

Projeto da Pós Graduação, disciplina Estatística para Cientista de Dados.

Antonio Vieira dos Santos Neto - Cpf 077.523.948-82

2023-03-10

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.

Objetivo do Projeto

Demonstrar os conhecimentos adquiridos na disciplina “Estatística para Cientista de Dados”, sendo que , através do uso da Linguagem “R”, serão feitas diversas análises em uma “Base de evolução de registro de ocorrências do Instituto de Segurança Pública do Rio de Janeiro”.

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 “PROJETO_ESTATISTICA.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 : Projeto_Estatistica_1

Importando bilbliotecas 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.
- `instal.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.

- instal.packages (“readxl”) - Este pacote permite a leitura de arquivos excel.
- instal.packages (“summarytools”) - Este pacote permite a análise de dados, como a frequência de uma determinada variável;
- instal.packages (“ggplot2”) - Este pacote permite o desenvolvimento de gráficos;
- instal.packages (“fitdistrplus”) - ** Atencao

```
library(tidyverse)
```

```
## -- 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()
```

```
library(rvest)
```

```
##
## Attaching package: 'rvest'
##
## The following object is masked from 'package:readr':
##
##     guess_encoding
```

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
##
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
##
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
library(robotstxt)
library(knitr)
library(dlookr)
```

```
##
## Attaching package: 'dlookr'
##
## The following object is masked from 'package:tidyr':
##
##     extract
##
## The following object is masked from 'package:base':
##
##     transform
```

```
library(readxl)
library(summarytools)
```

```
##
## Attaching package: 'summarytools'
##
## The following object is masked from 'package:tibble':
##
##     view

library(ggplot2)
library(fitdistrplus)

## Carregando pacotes exigidos: MASS
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
##
## Carregando pacotes exigidos: survival
library(readr)
```

Definindo o diretório de trabalho para o projeto Posteriormente, 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/work"
```

Importando dados A partir da definição do ambiente, o arquivo “BaseDPEvolucaoMensalCisp.csv”, contendo a base de evolução de registro de ocorrências do Instituto de Segurança Pública do Rio de Janeiro, será importado utilizando a biblioteca rvest, para ler o arquivo “CSV”.

```
## Bevolucao.tbl <- readr::read_csv2("BaseDPEvolucaoMensalCisp.csv")
Bevolucao.tbl <- readr::read_csv2("BaseDPEvolucaoMensalCisp.csv", locale=locale(encoding="latin1"))

## i Using "','" as decimal and '."' as grouping mark. Use `read_delim()` for more control.

## Rows: 32245 Columns: 63
## -- Column specification -----
## Delimiter: ";"
## chr (3): mes_ano, munic, Regiao
## dbl (60): CISP, mes, ano, AISP, RISP, mcirc, hom_doloso, lesao_corp_morte, l...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.

kable(head(Bevolucao.tbl))
```

CISP	mes	ano	mes_ano	AISP	RISP	munic	mcirc	Regiao	hom_doloso	lesao_corp_morte
1	1	2003	2003m01	5	1	Rio de Janeiro	3304557	Capital	0	0
4	1	2003	2003m01	5	1	Rio de Janeiro	3304557	Capital	3	0
5	1	2003	2003m01	5	1	Rio de Janeiro	3304557	Capital	3	0
6	1	2003	2003m01	1	1	Rio de Janeiro	3304557	Capital	6	0
7	1	2003	2003m01	1	1	Rio de Janeiro	3304557	Capital	4	0
9	1	2003	2003m01	2	1	Rio de Janeiro	3304557	Capital	1	1

Selecionando dados Uma vez que os dados foram importados, faremos a seleção das colunas importantes para a análise. Dessa forma, serão selecionados os dados referentes as ocorrências relacionadas ao evento de

apreensão de drogas versus roubos.

```
Bevolucao.roubos.tbl <- Bevolucao.tbl %>% dplyr::select(CISP,ano,mes_ano,munic,Regiao,apreensao_drogas,
kable(head(Bevolucao.roubos.tbl))
```

CISP	ano	mes_ano	munic	Regiao	apreensao_drogas	roubo_transeunte	roubo_celular	roubo_re
1	2003	2003m01	Rio de Janeiro	Capital	1	26	32	
4	2003	2003m01	Rio de Janeiro	Capital	35	25	14	
5	2003	2003m01	Rio de Janeiro	Capital	4	26	34	
6	2003	2003m01	Rio de Janeiro	Capital	20	14	20	
7	2003	2003m01	Rio de Janeiro	Capital	3	4	1	
9	2003	2003m01	Rio de Janeiro	Capital	11	18	16	

Selecionando os movimentos referentes aos três últimos anos Com base na coluna ano, selecionar o movimento dos três últimos anos.

```
Bevolucao.roubos.f1.tbl=Bevolucao.roubos.tbl %>% dplyr::filter (ano == '2022' | ano == '2021' | ano == '2020')
kable(head(Bevolucao.roubos.f1.tbl))
```

CISP	ano	mes_ano	munic	Regiao	apreensao_drogas	roubo_transeunte	roubo_celular	roubo_re
1	2020	2020m01	Rio de Janeiro	Capital	2	62	32	
4	2020	2020m01	Rio de Janeiro	Capital	7	59	19	
5	2020	2020m01	Rio de Janeiro	Capital	13	130	36	
6	2020	2020m01	Rio de Janeiro	Capital	3	63	15	
7	2020	2020m01	Rio de Janeiro	Capital	1	27	2	
9	2020	2020m01	Rio de Janeiro	Capital	6	69	27	

Agrupando por regioao, munic, ano e mes/ano. Vamos agrupar os dados por regioao, municipio e ano,com o objetivo de entender o movimento das ocorrencias de roubo no periodo.

```
Bevolucao.roubos.f2.tbl <- Bevolucao.roubos.f1.tbl %>% dplyr::group_by(Regiao, munic, ano, mes_ano) %>%
```

```
## `summarise()` has grouped output by 'Regiao', 'munic', 'ano'. You can override
## using the `.groups` argument.
```

```
kable(head(Bevolucao.roubos.f2.tbl))
```

Regiao	munic	ano	mes_ano	roubo_transeunte_sum	apreensao_drogas_sum	roubo_celu
Baixada Fluminense	Belford Roxo	2020	2020m01	198	17	
Baixada Fluminense	Belford Roxo	2020	2020m02	165	19	
Baixada Fluminense	Belford Roxo	2020	2020m03	87	17	
Baixada Fluminense	Belford Roxo	2020	2020m04	50	11	
Baixada Fluminense	Belford Roxo	2020	2020m05	106	4	
Baixada Fluminense	Belford Roxo	2020	2020m06	112	2	

Convertendo dados da coluna “mes_ano” e retirando caracter com acento. Os dados da coluna “mes_ano” encontram-se no formato “aaaammm” e deverão ser convertidos para o formato “aaaa-mm”.

```
Bevolucao.roubos.f3.tbl <-Bevolucao.roubos.f2.tbl %>% dplyr::mutate(mes_ano = stringr::str_replace_all(mes_ano, "m", "-"))
```

```
kable(head(Bevolucao.roubos.f3.tbl))
```

Regiao	munic	ano	mes_ano	roubo_transeunte_sum	apreensao_drogas_sum	roubo_celu
Baixada Fluminense	Belford Roxo	2020	2020-01	198	17	
Baixada Fluminense	Belford Roxo	2020	2020-02	165	19	
Baixada Fluminense	Belford Roxo	2020	2020-03	87	17	
Baixada Fluminense	Belford Roxo	2020	2020-04	50	11	
Baixada Fluminense	Belford Roxo	2020	2020-05	106	4	
Baixada Fluminense	Belford Roxo	2020	2020-06	112	2	

Iniciando a analise dos dados

Identificando os tipos de variaveis Para identificar os tipos de cada variável na base, vamos utilizar o pacote “dlookr”.

```
Bevolucao.roubos.f3.tbl %>% dlookr::diagnose()
```

```
## # A tibble: 9 x 6
##   variables      types      missing_count missing_percent unique_~1 uniqu~2
##   <chr>         <chr>          <int>          <dbl>         <int>    <dbl>
## 1 Regiao       character        0            0            4 0.00136
## 2 munic        character        0            0           82 0.0278
## 3 ano          numeric          0            0            3 0.00102
## 4 mes_ano      character        0            0           36 0.0122
## 5 roubo_transeunte_sum numeric          0            0          249 0.0843
## 6 apreensao_drogas_sum numeric          0            0          141 0.0478
## 7 roubo_celular_sum    numeric          0            0          145 0.0491
## 8 roubo_residencia_sum numeric          0            0           34 0.0115
## 9 roubo_rua_sum       numeric          0            0          311 0.105
## # ... with abbreviated variable names 1: unique_count, 2: unique_rate
```

Na tabela “Bevolucao.roubos.f3.tbl” é possível identificar as variáveis do tipo QUALITATIVAS NOMINAIS que são:

- Regiao = Regiao
- munic = Municipio
- ano = ano relacionado ao registro das ocorrencias.
- mes_ano= mes e ano relacionado ao registro das ocorrencias

Na tabela “Bevolucao.roubos.f4.tbl” é possível identificar as variáveis do tipo QUANTITATIVAS DISCRETAS que são:

- roubo_transeunte_sum
- apreensao_drogas_sum

Identificando a frequencia de variaveis Para o próximo passo irei analisar a frequencia das variáveis.

Dessa forma a base foi sumarizada pelas variáveis “Ano(ano)”, “Regiao(Regiao)” e “Municipio(munic)”, sendo que, para cada ano, foi verificada a frequencia da variável “Regiao”.

```
Bevolucao.roubos.f4.tbl <- Bevolucao.roubos.f3.tbl %>% dplyr::group_by(ano, Regiao, munic) %>% dplyr::summarise()
```

```
## Adding missing grouping variables: `munic`
## `summarise()` has grouped output by 'ano', 'Regiao'. You can override using the
## `.groups` argument.
```

```
kable(head(Bevolucao.roubos.f4.tbl))
```

ano	Regiao	munic	roubo_transeunte_sum	apreensao_drogas_sum	roubo_celular_sum
2020	Baixada Fluminense	Belford Roxo	1459	127	51
2020	Baixada Fluminense	Duque de Caxias	3981	330	103
2020	Baixada Fluminense	Guapimirim	47	65	2
2020	Baixada Fluminense	Itaguaí	183	66	6
2020	Baixada Fluminense	Japeri	94	60	3
2020	Baixada Fluminense	Magé	335	159	15

- Frequencia da variavel : Regiao

O que se observa com base na analise de frequencia, é que a maior ocorrencia dos eventos de roubos e apreensao de drogas, em termos de quantidade de municipios, se da na REGIAO DO INTERIOR DO RIO DE JANEIRO, com 79.27% (65 ocorrencias).

```
Bevolucao.roubos.f4.tbl %>% dplyr::group_by(ano) %>% dplyr::select(Regiao) %>% summarytools::freq()
```

```
## Adding missing grouping variables: `ano`
```

```
## Frequencies
```

```
## Bevolucao.roubos.f4.tbl$Regiao
```

```
## Type: Character
```

```
## Group: ano = 2020
```

```
##
```

	Freq	% Valid	% Valid Cum.	% Total	% Total Cum.
Baixada Fluminense	13	15.85	15.85	15.85	15.85
Capital	1	1.22	17.07	1.22	17.07
Grande Niterói	3	3.66	20.73	3.66	20.73
Interior	65	79.27	100.00	79.27	100.00
<NA>	0			0.00	100.00
Total	82	100.00	100.00	100.00	100.00

```
##
```

```
## Group: ano = 2021
```

```
##
```

	Freq	% Valid	% Valid Cum.	% Total	% Total Cum.
Baixada Fluminense	13	15.85	15.85	15.85	15.85
Capital	1	1.22	17.07	1.22	17.07
Grande Niterói	3	3.66	20.73	3.66	20.73
Interior	65	79.27	100.00	79.27	100.00
<NA>	0			0.00	100.00
Total	82	100.00	100.00	100.00	100.00

```
##
```

```
## Group: ano = 2022
```

```
##
```

	Freq	% Valid	% Valid Cum.	% Total	% Total Cum.
Baixada Fluminense	13	15.85	15.85	15.85	15.85
Capital	1	1.22	17.07	1.22	17.07
Grande Niterói	3	3.66	20.73	3.66	20.73
Interior	65	79.27	100.00	79.27	100.00
<NA>	0			0.00	100.00
Total	82	100.00	100.00	100.00	100.00

Fazendo analise descritiva e de histogramas de uma variavel quantitativa discreta Com base nas variaveis qualitativas (Ano e Regiao), foi feita uma análise considerando a centralidade dos dados para a

variavel roubo_transeunte_sum e apreensao_drogas_sum, através da funcao descr do pacote summarytools (descr).

-Variavel roubo_transeunte_sum : A conclusão é que na maioria das regiões existe uma divergência considerável entre os valores de média (Mean) e mediana (Median), o que sugere a presença de outliers.

```
Bevolucao.roubos.f4.tbl %>% dplyr::group_by(ano,Regiao) %>% dplyr::select(roubo_transeunte_sum) %>% summarise(
```

```
## Adding missing grouping variables: `ano`, `Regiao`
```

```
## Descriptive Statistics
```

```
## Bevolucao.roubos.f4.tbl$roubo_transeunte_sum
```

```
## Group: ano = 2020, Regiao = Baixada Fluminense
```

```
## N: 13
```

```
##
##              roubo_transeunte_sum
## -----
##              Mean              973.46
##              Std.Dev           1206.14
##              Min                20.00
##              Q1                112.00
##              Median            335.00
##              Q3               1459.00
##              Max              3981.00
##              MAD               467.02
##              IQR              1347.00
##              CV                 1.24
##              Skewness           1.21
##              SE.Skewness         0.62
##              Kurtosis           0.31
##              N.Valid            13.00
##              Pct.Valid          100.00
##
```

```
## Group: ano = 2020, Regiao = Capital
```

```
## N: 1
```

```
##
##              roubo_transeunte_sum
## -----
##              Mean            25356.00
##              Std.Dev           NA
##              Min            25356.00
##              Q1             25356.00
##              Median          25356.00
##              Q3             25356.00
##              Max            25356.00
##              MAD              0.00
##              IQR              0.00
##              CV              NA
##              Skewness         NA
##              SE.Skewness       0.00
##              Kurtosis         NA
##              N.Valid           1.00
##              Pct.Valid        100.00
##
```

```
## Group: ano = 2020, Regiao = Grande Niterói
```

```
## N: 3
```

```

##
##          roubo_transeunte_sum
## -----
##          Mean          1810.67
##          Std.Dev       1875.82
##          Min           274.00
##          Q1            274.00
##          Median        1257.00
##          Q3            3901.00
##          Max           3901.00
##          MAD            1457.40
##          IQR           1813.50
##          CV             1.04
##          Skewness       0.27
##          SE.Skewness     1.22
##          Kurtosis       -2.33
##          N.Valid        3.00
##          Pct.Valid      100.00
##
## Group: ano = 2020, Regiao = Interior
## N: 65
##
##          roubo_transeunte_sum
## -----
##          Mean          43.82
##          Std.Dev       95.90
##          Min           0.00
##          Q1            1.00
##          Median        5.00
##          Q3            30.00
##          Max           490.00
##          MAD            7.41
##          IQR           29.00
##          CV             2.19
##          Skewness       3.01
##          SE.Skewness     0.30
##          Kurtosis       8.77
##          N.Valid       65.00
##          Pct.Valid     100.00
##
## Group: ano = 2021, Regiao = Baixada Fluminense
## N: 13
##
##          roubo_transeunte_sum
## -----
##          Mean          874.15
##          Std.Dev       1115.75
##          Min           18.00
##          Q1            118.00
##          Median        411.00
##          Q3            1174.00
##          Max           3751.00
##          MAD            493.71
##          IQR           1056.00

```



```

##          CV          1.28
##      Skewness        1.35
##    SE.Skewness        0.62
##      Kurtosis        0.74
##      N.Valid        13.00
##      Pct.Valid       100.00
##
## Group: ano = 2021, Regiao = Capital
## N: 1
##
##          roubo_transeunte_sum
## -----
##          Mean          24004.00
##        Std.Dev          NA
##          Min          24004.00
##          Q1          24004.00
##        Median          24004.00
##          Q3          24004.00
##          Max          24004.00
##          MAD            0.00
##          IQR            0.00
##          CV            NA
##      Skewness          NA
##    SE.Skewness          0.00
##      Kurtosis          NA
##      N.Valid           1.00
##      Pct.Valid       100.00
##
## Group: ano = 2021, Regiao = Grande Niterói
## N: 3
##
##          roubo_transeunte_sum
## -----
##          Mean          1282.67
##        Std.Dev          1210.67
##          Min          259.00
##          Q1          259.00
##        Median          970.00
##          Q3          2619.00
##          Max          2619.00
##          MAD          1054.13
##          IQR          1180.00
##          CV            0.94
##      Skewness          0.24
##    SE.Skewness          1.22
##      Kurtosis         -2.33
##      N.Valid           3.00
##      Pct.Valid       100.00
##
## Group: ano = 2021, Regiao = Interior
## N: 65
##
##          roubo_transeunte_sum
## -----

```

```

##          Mean          40.00
##      Std.Dev       84.13
##          Min         0.00
##          Q1          1.00
##        Median         5.00
##          Q3         30.00
##          Max        423.00
##          MAD          7.41
##          IQR        29.00
##          CV          2.10
##      Skewness         2.93
##    SE.Skewness         0.30
##      Kurtosis         8.46
##      N.Valid        65.00
##      Pct.Valid     100.00
##
## Group: ano = 2022, Regiao = Baixada Fluminense
## N: 13
##
##          roubo_transeunte_sum
## -----
##          Mean          771.54
##      Std.Dev       962.72
##          Min         16.00
##          Q1        116.00
##        Median       343.00
##          Q3       1179.00
##          Max       3203.00
##          MAD       383.99
##          IQR      1063.00
##          CV         1.25
##      Skewness         1.30
##    SE.Skewness         0.62
##      Kurtosis         0.53
##      N.Valid         13.00
##      Pct.Valid      100.00
##
## Group: ano = 2022, Regiao = Capital
## N: 1
##
##          roubo_transeunte_sum
## -----
##          Mean       23378.00
##      Std.Dev          NA
##          Min       23378.00
##          Q1       23378.00
##        Median       23378.00
##          Q3       23378.00
##          Max       23378.00
##          MAD          0.00
##          IQR          0.00
##          CV          NA
##      Skewness          NA
##    SE.Skewness          0.00

```

```
##          Kurtosis          NA
##          N.Valid          1.00
##          Pct.Valid        100.00
##
## Group: ano = 2022, Regiao = Grande Niterói
## N: 3
##
##          roubo_transeunte_sum
## -----
##          Mean          1167.67
##          Std.Dev        1093.05
##          Min           213.00
##          Q1            213.00
##          Median         930.00
##          Q3            2360.00
##          Max            2360.00
##          MAD            1063.02
##          IQR            1073.50
##          CV             0.94
##          Skewness        0.21
##          SE.Skewness      1.22
##          Kurtosis        -2.33
##          N.Valid         3.00
##          Pct.Valid       100.00
##
```

```
## Group: ano = 2022, Regiao = Interior
## N: 65
##
##          roubo_transeunte_sum
## -----
##          Mean          33.34
##          Std.Dev        74.29
##          Min           0.00
##          Q1             1.00
##          Median         4.00
##          Q3            21.00
##          Max            386.00
##          MAD             5.93
##          IQR            20.00
##          CV             2.23
##          Skewness        3.09
##          SE.Skewness      0.30
##          Kurtosis         9.59
##          N.Valid        65.00
##          Pct.Valid       100.00
##
```

-Variável apreensao_drogas_sum : De forma análoga, à variável anterior, observa-se várias divergências na análise dos valores de média e mediana da variável apreensao_drogas_sum, também sugerindo a presença de outliers.

```
Bevolucao.roubos.f4.tbl %>% dplyr::group_by(ano,Regiao) %>% dplyr::select(apreensao_drogas_sum) %>% summarise(
```

```
## Adding missing grouping variables: `ano`, `Regiao`
```

```
## Descriptive Statistics
```

```
## Bevolucao.roubos.f4.tbl$apreensao_drogas_sum
```

```
## Group: ano = 2020, Regiao = Baixada Fluminense
## N: 13
```

```
##
##      apreensao_drogas_sum
## -----
##      Mean      120.77
##      Std.Dev   98.73
##      Min       28.00
##      Q1        60.00
##      Median    79.00
##      Q3       158.00
##      Max      330.00
##      MAD       60.79
##      IQR       98.00
##      CV        0.82
##      Skewness   1.14
##      SE.Skewness 0.62
##      Kurtosis   -0.15
##      N.Valid    13.00
##      Pct.Valid  100.00
```

```
## Group: ano = 2020, Regiao = Capital
## N: 1
```

```
##
##      apreensao_drogas_sum
## -----
##      Mean      4118.00
##      Std.Dev    NA
##      Min       4118.00
##      Q1        4118.00
##      Median    4118.00
##      Q3        4118.00
##      Max       4118.00
##      MAD        0.00
##      IQR        0.00
##      CV         NA
##      Skewness   NA
##      SE.Skewness 0.00
##      Kurtosis   NA
##      N.Valid     1.00
##      Pct.Valid  100.00
```

```
## Group: ano = 2020, Regiao = Grande Niterói
## N: 3
```

```
##
##      apreensao_drogas_sum
## -----
##      Mean      299.00
##      Std.Dev   175.58
##      Min      105.00
##      Q1       105.00
##      Median    345.00
##      Q3       447.00
##      Max      447.00
```

```

##          MAD          151.23
##          IQR          171.00
##          CV           0.59
##      Skewness         -0.24
##      SE.Skewness       1.22
##      Kurtosis         -2.33
##      N.Valid          3.00
##      Pct.Valid        100.00
##
## Group: ano = 2020, Regiao = Interior
## N: 65
##
##          apreensao_drogas_sum
## -----
##          Mean          218.54
##      Std.Dev          231.58
##          Min           5.00
##          Q1           67.00
##      Median          147.00
##          Q3           276.00
##          Max          1116.00
##          MAD           142.33
##          IQR           209.00
##          CV            1.06
##      Skewness          2.03
##      SE.Skewness        0.30
##      Kurtosis           4.41
##      N.Valid           65.00
##      Pct.Valid          100.00
##
## Group: ano = 2021, Regiao = Baixada Fluminense
## N: 13
##
##          apreensao_drogas_sum
## -----
##          Mean          133.00
##      Std.Dev           78.16
##          Min           30.00
##          Q1            82.00
##      Median          145.00
##          Q3           168.00
##          Max           270.00
##          MAD            90.44
##          IQR            86.00
##          CV             0.59
##      Skewness           0.40
##      SE.Skewness         0.62
##      Kurtosis           -1.17
##      N.Valid            13.00
##      Pct.Valid          100.00
##
## Group: ano = 2021, Regiao = Capital
## N: 1
##

```

```

##                                apreensao_drogas_sum
## -----
##              Mean                4408.00
##            Std.Dev                NA
##              Min                4408.00
##              Q1                4408.00
##            Median                4408.00
##              Q3                4408.00
##              Max                4408.00
##              MAD                 0.00
##              IQR                 0.00
##              CV                  NA
##            Skewness                NA
##          SE.Skewness                0.00
##            Kurtosis                NA
##            N.Valid                 1.00
##            Pct.Valid              100.00
##
## Group: ano = 2021, Regiao = Grande Niterói
## N: 3
##
##                                apreensao_drogas_sum
## -----
##              Mean                376.67
##            Std.Dev                229.99
##              Min                113.00
##              Q1                113.00
##            Median                481.00
##              Q3                536.00
##              Max                536.00
##              MAD                 81.54
##              IQR                211.50
##              CV                  0.61
##            Skewness               -0.36
##          SE.Skewness                1.22
##            Kurtosis               -2.33
##            N.Valid                 3.00
##            Pct.Valid              100.00
##
## Group: ano = 2021, Regiao = Interior
## N: 65
##
##                                apreensao_drogas_sum
## -----
##              Mean                221.77
##            Std.Dev                213.60
##              Min                 11.00
##              Q1                 62.00
##            Median                134.00
##              Q3                 318.00
##              Max                 969.00
##              MAD                 139.36
##              IQR                 256.00
##              CV                  0.96

```

```

##          Skewness          1.40
##      SE.Skewness          0.30
##          Kurtosis          1.59
##          N.Valid          65.00
##          Pct.Valid        100.00
##
## Group: ano = 2022, Regiao = Baixada Fluminense
## N: 13
##
##          apreensao_drogas_sum
## -----
##          Mean          125.92
##          Std.Dev        62.21
##          Min           44.00
##          Q1            87.00
##          Median        126.00
##          Q3           175.00
##          Max           259.00
##          MAD            72.65
##          IQR            88.00
##          CV             0.49
##          Skewness       0.45
##      SE.Skewness       0.62
##          Kurtosis      -0.71
##          N.Valid        13.00
##          Pct.Valid      100.00
##
## Group: ano = 2022, Regiao = Capital
## N: 1
##
##          apreensao_drogas_sum
## -----
##          Mean          3843.00
##          Std.Dev         NA
##          Min           3843.00
##          Q1            3843.00
##          Median         3843.00
##          Q3            3843.00
##          Max           3843.00
##          MAD             0.00
##          IQR             0.00
##          CV             NA
##          Skewness       NA
##      SE.Skewness       0.00
##          Kurtosis       NA
##          N.Valid         1.00
##          Pct.Valid      100.00
##
## Group: ano = 2022, Regiao = Grande Niterói
## N: 3
##
##          apreensao_drogas_sum
## -----
##          Mean          315.67

```

```
##          Std.Dev          198.90
##          Min           86.00
##          Q1            86.00
##          Median        430.00
##          Q3            431.00
##          Max           431.00
##          MAD            1.48
##          IQR           172.50
##          CV             0.63
##          Skewness       -0.38
##          SE.Skewness     1.22
##          Kurtosis        -2.33
##          N.Valid         3.00
##          Pct.Valid       100.00
##
## Group: ano = 2022, Regiao = Interior
## N: 65
##
##          apreensao_drogas_sum
## -----
##          Mean           218.68
##          Std.Dev        239.84
##          Min            10.00
##          Q1             62.00
##          Median         116.00
##          Q3             302.00
##          Max            1315.00
##          MAD            126.02
##          IQR            240.00
##          CV              1.10
##          Skewness        2.05
##          SE.Skewness     0.30
##          Kurtosis         5.29
##          N.Valid         65.00
##          Pct.Valid       100.00
```

Análise visual da variável Será realizada através da binarização dos dados. Para a análise, será considerada uma amostra do ano de 2022 e ocorrências da Região interior, onde verificou-se a maior frequência de cidades, com casos de roubo e apreensão de drogas.

Dado que não tenho conhecimento da binarização ideal, será considerado o intervalo interquartil, centralidade dos dados, dispersão, assimetria, para verificar a presença de outliers. Farei uso do pacote summarytools e funcao descr.

```
Bevolucao.roubos.f5.tbl <- Bevolucao.roubos.f4.tbl %>% dplyr::group_by(Regiao,ano,munic) %>% dplyr::filter(ano == 2022 & Regiao == "Interior")
```

```
## `summarise()` has grouped output by 'Regiao', 'ano'. You can override using the
## `.groups` argument.
```

```
kable(head(Bevolucao.roubos.f5.tbl))
```


Regiao	ano	munic	roubo_transeunte_sum	apreensao_drogas_sum	roubo_celular_sum	roubo
Interior	2022	Angra dos Reis	54	261	25	
Interior	2022	Araruama	107	289	71	
Interior	2022	Armação dos Búzios	38	107	8	
Interior	2022	Arraial do Cabo	17	152	7	
Interior	2022	Barra Mansa	63	301	20	
Interior	2022	Barra do Pirai	10	704	2	

```
Bevolucao.roubos.f5.tbl %>% dplyr::select(roubo_transeunte_sum) %>% summarytools::descr()
```

```
## Adding missing grouping variables: `Regiao`, `ano`
```

```
## Descriptive Statistics
```

```
## Bevolucao.roubos.f5.tbl$roubo_transeunte_sum
```

```
## Group: Regiao = Interior, ano = 2022
```

```
## N: 65
```

```
##
```

```
##          roubo_transeunte_sum
```

```
## -----
```

```
##          Mean          33.34
```

```
##          Std.Dev       74.29
```

```
##          Min           0.00
```

```
##          Q1            1.00
```

```
##          Median        4.00
```

```
##          Q3            21.00
```

```
##          Max          386.00
```

```
##          MAD           5.93
```

```
##          IQR           20.00
```

```
##          CV            2.23
```

```
##          Skewness       3.09
```

```
##          SE.Skewness     0.30
```

```
##          Kurtosis        9.59
```

```
##          N.Valid        65.00
```

```
##          Pct.Valid      100.00
```

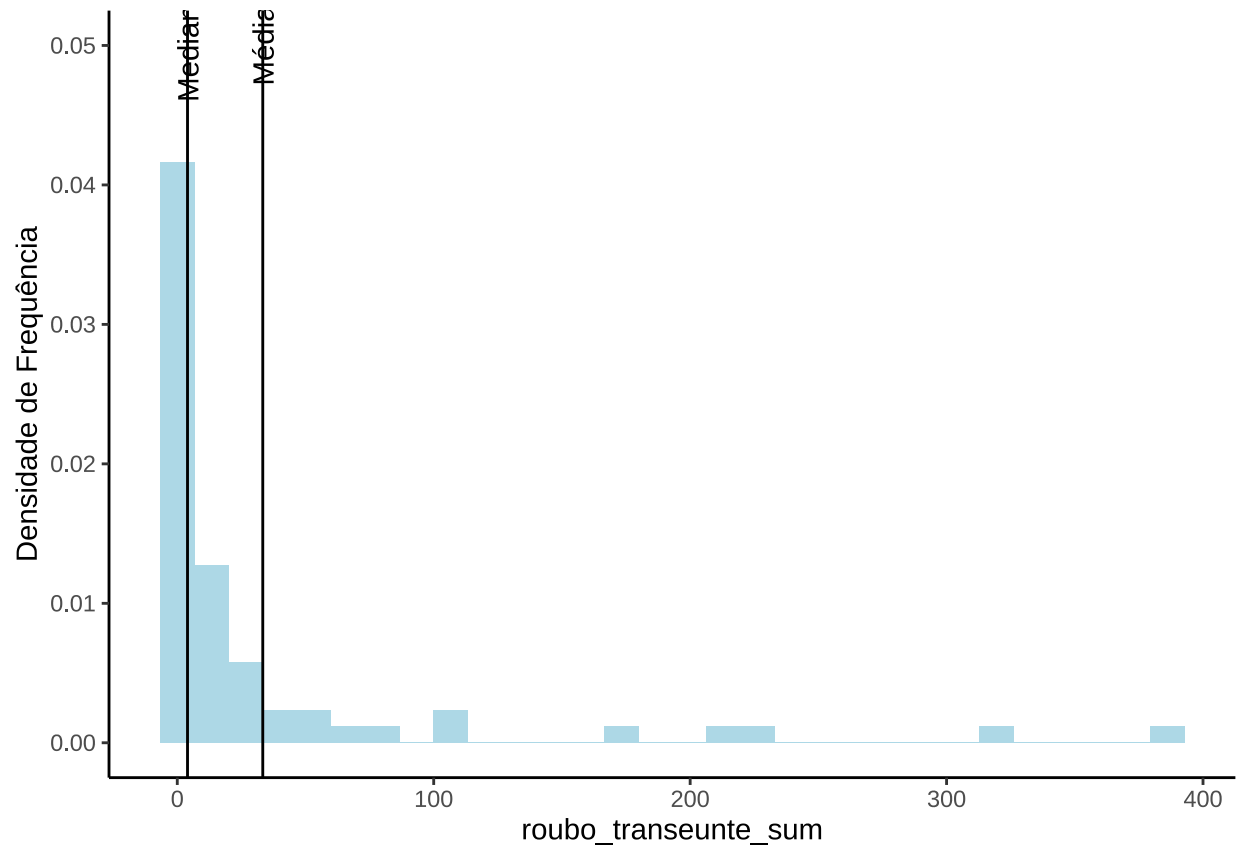
Apos a execução do da funcao descr, verifica-se que existe uma divergência considerável entre o valor da média e mediana. Também observa-se que o Skewness, possui um valor de assimetria superior a 1 (Skewness = 3.09), o que conclui-se que trata-se de um caso de ALTA ASSIMETRIA, COM PRESENÇA DE OUTLIERS.

Para definição do intervalo interquartil, vou considerar :

- intervalo interquartil buscando uma relacao entre a concentração dos dados;
- Considerando a presença de outliers, farei uso da regra de Freedman-Diaconis
- Também farei uma análise considerando a estimativa por kernel.

```
Bevolucao.roubos.f5.tbl %>% dplyr::select(roubo_transeunte_sum) %>% ggplot(aes(x=roubo_transeunte_sum)).
```

```
## Adding missing grouping variables: `Regiao`, `ano`
```

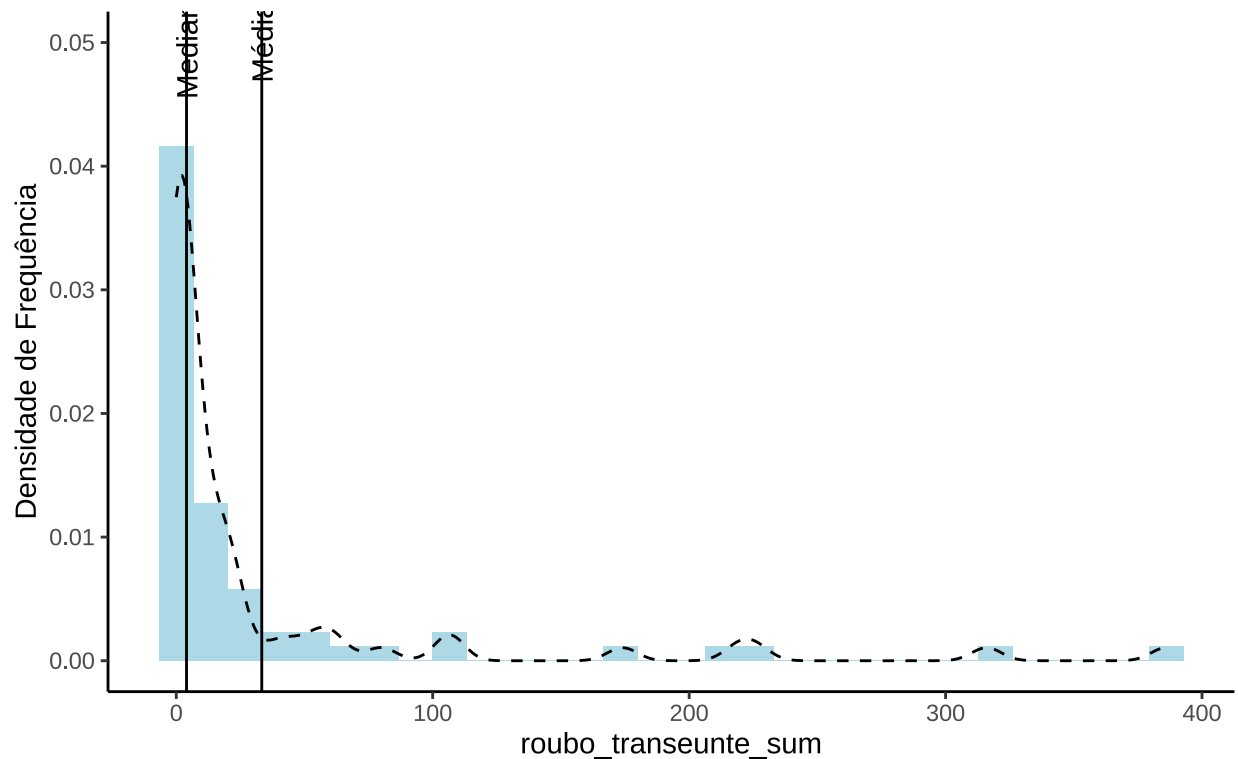


No Caso em questão, observação uma concentração de dados a esquerda, com calda a direita, reforçando o cenário de assimetria entre os dados.

```
Bevolucao.roubos.f5.tbl %>% dplyr::select(roubo_transeunte_sum) %>% ggplot(aes(x=roubo_transeunte_sum)).
## Adding missing grouping variables: `Regiao`, `ano`
```

Distribuição dos dados de crimes aproximada por Histograma

Binarização sugerida pelos detentores dos Dados



Neste caso, a amplitude da binarização foi ajustada para acompanhar a curva de pontos dos dados, onde cheguei a conclusão que a amplitude ideal para a binarização é considerar o valor de 30.

Poderíamos também considerar outras regra de binarização levando em consideração regras disponíveis na literatura, como a regra de Freedman-Diaconis, bem como a regra de Sturge, como segue:

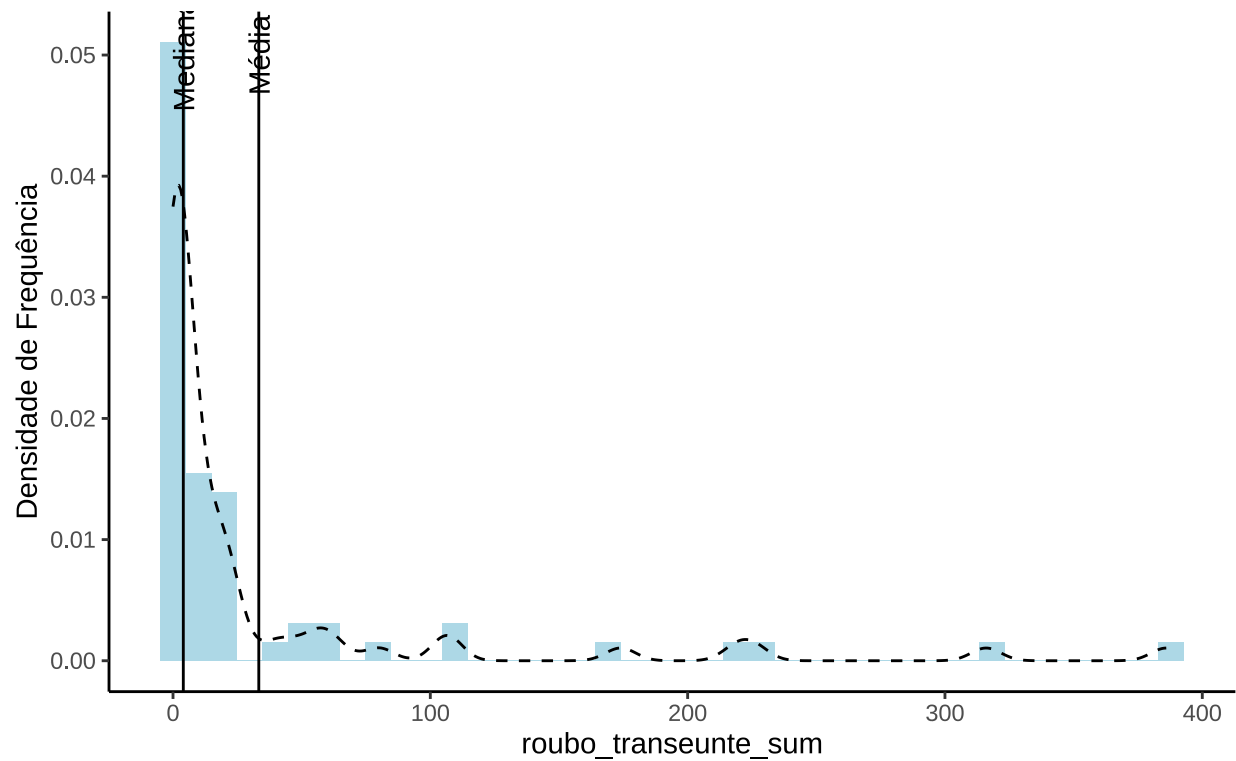
```
fd <- function(x) {  
  n <- length(x)  
  return((2*IQR(x))/n^(1/3))  
}  
  
sr <- function(x) {  
  n <- length(x)  
  return((3.49*sd(x))/n^(1/3))  
}
```

```
Bevolucao.roubos.f5.tbl %>% dplyr::select(roubo_transeunte_sum) %>% ggplot(aes(x=roubo_transeunte_sum)).
```

```
## Adding missing grouping variables: `Regiao`, `ano`
```

Distribuição dos dados de crimes aproximada por Histograma

Binarização pela Regra de FD

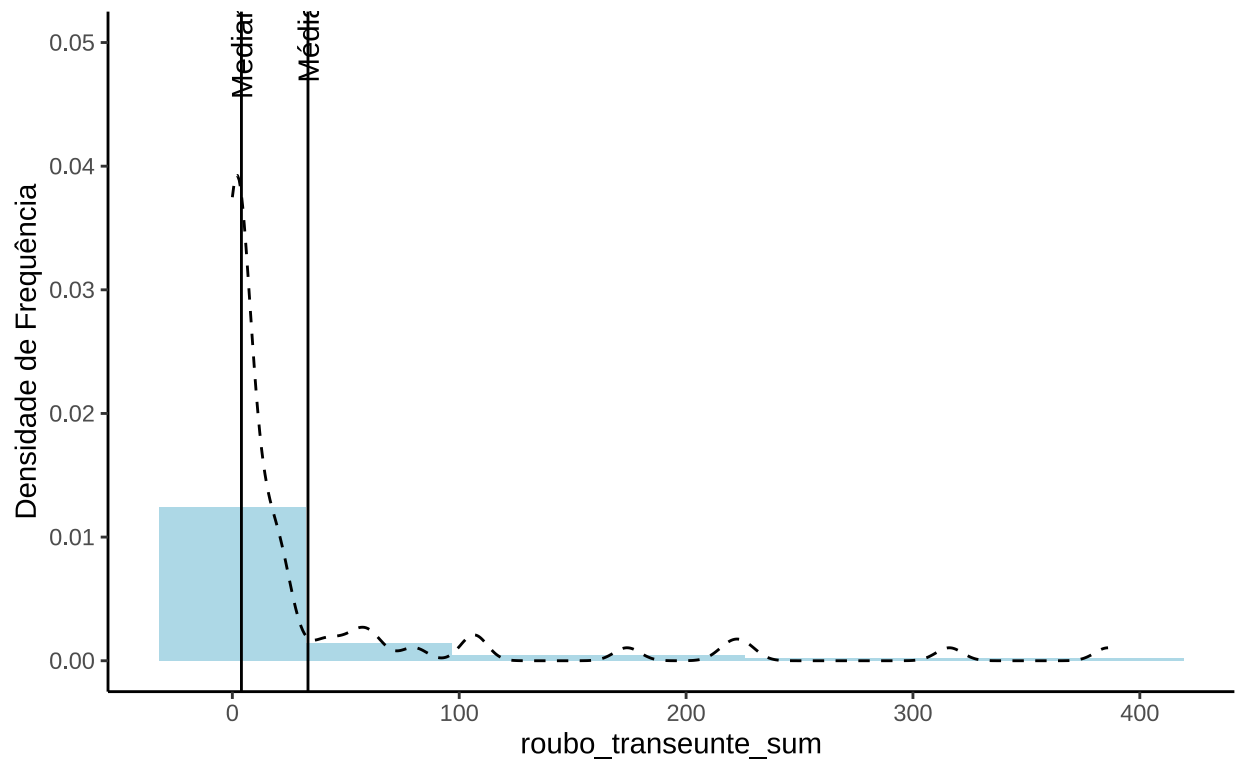


```
Bevolucao.roubos.f5.tbl %>% dplyr::select(roubo_transeunte_sum) %>% ggplot(aes(x=roubo_transeunte_sum)).
```

```
## Adding missing grouping variables: `Regiao`, `ano`
```

Distribuição dos dados de crimes aproximada por Histograma

Binarização pela Regra de Sturge



- CONCLUSÃO :

Para as características da base analisada, com uma assimetria acentuada, a regra de FD (Freedman Diaconis), se demonstra mais adequada para a análise.

Análise inter variáveis Neste caso farei o filtro para selecionando as localidades que tenham valores para apreensão de drogas (apreensao_drogas_sum) que deverá ser comparado com outras 4 variáveis da base.

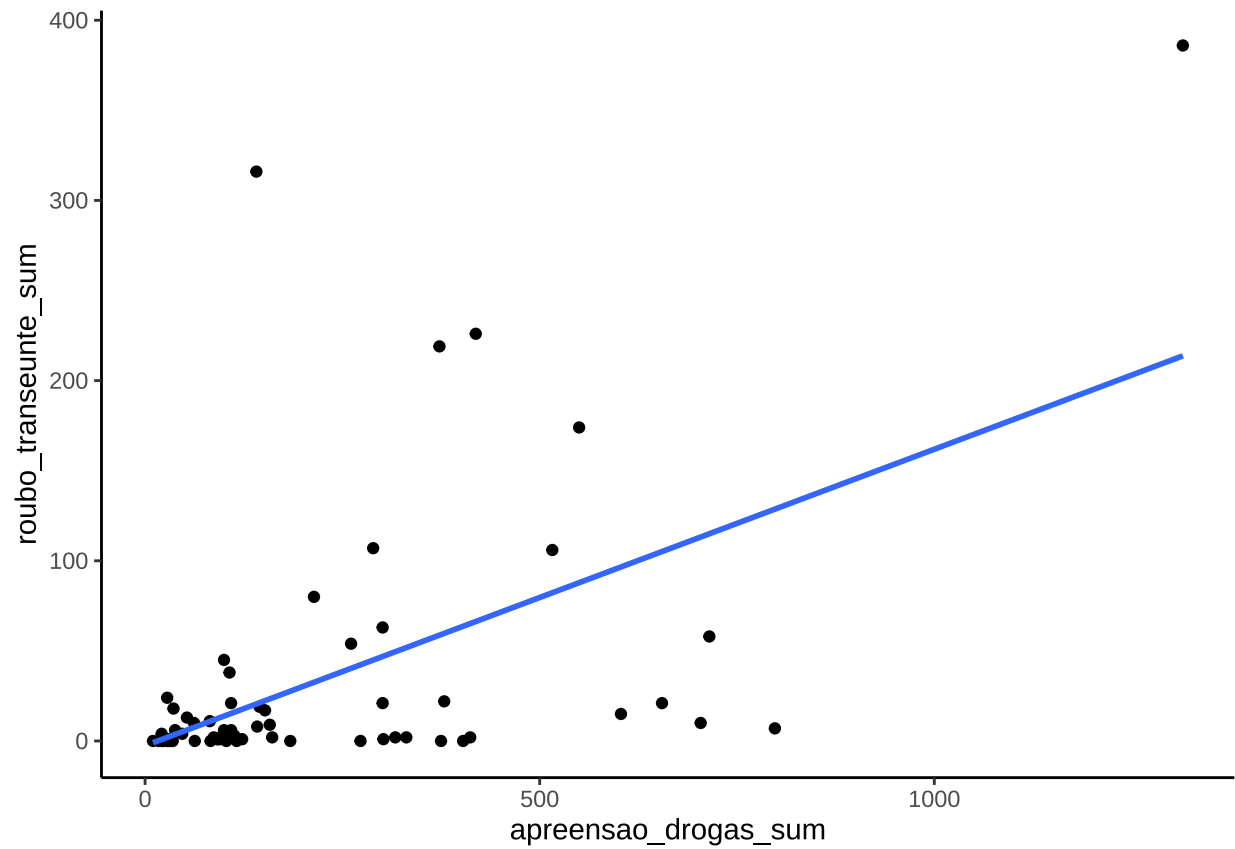
PAREI AQUI

```
{r Correlacao de variaveis , echo = TRUE} ### kable(cor(Bevolucao.roubos.f5.tbl %>%
dplyr::filter(!is.na(apreensao_drogas_sum)) %>% ### dplyr::select(apreensao_drogas_sum,
roubo_transeunte_sum,roubo_celular_sum,roubo_residencia_sum))) ###
```

- Scatterplot apreensao_drogas_sum x roubo_transeunte_sum

```
Bevolucao.roubos.f5.tbl %>% dplyr::filter(!is.na(apreensao_drogas_sum)) %>% dplyr::select(apreensao_drogas_sum, roubo_transeunte_sum)
```

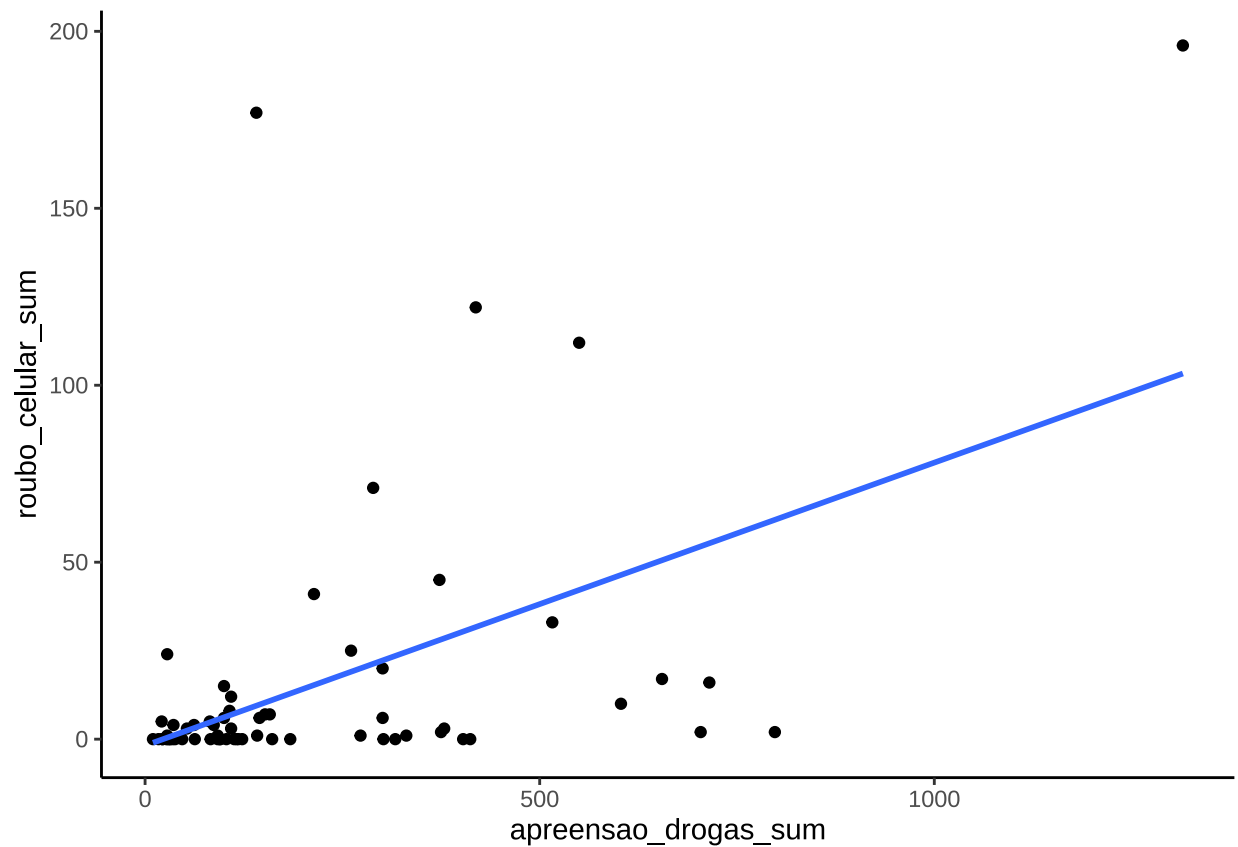
```
## Adding missing grouping variables: `Regiao`, `ano`
## `geom_smooth()` using formula = 'y ~ x'
```



- Scatterplot apreensao_drogas_sum x roubo_celular_sum

```
Bevolucao.roubos.f5.tbl %>% dplyr::filter(!is.na(apreensao_drogas_sum)) %>% dplyr::select(apreensao_drogas_sum, roubo_celular_sum)

## Adding missing grouping variables: `Regiao`, `ano`
## `geom_smooth()` using formula = 'y ~ x'
```

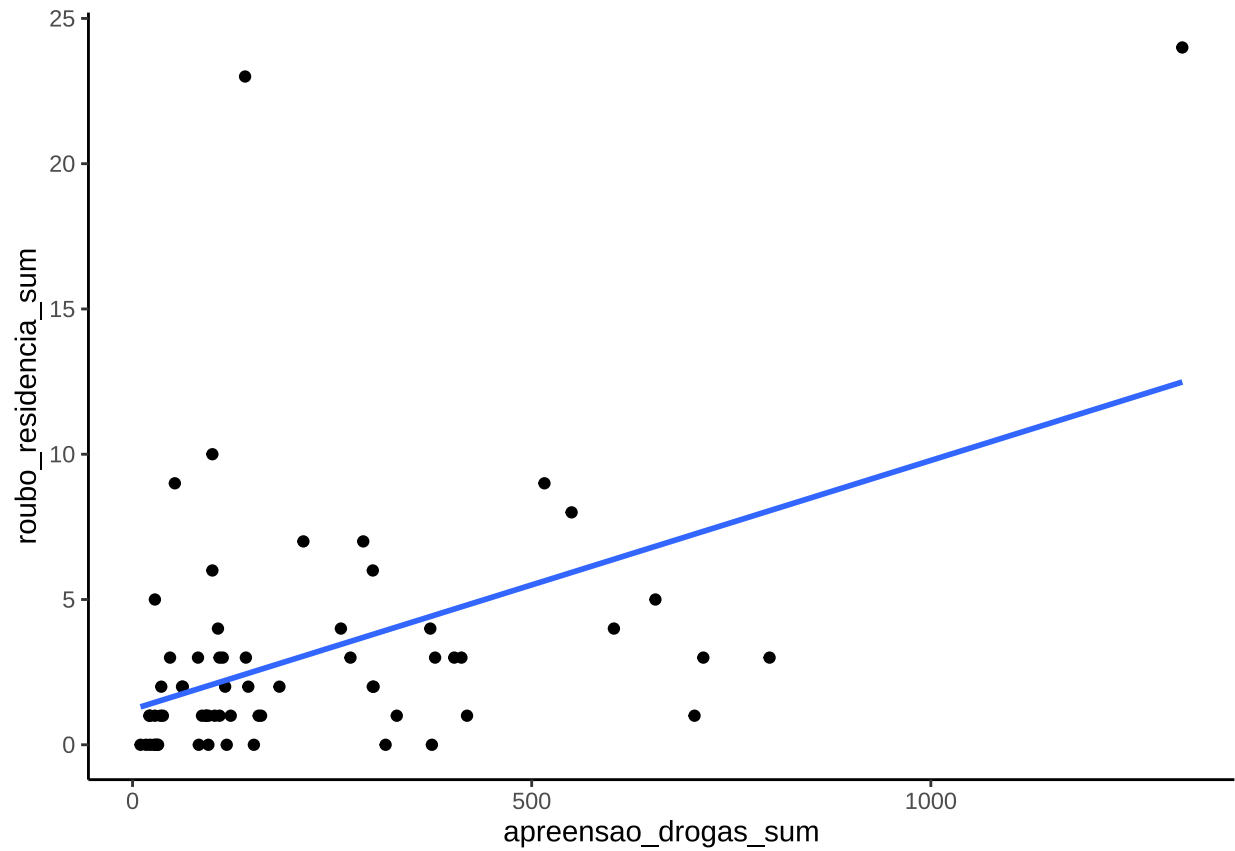


- Scatterplot `apreensao_drogas_sum` x `roubo_residencia_sum`

```
Bevolucao.roubos.f5.tbl %>% dplyr::filter(!is.na(apreensao_drogas_sum)) %>% dplyr::select(apreensao_drogas_sum, roubo_residencia_sum)
```

```
## Adding missing grouping variables: `Regiao`, `ano`
```

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



- Criando um grafico de barras, utilizando 2 variaveis

```
ggplot(Bevolucao.roubos.f5.tbl %>% filter(munic=="Angra dos Reis"|munic== "Araruama"| munic=="Cabo Frio"))
## Adding missing grouping variables: `Regiao`, `ano`
```