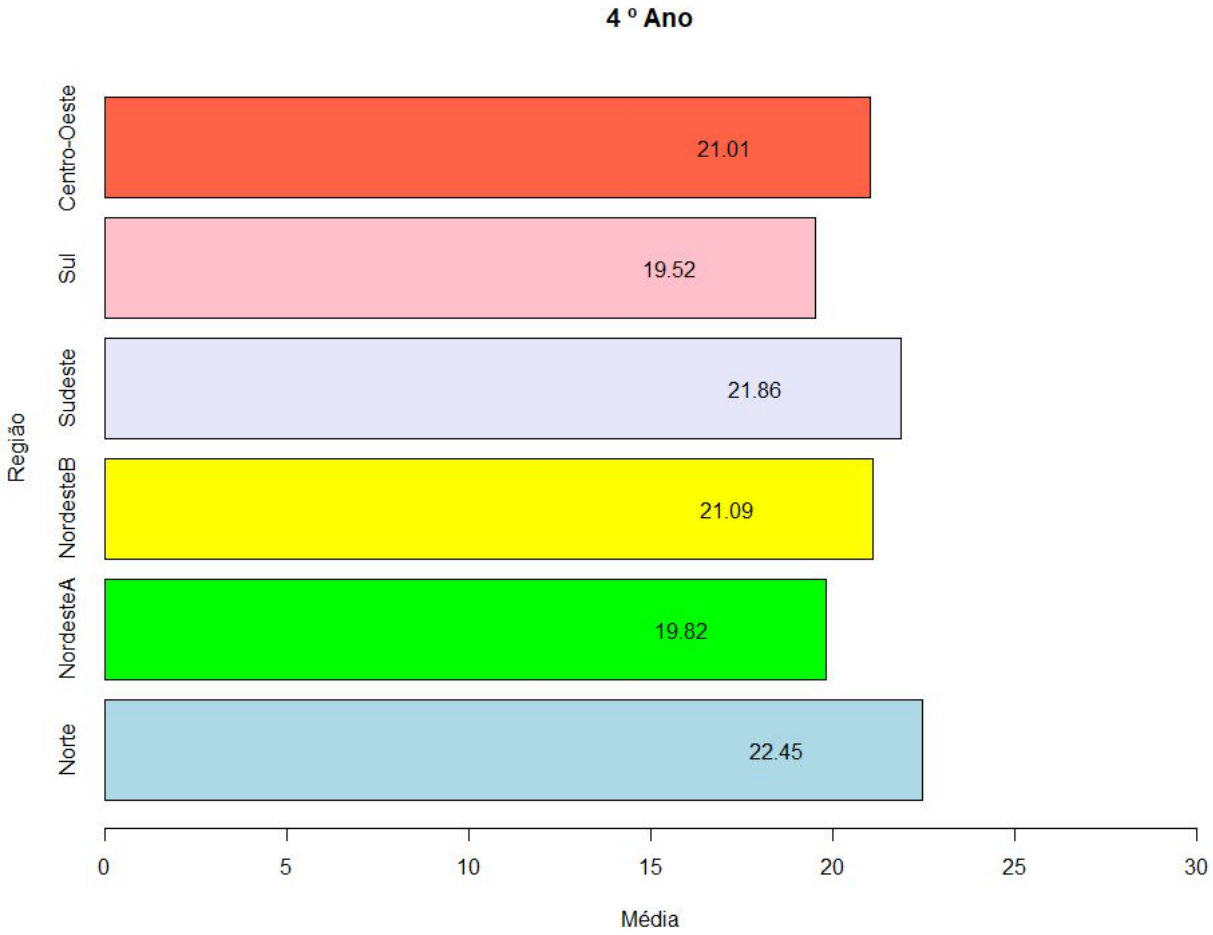


Gráfico de 4° ano

#Menor média de alunos:
#Sul
#19.52

#Maior média de alunos:
#Norte
#22.45

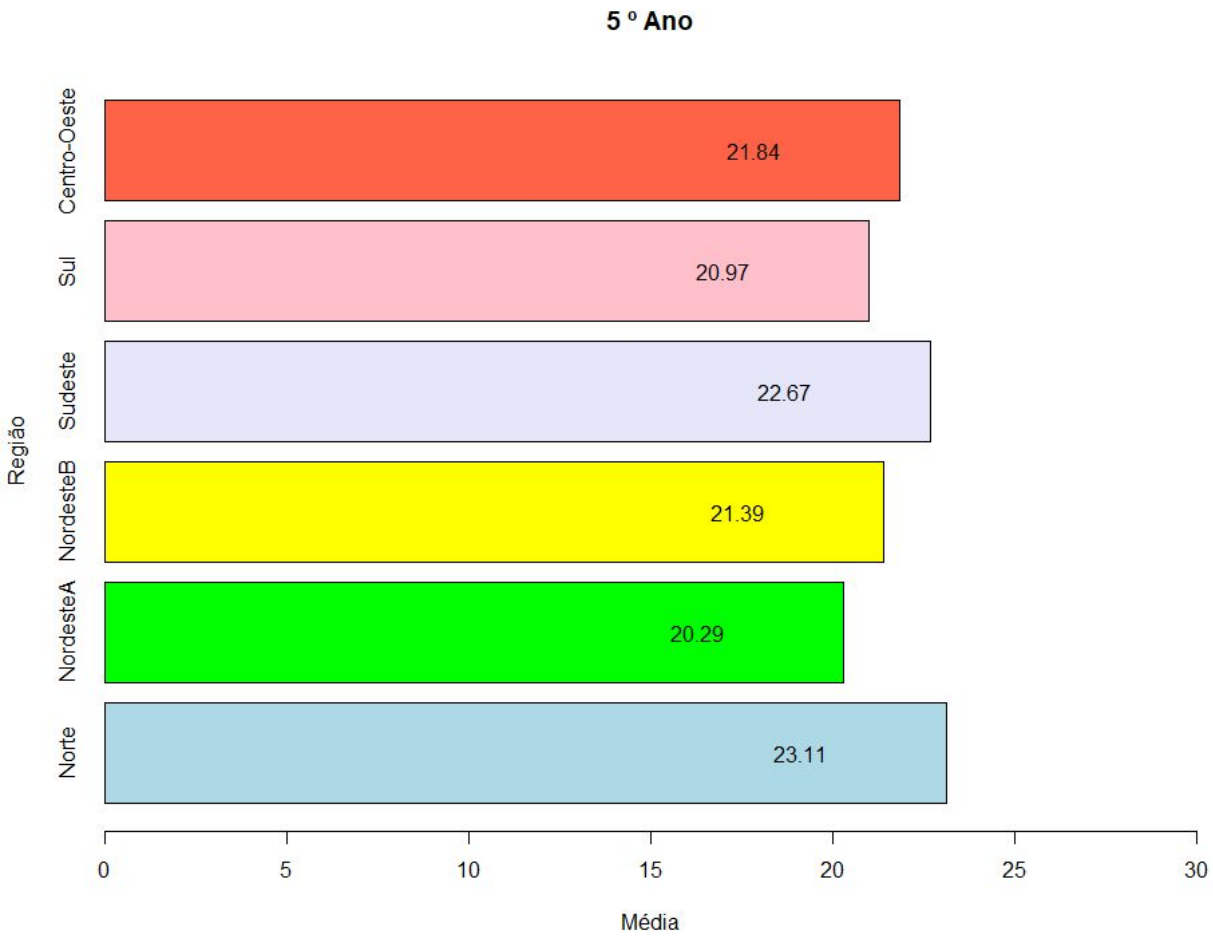


Gráfico de 5° ano

#Menor média de alunos:
#NordesteA
#20.29

#Maior média de alunos:
#Norte
#23.11

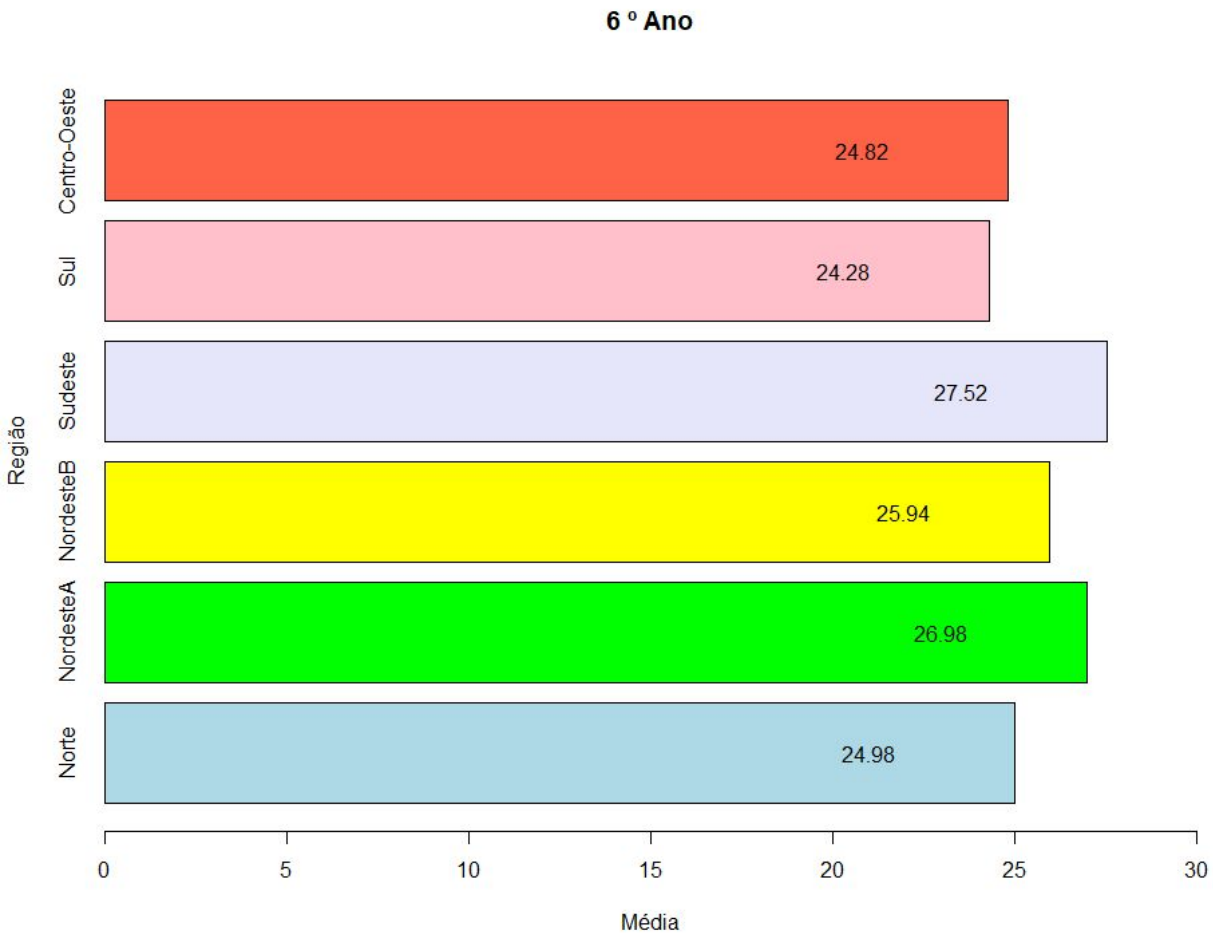


Gráfico de 6° ano

#Menor média de alunos:
#Sul
#24.28

#Maior média de alunos:
#Sudeste
#27.52

#Maiores médias de alunos

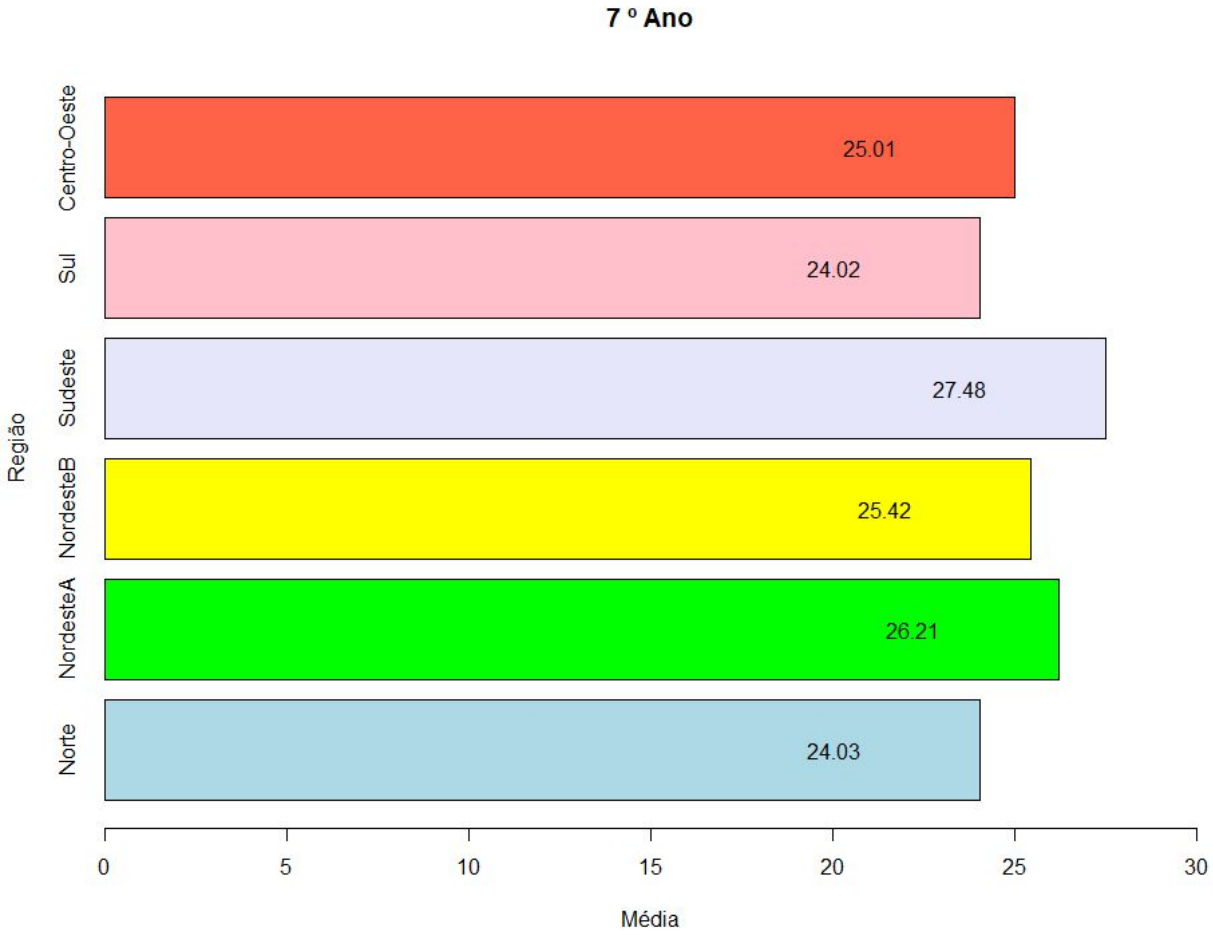


Gráfico de 7° ano

#Menor média de alunos:
#Sul
#24.02

#Maior média de alunos:
#Sudeste
#27.48

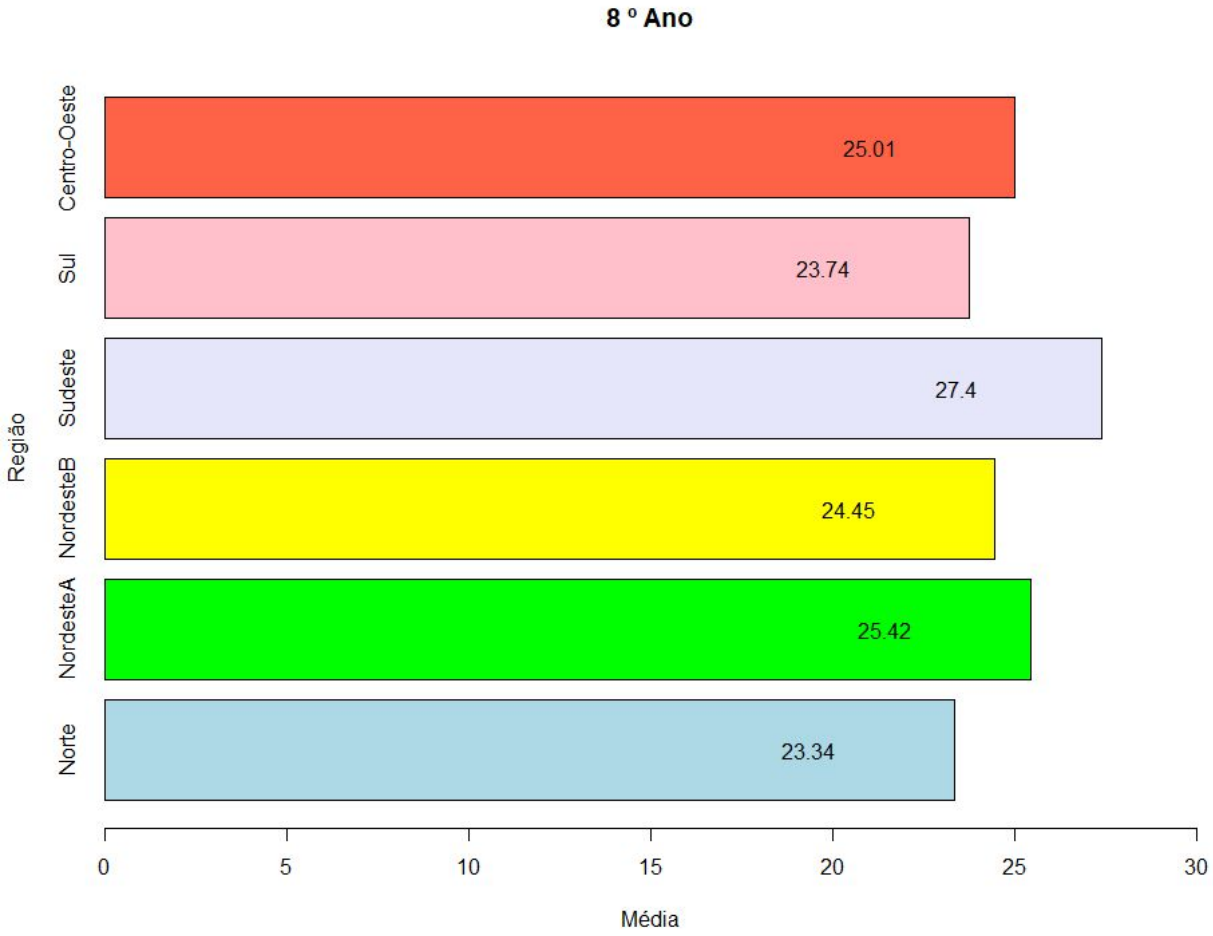


Gráfico de 8° ano

#Menor média de alunos:
#Norte
#23.34

#Maior média de alunos:
#Sudeste
#27.4

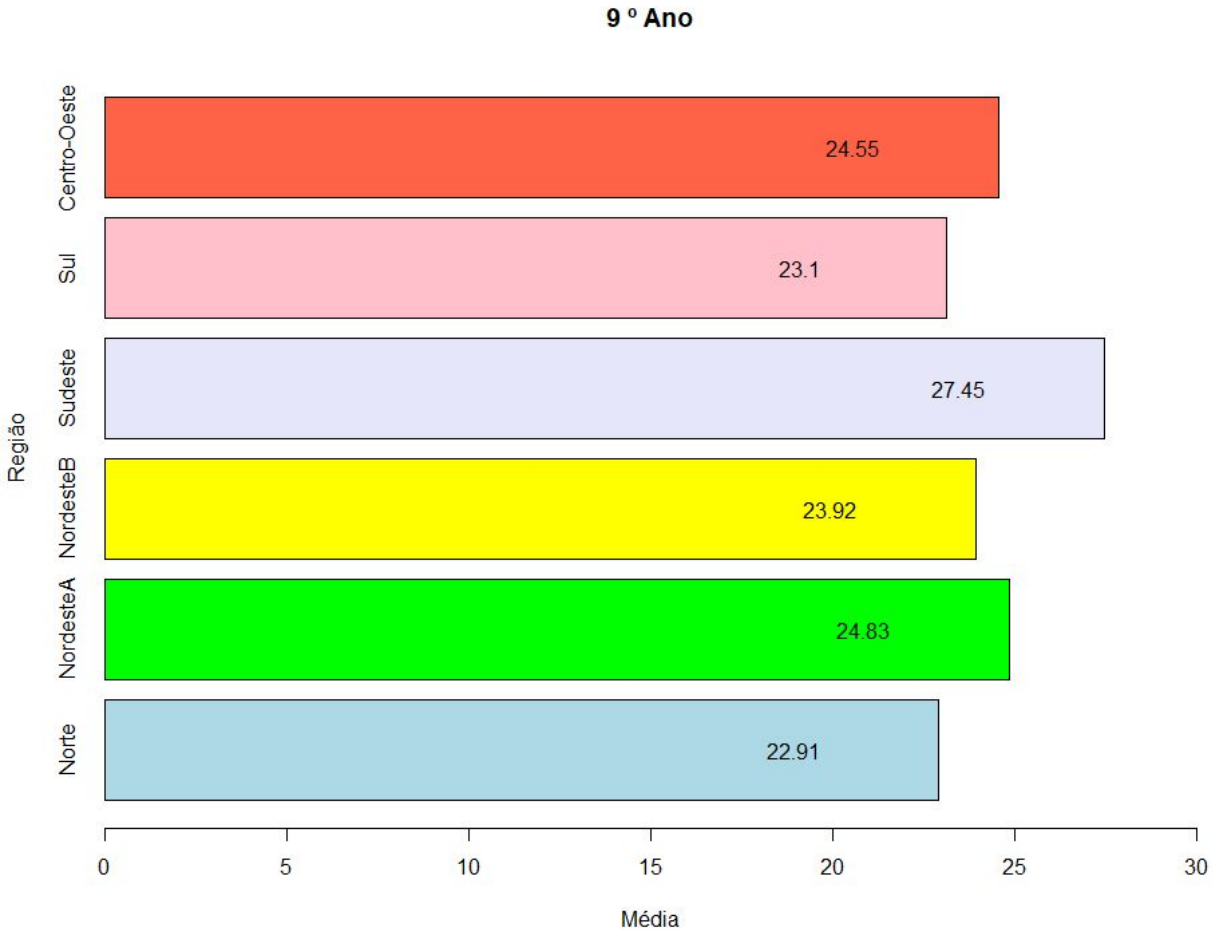


Gráfico de 9° ano

#Menor média de alunos:
#Norte
#22.91

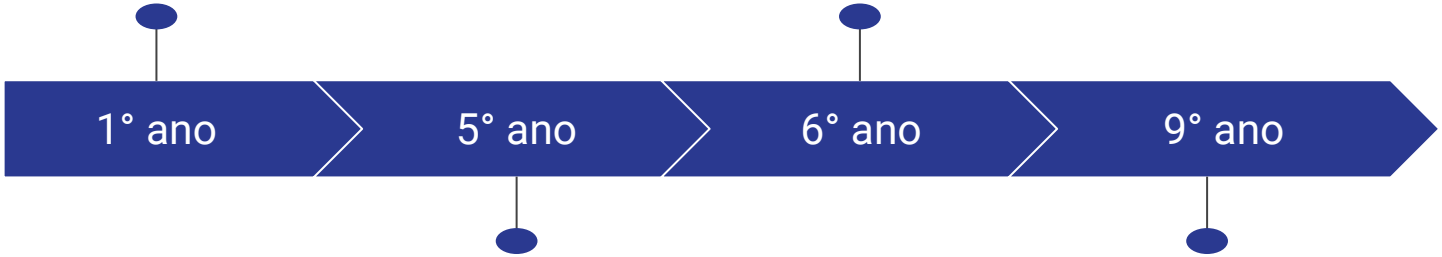
#Maior média de alunos:
#Sudeste
#27.45

Visão geral do gráfico


Comportamento médio observado do gráfico ao longo dos anos

Concentração das menores médias de alunos por turma.

Concentração das maiores médias de alunos por turma.



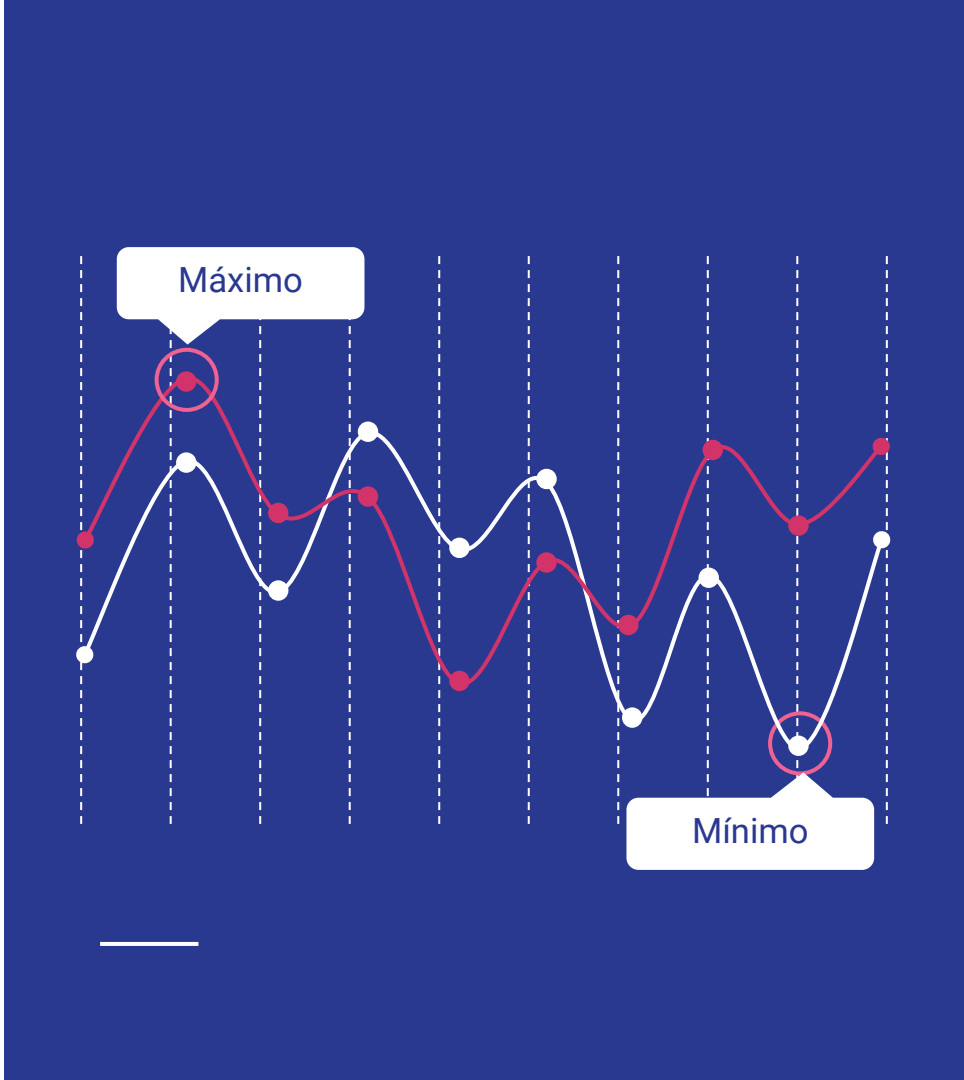
Fim do crescimento de forma linear do gráfico.

Fim do decrescimento do gráfico.

Gráfico total

Dados estatísticos gerais.

Avaliando a base inteira como uma única grande região.



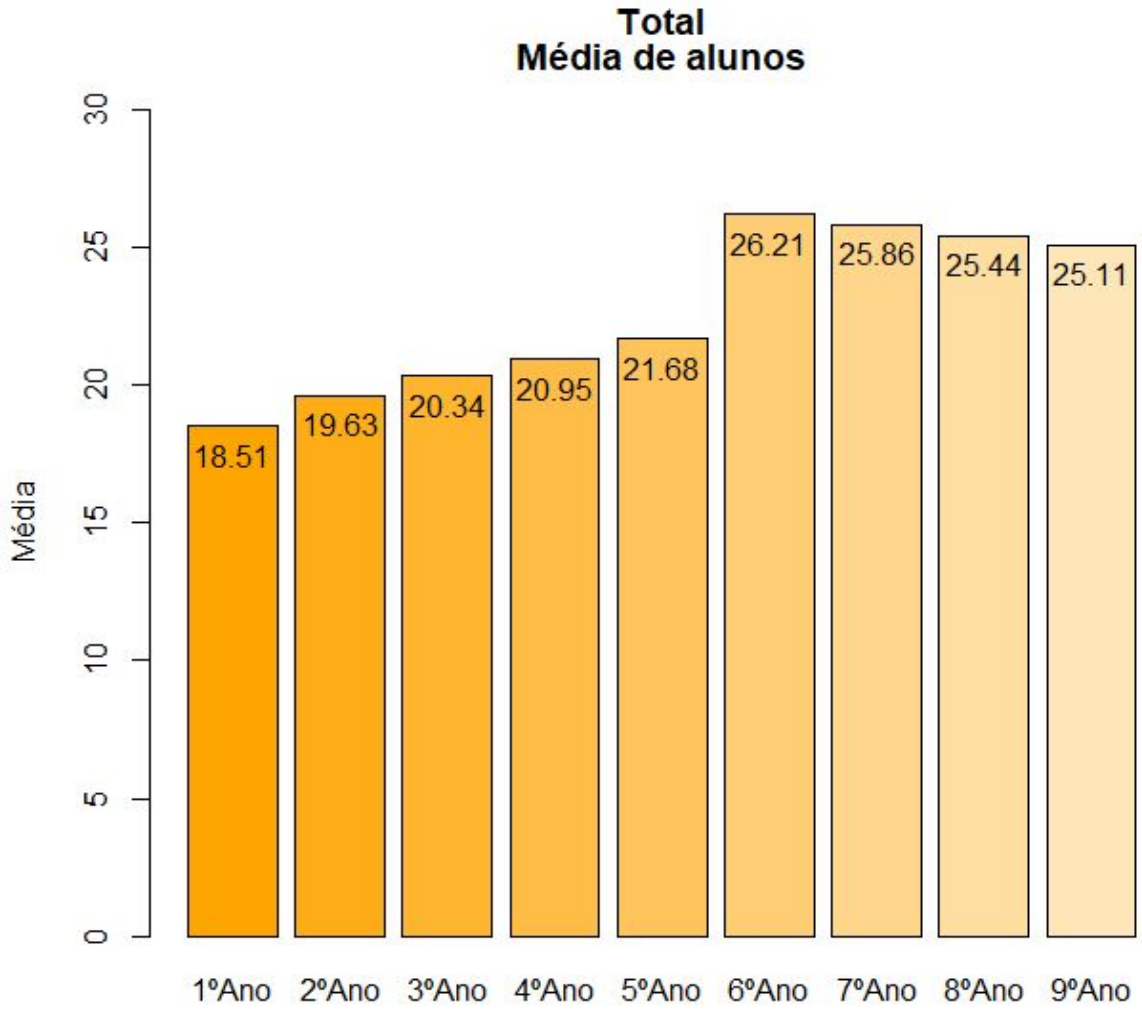


Gráfico total

Valores avaliados na menor coluna: 52446
Valores avaliados na maior coluna: 75375
Quantidade de linhas (incluindo NA): 190520

#Mantém o mesmo comportamento observad
anteriormente

#Menor valor:
#1°ano
#18.51

#Maior valor:
#6°ano
#26.21

Dados exploratórios totais

Total							
Ano	Media	Menor	Maior	Desvio	Mediana	Quartis	Outliers
1ºAno	18.51	1	73	7.375	19.0	13 19 24 73	41 41.5 41.7 42 42.3 42.5 42.8 43 43.3 43.5 44 45 45.5 46 46.5 47 48 49 50 51 52 52.5 53 54 55 56 57 59 61 66 72 73
2ºAno	19.63	1	101	7.674	20.0	14.3 20 25 101	41.5 41.7 42 42.3 42.5 42.6 42.7 42.8 43 43.5 43.7 44 45 45.5 45.7 46 46.5 47 47.5 48 48.7 49 50 51 52 53 54 55 57 58 58.5 60 61 69 71 77 79 101
3ºAno	20.34	1	97	8.146	21.0	15 21 26 97	42.7 42.8 43 43.5 44 44.2 44.5 45 45.5 45.6 46 46.5 46.7 47 47.5 48 48.3 49 50 50.5 51 52 53 53.5 54 55 56 57 60 61 65 76 80 83 97
4ºAno	20.95	1	122	8.549	21.5	15 21.5 27 122	45.3 45.5 45.6 46 46.5 47 47.5 48 48.5 49 50 51 52 53 54 56 57 58 59 60 61 64 66 80 87 116 122
5ºAno	21.68	1	126	8.917	22.0	15 22 28 126	48 48.3 48.5 49 49.5 50 50.3 50.5 51 51.5 52 53 54 55 55.3 56 57 58 59 60 62 63.5 66 69 71 72 73 80 81 88 109 126
6ºAno	26.21	1	289	9.380	27.0	20 27 33 289	53 53.5 54 54.5 55 55.5 55.7 56 56.8 57 58 58.5 59 60 60.8 61 61.5 62 63 64 65 66 67 68 69 70 71 72 75 77.5 79 80 82 84 86 89 90 91 100 102 110 116 137 165 289
7ºAno	25.86	1	292	9.610	26.9	19 26.9 33 292	54.2 54.3 54.5 55 55.1 55.4 55.5 56 56.5 56.7 57 57.5 57.7 58 58.5 59 60 61 61.5 62 62.5 62.7 63 63.5 64 64.8 65 65.5 66 67 68 70 73 74 75 77 78 80 84 85 87 88 90 98 100 106 147 207 292
8ºAno	25.44	1	372	9.871	26.0	18.5 26 32.6 372	54 54.8 55 56 57 57.5 58 58.5 59 60 61 62 63 64 65 66 67 69 70.5 72 73 73.7 74 75 76 79 84 91 96 98 102 103 104 133 209 372
9ºAno	25.11	1	102	10.127	26.0	18 26 32.8 102	55 55.5 55.7 56 57 58 59 60 60.5 61 62 62.5 63 63.5 64 65 66 66.5 67 70 73 74.5 75 78.5 79 81 88.5 89 91 96 99 100 102

Considerações finais

Foi observado durante o desenvolvimento deste trabalho, como a análise de dados utilizando-se da linguagem R pode ser feita de forma simplificada e intuitiva, podendo dessa forma auxiliar na tomada de decisões.

Através da análise desses dados, pode-se obter informações importantes das regiões Norte, Sul, Nordeste, Sudeste e Centro-Oeste, como suas estatísticas gerais e o seu comportamento esperado através do gráfico.

Ao se utilizar essas informações das regiões, é possível perceber detalhes importantes sobre as distribuições dos alunos pelos anos que não podem ser facilmente observáveis sem a devida manipulação desses dados, como por exemplo, a divisão que vai do 1° ano até 5° ano, e do 6° ano até 9° devido a grande diferença observada no gráfico de médias.

“Sem dados você é apenas mais uma pessoa com uma opinião.”

W. Edwards Deming

“Os erros causados por dados inadequados são muito menores do que aqueles devido à falta total de dados.”

Charles Babbage

Obrigado!