Projeto de Disciplna - Análise Exploratória de Dados

Vanessa Reis

abril/2024

Base de dados:

"Country-data.csv"

A base escolhida pode ser encontrada no link abaixo e é composta de dados socioeconômicos e de saúde de 166 países. https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data?select=Country-data.csv

Estes dados permitem diversas análises entre eles como: a correlação entre PIB e expectativa de vida, ou o impacto da inflacão nas importações e exportações, a relacao entre desenvolvimento econômico e taxa de mortalidade infantil, balança comercial... É possivel identificar, por exemplo, setores onde necessitam de maiores investimentos por parte de governo e empresas.

country - Nome do país

child_mort - Morte de crianças menores de 5 anos por 1.000 nascidos vivos

exports - Exportações de bens e serviços per capita. Dado como %idade do PIB per capita

helth - Gastos totais com saúde per capita. Dado como %idade do PIB per capita

imports - Importações de bens e serviços per capita. Dado como %idade do PIB per capita

income - Lucro líquido por pessoa

inflation - A medição da taxa de crescimento anual do PIB total

life_expec - O número médio de anos que uma criança recém-nascida viveria se os atuais padrões de mortalidade permanecessem os mesmos

total_fer - O número de filhos que nasceriam de cada mulher se as atuais taxas de fertilidade por idade permanecessem as mesmas.

gdpp - O PIB per capita. Calculado como o PIB total dividido pela população total.

```
## # A tibble: 167 x 10
##    country         child_mort exports health imports income inflation life_expec
##    <chr>                <dbl>   <dbl>  <dbl>   <dbl>  <dbl>     <dbl>      <dbl>
##  1 Afghanistan           90.2    10     7.58    44.9   1610     9.44        56.2
##  2 Albania               16.6    28     6.55    48.6   9930     4.49        76.3
##  3 Algeria               27.3    38.4   4.17    31.4  12900    16.1         76.5
##  4 Angola               119      62.3   2.85    42.9   5900    22.4         60.1
##  5 Antigua and Ba~       10.3    45.5   6.03    58.9  19100     1.44        76.8
##  6 Argentina             14.5    18.9   8.1     16    18700    20.9         75.8
##  7 Armenia               18.1    20.8   4.4     45.3   6700     7.77        73.3
##  8 Australia              4.8    19.8   8.73    20.9  41400     1.16        82
##  9 Austria                4.3    51.3  11       47.8  43200     0.873       80.5
```

```
## 10 Azerbaijan            39.2    54.3   5.88    20.7  16000    13.8         69.1
## # i 157 more rows
## # i 2 more variables: total_fer <dbl>, gdpp <dbl>
```

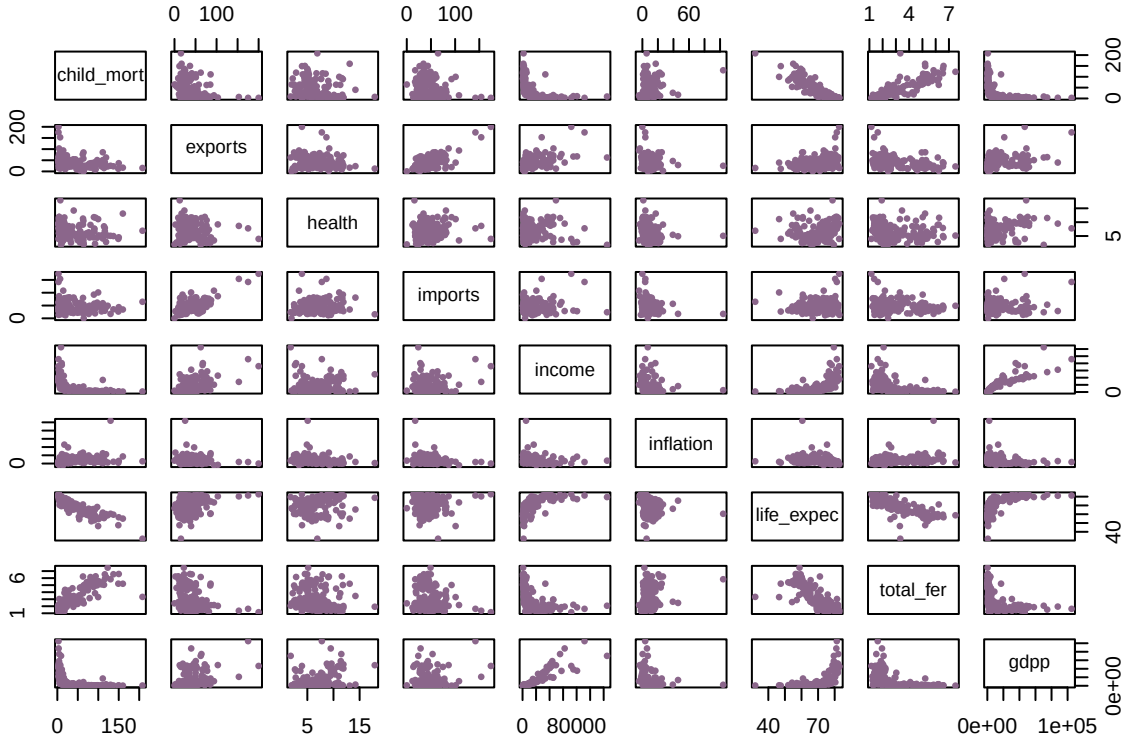
Descriptive Statistics
Country_data
N: 167
##

	child_mort	exports	gdpp	health	imports	income	inflation
Mean	38.27	41.11	12964.16	6.82	46.89	17144.69	7.78
Std.Dev	40.33	27.41	18328.70	2.75	24.21	19278.07	10.57
Min	2.60	0.11	231.00	1.81	0.07	609.00	-4.21
Q1	7.90	23.80	1310.00	4.91	30.00	3340.00	1.77
Median	19.30	35.00	4660.00	6.32	43.30	9960.00	5.39
Q3	62.20	51.40	14600.00	8.65	58.90	22900.00	10.90
Max	208.00	200.00	105000.00	17.90	174.00	125000.00	104.00
MAD	21.94	21.20	5814.76	2.64	21.05	11638.41	5.72
IQR	53.85	27.55	12720.00	3.68	28.55	19445.00	8.94
CV	1.05	0.67	1.41	0.40	0.52	1.12	1.36
Skewness	1.42	2.40	2.18	0.69	1.87	2.19	5.06
SE.Skewness	0.19	0.19	0.19	0.19	0.19	0.19	0.19
Kurtosis	1.62	9.65	5.23	0.59	6.41	6.67	39.95
N.Valid	167.00	167.00	167.00	167.00	167.00	167.00	167.00
Pct.Valid	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Table: Table continues below
##
##
##

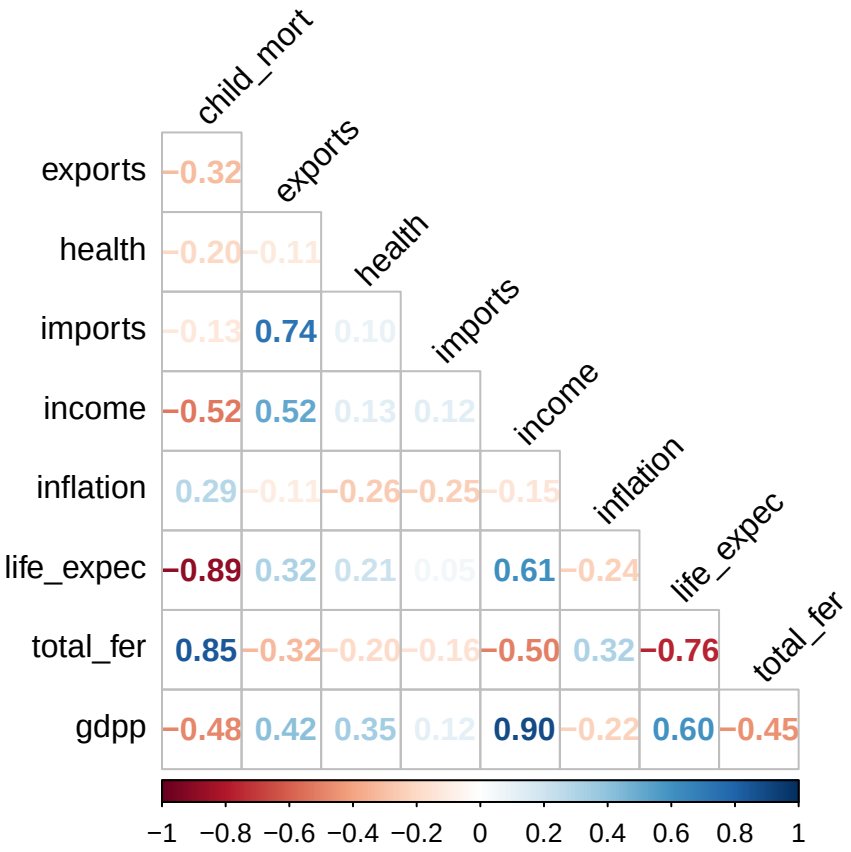
	life_expec	total_fer
Mean	70.56	2.95
Std.Dev	8.89	1.51
Min	32.10	1.15
Q1	65.30	1.79
Median	73.10	2.41
Q3	76.80	3.91
Max	82.80	7.49
MAD	8.90	1.22
IQR	11.50	2.08
CV	0.13	0.51
Skewness	-0.95	0.95
SE.Skewness	0.19	0.19
Kurtosis	1.03	-0.25
N.Valid	167.00	167.00
Pct.Valid	100.00	100.00

Correlações

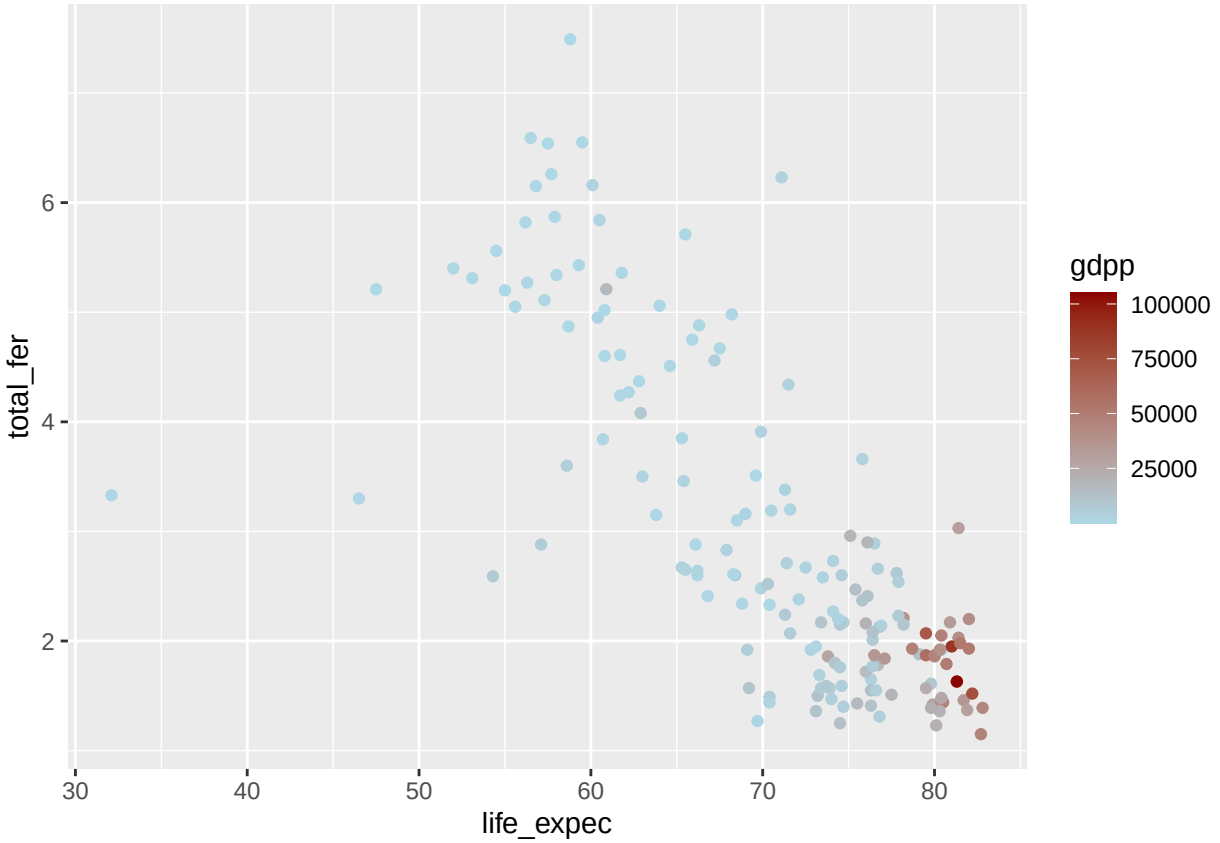


Visualmente, pela matriz de espalhamento, é possível identificarmos alguns pares de variáveis com alta correlação entre si. por exemplo, child-mort e life-xpec (inversamente correlacionadas), child-mort e total-fer, income e gdpp, life_expec e total_fer(inversamente correlacionadas).

Essa informação se confirma e é mais detalhada no gráfico de correlação abaixo.


child_mort
exports
health
imports
income
inflation
life_expec
total_fer
gdpp
exports −0.32
health −0.20 −0.11
imports −0.13 0.74 0.10
income −0.52 0.52 0.13 0.12
inflation 0.29 −0.11 −0.26 −0.25 −0.15
life_expec −0.89 0.32 0.21 0.05 0.61 −0.24
total_fer 0.85 −0.32 −0.20 −0.16 −0.50 0.32 −0.76
gdpp −0.48 0.42 0.35 0.12 0.90 −0.22 0.60 −0.45
−1 −0.8 −0.6 −0.4 −0.2 0 0.2 0.4 0.6 0.8 1

```
## NULL
```



O gráfico de dispersão acima indica a alta correlação (inversa) entre a expectativa de vida e taxa de fertilidade total. Além disso, o PIB per capita está correlacionado com as variáveis anteriores. Quanto maior o PIB per capita, maior a expectativa de vida e menor a taxa de fertilidade total.

Distribuição Normal

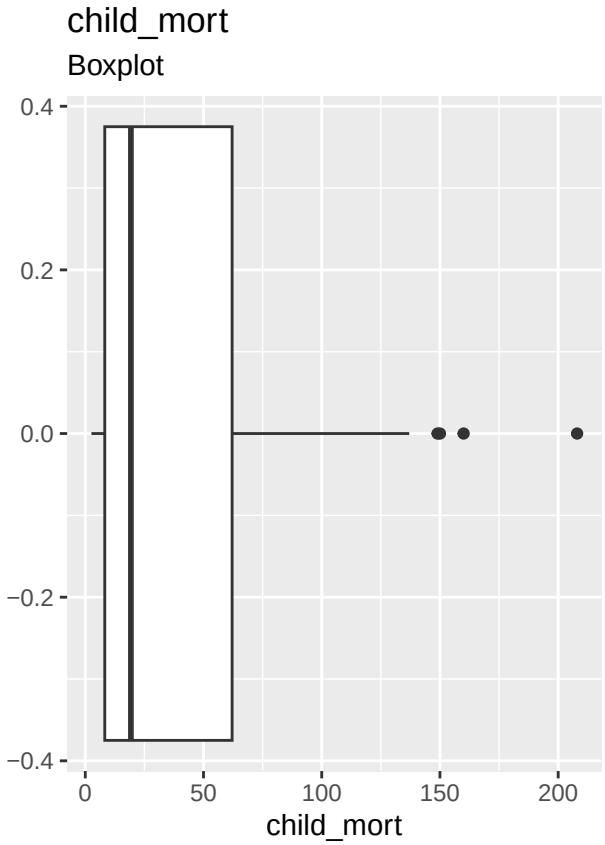
Uma distribuição normal é um tipo de distribuição estatística utilizada para descrever um determinado comportamento de variáveis e tem uma forma simétrica de sino quando representada graficamente. Pode ser caracterizada por dois parâmetros principais: a média (média, mediana e moda são iguais e representam o ponto central) e o desvio padrão (indica a dispersão dos dados em torno da média. Em uma distribuição normal, também é possível saber o quão provável é que um evento ocorra dentro de um intervalo específico, através da área sob a curva.

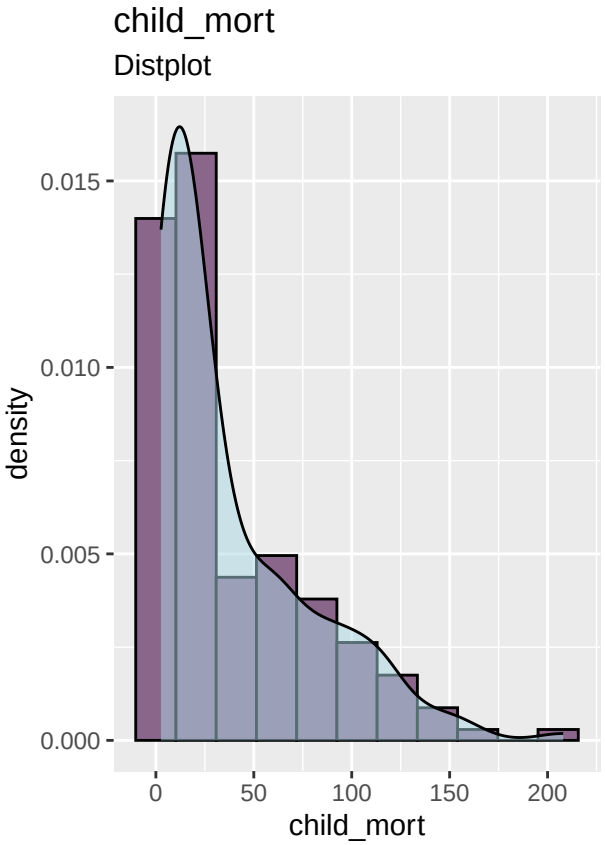
Histogramas

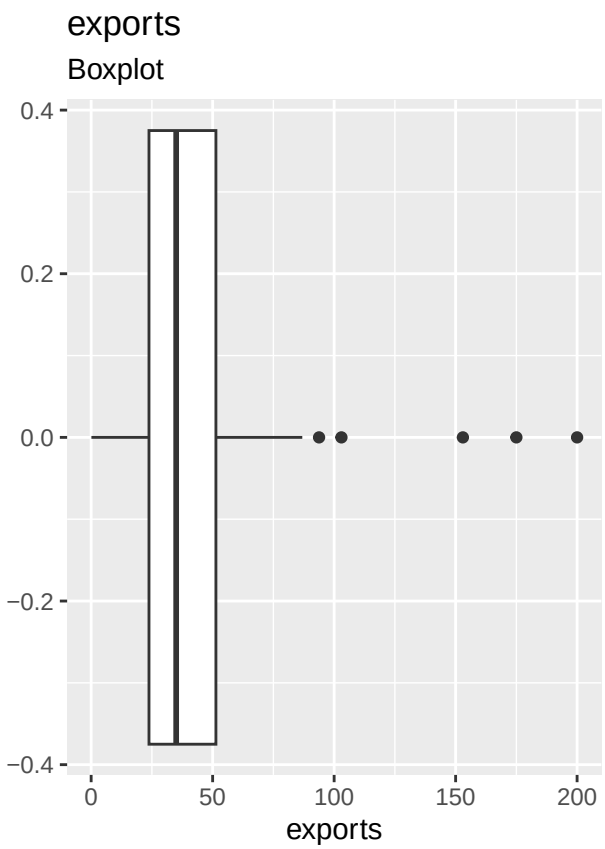
Para cada variável, foi calculado o número ideal de bins de acordo com a regra de Freedman-Diaconis.

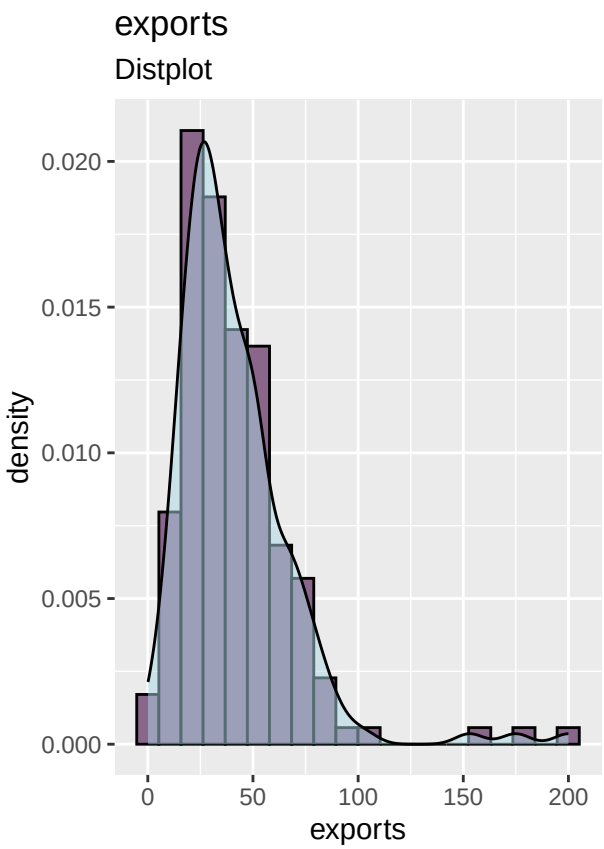
```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

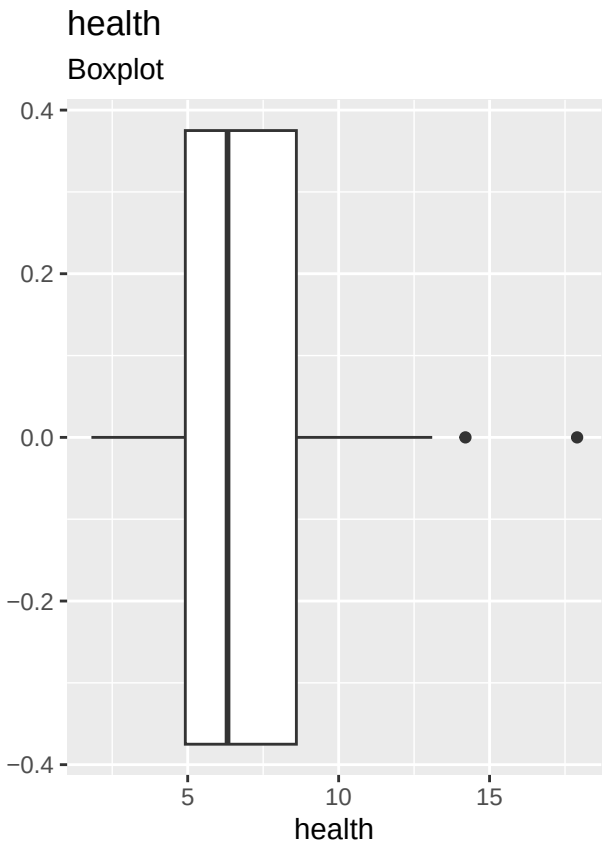
```
## Warning: The dot-dot notation (‘..density..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(density)‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

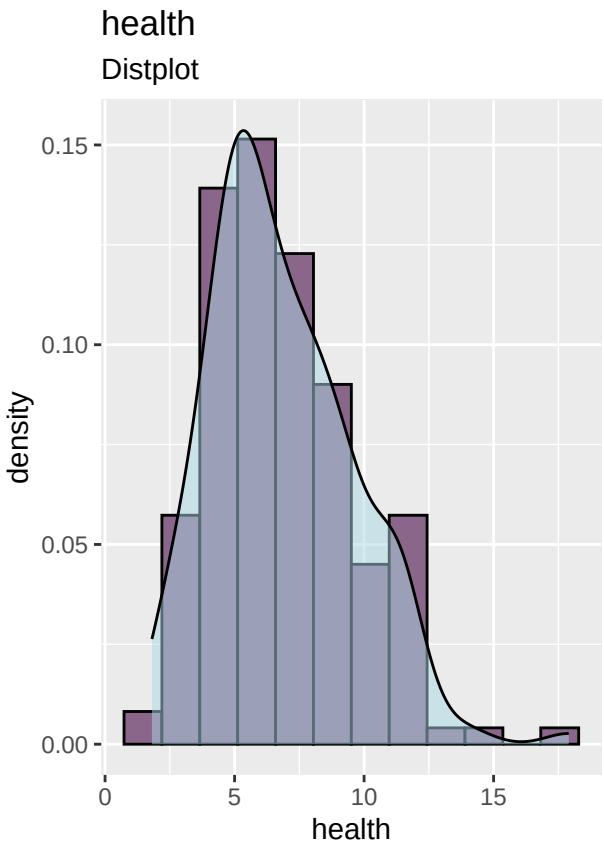


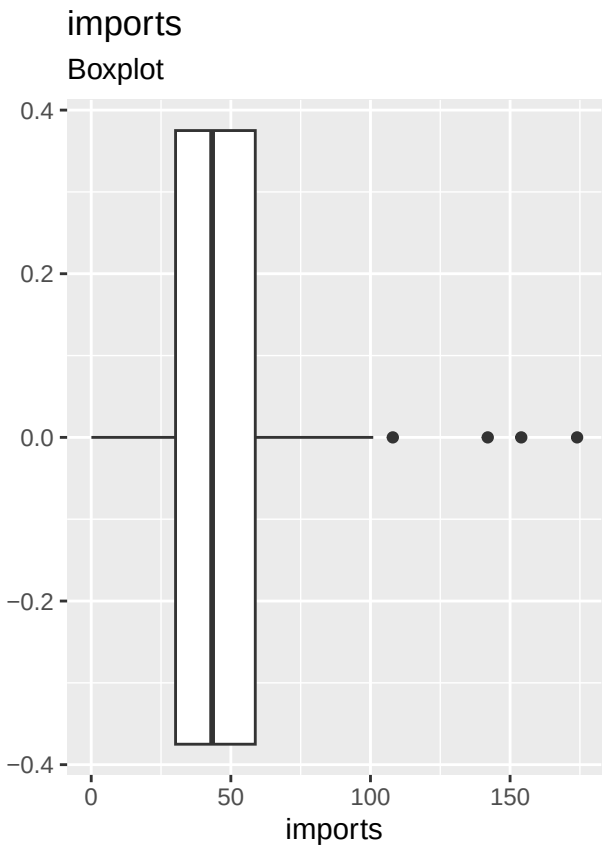


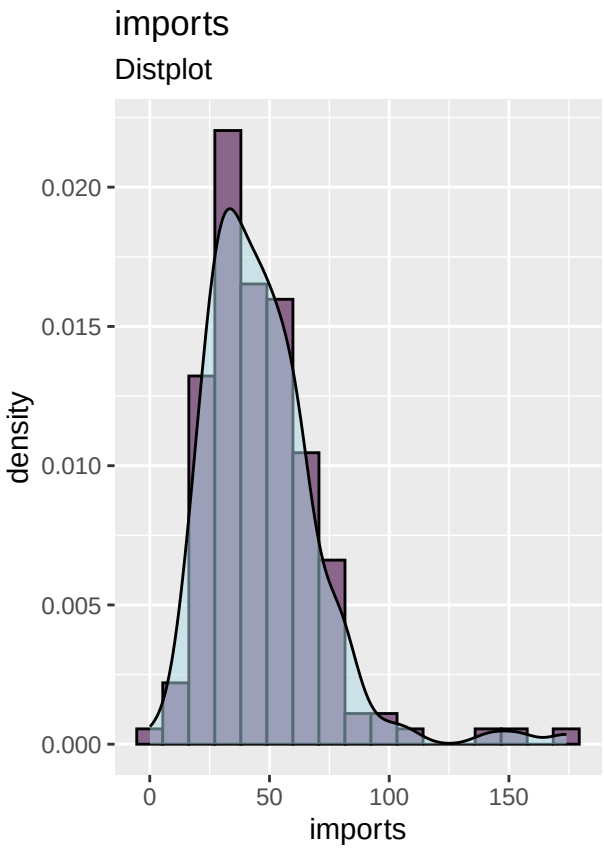

exports
Boxplot
0.4
0.2
0.0
−0.2
−0.4
0
50
100
150
200
exports

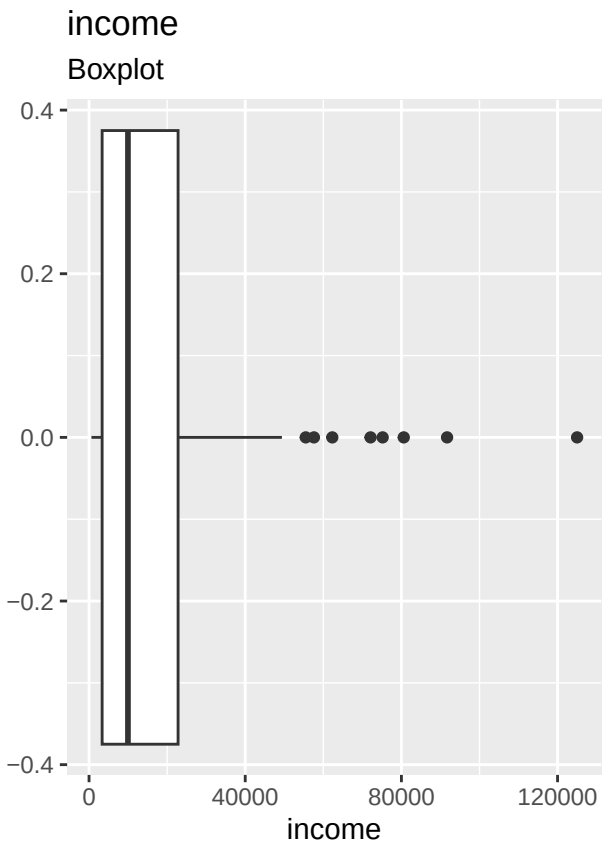

exports
Distplot
0.020
0.015
0.010
0.005
0.000
density
0
50
100
150
200
exports

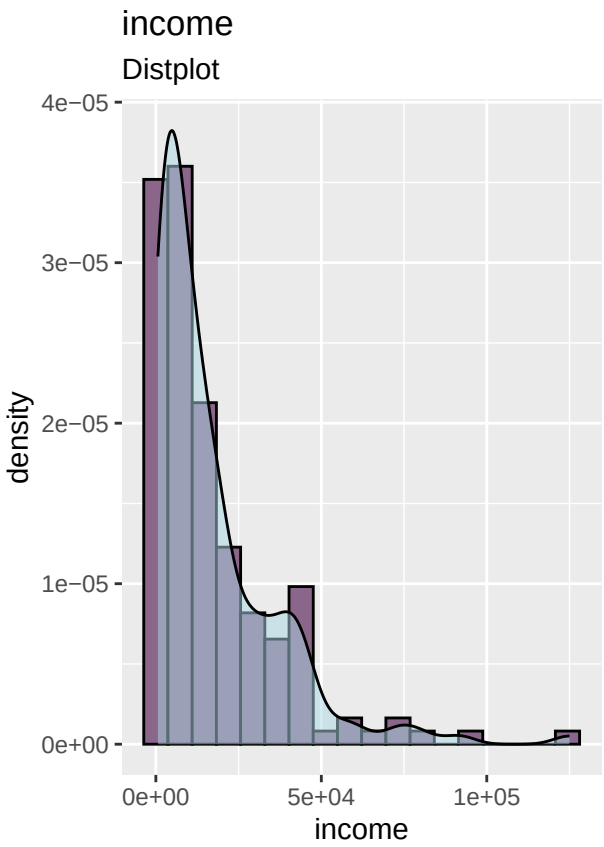

health
Boxplot
0.4
0.2
0.0
−0.2
−0.4
5
10
15
health

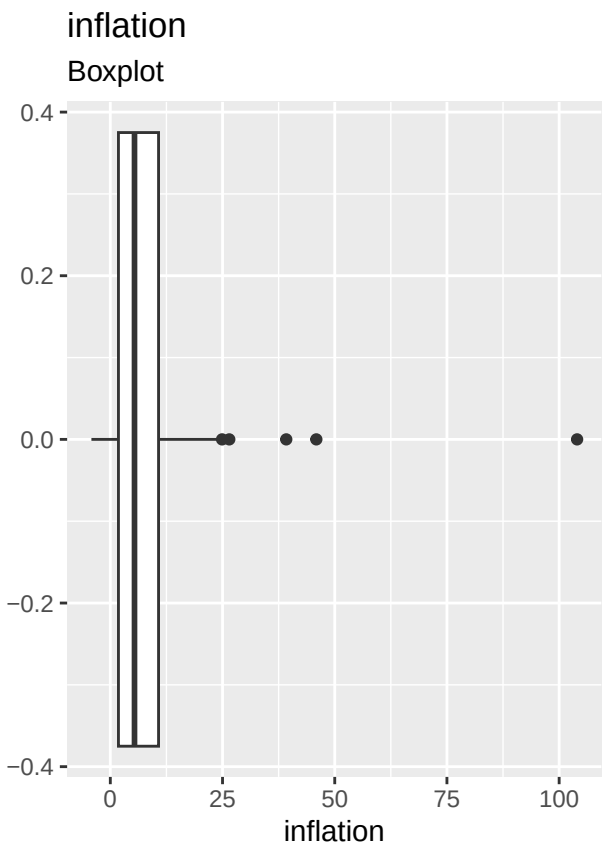

health
Distplot
0.15
0.10
0.05
0.00
density
0
5
10
15
health

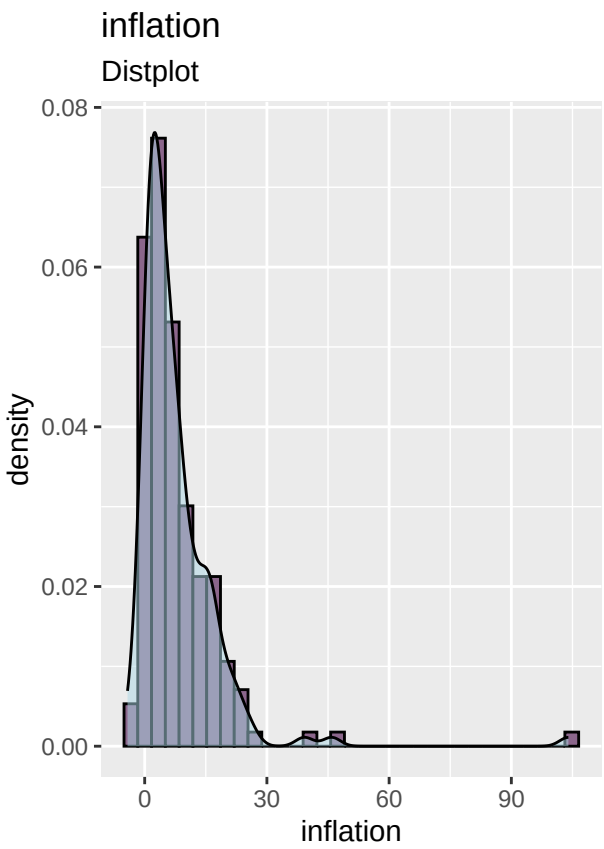

imports
Boxplot
0.4
0.2
0.0
−0.2
−0.4
0
50
100
150
imports

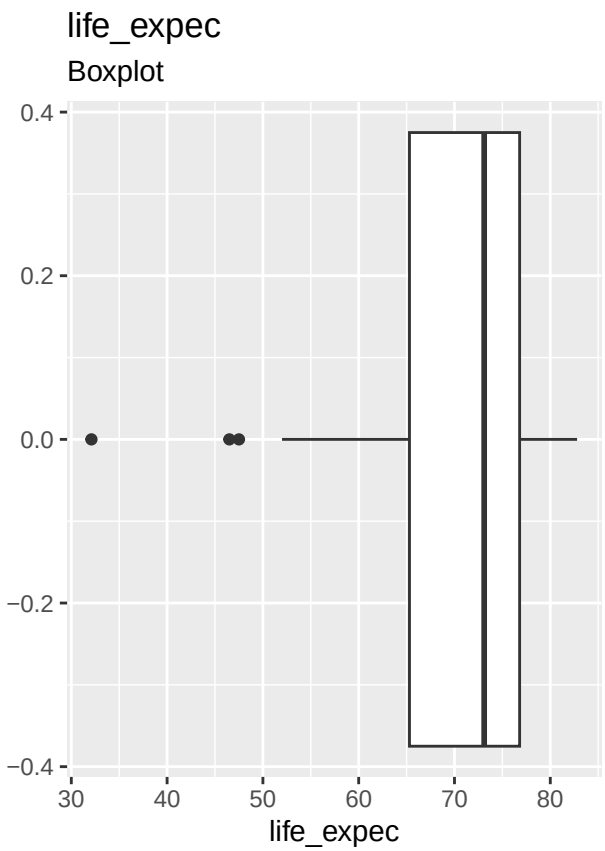

imports
Distplot
0.020
0.015
0.010
0.005
0.000
density
0
50
100
150
imports

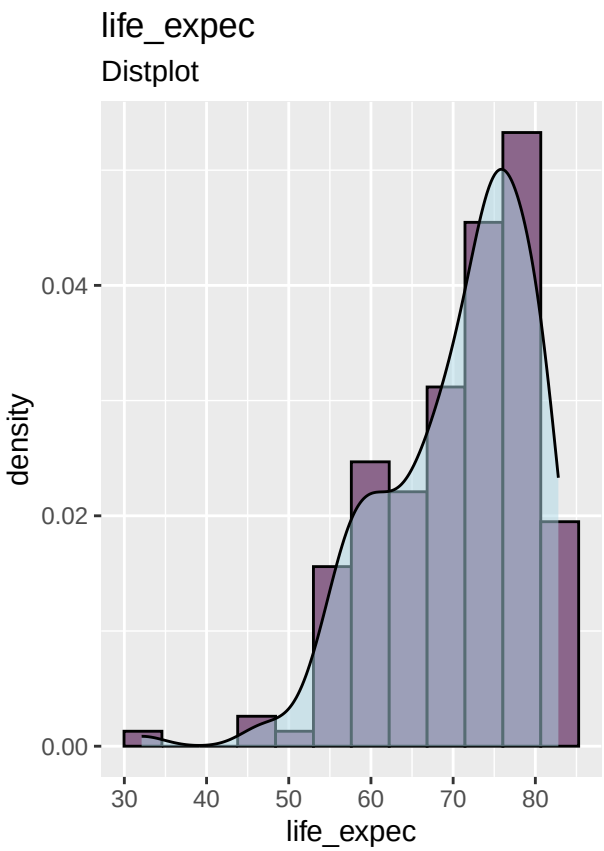

income
Boxplot
0.4
0.2
0.0
−0.2
−0.4
0
40000
80000
120000
income

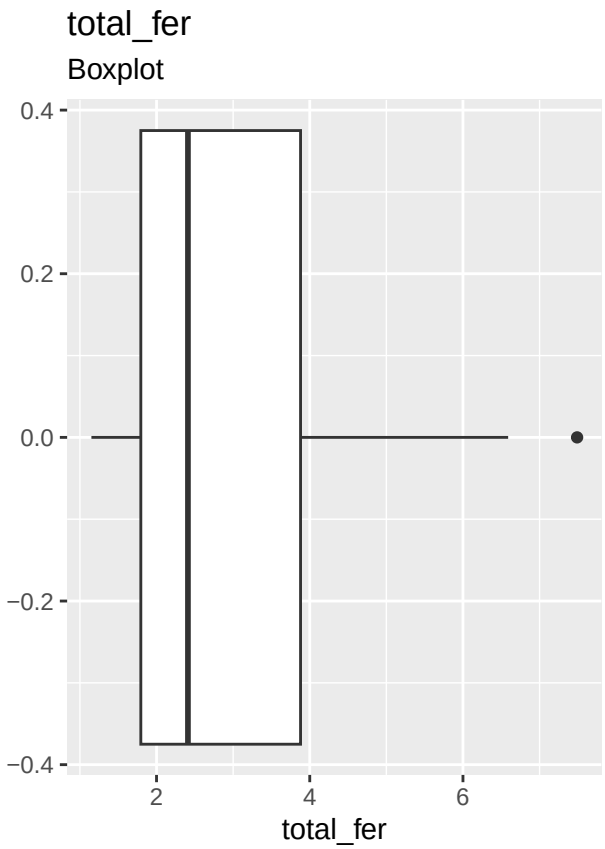

income
Distplot
4e−05
3e−05
2e−05
1e−05
0e+00
density
0e+00
5e+04
1e+05
income

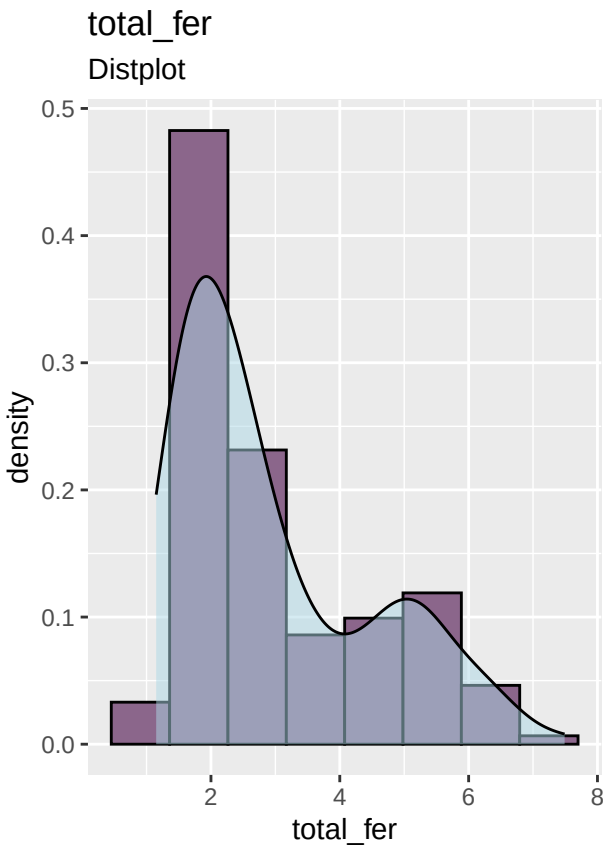

inflation
Boxplot
0.4
0.2
0.0
−0.2
−0.4
0
25
50
75
100
inflation

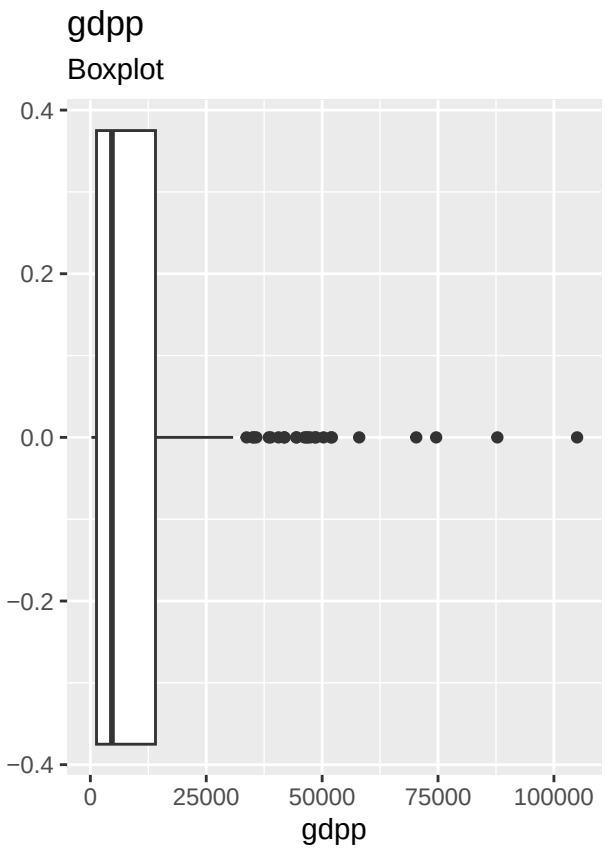

inflation
Distplot
0.08
0.06
0.04
0.02
0.00
density
0
30
60
90
inflation

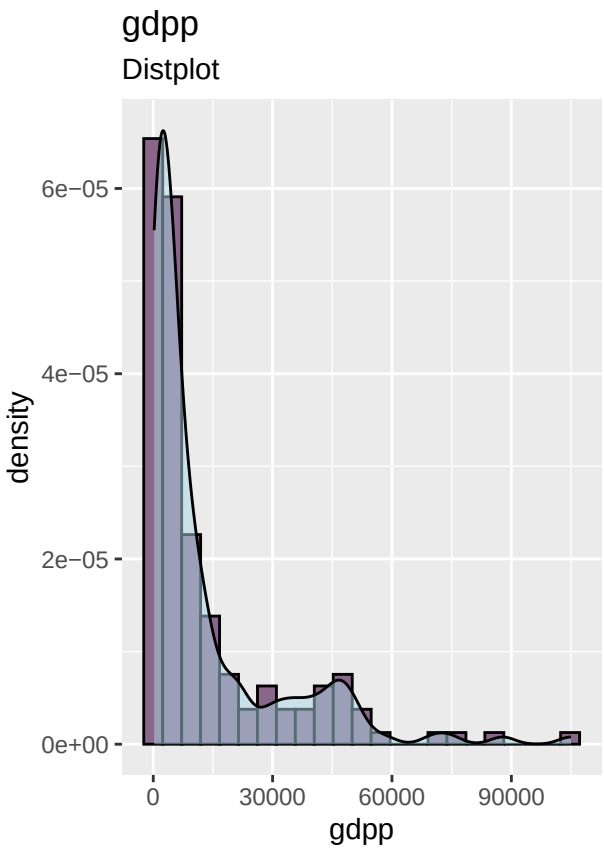

life_expec
Boxplot
0.4
0.2
0.0
−0.2
−0.4
30
40
50
60
70
80
life_expec


life_expec
Distplot
density
0.04
0.02
0.00
30
40
50
60
70
80
life_expec

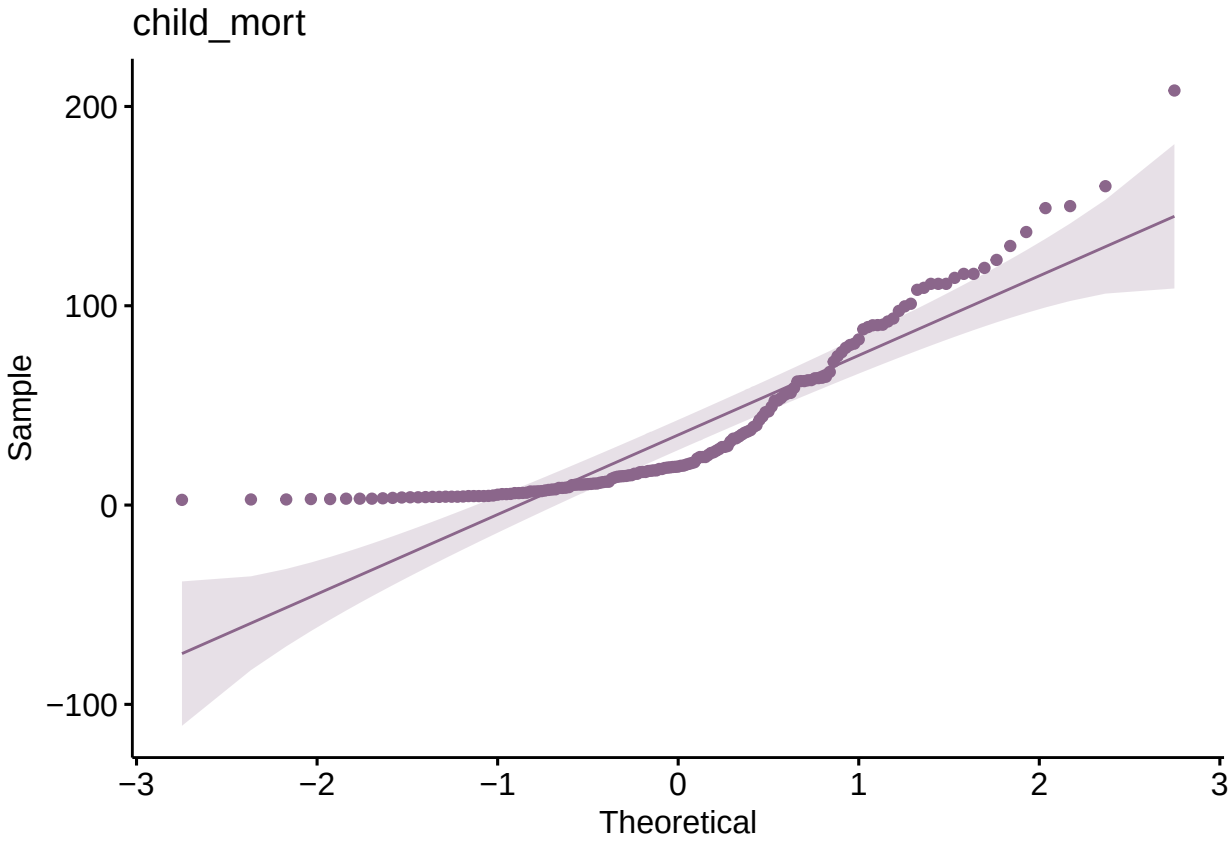

total_fer
Boxplot
0.4
0.2
0.0
−0.2
−0.4
2
4
6
total_fer

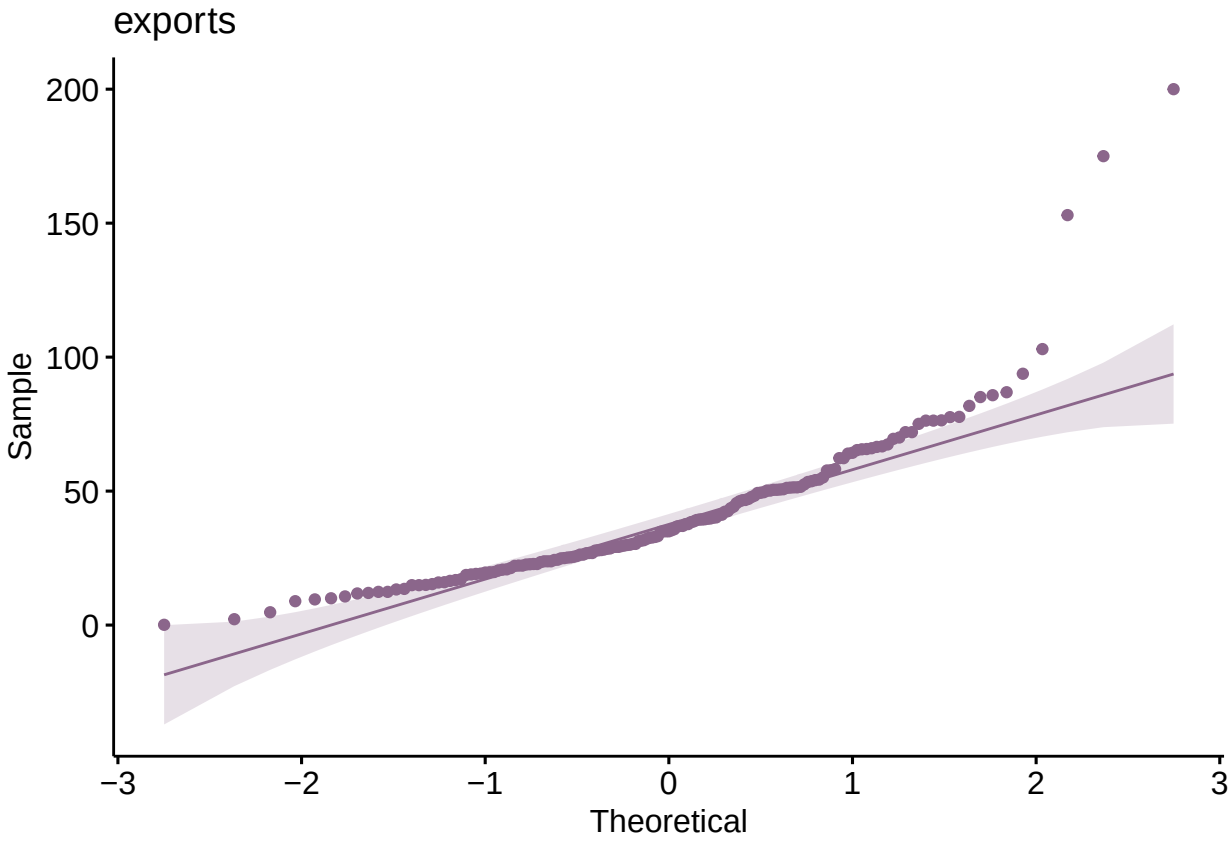

total_fer
Distplot
density
0.5
0.4
0.3
0.2
0.1
0.0
2
4
6
8
total_fer

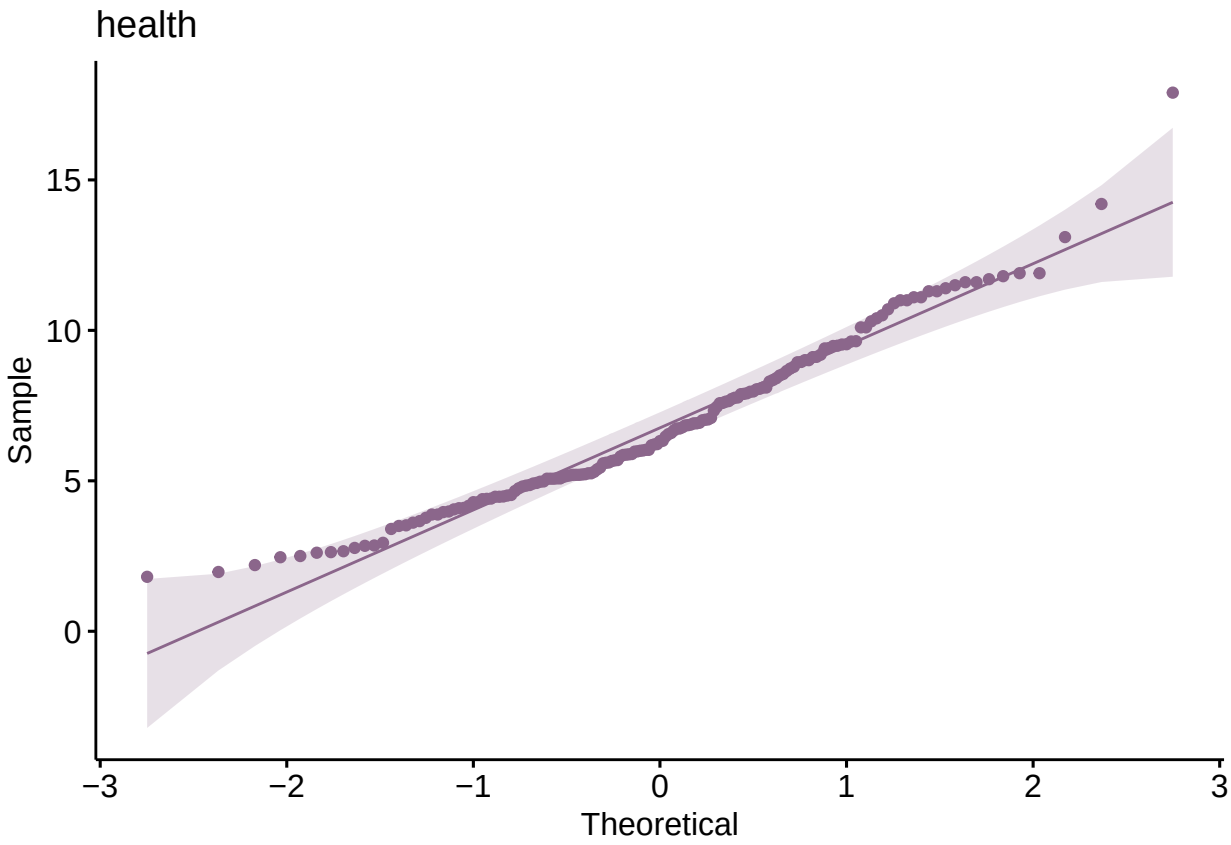

gdpp
Boxplot
0.4
0.2
0.0
−0.2
−0.4
0
25000
50000
75000
100000
gdpp

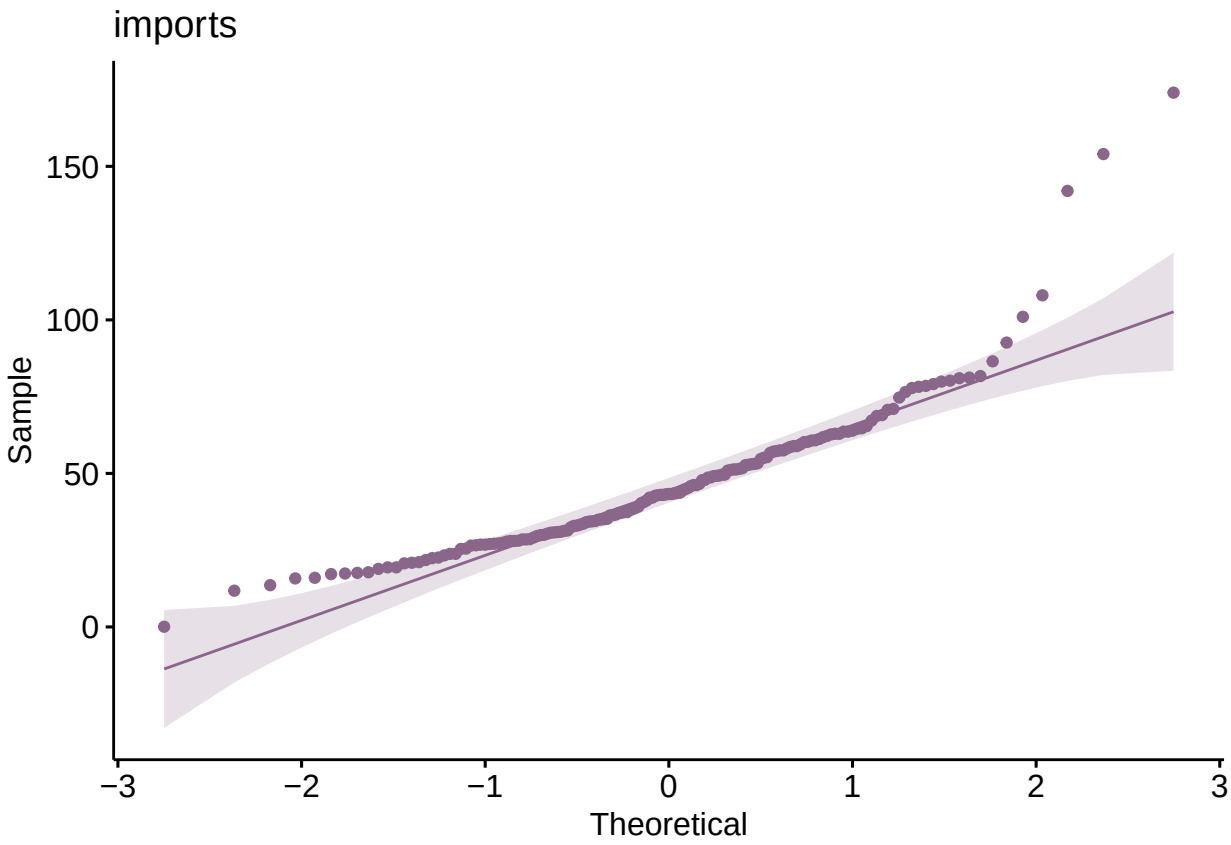

gdpp
Distplot
6e−05
4e−05
2e−05
0e+00
density
0
30000
60000
90000
gdpp

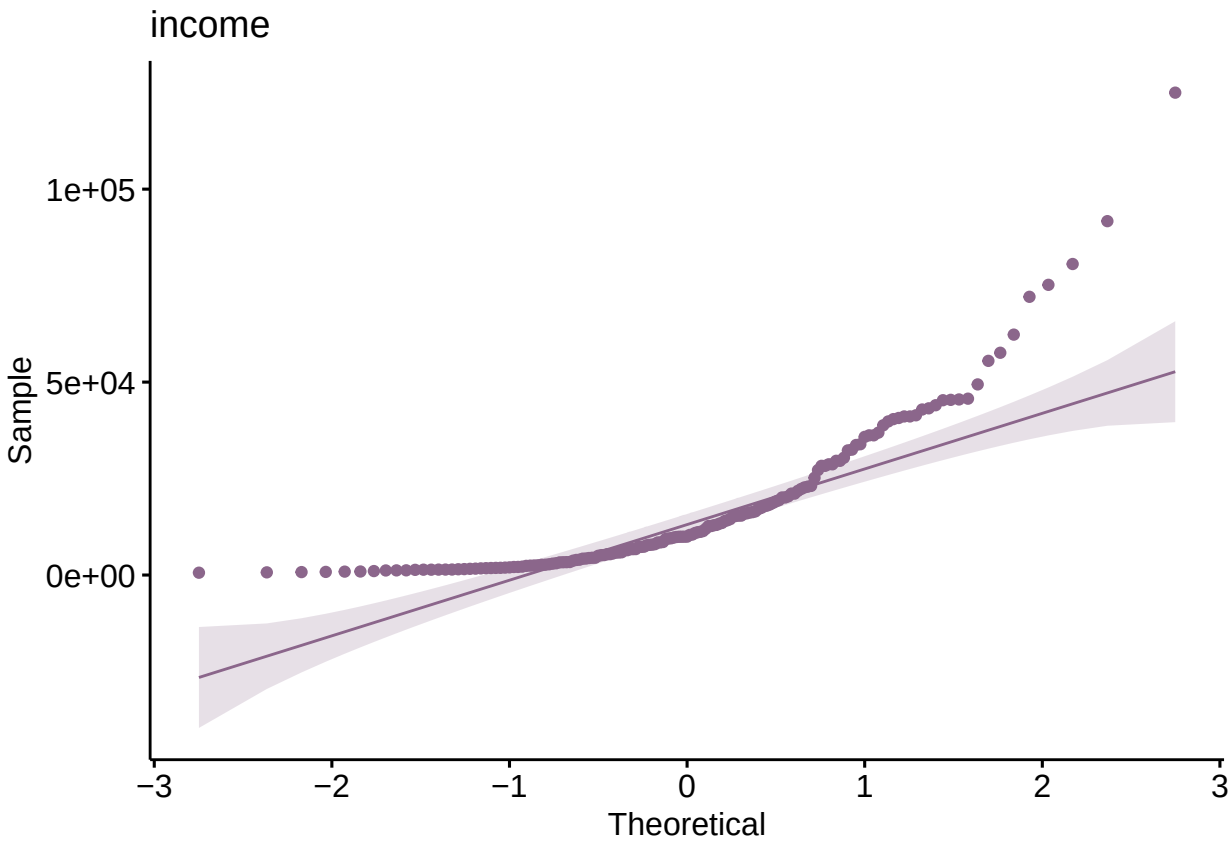
Gráficos Q-Q

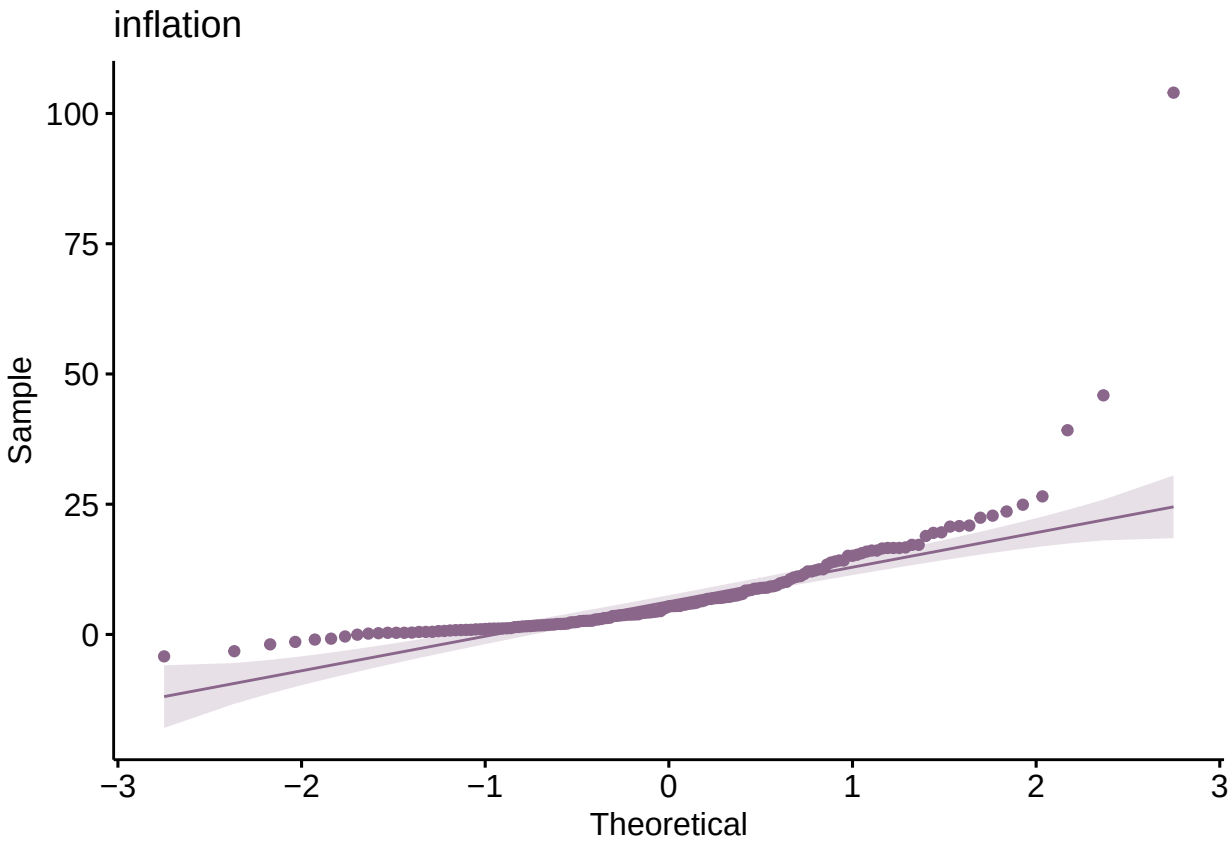

child_mort
Sample
Theoretical
200
100
0
−100
−3
−2
−1
0
1
2
3

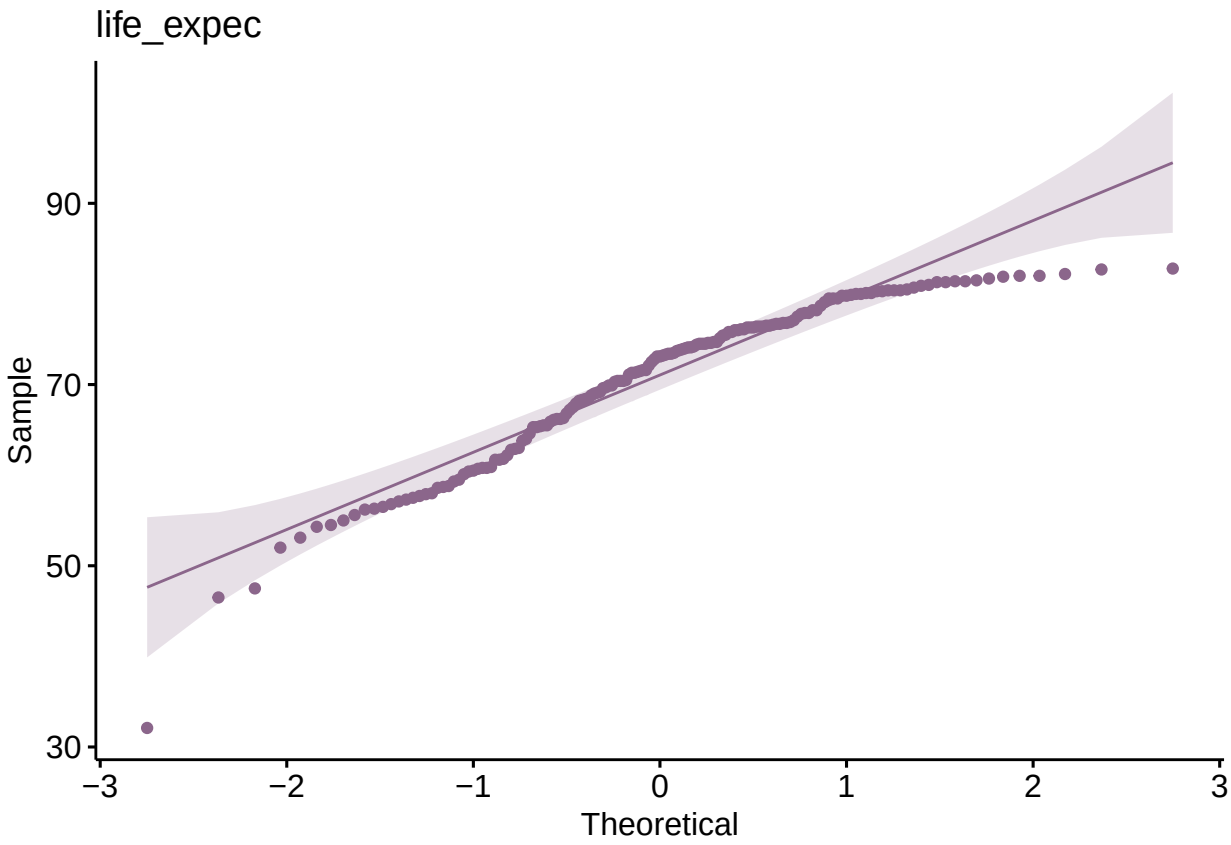

exports
Sample
Theoretical
200
150
100
50
0
−3
−2
−1
0
1
2
3

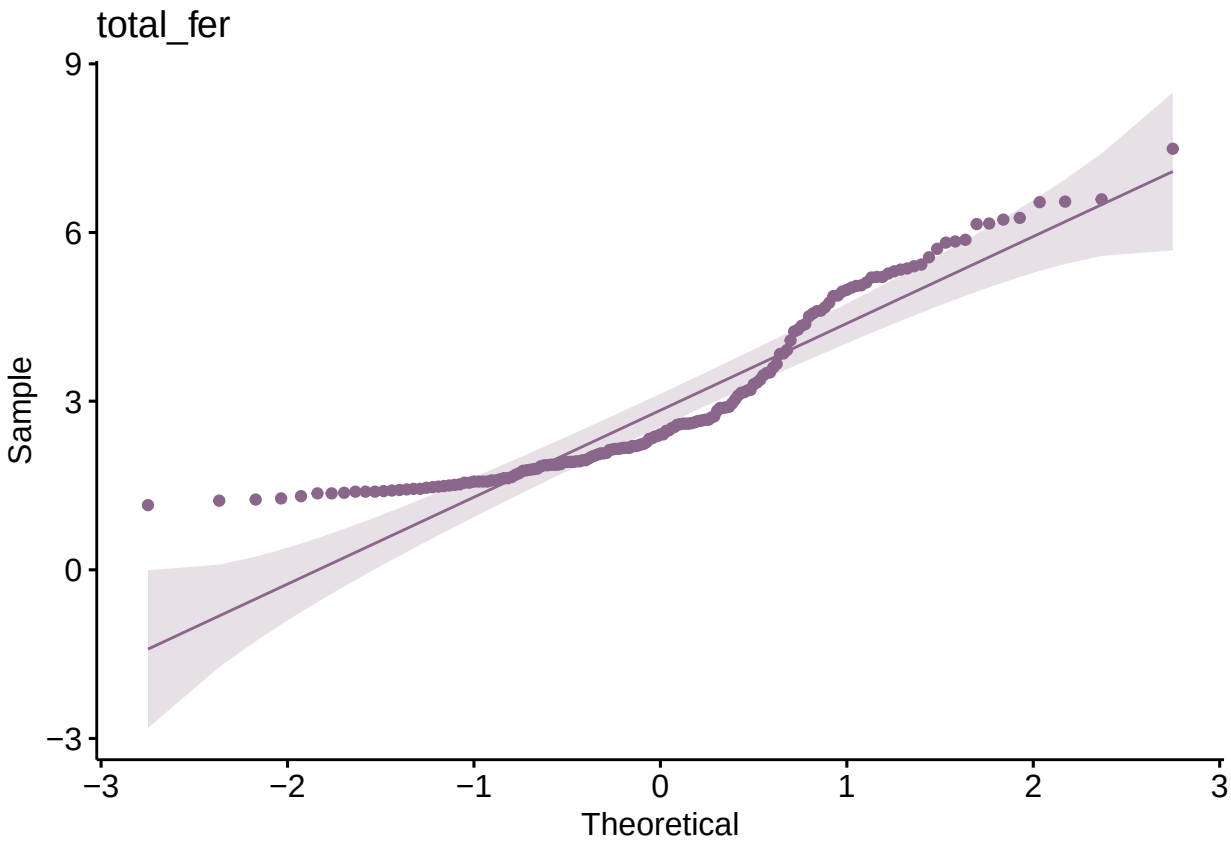

health
Sample
15
10
5
0
−3
−2
−1
0
1
2
3
Theoretical

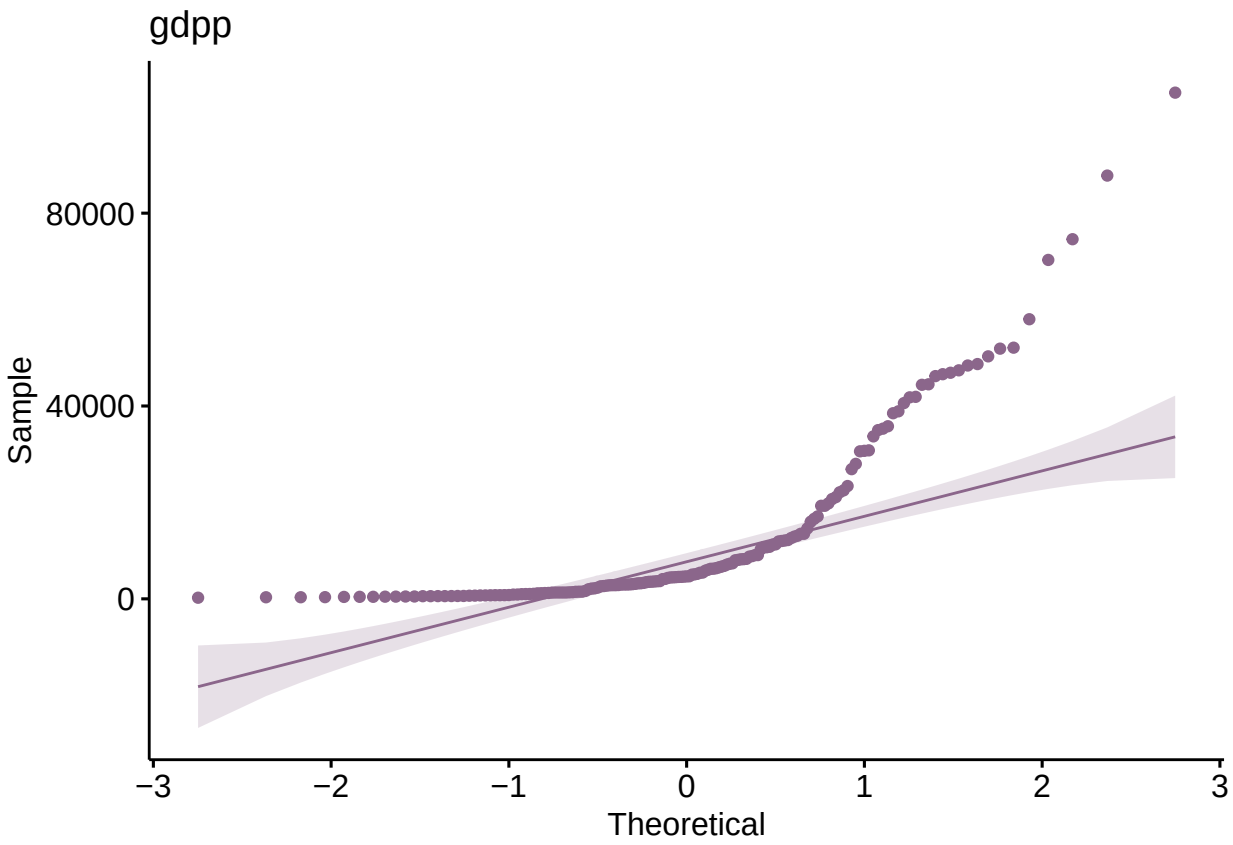

imports
150
100
50
0
Sample
−3
−2
−1
0
1
2
3
Theoretical


income
1e+05
5e+04
0e+00
Sample
−3
−2
−1
0
1
2
3
Theoretical


inflation
Sample
100
75
50
25
0
−3
−2
−1
0
1
2
3
Theoretical


life_expec
Sample
90
70
50
30
−3
−2
−1
0
1
2
3
Theoretical


total_fer
9
6
3
0
−3
Sample
−3
−2
−1
0
1
2
3
Theoretical



Teste de Shapiro-Wilk

No teste de Shapiro-Wilk, uma estatística alta (valor de W próximo a 1), indica que a amostra se aproxima de uma distribuição normal. Porém, o p-valor também deve ser validado. Para que a hipótese nula (normalidade) não seja rejeitada, o valor de p deve ser maior que 0,005.

Resultados do Teste de Normalidade (Shapiro-Wilk)

statistic

p.value

child_mort

0.8119473

2.165134e-13

exports

0.8137532

2.546082e-13

health

0.9641355

0.0002628207

imports

0.8688098

6.639577e-11

income

0.7712606

7.280748e-15

inflation

0.616314

3.640728e-19

life_expec

0.9263998

1.643109e-07

total_fer

0.8722122

9.826096e-11

gdpp

0.6964743

3.834858e-17

Com base nas análises realizadas, não foram encontrados indícios de que as variáveis do conjunto de dados sigam uma distribuição normal. Isso é evidenciado pela ausência de padrões simétricos nos histogramas, bem como pela divergência dos pontos do gráfico Q-Q da linha de referência. Além disso, os resultados do teste de Shapiro-Wilk corroboram essa conclusão, uma vez que os valores-p obtidos foram significativamente baixos, indicando uma rejeição da hipótese nula de normalidade.

Completude dos dados

A completude dos dados refere-se à proporção de dados presentes em relação ao total de dados esperados. É a medida de quão completos são os registros ou observações em um conjunto de dados. A completude dos dados é fundamental na análise exploratória de dados, pois dados incompletos podem distorcer as conclusões e limitar a eficácia das análises. Podem ocorrer viéses na interpretação dos resultados, pois as informações ausentes podem afetar a representatividade da amostra. É fundamental identificar e lidar com dados ausentes de maneira adequada durante a análise exploratória, utilizando técnicas como imputação de dados ou exclusão de registros incompletos, para garantir resultados robustos e confiáveis.

```
##            Completude
## country           100
## child_mort        100
## exports           100
## health            100
## imports           100
## income            100
## inflation         100
## life_expec        100
## total_fer         100
## gdpp              100
```

Simulação de dados faltantes

Para efeitos de estudo, uma vez que a base de dados utilizada possui uma completude de 100%, foi feita uma simulação de dados faltantes. A proporção máxima definida para os dados faltantes foi de 10% para cada variável.

```
##    country            child_mort        exports            health
##  Length:167         Min.   :  2.60   Min.   :  0.109   Min.   : 1.810
##  Class :character   1st Qu.:  7.85   1st Qu.: 24.350   1st Qu.: 4.885
##  Mode  :character   Median : 18.90   Median : 35.800   Median : 6.210
##                     Mean   : 37.56   Mean   : 42.441   Mean   : 6.753
##                     3rd Qu.: 60.40   3rd Qu.: 52.050   3rd Qu.: 8.480
##                     Max.   :208.00   Max.   :200.000   Max.   :17.900
##                     NA's   :16       NA's   :16        NA's   :16
##     imports             income        inflation         life_expec
##  Min.   :  0.0659   Min.   :  609   Min.   : -4.210   Min.   :32.10
##  1st Qu.: 29.9500   1st Qu.: 3355   1st Qu.:  1.690   1st Qu.:64.95
##  Median : 43.3000   Median :10400   Median :  5.440   Median :73.20
##  Mean   : 47.0223   Mean   :16460   Mean   :  7.693   Mean   :70.43
##  3rd Qu.: 58.3500   3rd Qu.:22050   3rd Qu.: 10.750   3rd Qu.:76.80
##  Max.   :174.0000   Max.   :91700   Max.   :104.000   Max.   :82.80
##  NA's   :16         NA's   :16      NA's   :16        NA's   :16
##    total_fer          gdpp
##  Min.   :1.150   Min.   :   231
##  1st Qu.:1.850   1st Qu.:  1365
##  Median :2.410   Median :  4560
##  Mean   :2.997   Mean   : 13411
##  3rd Qu.:4.160   3rd Qu.: 16300
##  Max.   :7.490   Max.   :105000
##  NA's   :16      NA's   :16
```

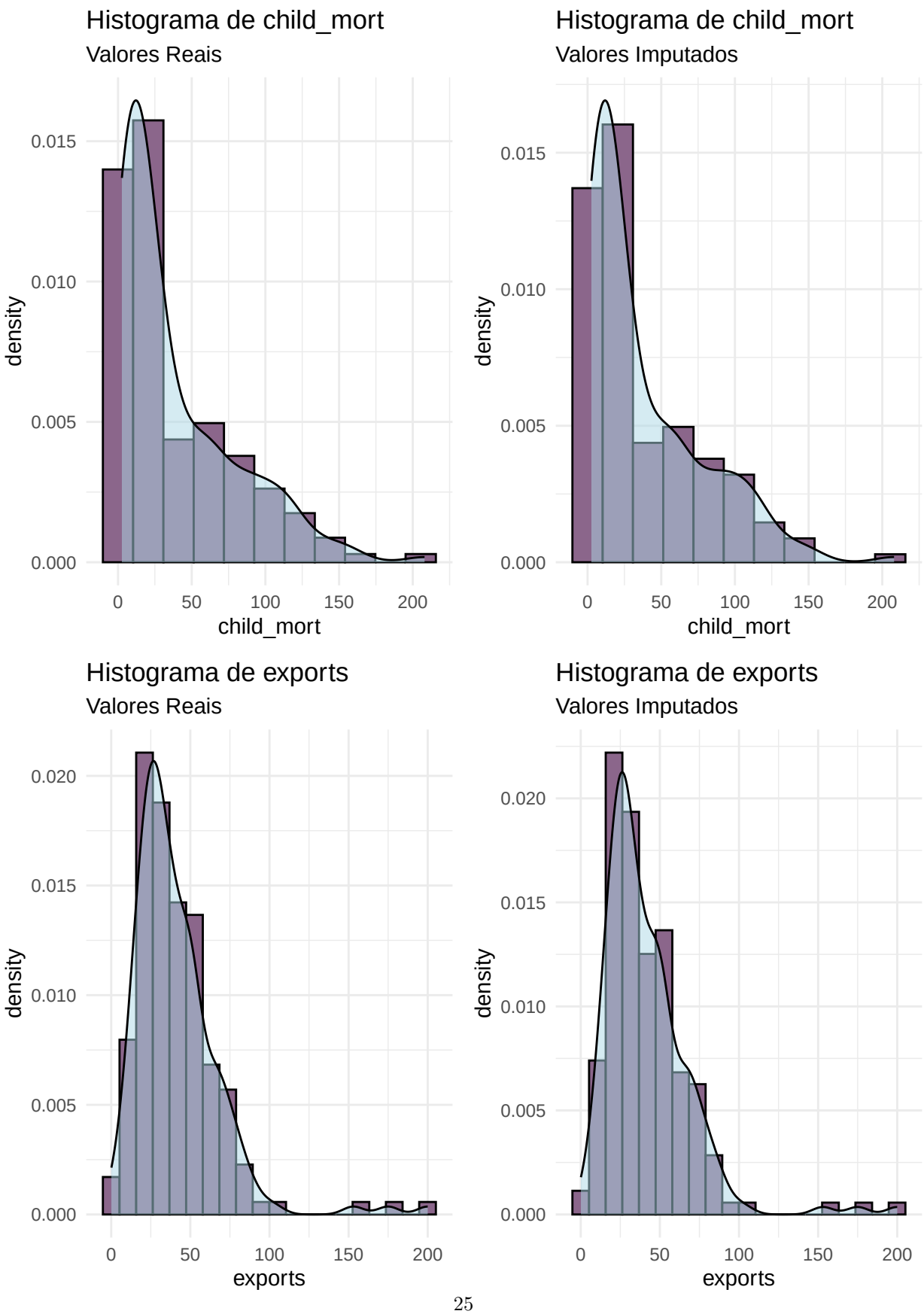
```
##    country child_mort    exports     health    imports     income  inflation
##         16         16         16         16         16         16         16
## life_expec  total_fer       gdpp
##         16         16         16
```

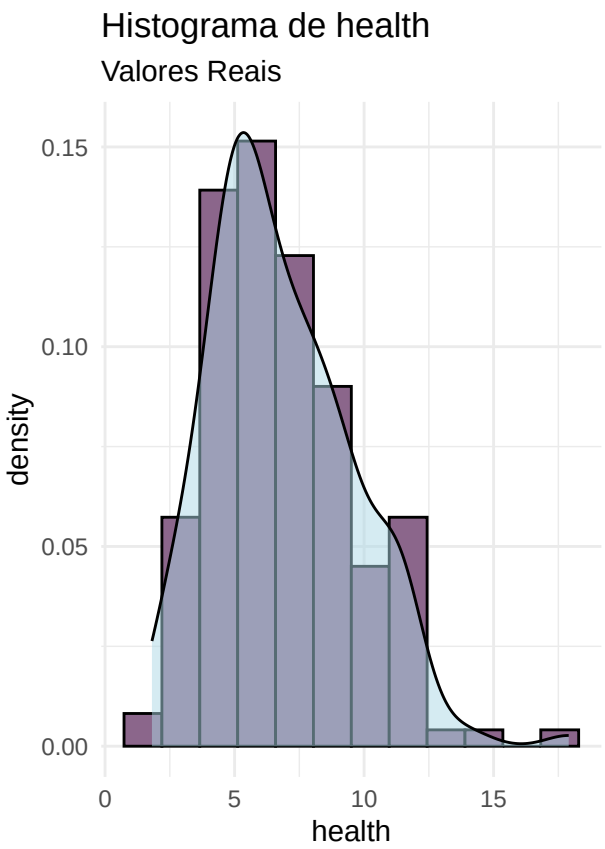
```
##            Completude
## country      90.41916
## child_mort   90.41916
## exports      90.41916
## health       90.41916
## imports      90.41916
## income       90.41916
## inflation    90.41916
## life_expec   90.41916
## total_fer    90.41916
## gdpp         90.41916
```

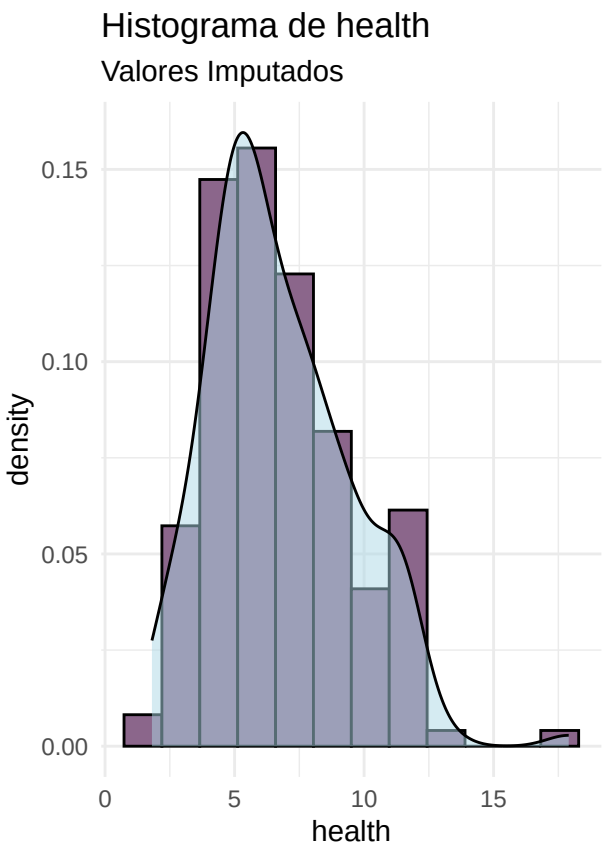
Imputação dos dados faltantes

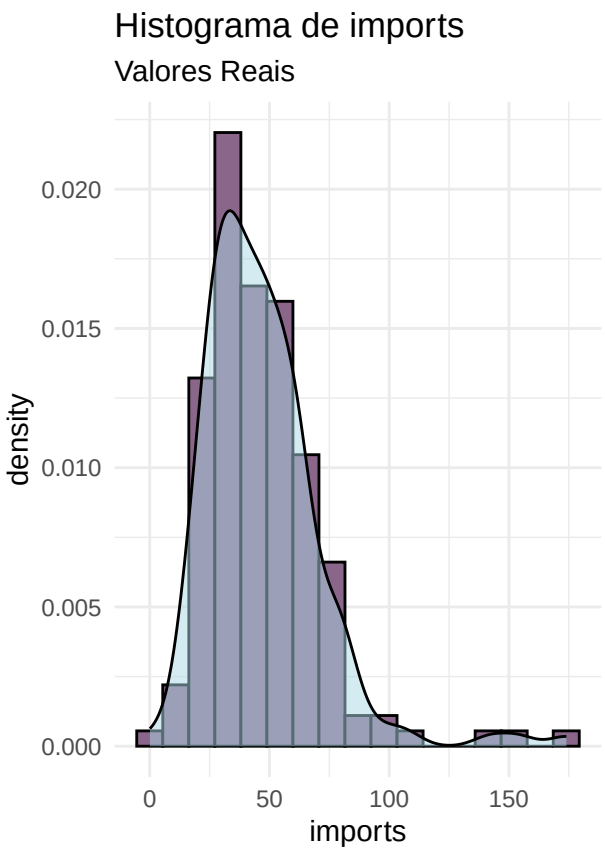
Com a base de dados simulada, foi utilizado o pacote mice para imputação dos dados faltantes.

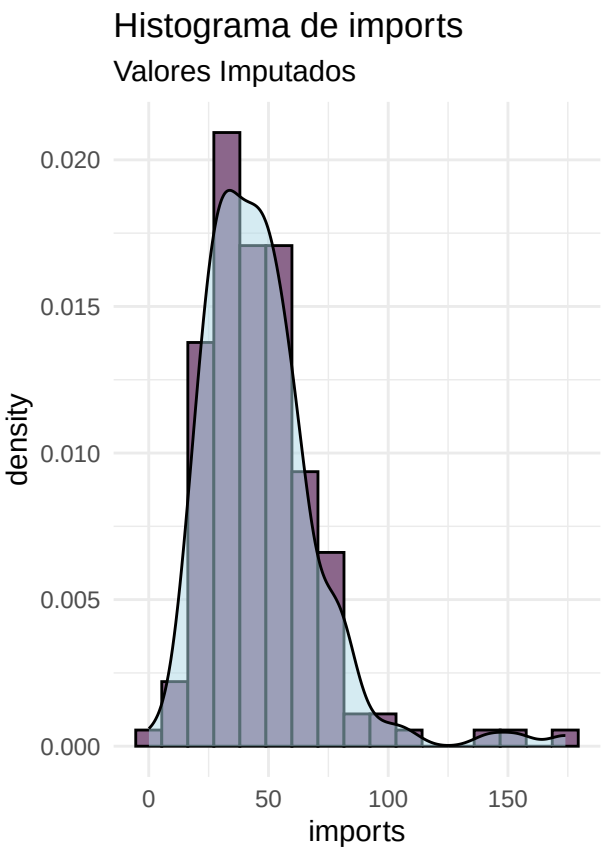
Base de dados original Vs Base com dados imputados

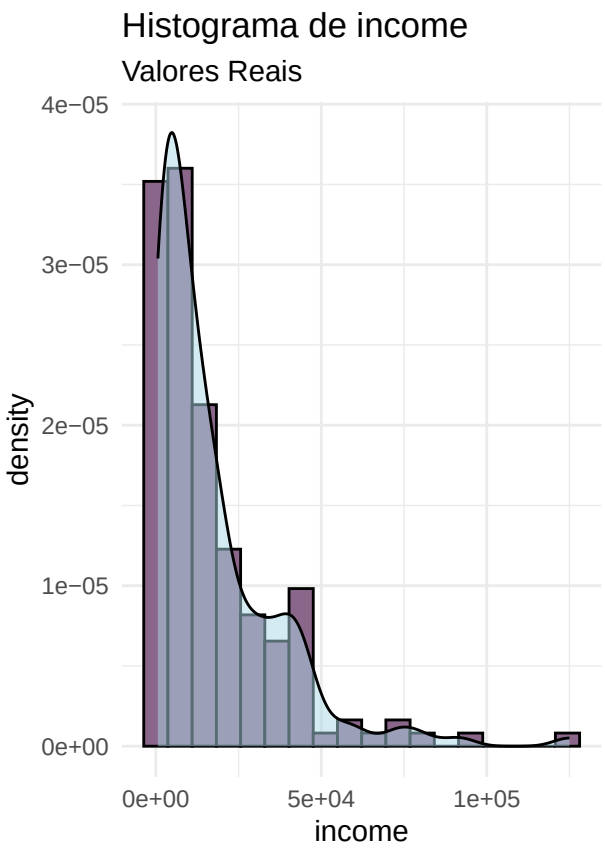


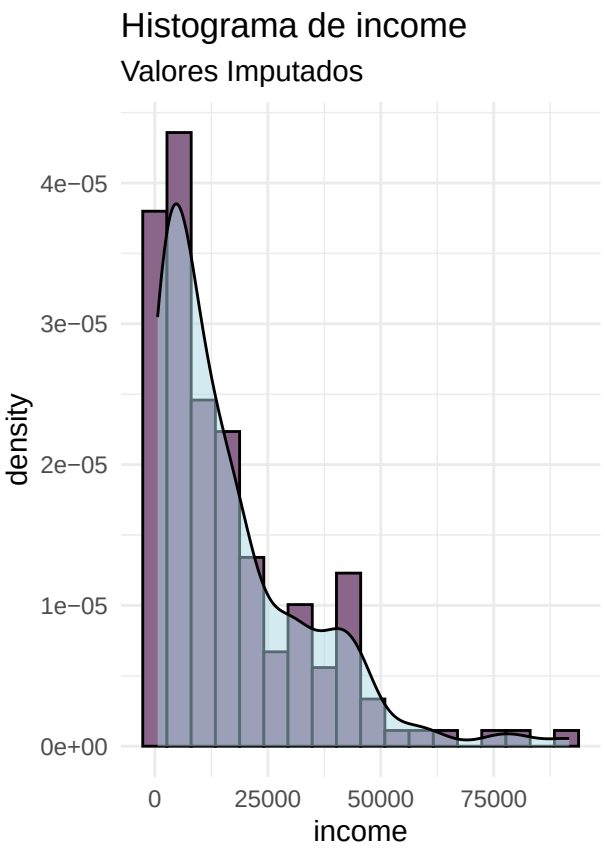

Histograma de health
Valores Reais
density
0.15
0.10
0.05
0.00
0
5
10
15
health

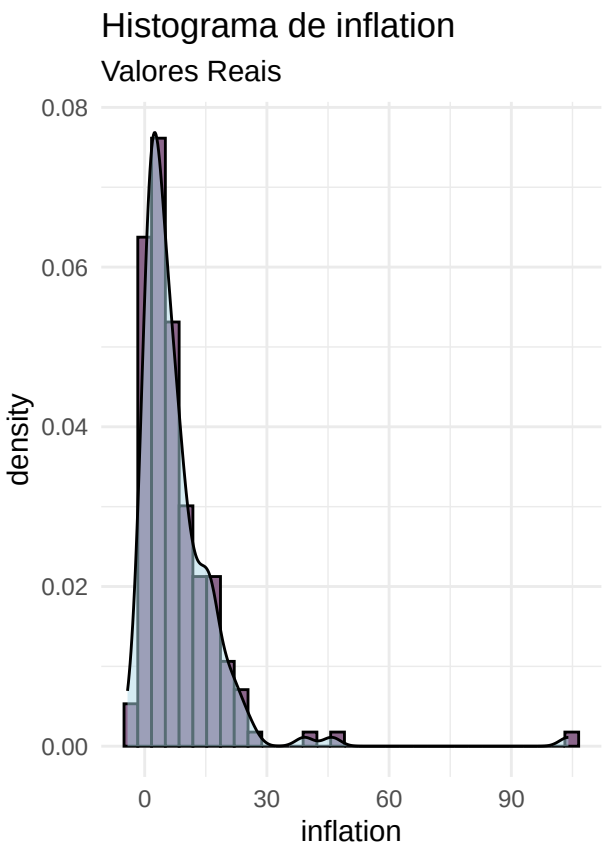

Histograma de health
Valores Imputados
density
0.15
0.10
0.05
0.00
0
5
10
15
health

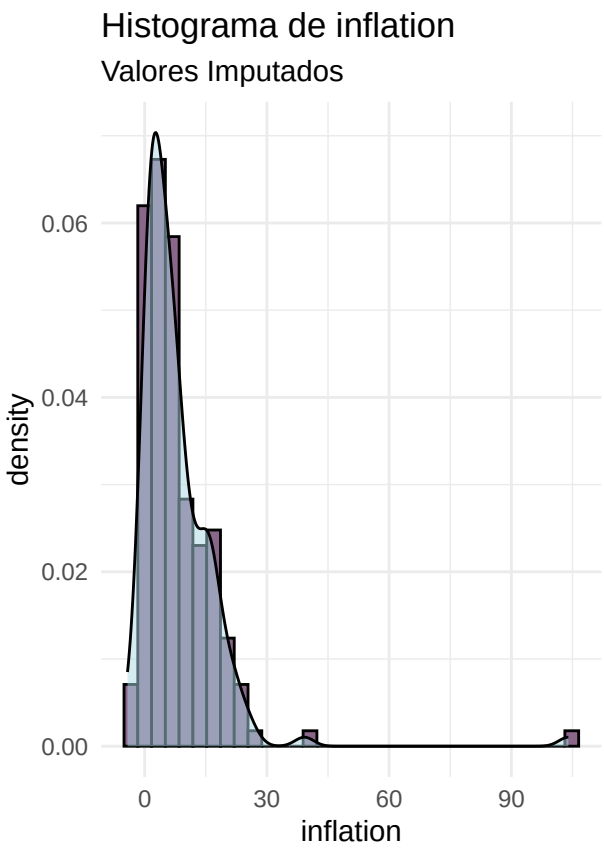

Histograma de imports
Valores Reais
density
0.020
0.015
0.010
0.005
0.000
0
50
100
150
imports

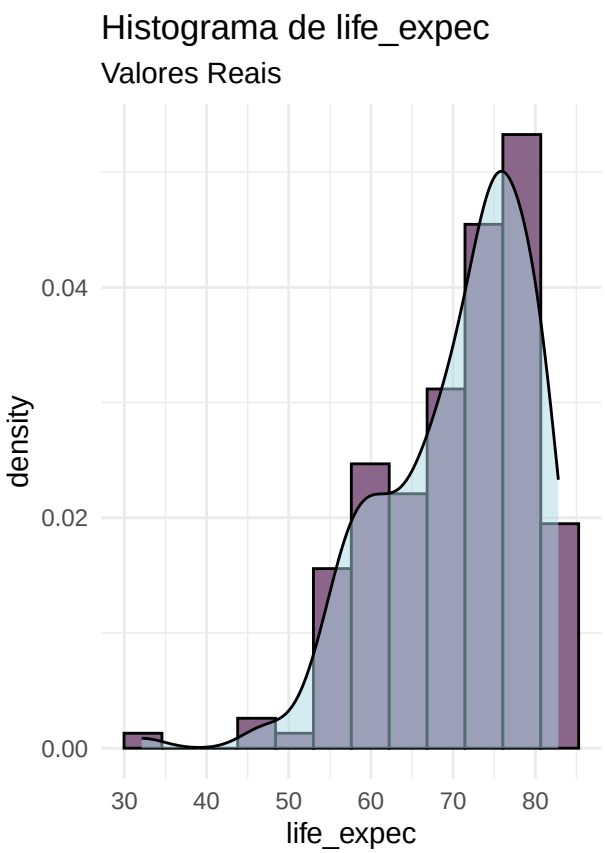

Histograma de imports
Valores Imputados
density
0.020
0.015
0.010
0.005
0.000
0
50
100
150
imports

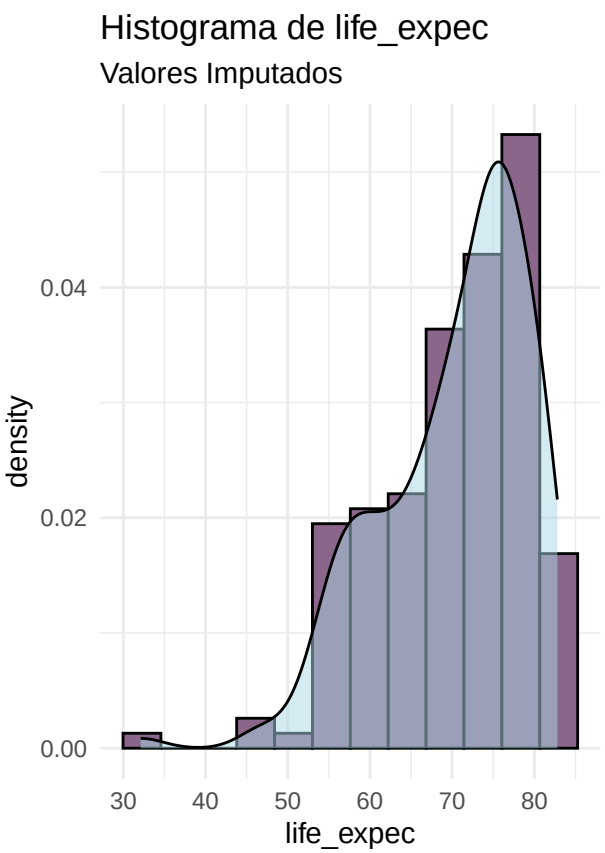

Histograma de income
Valores Reais
density
4e−05
3e−05
2e−05
1e−05
0e+00
0e+00
5e+04
1e+05
income

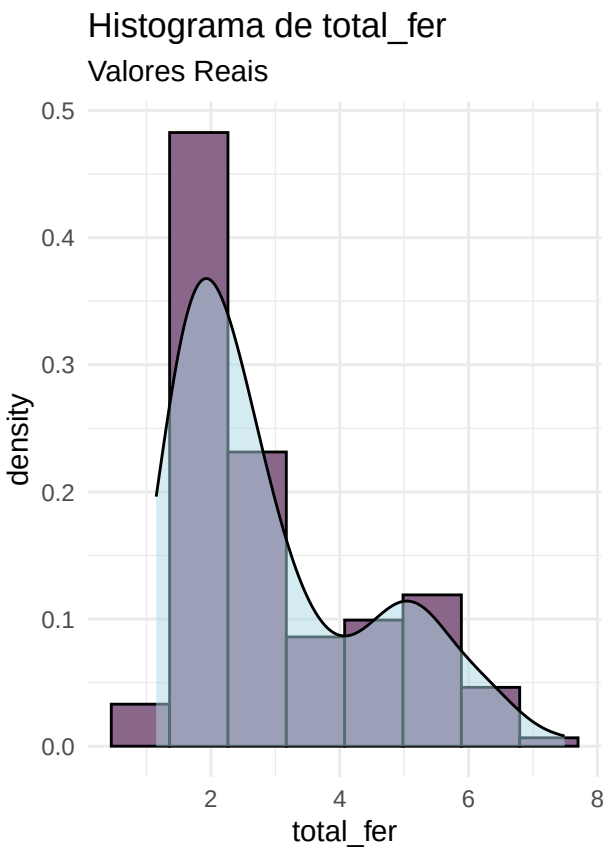

Histograma de income
Valores Imputados
density
4e−05
3e−05
2e−05
1e−05
0e+00
0
25000
50000
75000
income

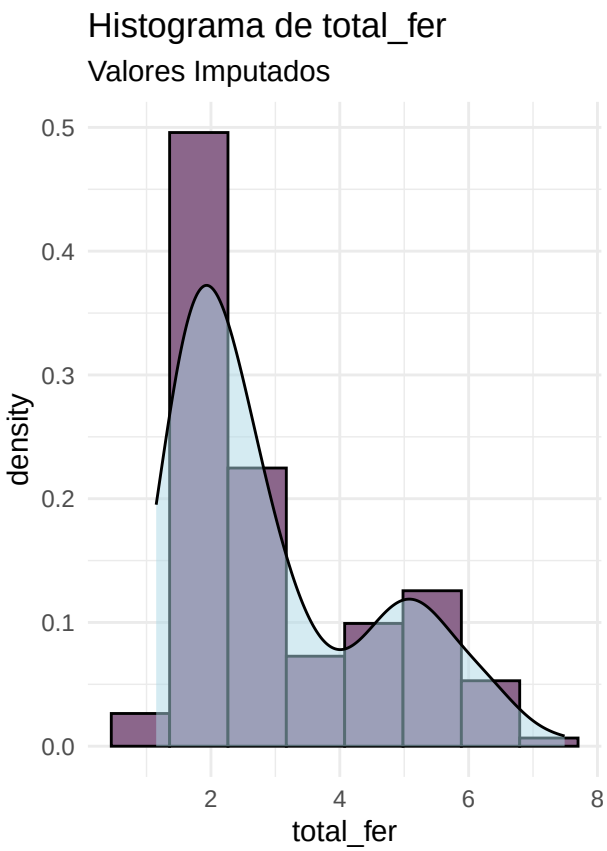

Histograma de inflation
Valores Reais
density
0.08
0.06
0.04
0.02
0.00
0
30
60
90
inflation

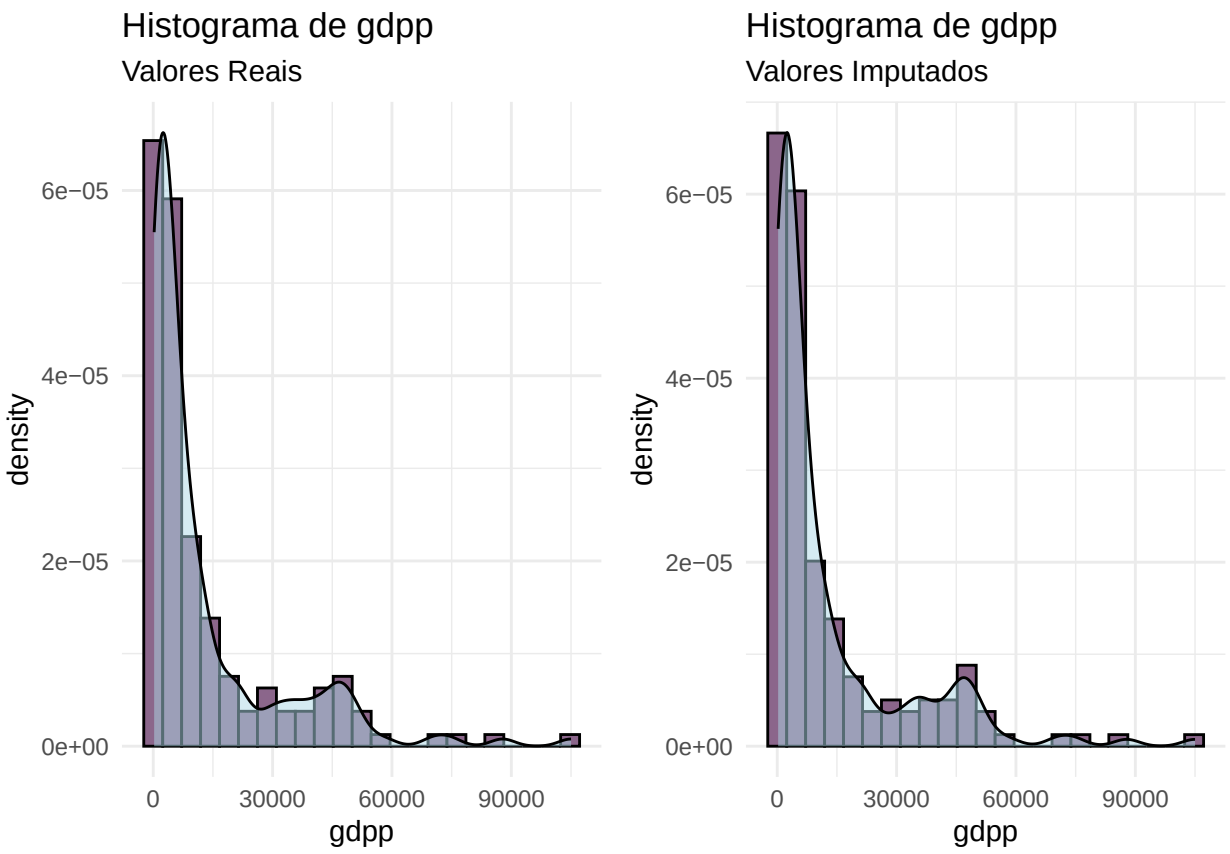

Histograma de inflation
Valores Imputados
density
0.06
0.04
0.02
0.00
0
30
60
90
inflation


Histograma de life_expec
Valores Reais
density
0.04
0.02
0.00
30
40
50
60
70
80
life_expec


Histograma de life_expec
Valores Imputados
density
0.04
0.02
0.00
30
40
50
60
70
80
life_expec


Histograma de total_fer
Valores Reais
density
total_fer
0.0
0.1
0.2
0.3
0.4
0.5
2
4
6
8


Histograma de total_fer
Valores Imputados
density
total_fer
0.0
0.1
0.2
0.3
0.4
0.5
2
4
6
8



A comparação entre os dados originais e os dados imputados, obtidos após a imputação dos valores ausentes utilizando o pacote MICE com o método PMM, revelou uma notável semelhança nos padrões visuais das distribuições. Os histogramas lado a lado para cada variável indicaram que as características essenciais dos dados originais foram preservadas no processo de imputação.

App Shiny

GitHub

Os arquivos RMarkdown e Shiny estão disponibilizados no repositório a seguir:

https://github.com/VanessaFerreiraReis/PD-VanessaReis-AnaliseExploratoriadeDados.git

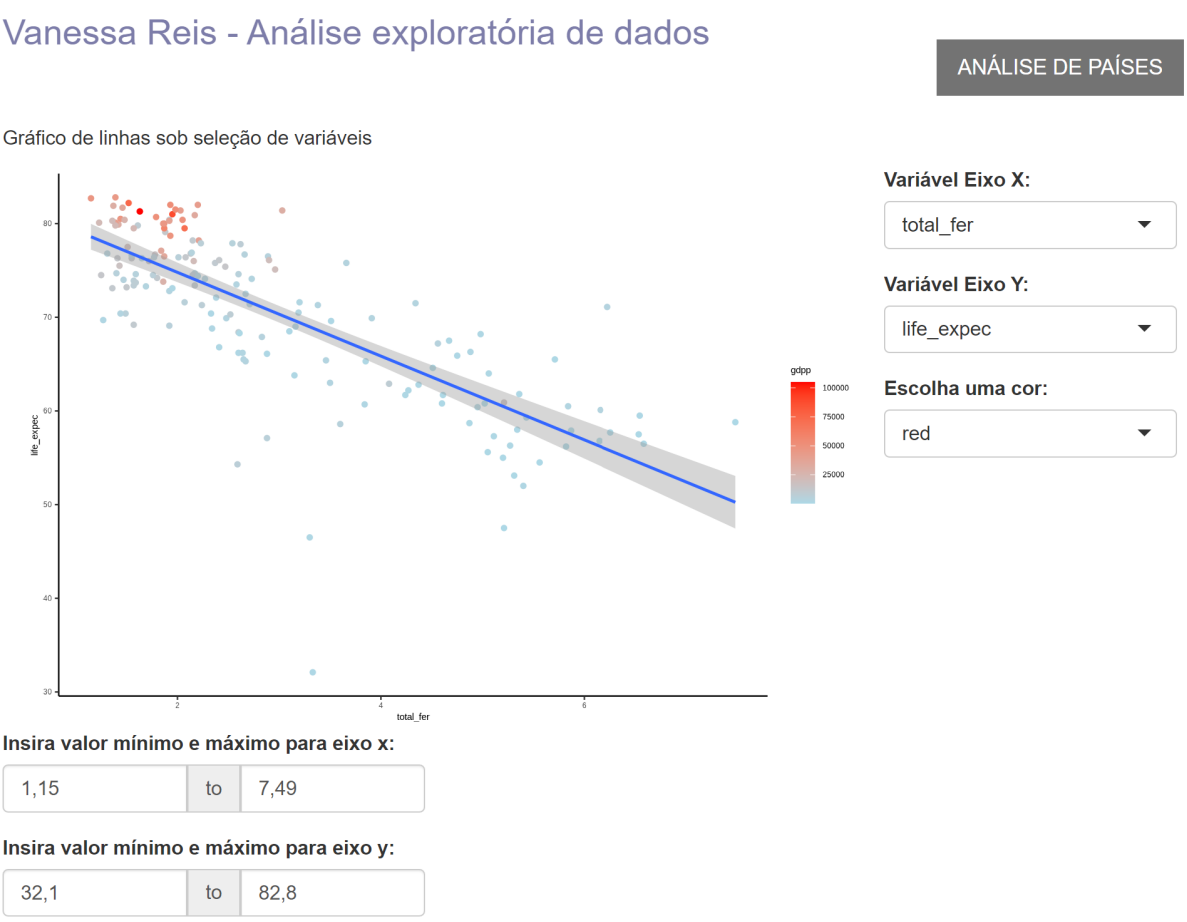


Figure 1: Print App Shiny