Projeto-Pós Graduação, Analise Exploratória de Dados.

Antonio Vieira dos Santos Neto - Cpf 077.523.948-82

2023-04-11

Contents

1 Introdução

O presente projeto visa demonstrar os conhecimentos nos fundamentos básicos na utilização da linguagem "R", bem como, nos conhecimentos de Estatística.

Neste documento, segue o passo a passo, onde são demonstrados os conhecimentos adquiridos na disciplina, “ANALISE EXPLORATÓRIA DE DADOS”.

2 Objetivo do Projeto

Demonstrar os conhecimentos adquiridos na disciplina “ANALISE EXPLORATÓRIA DE DADOS”, sendo que , através do uso da Linguagem “R”, serão feitas diversas análises em uma “Base de PIB real por países, extraida do livro do Stock & Watson, apresentada em papaer do Acemoglu. - A análise visa, demonstrar o crescimento do PIB nos 3 ultimos anos de 5 paises europeus e 5 paises da América do Sul.

3 Preparando o ambiente de análise

Para a criação do ambiente de trabalho foram adotados os seguinte passos :

- 1-Instalação da linguagem “R” na máquina do aluno.
- 2-Instalação do Studio “R” na máquina do aluno.
- 3-Instalação do “GIT” na máquina do aluno. Esta aplicação permite o controle dos versionamentos dos aplicações desenvolvidas.
- 4-Configuração na nuvem do “GITHUB, criando um repositório”WORK”, para arquivo das aplicações desenvolvidas, bem como, controle do versionamento destas.
- Atenção:
- 1 - As evidências da configuração do ambiente, segue no documento “Antonio VSNeto_Analise_Expl_Dados_evidencias.pdf”
- 2 - O caminho para o Github do aluno e´: https://github.com/avsneto2/work.git
- 3- Os arquivos do projeto se encontram na branch : Anal_Exploratoria

4 Importando bibliotecas necessárias ao desenvolvimento. -

Para suporte no tratamento da base de dados, foram instaladas as bibliotecas a seguir a partir do comando “install.packages”.

- install.packages(“tidyverse”) - Pacote de ferramentas que tem por objetivo manipulação, exploração e visualização de dados.
- install.packages(“data.table”) - Pacote que também tem a função de manipular dados, porém, em algumas situações, permite o tratamento de dados com maior velocidade
- install.packages(“rvest”) - Uma das funções do Pacote “rvest” e permitir a leitura de dados a partir de código html, dessa forma o “R” poderá mapear e navegar pela arvore do html.
- install.packages(“robotstxt”) - Este pacote fornece funcoes para baixar e analisar arquivos ‘robots.txt’.
- install.packages(“knitr”) - Este pacote tem a funcao de gerar relatorios dinamicos com R.
- install.packages (“dlookr”) - Este pacote informações estatísticas sobre dados como visualização, valores ausentes, discrepantes e valores exclusivos e negativos, com o objetivo de entender a distribuição e qualidade dos dados.
- install.packages (“readxl”) - Este pacote permite a leitura de arquivos excel.
- install.packages (“summarytools”) - Este pacote permite a análise de dados, como a frequencia de uma determinada variavel;
- install.packages (“ggplot2”) - Este pacote permite o desenvolvimento de graficos;

- install.packages (“fitdistrplus”) - Pacote com várias funções para ajudar no ajuste de uma distribuição paramétrica a dados não censurados ou censurados.
- install.packages (“Amelia”) - Pacote R para a imputação múltipla de dados incompletos multivariados. Ele usa um algoritmo que combina bootstrapping e o algoritmo EM para extrair dados posteriores dos dados ausentes. O pacote Amelia inclui transformações de normalização, priorizações em nível de célula e métodos para lidar com dados transversais de séries temporais.
- install.packages (“mice”) - Pacote R que ajuda a imputar valores ausentes com valores de dados plausíveis. Esses valores plausíveis são extraídos de uma distribuição projetada especificamente para cada ponto de dados ausente
- install.packages (“corrplot”) - Pacote R que fornece uma ferramenta exploratória visual na matriz de correlação que oferece suporte à reordenação automática de variáveis para ajudar a detectar padrões ocultos entre as variáveis.
- install.packages (“ggpubr”) - Pacote R que facilita a criação de belos gráficos baseados em ggplot2 para pesquisadores com experiência em programação não avançada.
- install.packages (“data.table”) - Amplamente usado para agregação rápida de grandes conjuntos de dados, adição/atualização/remoção de colunas de baixa latência, junções ordenadas mais rápidas e um leitor de arquivos rápido.
- install.packages (“plm”) - Pacote para R que pretende simplificar a estimação de modelos de painéis lineares. O PLM fornece funções para estimar uma ampla variedade de modelos e fazer inferências (robustas).
- install.packages (“shiny”) - pacote R que facilita a criação de aplicativos da Web interativos diretamente do R. Você pode hospedar aplicativos independentes em uma página da Web ou incorporá-los em documentos R Markdown ou criar painéis.
- install.packages (“flexdashboard”) - pode criar um documento a partir do RStudio usando a caixa de diálogo Arquivo -> Novo arquivo -> R Markdown e escolhendo um modelo “Flex Dashboard”

5 Carregando bibliotecas

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.1     v purrr   1.0.1
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.4     v forcats 1.0.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
##
## Attaching package: 'rvest'
##
##
## The following object is masked from 'package:readr':
##
##     guess_encoding
## Warning: package 'data.table' was built under R version 4.2.3
##
## Attaching package: 'data.table'
##
## The following objects are masked from 'package:dplyr':
```

```
##
##     between, first, last
##
## The following object is masked from 'package:purrr':
##
##     transpose
## Warning in !is.null(rmarkdown::metadata$output) && rmarkdown::metadata$output
## %in% : 'length(x) = 2 > 1' in coercion to 'logical(1)'
##
## Attaching package: 'dlookr'
##
## The following object is masked from 'package:tidyr':
##
##     extract
##
## The following object is masked from 'package:base':
##
##     transform
##
##
## Attaching package: 'summarytools'
##
## The following object is masked from 'package:tibble':
##
##     view
##
## Carregando pacotes exigidos: MASS
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
##
## Carregando pacotes exigidos: survival
##
## Attaching package: 'kableExtra'
##
## The following object is masked from 'package:dplyr':
##
##     group_rows
## Warning: package 'Amelia' was built under R version 4.2.3
## Carregando pacotes exigidos: Rcpp
## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.8.1, built: 2022-11-18)
## ## Copyright (C) 2005-2023 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##
## Warning: package 'mice' was built under R version 4.2.3
##
```

```
## Attaching package: 'mice'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
## Warning: package 'corrplot' was built under R version 4.2.3
```

```
## corrplot 0.92 loaded
```

```
## Warning: package 'ggpubr' was built under R version 4.2.3
```

```
## Warning: package 'plm' was built under R version 4.2.3
```

```
##
## Attaching package: 'plm'
##
## The following object is masked from 'package:data.table':
##
##     between
##
## The following objects are masked from 'package:dplyr':
##
##     between, lag, lead
```

```
## Warning: package 'shiny' was built under R version 4.2.3
```

```
## Warning: package 'flexdashboard' was built under R version 4.2.3
```

6 Definindo o diretório de trabalho para o projeto

Posteriomente, vamos trocar o diretório de referência para o trabalho, mas não vamos deixar essa informação pública para o usuário.

```
## [1] "E:/DADOS/VIEIRA/POS GRADUACAO/INFINET/CURSO/WORK_TRABALHO_2"
```

7 Importando dados

A partir da definição do ambiente, o arquivo "income_democracy.xlsx", contendo a base de pib real por países, será importada utilizando a biblioteca rvest, para ler o arquivo xlsx.

```
## Bevolucao.tbl <- readr::read_csv2("BaseDPEvolucaoMensalCisp.csv")
## Bevolucao.tbl <- readr::read_csv2("BaseDPEvolucaoMensalCisp.csv", locale=locale(encoding="latin1"))
## kable(head(Bevolucao.tbl, 5), booktabs = TRUE) %>%  kable_styling(font_size = 10)

income <- readxl::read_excel("income_democracy.xlsx")
names(income)[c(1:2, 4:length(names(income)))] <- c("pais", "ano", "Log.pib.real", "Log.populacao", "fra
```

pais.idx	pais	ano	Log.pib.real	Log.populacao	educ.adultos	fracao.pop.0_14	fracao.pop.15_19	fracao.p
1	Andorra	1960	NA	NA	NA	NA	NA	
1	Andorra	1965	NA	NA	NA	NA	NA	
1	Andorra	1970	NA	NA	NA	NA	NA	
1	Andorra	1975	NA	NA	NA	NA	NA	
1	Andorra	1980	NA	NA	NA	NA	NA	

pais.idx	pais	ano	Log.pib.real	Log.populacao	educ.adultos	fracao.pop.0_14	fracao.pop.15_
1	Andorra	1995	NA	NA	NA	NA	
2	Afghanistan	1995	NA	9.780493	0.91	0.4398145	0.2649
3	Angola	1995	7.57377	9.166389	NA	0.4633337	0.2540
4	Albania	1995	NA	8.094684	NA	0.3275600	0.2895
5	United Arab Emirates	1995	NA	7.519692	NA	0.2848008	0.2616

8 Selecionando dados

8.1 Selecionando - variaveis

Considerando o espaço amostral importado, para fins do trabalho será selecionado um evento contendo 5 (cinco) países europeus e 5 (cinco) paises da America do Sul, considerando os anos de 1995 e 2000.

```
income.arq1 <- income %>% dplyr::select(pais.idx, pais,ano,Log.pib.real,Log.populacao, educ.adultos, fra
kable(head(income.arq1, 5), booktabs = TRUE) %>%  kable_styling(font_size = 10)
```

8.2 Selecionando - dados referentes aos anos 1995 e 2000.

Com base na coluna ano, selecionar o movimento dos dois últimos anos(1995 e 2000).

```
income.arq2.1995 <- income.arq1  %>% dplyr::filter (ano =='1995')
kable(head(income.arq2.1995, 5), booktabs = TRUE) %>%  kable_styling(font_size = 10)
```

```
income.arq2.2000 <- income.arq1  %>% dplyr::filter (ano =='2000')
kable(head(income.arq2.2000, 5), booktabs = TRUE) %>%  kable_styling(font_size = 10)
```

8.3 Selecionando 1995/2000 + Paises Europa/America Suls

```
income.arq3 <- income.arq1  %>% dplyr::filter (ano =='1995'| ano=='2000' )
kable(head(income.arq3, 5), booktabs = TRUE) %>%  kable_styling(font_size = 10)
```

```
income.arq3 <- income.arq3  %>% dplyr::filter (ano =='1995'| ano=='2000' )%>% dplyr::filter (pais =='Ita
kable(head(income.arq3, 5), booktabs = TRUE) %>%  kable_styling(font_size = 10)
```

pais.idx	pais	ano	Log.pib.real	Log.populacao	educ.adultos	fracao.pop.0_14	fracao.pop.15_
1	Andorra	2000	NA	NA	NA	NA	
2	Afghanistan	2000	NA	9.995604	1.126	0.4379684	0.2660
3	Angola	2000	7.132994	9.336003	NA	0.4678443	0.2573
4	Albania	2000	7.947575	8.066208	NA	0.3158556	0.2706
5	United Arab Emirates	2000	NA	7.758334	NA	0.2791350	0.2903

pais.idx	pais	ano	Log.pib.real	Log.populacao	educ.adultos	fracao.pop.0_14	fracao.pop.15_19	frac
1	Andorra	1995	NA	NA	NA	NA	NA	
1	Andorra	2000	NA	NA	NA	NA	NA	
2	Afghanistan	1995	NA	9.780493	0.910	0.4398145	0.2649467	
2	Afghanistan	2000	NA	9.995604	1.126	0.4379684	0.2660803	
3	Angola	1995	7.57377	9.166389	NA	0.4633337	0.2540413	

pais.idx	pais	ano	Log.pib.real	Log.populacao	educ.adultos	fracao.pop.0_14	fracao.pop.15_19	fracao
6	Argentina	1995	8.884429	10.389830	7.767	0.3062166	0.2339052	
6	Argentina	2000	9.236610	10.456451	8.119	0.2890014	0.2483893	
25	Bolivia	1995	7.802203	8.790726	4.738	0.4093567	0.2717049	
25	Bolivia	2000	7.868782	8.911125	5.183	0.4058029	0.2708918	
26	Brazil	1995	8.735161	11.904680	3.762	0.3516679	0.2847025	

8.4 Selecionando 1995/2000 - dados referentes aos paises : Europeus = Italy, France, Germany, Germany.West, Spain / Am.Sul = Brazil, Argentina, Paraguay, Chile, Uruguay

```
income.arq3.1995 <- income.arq2.1995  %>% dplyr::filter (pais =='Italy'| pais =='France' | pais =='Germa
kable(head(income.arq3.1995, 5), booktabs = TRUE) %>%  kable_styling(font_size = 10)
```

```
income.arq3.2000 <- income.arq2.2000  %>% dplyr::filter (pais =='Italy'| pais =='France' | pais =='Germa
kable(head(income.arq3.2000, 5), booktabs = TRUE) %>%  kable_styling(font_size = 10)
```

8.5 Separando bases -eur - 1995/2000

```
income.arq3.1995.eur <- income.arq3.1995  %>% dplyr::filter (pais =='Italy'| pais =='France' | pais ='
kable(head(income.arq3.1995.eur , 5), booktabs = TRUE) %>%  kable_styling(font_size = 10)
```

```
income.arq3.2000.eur <- income.arq3.2000  %>% dplyr::filter (pais =='Italy'| pais =='France' | pais ='
kable(head(income.arq3.2000.eur , 5), booktabs = TRUE) %>%  kable_styling(font_size = 10)
```

8.6 Separando bases Am.Sul - 1995/2000

```
income.arq3.1995.amsul <- income.arq3.1995  %>% dplyr::filter (pais =='Brazil' | pais == 'Argentina' | p
kable(head(income.arq3.1995.amsul, 5), booktabs = TRUE) %>%  kable_styling(font_size = 10)
```

```
income.arq3.2000.amsul <- income.arq3.2000  %>% dplyr::filter (pais =='Brazil' | pais == 'Argentina' | p
kable(head(income.arq3.1995.amsul, 5), booktabs = TRUE) %>%  kable_styling(font_size = 10)
```

pais.idx	pais	ano	Log.pib.real	Log.populacao	educ.adultos	fracao.pop.0_14	fracao.pop.15_19	fracao
6	Argentina	1995	8.884429	10.389830	7.767	0.3062166	0.2339052	
25	Bolivia	1995	7.802203	8.790726	4.738	0.4093567	0.2717049	
26	Brazil	1995	8.735161	11.904680	3.762	0.3516679	0.2847025	
34	Chile	1995	8.723846	9.480291	7.138	0.3006566	0.2824859	
47	Germany	1995	9.881375	11.282670	NA	0.1608274	0.2239497	

pais.idx	pais	ano	Log.pib.real	Log.populacao	educ.adultos	fracao.pop.0_14	fracao.pop.15_19	fracao
6	Argentina	2000	9.236610	10.456451	8.119	0.2890014	0.2483893	
25	Bolivia	2000	7.868782	8.911125	5.183	0.4058029	0.2708918	
26	Brazil	2000	8.819525	11.979680	4.175	0.3240546	0.2815064	
34	Chile	2000	9.046407	9.561729	7.531	0.2945316	0.2582870	
47	Germany	2000	9.953698	11.310100	NA	0.1624316	0.1963409	

pais.idx	pais	ano	Log.pib.real	Log.populacao	educ.adultos	fracao.pop.0_14	fracao.pop.15_19	f
47	Germany	1995	9.881375	11.28267	NA	0.1608274	0.2239497	
49	Germany, West	1995	NA	NA	9.059	NA	NA	
58	Spain	1995	9.580330	10.56710	6.085	0.1937815	0.2490713	
65	France	1995	9.904653	10.94615	7.556	0.2024676	0.2261391	
90	Italy	1995	9.868263	10.94586	6.162	0.1586566	0.2373766	

pais.idx	pais	ano	Log.pib.real	Log.populacao	educ.adultos	fracao.pop.0_14	fracao.pop.15_19	f
47	Germany	2000	9.953698	11.31010	NA	0.1624316	0.1963409	
49	Germany, West	2000	NA	NA	9.571	NA	NA	
58	Spain	2000	9.699554	10.57669	6.616	0.1624308	0.2496432	
65	France	2000	9.909256	10.96550	7.944	0.1952080	0.2120300	
90	Italy	2000	9.918014	10.95438	6.600	0.1495436	0.2213399	

pais.idx	pais	ano	Log.pib.real	Log.populacao	educ.adultos	fracao.pop.0_14	fracao.pop.15_19	fracao
6	Argentina	1995	8.884429	10.389830	7.767	0.3062166	0.2339052	
25	Bolivia	1995	7.802203	8.790726	4.738	0.4093567	0.2717049	
26	Brazil	1995	8.735161	11.904680	3.762	0.3516679	0.2847025	
34	Chile	1995	8.723846	9.480291	7.138	0.3006566	0.2824859	
191	Uruguay	1995	8.890516	8.041091	6.687	0.2603798	0.2298037	

pais.idx	pais	ano	Log.pib.real	Log.populacao	educ.adultos	fracao.pop.0_14	fracao.pop.15_19	fracao
6	Argentina	1995	8.884429	10.389830	7.767	0.3062166	0.2339052	
25	Bolivia	1995	7.802203	8.790726	4.738	0.4093567	0.2717049	
26	Brazil	1995	8.735161	11.904680	3.762	0.3516679	0.2847025	
34	Chile	1995	8.723846	9.480291	7.138	0.3006566	0.2824859	
191	Uruguay	1995	8.890516	8.041091	6.687	0.2603798	0.2298037	

9 Iniciando a analise dos dados

9.1 Identificando os tipos de variaveis

Para identificar os tipos de cada variável na base, vamos utilizar o pacote "dlookr"e a função "diagnose".

```
income.arq3 %>% dlookr::diagnose()  %>% kable()
```

variables	types	missing_count	missing_percent	unique_count	unique_rate
pais.idx	numeric	0	0	10	0.50
pais	character	0	0	10	0.50
ano	numeric	0	0	2	0.10
Log.pib.real	numeric	2	10	19	0.95
Log.populacao	numeric	2	10	19	0.95
educ.adultos	numeric	2	10	19	0.95
fracao.pop.0_14	numeric	2	10	19	0.95
fracao.pop.15_19	numeric	2	10	19	0.95
fracao.pop.30_44	numeric	2	10	19	0.95
fracao.pop.45_59	numeric	2	10	19	0.95
fracao.pop.60_mais	numeric	2	10	19	0.95
idade.mediana	numeric	2	10	19	0.95

A partir do função diagnose é possível identificar o tipo de cada variável do espaço amostral, sendo:

- pais.idx - Variavel ordinal
- pais - Variavel qualitativa

- ano - variavel ordinal
- log.pib.real - variavel quantitativa continua
- log.populacao - variavel quantitativa continua
- educ.adultos - variavel quantitativa continua
- fracao.pop.0_14 - variavel quantitativa continua
- fracao.pop.15_19 - variavel quantitativa continua
- fracao.pop.30_44 - variavel quantitativa continua
- fracao.pop.45_59 - variavel quantitativa continua
- fracao.pop.60_mais - variavel quantitativa continua
- idade mediana - variavel quantitativa continua

9.2 Análise descritiva - 1995

Considerando as variaveis continuas, vamos analisar a centralidade dos dados, bem como a presença de outliers. Será utilizada a função descr do pacote summarytools, para esta análise, considerando o dados referente ao ano de 1995.

```
income.arq3.1995 %>% dplyr::select(Log.pib.real) %>% summarytools::descr()
```

Descriptive Statistics
income.arq3.1995$Log.pib.real
N: 10

	Log.pib.real
Mean	9.14
Std.Dev	0.72
Min	7.80
Q1	8.74
Median	8.89

	Log.pib.real
Q3	9.87
Max	9.90
MAD	1.02
IQR	1.13
CV	0.08
Skewness	-0.38
SE.Skewness	0.72
Kurtosis	-1.22
N.Valid	9.00
Pct.Valid	90.00

Com base o indicador skewness acima, em -0.38 e sabendo-se que valores de skew na faixa de -0.5 < skew < 0.5, podemos afirmar que trata-se de um caso fracamente simetrico

Pode-se observar uma leve diferença entre o valor da media (Mean = 9.14) e Mediana (Median =8.89), o que também remete a uma simetria entre os valores, sem a presença de outliers significativos.

```
income.arq3.1995 %>% dplyr::select(Log.populacao) %>% summarytools::descr()
```

Descriptive Statistics
income.arq3.1995$Log.populacao
N: 10

	Log.populacao
Mean	10.26
Std.Dev	1.25
Min	8.04
Q1	9.48
Median	10.57
Q3	10.95
Max	11.90
MAD	1.06
IQR	1.47
CV	0.12
Skewness	-0.48
SE.Skewness	0.72
Kurtosis	-1.26
N.Valid	9.00
Pct.Valid	90.00

Com base o indicador skewness acima, em -0.48 e sabendo-se que valores de skew na faixa de -0.5 < skew < 0.5, podemos afirmar que trata-se de um caso fracamente simetrico

Pode-se observar uma leve diferença entre o valor da media (Mean = 10.26) e Mediana (Median =10.57), o que também remete a uma simetria entre os valores, sem a presença de outliers signficativos.

```
income.arq3.1995 %>% dplyr::select(educ.adultos) %>% summarytools::descr()
```

Descriptive Statistics
income.arq3.1995$educ.adultos
N: 10

	educ.adultos
Mean	6.55
Std.Dev	1.60
Min	3.76
Q1	6.09
Median	6.69
Q3	7.56
Max	9.06
MAD	1.29
IQR	1.47
CV	0.24
Skewness	-0.24
SE.Skewness	0.72
Kurtosis	-1.08
N.Valid	9.00
Pct.Valid	90.00

Com base o indicador skewness acima, em -0.24 e sabendo-se que valores de skew na faixa de -0.5 < skew < 0.5, podemos afirmar que trata-se de um caso fracamente simetrico

Pode-se observar uma leve diferença entre o valor da media (Mean = 6.55) e Mediana (Median =6.69), o que também remete a uma simetria entre os valores, sem a presença de outliers signficativos.

```
income.arq3.1995 %>% dplyr::select(fracao.pop.0_14) %>% summarytools::descr()
```

Descriptive Statistics
income.arq3.1995$fracao.pop.0_14
N: 10

	fracao.pop.0_14
Mean	0.26
Std.Dev	0.09
Min	0.16
Q1	0.19
Median	0.26
Q3	0.31
Max	0.41
MAD	0.10
IQR	0.11
CV	0.34
Skewness	0.27
SE.Skewness	0.72
Kurtosis	-1.53
N.Valid	9.00
Pct.Valid	90.00

Com base o indicador skewness acima, em 0.27 e sabendo-se que valores de skew na faixa de -0.5 < skew < 0.5, podemos afirmar que trata-se de um caso fracamente simetrico

Pode-se observar uma leve diferença entre o valor da media (Mean = 0.26) e Mediana (Median =0.26), o que também remete a uma simetria entre os valores, sem a presença de outliers.

```
income.arq3.1995 %>% dplyr::select(fracao.pop.15_19) %>% summarytools::descr()
```

Descriptive Statistics
income.arq3.1995$fracao.pop.15_19
N: 10

	fracao.pop.15_19
Mean	0.25
Std.Dev	0.02
Min	0.22
Q1	0.23
Median	0.24
Q3	0.27
Max	0.28
MAD	0.02
IQR	0.04
CV	0.10
Skewness	0.44
SE.Skewness	0.72
Kurtosis	-1.74
N.Valid	9.00
Pct.Valid	90.00

Com base o indicador skewness acima, em 0.44 e sabendo-se que valores de skew na faixa de -0.5 < skew < 0.5, podemos afirmar que trata-se de um caso fracamente simetrico

Pode-se observar uma leve diferença entre o valor da media (Mean = 0.25) e Mediana (Median =0.24), o que também remete a uma simetria entre os valores, sem a presença de outliers significativos.

```
income.arq3.1995 %>% dplyr::select(fracao.pop.30_44) %>% summarytools::descr()
```

Descriptive Statistics
income.arq3.1995$fracao.pop.30_44
N: 10

	fracao.pop.30_44
Mean	0.20
Std.Dev	0.02
Min	0.16
Q1	0.19
Median	0.20
Q3	0.21
Max	0.22
MAD	0.01
IQR	0.02
CV	0.09
Skewness	-0.47
SE.Skewness	0.72
Kurtosis	-0.56
N.Valid	9.00
Pct.Valid	90.00

Com base o indicador skewness acima, em -0.47 e sabendo-se que valores de skew na faixa de -0.5 < skew < 0.5, podemos afirmar que trata-se de um caso fracamente simetrico

Pode-se observar uma leve diferença entre o valor da media (Mean = 0.20) e Mediana (Median =0.20), o que também remete a uma simetria entre os valores, sem a presença de outliers.

```
income.arq3.1995 %>% dplyr::select(fracao.pop.45_59) %>% summarytools::descr()
```

Descriptive Statistics
income.arq3.1995$fracao.pop.45_59
N: 10

	fracao.pop.45_59
Mean	0.15
Std.Dev	0.04
Min	0.09
Q1	0.12
Median	0.16
Q3	0.16
Max	0.20
MAD	0.04
IQR	0.04
CV	0.24
Skewness	-0.06
SE.Skewness	0.72
Kurtosis	-1.47
N.Valid	9.00
Pct.Valid	90.00

Com base o indicador skewness acima, em -0.06 pode-se afirmar que a distribuição neste caso é normal.

Pode-se observar uma leve diferença entre o valor da media (Mean = 0.15) e Mediana (Median =0.16), o que também remete a uma simetria entre os valores, sem a presença de outliers significativos.

```
income.arq3.1995 %>% dplyr::select(fracao.pop.60_mais) %>% summarytools::descr()
```

Descriptive Statistics
income.arq3.1995$fracao.pop.60_mais
N: 10

	fracao.pop.60_mais
Mean	0.15
Std.Dev	0.06
Min	0.06
Q1	0.09
Median	0.16
Q3	0.19
Max	0.21
MAD	0.06
IQR	0.10
CV	0.42
Skewness	-0.30
SE.Skewness	0.72
Kurtosis	-1.83

	fracao.pop.60_mais
N.Valid	9.00
Pct.Valid	90.00

Com base o indicador skewness acima, em -0.37 e sabendo-se que valores de skew na faixa de -0.5 < skew < 0.5, podemos afirmar que trata-se de um caso fracamente simetrico

Pode-se observar uma leve diferença entre o valor da media (Mean = 0.15) e Mediana (Median =0.16), o que também remete a uma simetria entre os valores, sem a presença de outliers.

```
income.arq3.1995 %>% dplyr::select(idade.mediana) %>% summarytools::descr()
```

Descriptive Statistics
income.arq3.1995$idade.mediana
N: 10

	idade.mediana
Mean	29.89
Std.Dev	6.62
Min	19.20
Q1	25.60
Median	30.70
Q3	34.70
Max	37.70
MAD	7.56
IQR	9.10
CV	0.22
Skewness	-0.24
SE.Skewness	0.72
Kurtosis	-1.61
N.Valid	9.00
Pct.Valid	90.00

Com base o indicador skewness acima, em -0.27 e sabendo-se que valores de skew na faixa de -0.5 < skew < 0.5, podemos afirmar que trata-se de um caso fracamente simetrico

Pode-se observar uma diferença entre o valor da media (Mean = 29.89) e Mediana (Median =30.70), o que também remete a uma simetria entre os valores, sem a presença significativa de outliers.

Conclusão : Para os dados analisados em 1995, de forma geral, os dados tendem a uma distribuição normal, sem a presença significativa de outliers

9.3 Análise descritiva - 2000

Considerando as variaveis continuas, vamos analisar a centralidade dos dados, bem como a presença de outliers. Será utilizada a função descr do pacote summarytools, para esta análise, considerando o dados referente ao ano de 2000.

```
income.arq3.2000 %>% dplyr::select(Log.pib.real) %>% summarytools::descr()
```

Descriptive Statistics
income.arq3.2000$Log.pib.real
N: 10

	Log.pib.real
Mean	9.28
Std.Dev	0.68
Min	7.87
Q1	9.05
Median	9.24
Q3	9.91
Max	9.95
MAD	0.69
IQR	0.86
CV	0.07
Skewness	-0.72
SE.Skewness	0.72
Kurtosis	-0.66
N.Valid	9.00
Pct.Valid	90.00

Com base o indicador skewness acima, em -0.72 e sabendo-se que valores de skew na faixa de $0.5 < \text{skew} < 1$, podemos afirmar que trata-se de um caso assimetria moderada.

Pode-se observar uma leve diferença entre o valor da media (Mean = 9.28) e Mediana (Median = 9.24), o que também remete a uma simetria entre os valores, sem a presença de outliers significativos.

```
income.arq3.2000 %>% dplyr::select(Log.populacao) %>% summarytools::descr()
```

Descriptive Statistics
income.arq3.2000$Log.populacao
N: 10

	Log.populacao
Mean	10.31
Std.Dev	1.24
Min	8.08
Q1	9.56
Median	10.58
Q3	10.97
Max	11.98
MAD	1.09
IQR	1.40
CV	0.12
Skewness	-0.47
SE.Skewness	0.72
Kurtosis	-1.19
N.Valid	9.00
Pct.Valid	90.00

Com base o indicador skewness acima, em -0.47 e sabendo-se que valores de skew na faixa de $-0.5 < \text{skew} < 0.5$, podemos afirmar que trata-se de um caso fracamente simetrico

Pode-se observar uma leve diferença entre o valor da media (Mean = 10.31) e Mediana (Median =10.58), o que também remete a uma simetria entre os valores, sem a presença de outliers signficativos.

```
income.arq3.2000 %>% dplyr::select(educ.adultos) %>% summarytools::descr()
```

Descriptive Statistics
income.arq3.2000$educ.adultos
N: 10

	educ.adultos
Mean	6.96
Std.Dev	1.61
Min	4.18
Q1	6.60
Median	6.88
Q3	7.94
Max	9.57
MAD	1.58
IQR	1.34
CV	0.23
Skewness	-0.18
SE.Skewness	0.72
Kurtosis	-1.00
N.Valid	9.00
Pct.Valid	90.00

Com base o indicador skewness acima, em -0.18 e sabendo-se que valores de skew na faixa de -0.5 < skew < 0.5, podemos afirmar que trata-se de um caso fracamente simetrico

Pode-se observar uma leve diferença entre o valor da media (Mean = 6.96) e Mediana (Median =6.88), o que também remete a uma simetria entre os valores, sem a presença de outliers signficativos.

```
income.arq3.2000 %>% dplyr::select(fracao.pop.0_14) %>% summarytools::descr()
```

Descriptive Statistics
income.arq3.2000$fracao.pop.0_14
N: 10

	fracao.pop.0_14
Mean	0.25
Std.Dev	0.09
Min	0.15
Q1	0.16
Median	0.25
Q3	0.29
Max	0.41
MAD	0.11
IQR	0.13
CV	0.35
Skewness	0.35
SE.Skewness	0.72
Kurtosis	-1.38
N.Valid	9.00
Pct.Valid	90.00

Com base o indicador skewness acima, em 0.35 e sabendo-se que valores de skew na faixa de -0.5 < skew < 0.5, podemos afirmar que trata-se de um caso fracamente simetrico

Pode-se observar uma leve diferença entre o valor da media (Mean = 0.25) e Mediana (Median =0.25), o que também remete a uma simetria entre os valores, sem a presença de outliers.

```
income.arq3.2000 %>% dplyr::select(fracao.pop.15_19) %>% summarytools::descr()
```

Descriptive Statistics
income.arq3.2000$fracao.pop.15_19
N: 10

	fracao.pop.15_19
Mean	0.24
Std.Dev	0.03
Min	0.20
Q1	0.22
Median	0.25
Q3	0.26
Max	0.28
MAD	0.03
IQR	0.04
CV	0.12
Skewness	-0.16
SE.Skewness	0.72
Kurtosis	-1.44
N.Valid	9.00
Pct.Valid	90.00

Com base o indicador skewness acima, em -0.16 e sabendo-se que valores de skew na faixa de -0.5 < skew < 0.5, podemos afirmar que trata-se de um caso fracamente simetrico

Pode-se observar uma leve diferença entre o valor da media (Mean = 0.24) e Mediana (Median =0.25), o que também remete a uma simetria entre os valores, sem a presença de outliers significativos.

```
income.arq3.2000 %>% dplyr::select(fracao.pop.30_44) %>% summarytools::descr()
```

Descriptive Statistics
income.arq3.2000$fracao.pop.30_44
N: 10

	fracao.pop.30_44
Mean	0.21
Std.Dev	0.02
Min	0.17
Q1	0.19
Median	0.22
Q3	0.22
Max	0.24
MAD	0.01
IQR	0.03
CV	0.10
Skewness	-0.64
SE.Skewness	0.72

	fracao.pop.30_44
Kurtosis	-0.84
N.Valid	9.00
Pct.Valid	90.00

Com base o indicador skewness acima, em -0.64 e sabendo-se que valores de skew na faixa de -0.5 < skew < 0.5, podemos afirmar que trata-se de um caso fracamente simetrico

Pode-se observar uma leve diferença entre o valor da media (Mean = 0.21) e Mediana (Median =0.22), o que também remete a uma simetria entre os valores, sem a presença de outliers.

```
income.arq3.2000 %>% dplyr::select(fracao.pop.45_59) %>% summarytools::descr()
```

Descriptive Statistics
income.arq3.2000$fracao.pop.45_59
N: 10

	fracao.pop.45_59
Mean	0.15
Std.Dev	0.03
Min	0.10
Q1	0.13
Median	0.15
Q3	0.17
Max	0.20
MAD	0.03
IQR	0.04
CV	0.23
Skewness	-0.12
SE.Skewness	0.72
Kurtosis	-1.47
N.Valid	9.00
Pct.Valid	90.00
Com base o indicad	or skewness acima, em -0.12 e sabendo-se que valores
de skew na faixa d	e -0.5 < skew < 0.5, podemos afirmar que trata-se de
um caso fracamente	simetrico

Pode-se observar uma leve diferença entre o valor da media (Mean = 0.15) e Mediana (Median =0.15), o que também remete a uma simetria entre os valores, sem a presença de outliers.

```
income.arq3.2000 %>% dplyr::select(fracao.pop.60_mais) %>% summarytools::descr()
```

Descriptive Statistics
income.arq3.2000$fracao.pop.60_mais
N: 10

	fracao.pop.60_mais
Mean	0.15
Std.Dev	0.06
Min	0.06
Q1	0.10
Median	0.17

	fracao.pop.60_mais
Q3	0.21
Max	0.22
MAD	0.06
IQR	0.11
CV	0.42
Skewness	-0.28
SE.Skewness	0.72
Kurtosis	-1.83
N.Valid	9.00
Pct.Valid	90.00

Com base o indicador skewness acima, em -0.28 e sabendo-se que valores de skew na faixa de -0.5 < skew < 0.5, podemos afirmar que trata-se de um caso fracamente simetrico

Pode-se observar uma leve diferença entre o valor da media (Mean = 0.15) e Mediana (Median =0.17), o que também remete a uma simetria entre os valores, sem a presença de outliers.

```
income.arq3.2000 %>% dplyr::select(idade.mediana) %>% summarytools::descr()
```

Descriptive Statistics
income.arq3.2000$idade.mediana
N: 10

	idade.mediana
Mean	30.86
Std.Dev	6.79
Min	19.60
Q1	27.00
Median	31.10
Q3	36.20
Max	38.50
MAD	7.56
IQR	9.20
CV	0.22
Skewness	-0.26
SE.Skewness	0.72
Kurtosis	-1.60
N.Valid	9.00
Pct.Valid	90.00

Com base o indicador skewness acima, em -0.26 e sabendo-se que valores de skew na faixa de -0.5 < skew < 0.5, podemos afirmar que trata-se de um caso fracamente simetrico

Pode-se observar uma diferença entre o valor da media (Mean = 30.86) e Mediana (Median =31.10), o que também remete a uma simetria entre os valores, sem a presença significativa de outliers.

Conclusão : Para os dados analisados em 2000, de forma geral, os dados tendem a uma distribuição normal, sem a presença significativa de outlier.

9.4 Analisando a matriz de correlação - Grafico de scatter matrix - 1995

```
kable(cor(income.arq3.1995 %>% dplyr::filter(!is.na(Log.pib.real) ) %>% dplyr::select(ano, Log.pib.real
```

```
## Warning in cor(income.arq3.1995 %>% dplyr::filter(!is.na(Log.pib.real)) %>% : o
## desvio padrão é zero
```

	ano	Log.pib.real	Log.populacao	educ.adultos	fracao.pop.0_14	fracao.pop.15_19	fracao.p
ano	1	NA	NA	NA	NA	NA	
Log.pib.real	NA	1.0000000	0.5666880	NA	-0.9605685	-0.6737943	
Log.populacao	NA	0.5666880	1.0000000	NA	-0.3694837	-0.0382808	
educ.adultos	NA	NA	NA	1	NA	NA	
fracao.pop.0_14	NA	-0.9605685	-0.3694837	NA	1.0000000	0.7114847	-
fracao.pop.15_19	NA	-0.6737943	-0.0382808	NA	0.7114847	1.0000000	-
fracao.pop.30_44	NA	0.8759909	0.5538517	NA	-0.7815898	-0.4144241	
fracao.pop.45_59	NA	0.8778511	0.2664224	NA	-0.9532127	-0.8164068	
fracao.pop.60_mais	NA	0.9103405	0.2420080	NA	-0.9651454	-0.8446669	
idade.mediana	NA	0.9543536	0.3413811	NA	-0.9934697	-0.7826979	

Na analise acima é possivel verificar maior correlação entre as seguintes variaveis:

- Log.pib.real x fracao.pop.30_44 = 0.87
- log.pib.real x fracao.pop.45_59 = 0.87
- log.pib.real x fracao.pop.60_mais = 0.91
- log.pib.real x idade.mediana = 0.95
- fracao.pop.60_mais x idade.mediana = 0.98

9.5 Analisando a matriz de correlação - Grafico de scatter matrix - 2000

```
kable(cor(income.arq3.2000 %>% dplyr::filter(!is.na(Log.pib.real) ) %>% dplyr::select(ano, Log.pib.real
```

```
## Warning in cor(income.arq3.2000 %>% dplyr::filter(!is.na(Log.pib.real)) %>% : o
## desvio padrão é zero
```

	ano	Log.pib.real	Log.populacao	educ.adultos	fracao.pop.0_14	fracao.pop.15_19	fracao.p
ano	1	NA	NA	NA	NA	NA	
Log.pib.real	NA	1.0000000	0.5070637	NA	-0.9597290	-0.7961870	
Log.populacao	NA	0.5070637	1.0000000	NA	-0.3848088	-0.1699621	
educ.adultos	NA	NA	NA	1	NA	NA	
fracao.pop.0_14	NA	-0.9597290	-0.3848088	NA	1.0000000	0.7885714	-
fracao.pop.15_19	NA	-0.7961870	-0.1699621	NA	0.7885714	1.0000000	-
fracao.pop.30_44	NA	0.8120684	0.6062413	NA	-0.7412838	-0.5621185	
fracao.pop.45_59	NA	0.9361182	0.3374141	NA	-0.9605310	-0.9096630	
fracao.pop.60_mais	NA	0.8991904	0.2205028	NA	-0.9623150	-0.8510550	
idade.mediana	NA	0.9547972	0.3457033	NA	-0.9886892	-0.8676642	

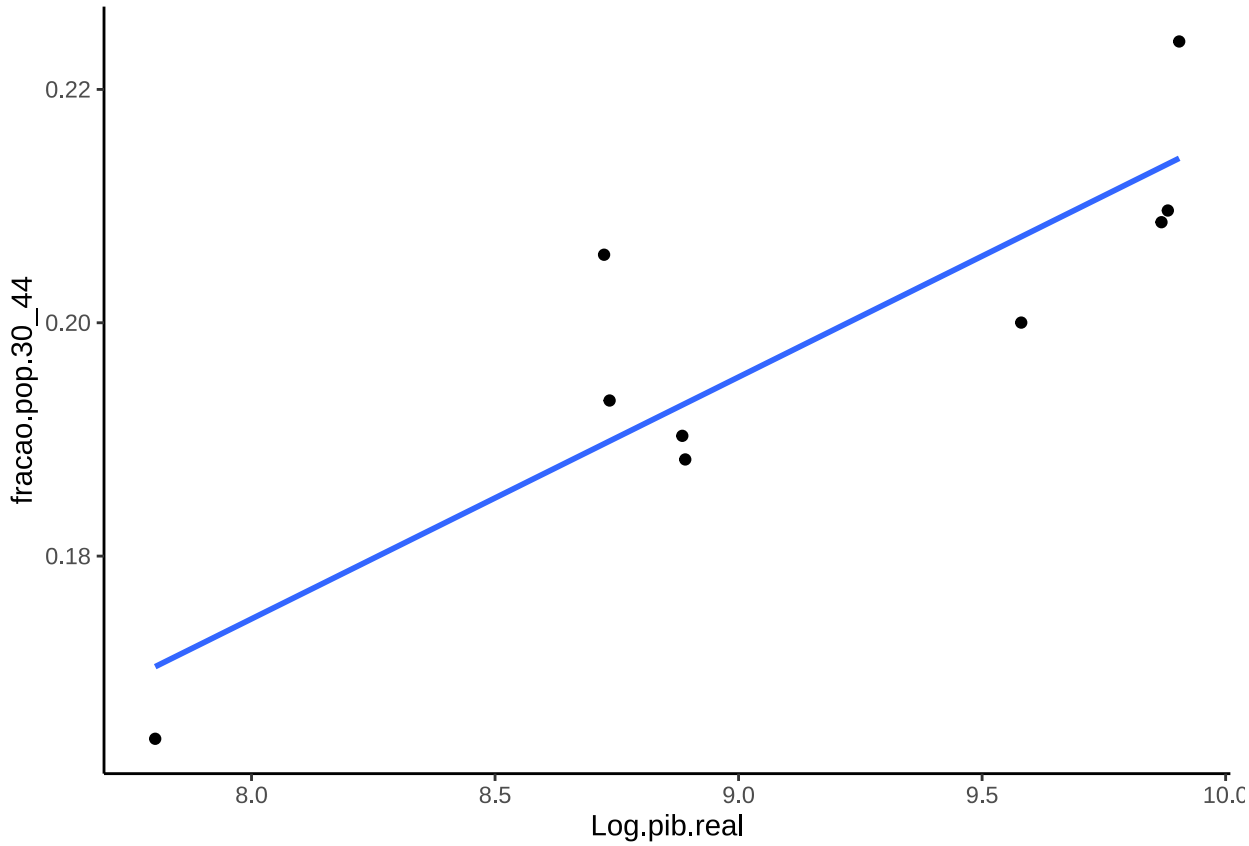
Na analise acima é possivel verificar maior correlação entre as seguintes variaveis:

- Log.pib.real x fracao.pop.30_44 = 0.87
- log.pib.real x fracao.pop.45_59 = 0.87
- log.pib.real x fracao.pop.60_mais = 0.91
- log.pib.real x idade.mediana = 0.95
- fracao.pop.60_mais x idade.mediana = 0.97

Analisando Scatter - 1995

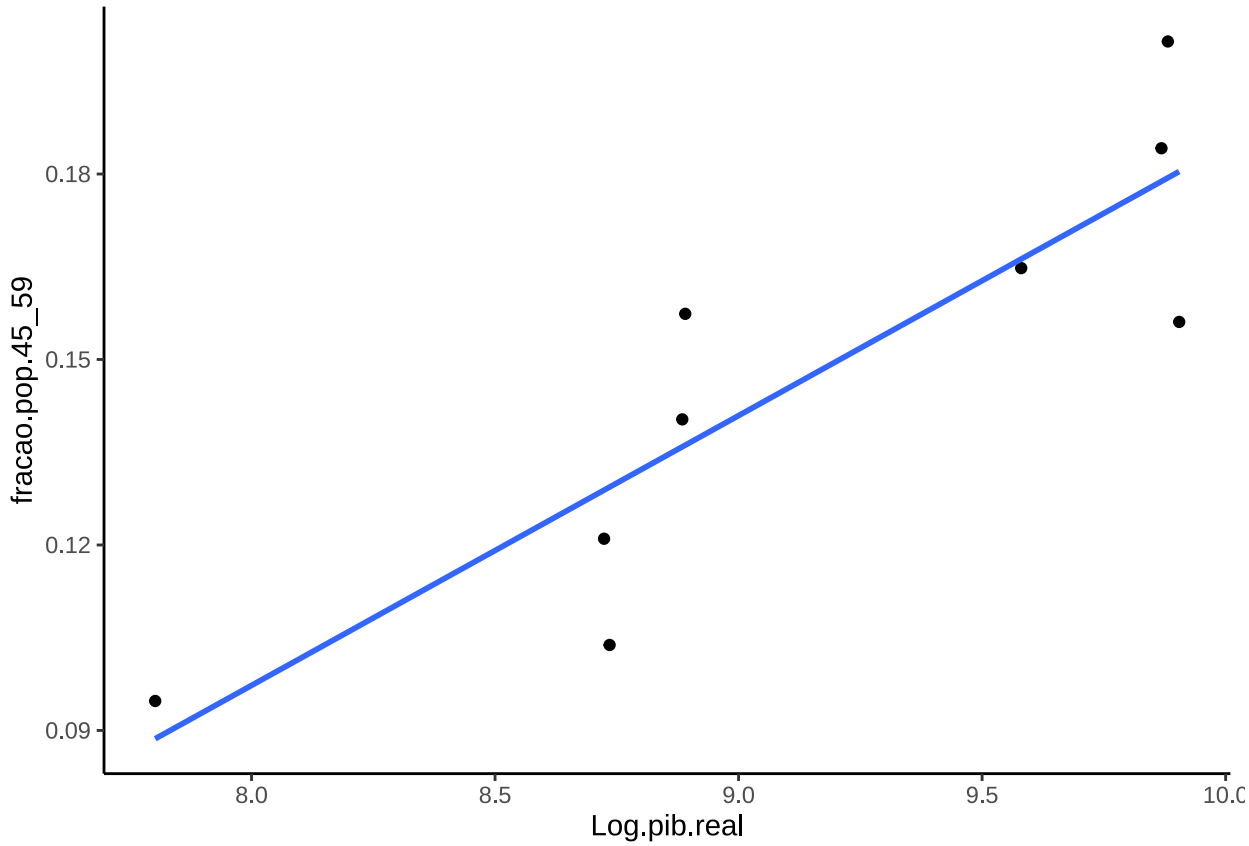
```
income.arq3.1995 %>% dplyr::filter(!is.na(Log.pib.real)) %>% dplyr::select(Log.pib.real, fracao.pop.30_
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



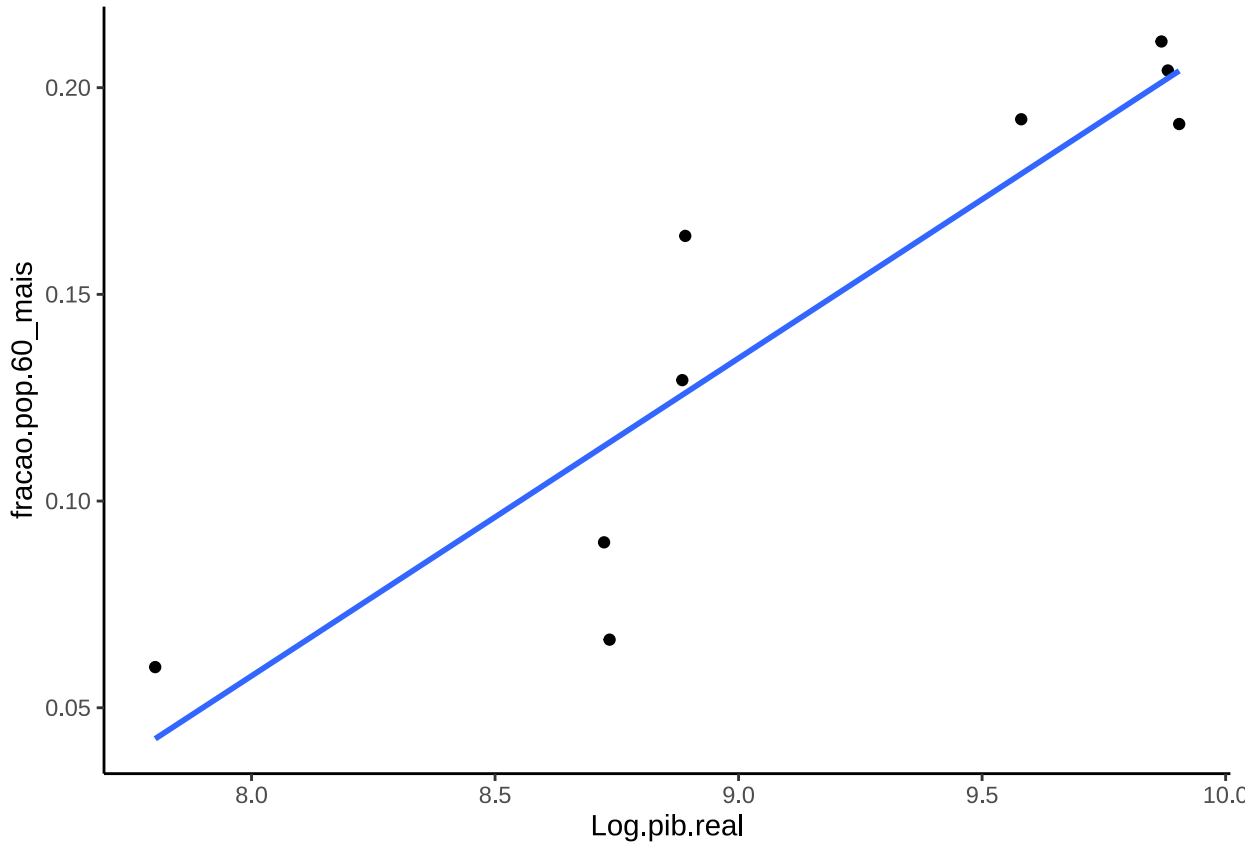

```
income.arq3.1995 %>% dplyr::filter(!is.na(Log.pib.real)) %>% dplyr::select(Log.pib.real, fracao.pop.45_5
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



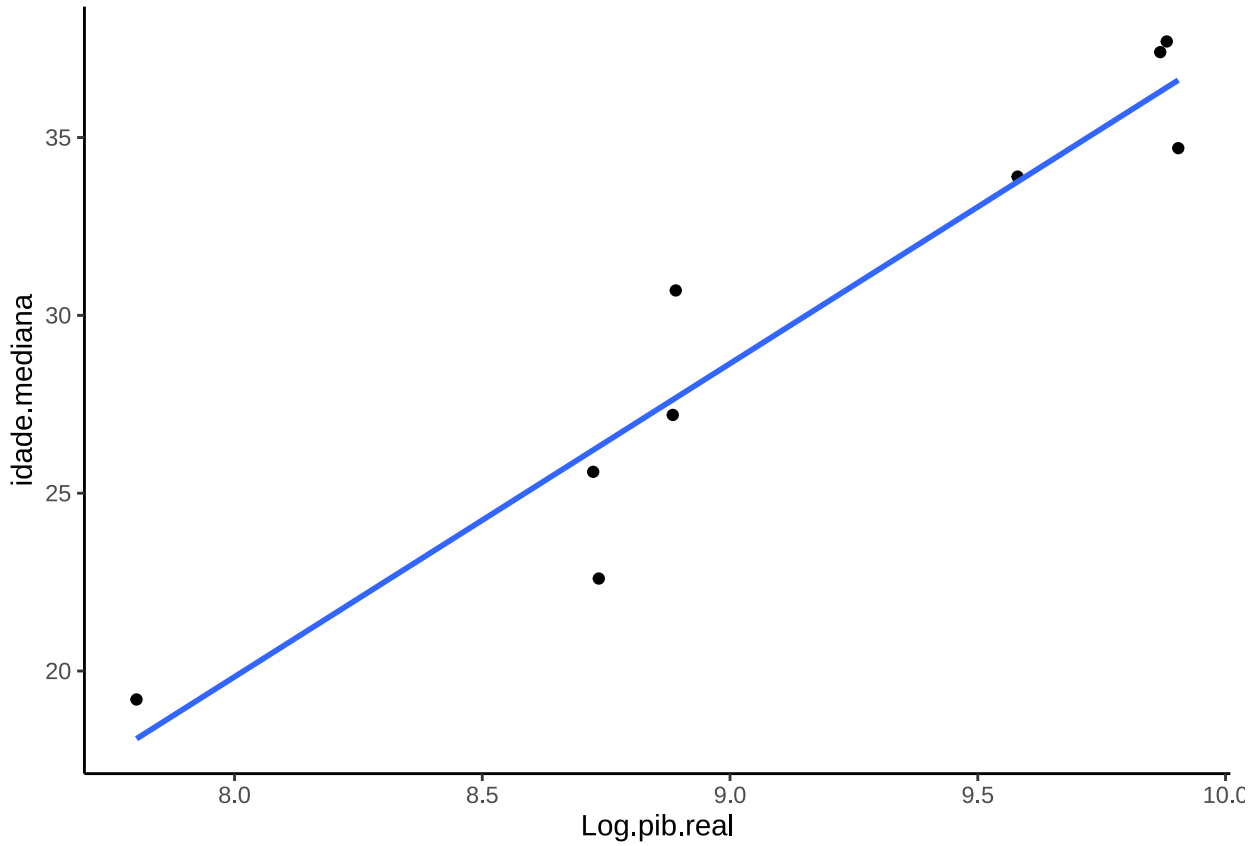

```
income.arq3.1995 %>% dplyr::filter(!is.na(Log.pib.real)) %>% dplyr::select(Log.pib.real, fracao.pop.60_
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



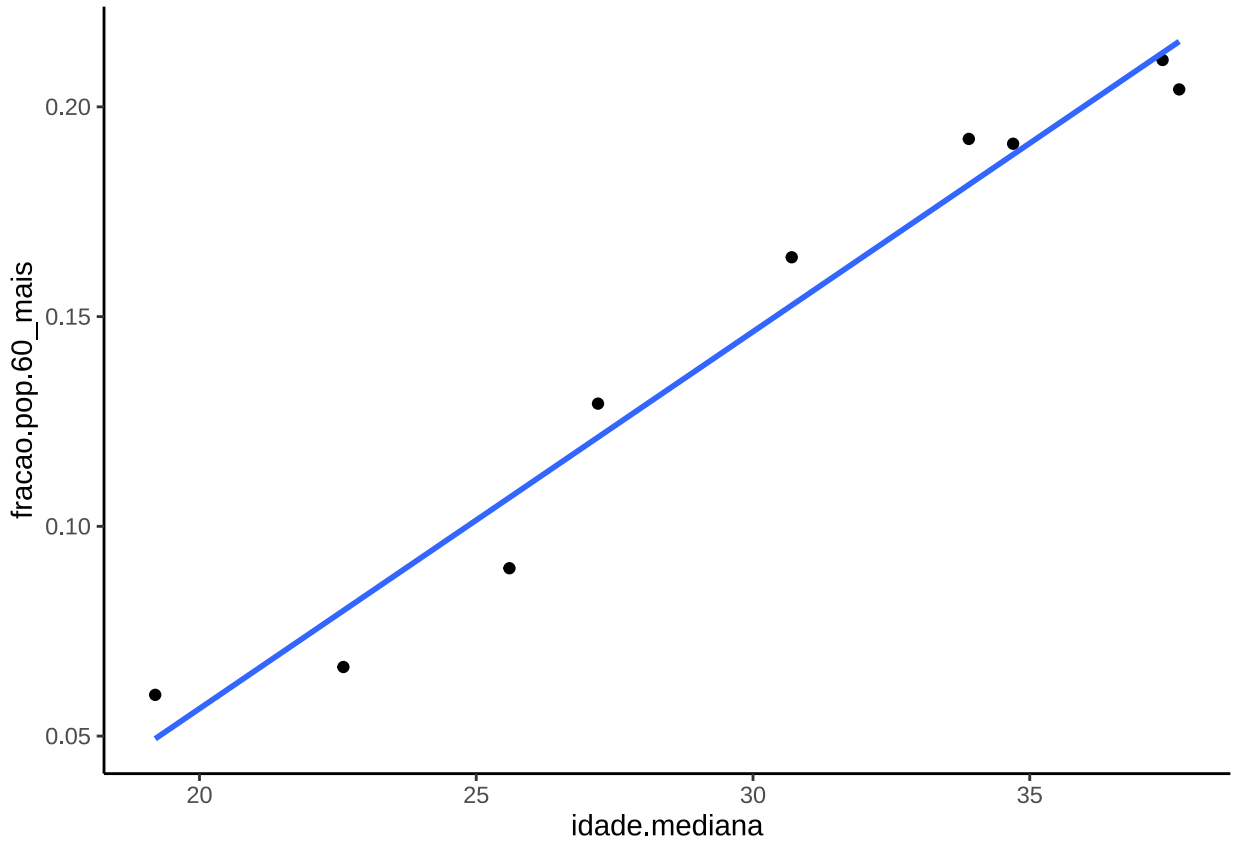

```
income.arq3.1995 %>% dplyr::filter(!is.na(Log.pib.real)) %>% dplyr::select(Log.pib.real, idade.mediana)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
income.arq3.1995 %>% dplyr::filter(!is.na(idade.mediana)) %>% dplyr::select(idade.mediana,fracao.pop.60
```

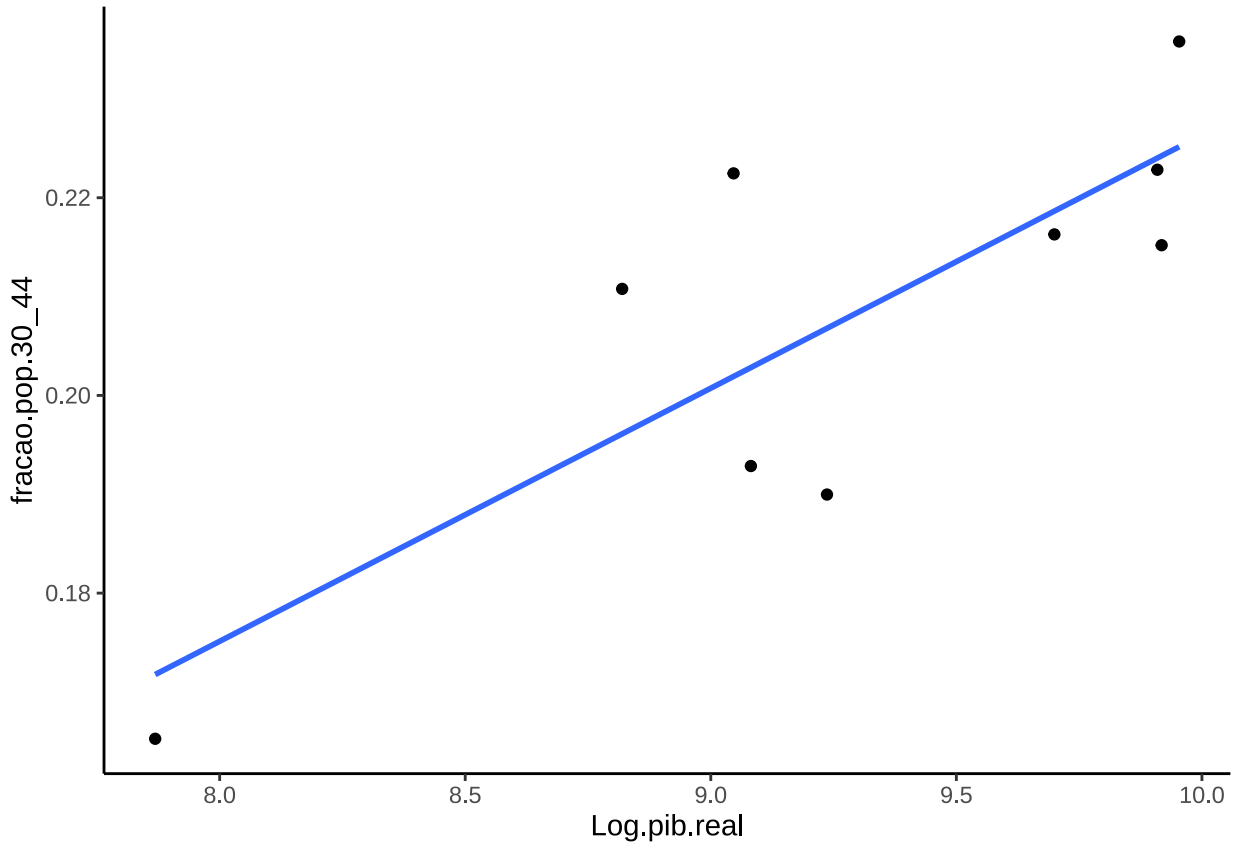
```
## `geom_smooth()` using formula = 'y ~ x'
```



Analisando Scatter - 2000

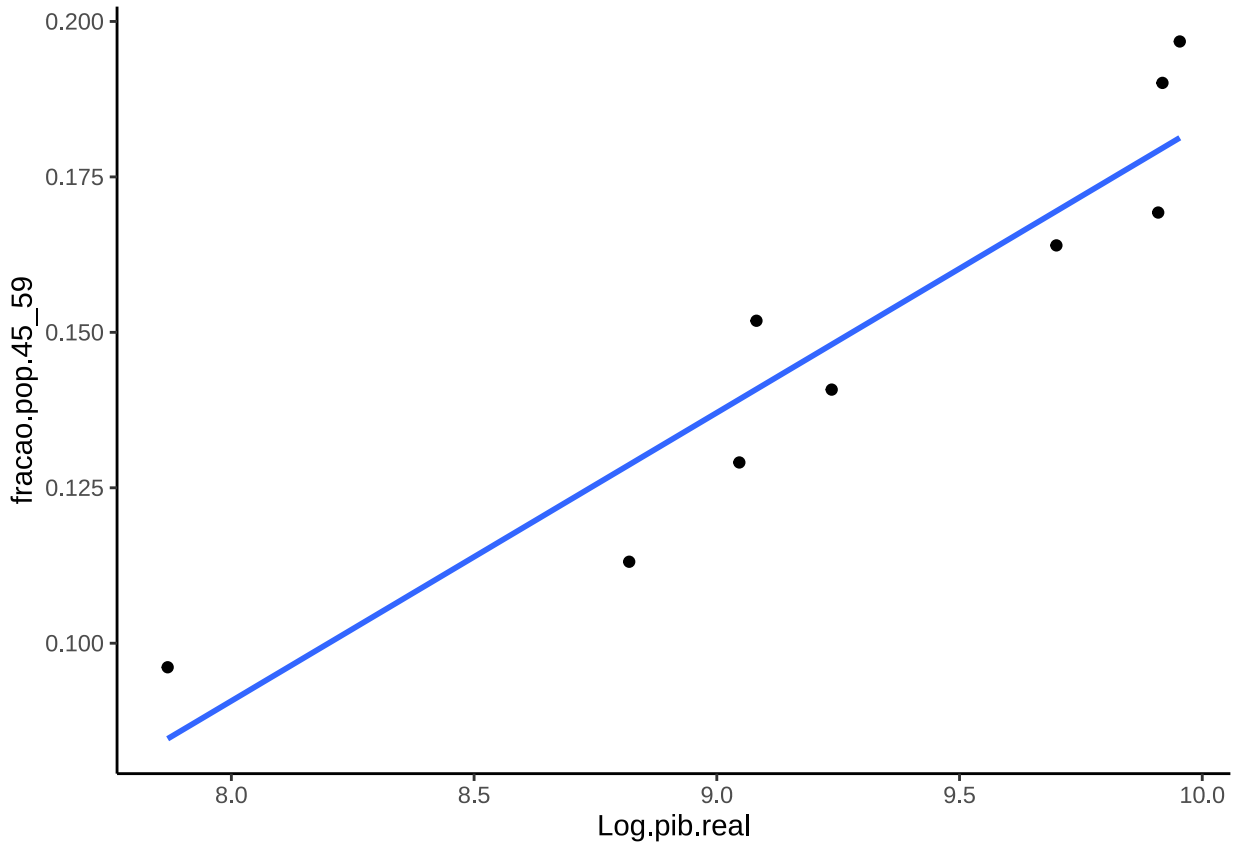
```
income.arq3.2000 %>% dplyr::filter(!is.na(Log.pib.real)) %>% dplyr::select(Log.pib.real, fracao.pop.30_
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



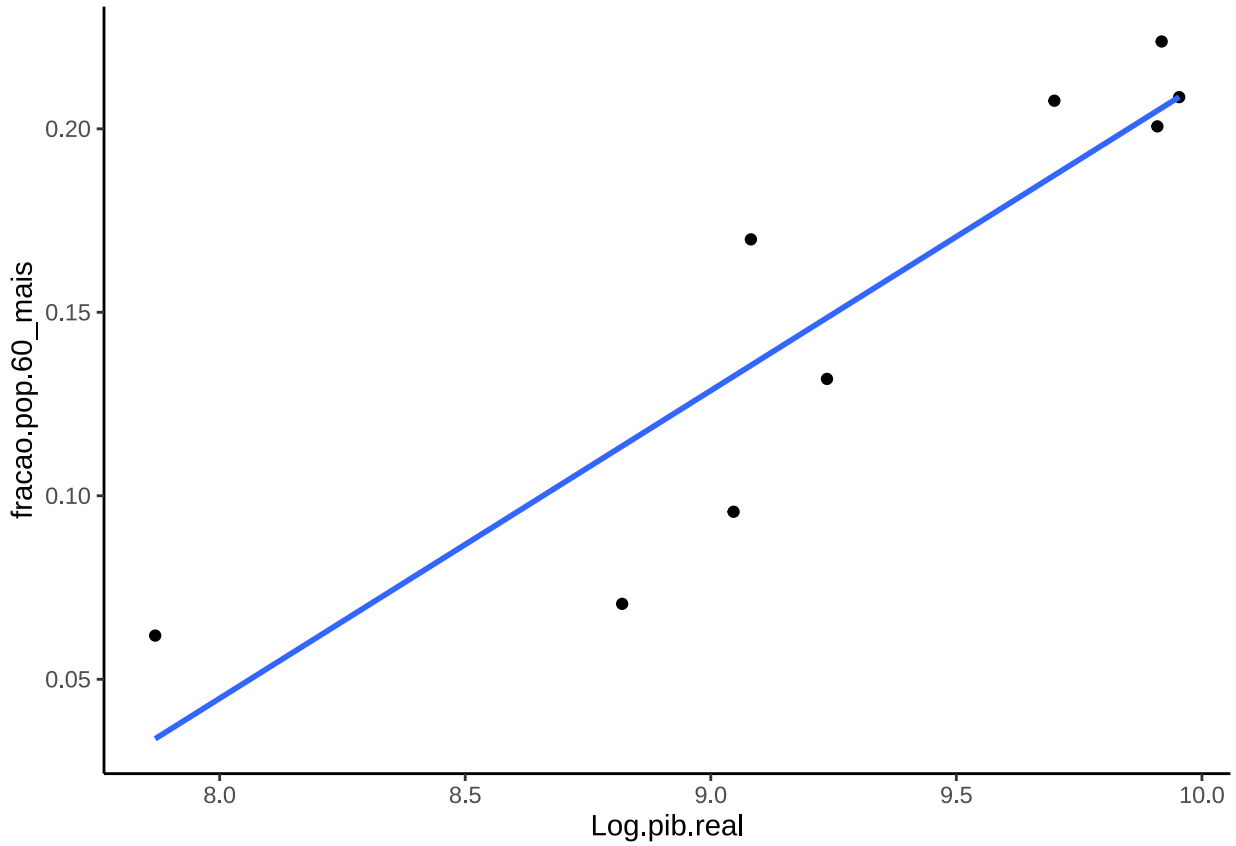

```
income.arq3.2000 %>% dplyr::filter(!is.na(Log.pib.real)) %>% dplyr::select(Log.pib.real, fracao.pop.45_5
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



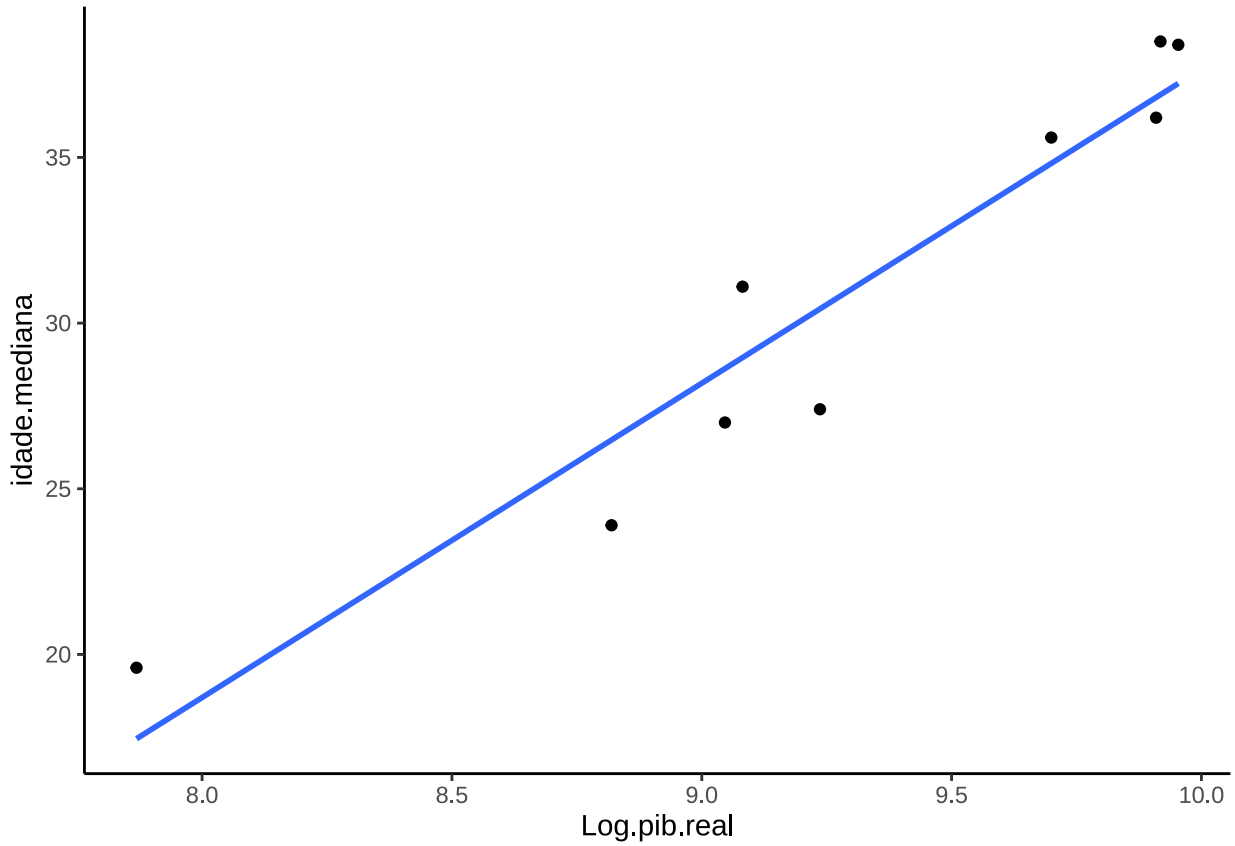

```
income.arq3.2000 %>% dplyr::filter(!is.na(Log.pib.real)) %>% dplyr::select(Log.pib.real, fracao.pop.60_
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



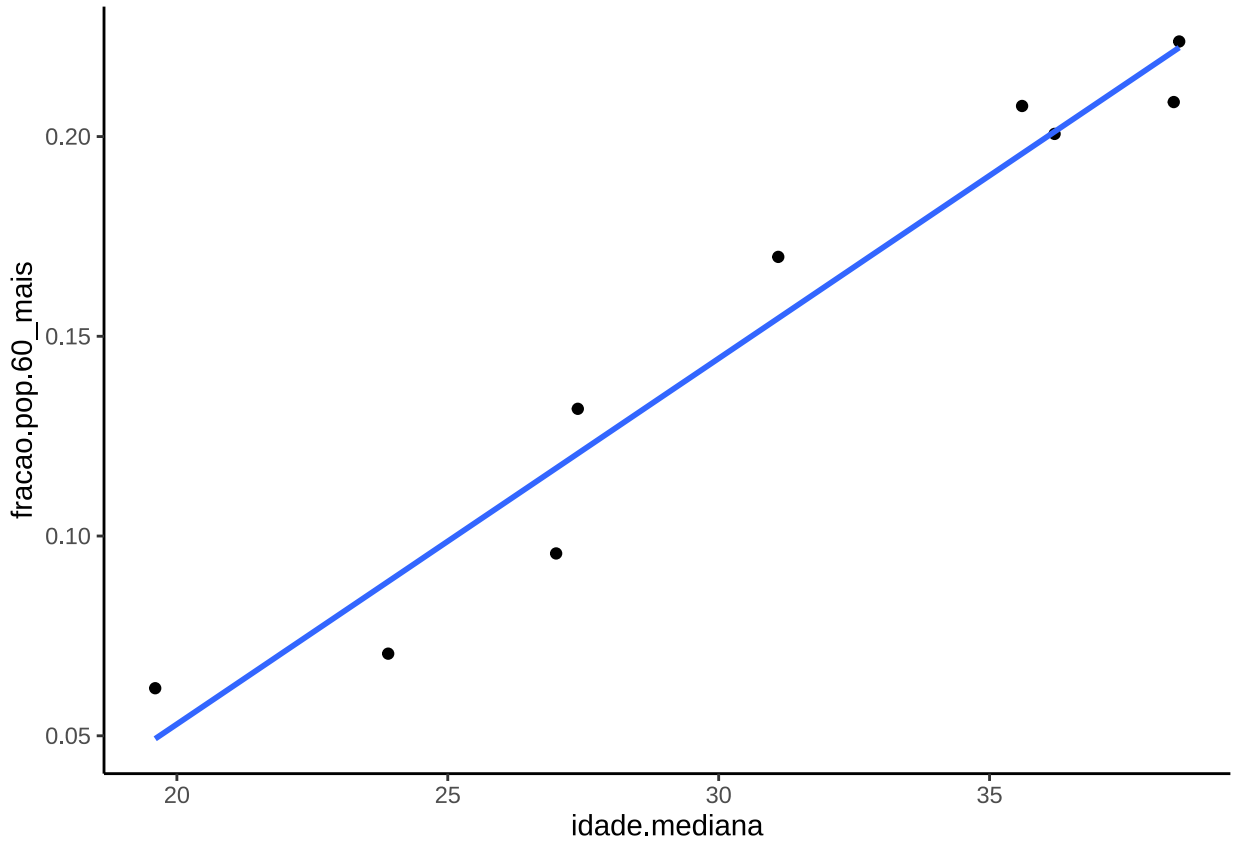

```
income.arq3.2000 %>% dplyr::filter(!is.na(Log.pib.real)) %>% dplyr::select(Log.pib.real, idade.mediana)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
income.arq3.2000 %>% dplyr::filter(!is.na(idade.mediana)) %>% dplyr::select(idade.mediana,fracao.pop.60
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



9.6 Verificando distribuição normal

Deve-se entender como DISTRIBUIÇÃO NORMAL, uma função de probabilidade cujo gráfico descreve uma curva no formato de sino. Também conhecida como GAUSSIANA. Dessa forma, podemos observar uma simetria no formato da curva do gráfico, bem como, valores muito próximos entre MÉDIA e MEDIANA.

9.7 Criando um Histograma

Através do histrograma abaixo, será possível ilustrar a distribuição do evento amostral considerado neste trabalho, observando a concentração de dados, bem como, a eventual presença de outliers.

A escolha do numero de bins foi livre e considerando que os eventos amostrais avaliados, valores de media/media e 1 e 3 quartil.

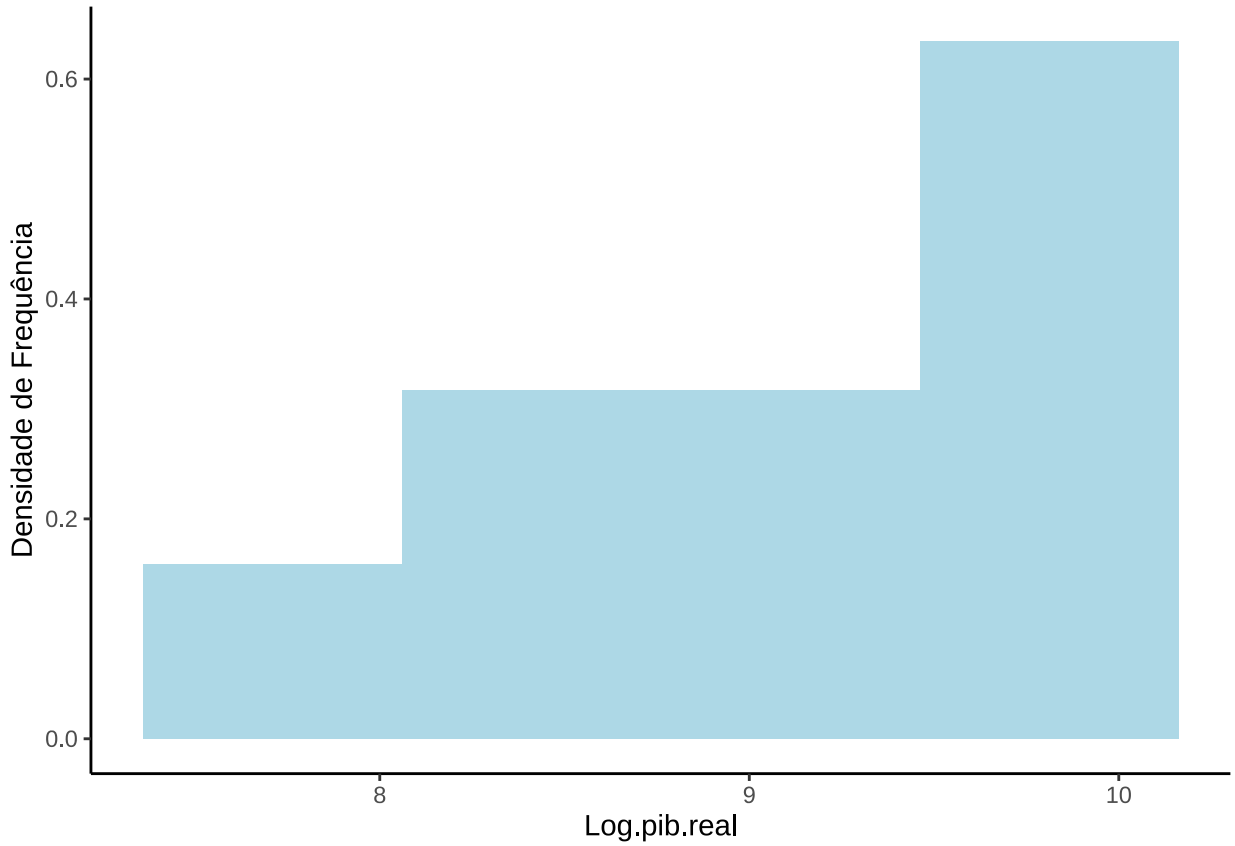
Até o momento tem um simetria moderada, Também no gráfico, não observa-se a presença significativa de outliers, confirmando o resultado das funções summarytools::descr(), executadas acima.

histrograma - ano 1995

```
income.arq3.1995 %>% dplyr::filter(!is.na(Log.pib.real)) %>% dplyr::select(Log.pib.real) %>% ggplot(aes
```

```
## Warning: Removed 2 rows containing missing values (`geom_vline()`).
```

```
## Warning: Removed 1 rows containing missing values (`geom_text()`).
## Removed 1 rows containing missing values (`geom_text()`).
```

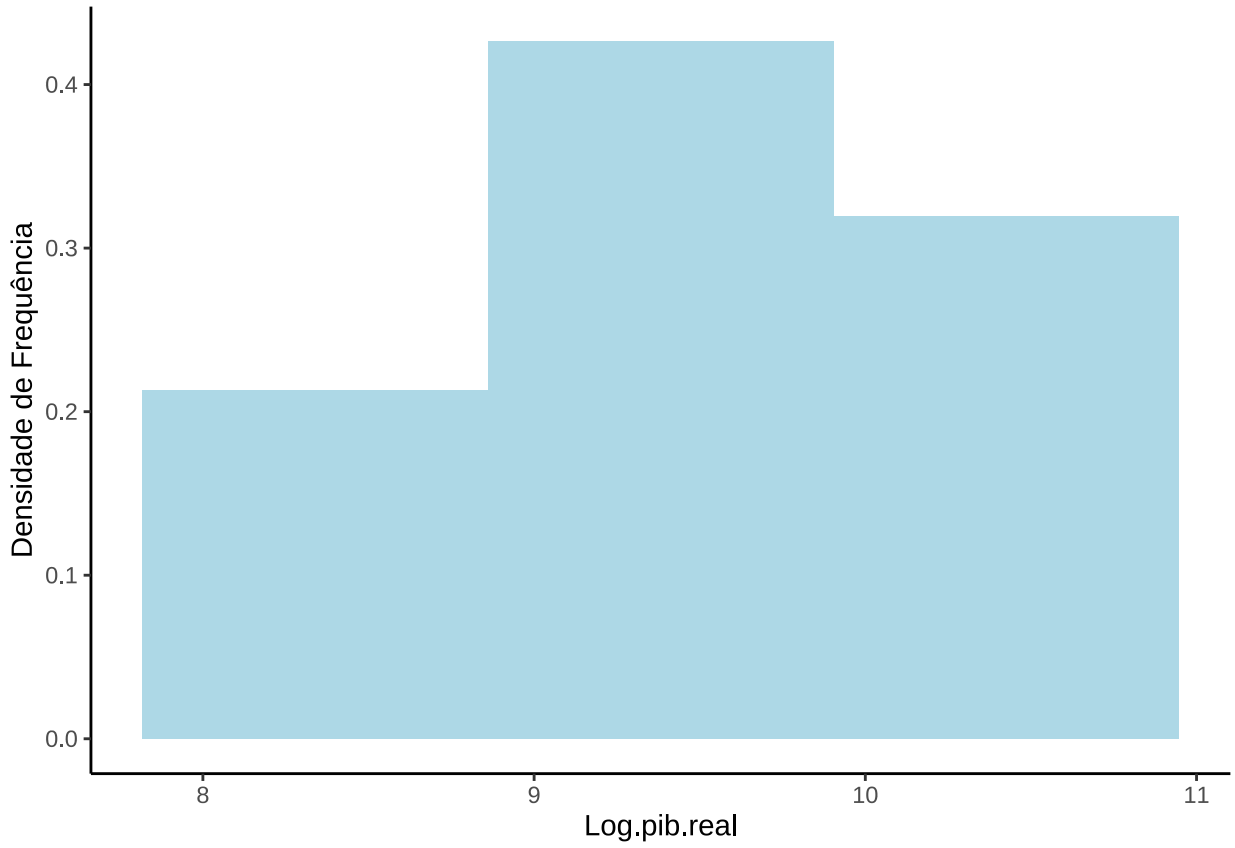


histograma - ano 2000

```
income.arq3.2000 %>% dplyr::filter(!is.na(Log.pib.real)) %>% dplyr::select(Log.pib.real) %>% ggplot(aes
```

```
## Warning: Removed 2 rows containing missing values (`geom_vline()`).
```

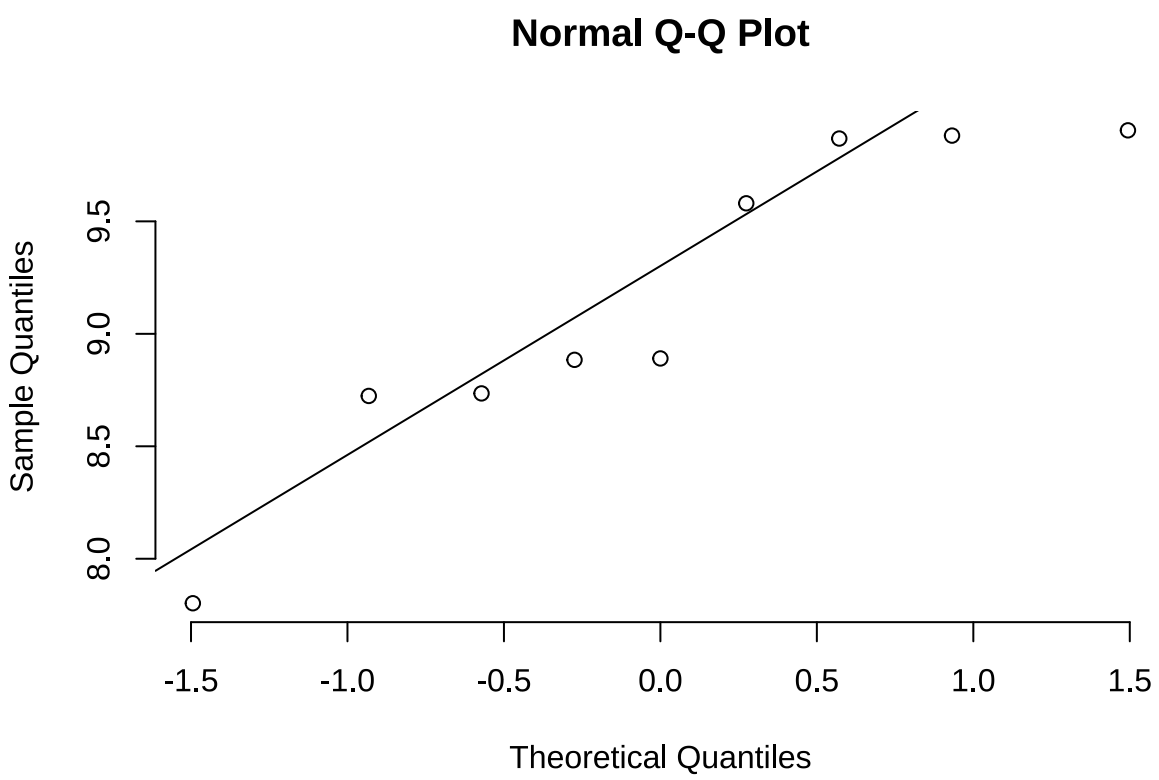
```
## Warning: Removed 1 rows containing missing values (`geom_text()`).
## Removed 1 rows containing missing values (`geom_text()`).
```



9.8 Criando um grafico Q-Q (qqplot)

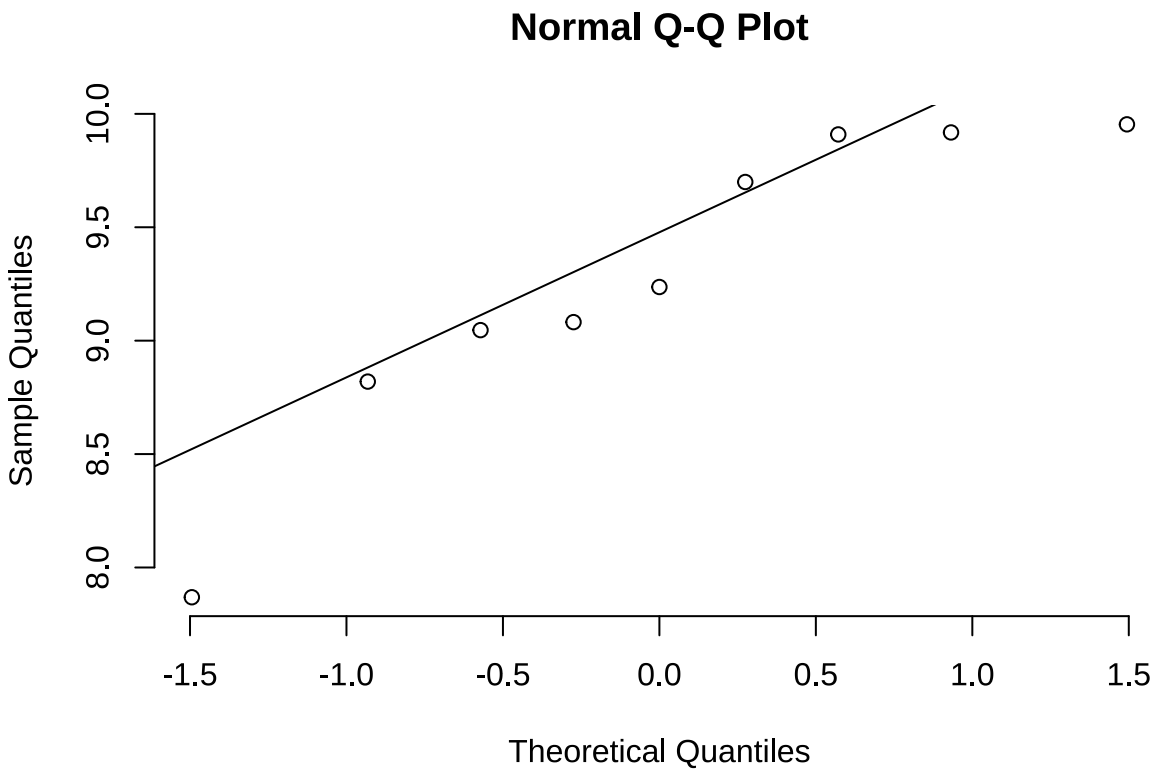
QQ-plot - 1995

```
qqnorm ( income.arq3.1995$Log.pib.real, pch = 1 , frame = FALSE )
qqline (income.arq3.1995$Log.pib.real)
```



QQ-plot - 2000

```
qqnorm ( income.arq3.2000$Log.pib.real, pch = 1 , frame = FALSE )
qqline (income.arq3.2000$Log.pib.real)
```



9.9 Executando teste Shapiro-Wilk

Shapiro - 1995

```
##
##  Shapiro-Wilk normality test
##
## data:  income.arq3.1995$Log.pib.real
## W = 0.88117, p-value = 0.1615
```

Shapiro - 2000

```
##
##  Shapiro-Wilk normality test
##
## data:  income.arq3.2000$Log.pib.real
## W = 0.88097, p-value = 0.1608
```

9.10 Conclusão

Os dados avaliados tanto para o ano de 1995 quanto para ano 2000, considerando paises da Europa e Ameri
se a normalidade dos dados, observou-se na media e medianas com valores muito próximos, sem a presença d

Também na análise dos graficos, seja de histrograma observa-se uma distribuição concentrada dos dados, c

9.11 Completude

A qualidade de dados é um quesito de extrema relevancia nas ações de negócios. Desta forma a coleta, organização, tratamento dos dados, é um processo delicado que requer especialização dos profissionais envolvidos.

A completude dos dados, esta relacionada ao devido preenchimento de dados nas variáveis, evitando que a ausência de informações que levem a conclusões incorretas nas respectivas análises.

Vamor utilizar a função freq do summaytools para determinar a completude das variáveis do banco:

- Variavel pais.idx possui uma completude de 100.000 %
- Variavel pais possui uma completude de 100.000 %
- Variavel ano possui uma completude de 100.000 %
- Variavel log.pib.real possui uma completude de 90.00 %
- Variavel log.populacao possui uma completude de 90.00%
- Variavel educ.adultos possui uma completude de 90.00%
- Variavel fracao.pop.0_14 possui uma completude de 90.00%
- Variavel fracao.pop.15_19 possui uma completude de 90.00%
- Variavel fracao.pop.30_44 possui uma completude de 90.00%
- Variavel fracao.pop. 45_59 possui uma completude de 90.00%
- Variavel fracao.pop. 60_mais possui uma completude de 90.00%
- Variavel idade.mediana possui uma completude de 90.00%

```
income.arq3 %>%  dplyr::select(pais.idx) %>% summarytools::freq()
```

```
## Frequencies
## income.arq3$pais.idx
## Type: Numeric
##
##               Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##           6      2     10.00          10.00     10.00          10.00
##          25      2     10.00          20.00     10.00          20.00
##          26      2     10.00          30.00     10.00          30.00
##          34      2     10.00          40.00     10.00          40.00
##          47      2     10.00          50.00     10.00          50.00
##          49      2     10.00          60.00     10.00          60.00
##          58      2     10.00          70.00     10.00          70.00
##          65      2     10.00          80.00     10.00          80.00
##          90      2     10.00          90.00     10.00          90.00
##         191      2     10.00         100.00     10.00         100.00
##        <NA>      0                               0.00         100.00
##       Total     20    100.00         100.00    100.00         100.00
```

```
income.arq3 %>%  dplyr::select(pais) %>% summarytools::freq()
```

```
## Frequencies
## income.arq3$pais
## Type: Character
##
##                      Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ------------------ ------ --------- -------------- --------- --------------
##          Argentina      2     10.00          10.00     10.00          10.00
##            Bolivia      2     10.00          20.00     10.00          20.00
##             Brazil      2     10.00          30.00     10.00          30.00
##              Chile      2     10.00          40.00     10.00          40.00
```

```
##              France      2     10.00          50.00     10.00          50.00
##             Germany      2     10.00          60.00     10.00          60.00
##       Germany, West      2     10.00          70.00     10.00          70.00
##               Italy      2     10.00          80.00     10.00          80.00
##               Spain      2     10.00          90.00     10.00          90.00
##             Uruguay      2     10.00         100.00     10.00         100.00
##                <NA>      0                                0.00         100.00
##               Total     20    100.00         100.00    100.00         100.00
```

```
income.arq3 %>%  dplyr::select(ano) %>% summarytools::freq()
```

```
## Frequencies
## income.arq3$ano
## Type: Numeric
##
##               Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##        1995     10     50.00          50.00     50.00          50.00
##        2000     10     50.00         100.00     50.00         100.00
##        <NA>      0                                0.00         100.00
##       Total     20    100.00         100.00    100.00         100.00
```

```
income.arq3 %>%  dplyr::select(Log.pib.real) %>% summarytools::freq()
```

```
## Frequencies
## income.arq3$Log.pib.real
## Type: Numeric
##
##                         Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## --------------------- ------ --------- -------------- --------- --------------
##       7.80220317840576      1      5.56           5.56      5.00           5.00
##       7.86878204345703      1      5.56          11.11      5.00          10.00
##       8.72384643554688      1      5.56          16.67      5.00          15.00
##       8.73516082763672      1      5.56          22.22      5.00          20.00
##       8.81952476501465      1      5.56          27.78      5.00          25.00
##       8.88442897796631      1      5.56          33.33      5.00          30.00
##       8.89051628112793      1      5.56          38.89      5.00          35.00
##       9.04640674591064      1      5.56          44.44      5.00          40.00
##       9.08168506622314      1      5.56          50.00      5.00          45.00
##       9.23660945892334      1      5.56          55.56      5.00          50.00
##       9.58032989501953      1      5.56          61.11      5.00          55.00
##       9.69955444335938      1      5.56          66.67      5.00          60.00
##       9.86826324462891      1      5.56          72.22      5.00          65.00
##       9.88137531280518      1      5.56          77.78      5.00          70.00
##       9.90465259552002      1      5.56          83.33      5.00          75.00
##       9.90925598144531      1      5.56          88.89      5.00          80.00
##       9.91801357269287      1      5.56          94.44      5.00          85.00
##       9.95369815826416      1      5.56         100.00      5.00          90.00
##                   <NA>      2                                10.00         100.00
##                  Total     20    100.00         100.00    100.00         100.00
```

```
income.arq3 %>%  dplyr::select(Log.populacao) %>% summarytools::freq()
```

```
## Frequencies
## income.arq3$Log.populacao
## Type: Numeric
```

```
##
##                         Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ---------------------- ------ --------- -------------- --------- --------------
##          8.041090965271      1      5.56           5.56      5.00           5.00
##       8.07651519775391      1      5.56          11.11      5.00          10.00
##       8.79072570800781      1      5.56          16.67      5.00          15.00
##       8.91112518310547      1      5.56          22.22      5.00          20.00
##       9.48029136657715      1      5.56          27.78      5.00          25.00
##       9.56172943115234      1      5.56          33.33      5.00          30.00
##       10.3898296356201      1      5.56          38.89      5.00          35.00
##       10.4564504623413      1      5.56          44.44      5.00          40.00
##        10.567099571228      1      5.56          50.00      5.00          45.00
##       10.5766897201538      1      5.56          55.56      5.00          50.00
##       10.9458599090576      1      5.56          61.11      5.00          55.00
##       10.9461498260498      1      5.56          66.67      5.00          60.00
##       10.9543800354004      1      5.56          72.22      5.00          65.00
##       10.9654998779297      1      5.56          77.78      5.00          70.00
##       11.2826700210571      1      5.56          83.33      5.00          75.00
##       11.3100996017456      1      5.56          88.89      5.00          80.00
##       11.9046802520752      1      5.56          94.44      5.00          85.00
##       11.9796800613403      1      5.56         100.00      5.00          90.00
##                   <NA>      2                                10.00         100.00
##                  Total     20    100.00         100.00    100.00         100.00
```

```
income.arq3 %>%  dplyr::select(educ.adultos) %>% summarytools::freq()
```

```
## Frequencies
## income.arq3$educ.adultos
## Type: Numeric
##
##                         Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ---------------------- ------ --------- -------------- --------- --------------
##       3.76200008392334      1      5.56           5.56      5.00           5.00
##       4.17500019073486      1      5.56          11.11      5.00          10.00
##       4.73799991607666      1      5.56          16.67      5.00          15.00
##       5.18300008773804      1      5.56          22.22      5.00          20.00
##       6.08500003814697      1      5.56          27.78      5.00          25.00
##       6.16200017929077      1      5.56          33.33      5.00          30.00
##       6.59999990463257      1      5.56          38.89      5.00          35.00
##       6.61600017547607      1      5.56          44.44      5.00          40.00
##       6.68699979782104      1      5.56          50.00      5.00          45.00
##        6.8769998550415      1      5.56          55.56      5.00          50.00
##       7.13800001144409      1      5.56          61.11      5.00          55.00
##        7.5310001373291      1      5.56          66.67      5.00          60.00
##       7.55600023269653      1      5.56          72.22      5.00          65.00
##       7.76700019836426      1      5.56          77.78      5.00          70.00
##       7.94399976730347      1      5.56          83.33      5.00          75.00
##       8.11900043487549      1      5.56          88.89      5.00          80.00
##       9.05900001525879      1      5.56          94.44      5.00          85.00
##       9.57100009918213      1      5.56         100.00      5.00          90.00
##                   <NA>      2                                10.00         100.00
##                  Total     20    100.00         100.00    100.00         100.00
```

```
income.arq3 %>%  dplyr::select(fracao.pop.0_14) %>% summarytools::freq()
```

```
## Frequencies
## income.arq3$fracao.pop.0_14
## Type: Numeric
##
##                         Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ---------------------- ------ --------- -------------- --------- --------------
##       0.149543598294258      1      5.56           5.56      5.00           5.00
##       0.158656597137451      1      5.56          11.11      5.00          10.00
##       0.160827398300171      1      5.56          16.67      5.00          15.00
##       0.162430793046951      1      5.56          22.22      5.00          20.00
##       0.162431597709656      1      5.56          27.78      5.00          25.00
##       0.193781495094299      1      5.56          33.33      5.00          30.00
##       0.195207998156548      1      5.56          38.89      5.00          35.00
##       0.202467605471611      1      5.56          44.44      5.00          40.00
##       0.250310599803925      1      5.56          50.00      5.00          45.00
##       0.260379791259766      1      5.56          55.56      5.00          50.00
##       0.289001405239105      1      5.56          61.11      5.00          55.00
##       0.294531613588333      1      5.56          66.67      5.00          60.00
##       0.300656586885452      1      5.56          72.22      5.00          65.00
##       0.306216597557068      1      5.56          77.78      5.00          70.00
##       0.324054598808289      1      5.56          83.33      5.00          75.00
##       0.351667910814285      1      5.56          88.89      5.00          80.00
##        0.40580290555954      1      5.56          94.44      5.00          85.00
##       0.409356713294983      1      5.56         100.00      5.00          90.00
##                   <NA>      2                                10.00         100.00
##                  Total     20    100.00         100.00    100.00         100.00
```

```
income.arq3 %>%  dplyr::select(fracao.pop.15_19) %>% summarytools::freq()
```

```
## Frequencies
## income.arq3$fracao.pop.15_19
## Type: Numeric
##
##                         Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ---------------------- ------ --------- -------------- --------- --------------
##       0.196340903639793      1      5.56           5.56      5.00           5.00
##       0.212029993534088      1      5.56          11.11      5.00          10.00
##       0.221339896321297      1      5.56          16.67      5.00          15.00
##       0.223949700593948      1      5.56          22.22      5.00          20.00
##       0.226139098405838      1      5.56          27.78      5.00          25.00
##       0.229803696274757      1      5.56          33.33      5.00          30.00
##        0.23390519618988      1      5.56          38.89      5.00          35.00
##       0.235093206167221      1      5.56          44.44      5.00          40.00
##       0.237376600503922      1      5.56          50.00      5.00          45.00
##       0.248389303684235      1      5.56          55.56      5.00          50.00
##       0.249071300029755      1      5.56          61.11      5.00          55.00
##       0.249643206596375      1      5.56          66.67      5.00          60.00
##       0.258287012577057      1      5.56          72.22      5.00          65.00
##       0.270891785621643      1      5.56          77.78      5.00          70.00
##       0.271704912185669      1      5.56          83.33      5.00          75.00
##       0.281506389379501      1      5.56          88.89      5.00          80.00
##       0.282485902309418      1      5.56          94.44      5.00          85.00
```

```
##       0.284702509641647      1      5.56         100.00      5.00          90.00
##                    <NA>      2                                10.00         100.00
##                   Total     20    100.00         100.00    100.00         100.00
```

```
income.arq3 %>%  dplyr::select(fracao.pop.30_44) %>% summarytools::freq()
```

```
## Frequencies
## income.arq3$fracao.pop.30_44
## Type: Numeric
##
##                          Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------------------- ------ --------- -------------- --------- --------------
##       0.164342492818832      1      5.56           5.56      5.00           5.00
##       0.165262699127197      1      5.56          11.11      5.00          10.00
##       0.188284501433372      1      5.56          16.67      5.00          15.00
##       0.189973503351212      1      5.56          22.22      5.00          20.00
##       0.190309301018715      1      5.56          27.78      5.00          25.00
##       0.192857101559639      1      5.56          33.33      5.00          30.00
##       0.193336397409439      1      5.56          38.89      5.00          35.00
##       0.200015306472778      1      5.56          44.44      5.00          40.00
##       0.205833002924919      1      5.56          50.00      5.00          45.00
##       0.208621293306351      1      5.56          55.56      5.00          50.00
##       0.209623202681541      1      5.56          61.11      5.00          55.00
##       0.210779398679733      1      5.56          66.67      5.00          60.00
##       0.215196892619133      1      5.56          72.22      5.00          65.00
##       0.216290697455406      1      5.56          77.78      5.00          70.00
##       0.222464606165886      1      5.56          83.33      5.00          75.00
##       0.222831904888153      1      5.56          88.89      5.00          80.00
##       0.224112093448639      1      5.56          94.44      5.00          85.00
##       0.235809907317162      1      5.56         100.00      5.00          90.00
##                    <NA>      2                                10.00         100.00
##                   Total     20    100.00         100.00    100.00         100.00
```

```
income.arq3 %>%  dplyr::select(fracao.pop.45_59) %>% summarytools::freq()
```

```
## Frequencies
## income.arq3$fracao.pop.45_59
## Type: Numeric
##
##                          Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------------------- ------ --------- -------------- --------- --------------
##        0.094766803085804      1      5.56           5.56      5.00           5.00
##       0.0961358025670052      1      5.56          11.11      5.00          10.00
##        0.103825099766254      1      5.56          16.67      5.00          15.00
##        0.113107100129128      1      5.56          22.22      5.00          20.00
##        0.121010802686214      1      5.56          27.78      5.00          25.00
##        0.129073098301888      1      5.56          33.33      5.00          30.00
##        0.140318498015404      1      5.56          38.89      5.00          35.00
##        0.140790402889252      1      5.56          44.44      5.00          40.00
##        0.151863396167755      1      5.56          50.00      5.00          45.00
##         0.15607650578022      1      5.56          55.56      5.00          50.00
##        0.157386496663094      1      5.56          61.11      5.00          55.00
##        0.163983300328255      1      5.56          66.67      5.00          60.00
##        0.164775297045708      1      5.56          72.22      5.00          65.00
##        0.169269695878029      1      5.56          77.78      5.00          70.00
```

```
##        0.184167802333832      1      5.56          83.33      5.00          75.00
##        0.190118804574013      1      5.56          88.89      5.00          80.00
##        0.196781694889069      1      5.56          94.44      5.00          85.00
##        0.201440200209618      1      5.56         100.00      5.00          90.00
##                     <NA>      2                                10.00         100.00
##                    Total     20    100.00         100.00    100.00         100.00
```

```
income.arq3 %>%  dplyr::select(fracao.pop.60_mais) %>% summarytools::freq()
```

```
## Frequencies
## income.arq3$fracao.pop.60_mais
## Type: Numeric
##
##                          Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ------------------------ ------ --------- -------------- --------- --------------
##       0.0598291009664536      1      5.56           5.56      5.00           5.00
##       0.0619066990911961      1      5.56          11.11      5.00          10.00
##       0.0664682015776634      1      5.56          16.67      5.00          15.00
##       0.0705526024103165      1      5.56          22.22      5.00          20.00
##       0.0900136977434158      1      5.56          27.78      5.00          25.00
##       0.0956436023116112      1      5.56          33.33      5.00          30.00
##          0.1292504966259      1      5.56          38.89      5.00          35.00
##        0.131845399737358      1      5.56          44.44      5.00          40.00
##         0.16414549946785      1      5.56          50.00      5.00          45.00
##        0.169875800609589      1      5.56          55.56      5.00          50.00
##        0.191204696893692      1      5.56          61.11      5.00          55.00
##         0.19235660135746      1      5.56          66.67      5.00          60.00
##         0.20066049695015      1      5.56          72.22      5.00          65.00
##        0.204159498214722      1      5.56          77.78      5.00          70.00
##        0.207652002573013      1      5.56          83.33      5.00          75.00
##        0.208635896444321      1      5.56          88.89      5.00          80.00
##        0.211177706718445      1      5.56          94.44      5.00          85.00
##        0.223800599575043      1      5.56         100.00      5.00          90.00
##                     <NA>      2                                10.00         100.00
##                    Total     20    100.00         100.00    100.00         100.00
```

```
income.arq3 %>%  dplyr::select(idade.mediana) %>% summarytools::freq()
```

```
## Frequencies
## income.arq3$idade.mediana
## Type: Numeric
##
##                        Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ---------------------- ------ --------- -------------- --------- --------------
##       19.2000007629395      1      5.56           5.56      5.00           5.00
##       19.6000003814697      1      5.56          11.11      5.00          10.00
##       22.6000003814697      1      5.56          16.67      5.00          15.00
##       23.8999996185303      1      5.56          22.22      5.00          20.00
##       25.6000003814697      1      5.56          27.78      5.00          25.00
##                     27      1      5.56          33.33      5.00          30.00
##       27.2000007629395      1      5.56          38.89      5.00          35.00
##       27.3999996185303      1      5.56          44.44      5.00          40.00
##       30.7000007629395      1      5.56          50.00      5.00          45.00
##       31.1000003814697      1      5.56          55.56      5.00          50.00
##       33.9000015258789      1      5.56          61.11      5.00          55.00
```

```
##       34.7000007629395      1      5.56          66.67      5.00          60.00
##       35.5999984741211      1      5.56          72.22      5.00          65.00
##       36.2000007629395      1      5.56          77.78      5.00          70.00
##       37.4000015258789      1      5.56          83.33      5.00          75.00
##       37.7000007629395      1      5.56          88.89      5.00          80.00
##       38.4000015258789      1      5.56          94.44      5.00          85.00
##                   38.5      1      5.56         100.00      5.00          90.00
##                   <NA>      2                                10.00         100.00
##                  Total     20    100.00         100.00    100.00         100.00
```

9.12 Imputando dados pelo Mice

```
summary(lm(Log.pib.real ~ idade.mediana, data = income.arq3))
```

```
##
## Call:
## lm(formula = Log.pib.real ~ idade.mediana, data = income.arq3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35359 -0.09472 -0.02241  0.15792  0.32227
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.176255   0.243926   25.32 2.45e-14 ***
## idade.mediana 0.099930   0.007862   12.71 8.90e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2114 on 16 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.9099, Adjusted R-squared:  0.9043
## F-statistic: 161.6 on 1 and 16 DF,  p-value: 8.901e-10
```

```
summary(lm(Log.pib.real ~ fracao.pop.60_mais, data = income.arq3))
```

```
##
## Call:
## lm(formula = Log.pib.real ~ fracao.pop.60_mais, data = income.arq3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49971 -0.12322  0.06247  0.19132  0.40805
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          7.6907     0.1942  39.606  < 2e-16 ***
## fracao.pop.60_mais  10.2167     1.2136   8.418 2.84e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3022 on 16 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.8158, Adjusted R-squared:  0.8043
## F-statistic: 70.87 on 1 and 16 DF,  p-value: 2.845e-07
```

```
summary(lm(Log.pib.real ~ idade.mediana + fracao.pop.60_mais, data = income.arq3))
```

```
##
## Call:
## lm(formula = Log.pib.real ~ idade.mediana + fracao.pop.60_mais,
##     data = income.arq3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25610 -0.14583 -0.03148  0.10571  0.40079
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.17742    0.51797   9.996 5.03e-08 ***
## idade.mediana       0.17156    0.03435   4.995  0.00016 ***
## fracao.pop.60_mais -7.90513    3.70861  -2.132  0.04999 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1913 on 15 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.9308, Adjusted R-squared:  0.9216
## F-statistic: 100.9 on 2 and 15 DF,  p-value: 1.99e-09
```

Considerando as regressões acima, o modelo que utiliza as variaveis log.pib.real, idade.mediana e fracao.pop.60_mais é o mais satisfatório. Dessa forma será feito a imput de dados conforme abaixo, utilizando o pacote mice.

```
imp <- mice(income.arq3 %>% dplyr::mutate(Log.pib.real = Log.pib.real, Log.populacao = Log.populacao, e
```

```
## Warning: Number of logged events: 95
```

```
fit <- with(data = imp, exp = lm(Log.pib.real ~ fracao.pop.60_mais, data = income.arq3))
est <- pool(fit)
```