

Análise Exploratória de Dados - Projeto Final

Tatiana Sant'Anna

2025-04-12

Contents

1	Introdução	2
2	Análise Exploratória	5
3	Testes de imputação de dados - knn	10
4	Teste de imputação dos dados - método de imputação múltipla (pmm)	12
5	Teste de imputação de dados - método midastouch (pmm ponderado)	15
6	Sobre a normalidade das variáveis	19
7	Descrição via boxplot e tabelas de contingência	28
8	Filtrar o espaço amostral	40
9	Discussão da relação entre variáveis quantitativas e qualitativas	41
10	Calculando a dispersão e as correlações de Pearson e de Spearman com duas variáveis	44
11	Calculando a probabilidade do evento que estou investigando	48
12	Teste de hipóteses	48
13	Gráfico com a matriz de espalhamento (scatter matrix plot)	53
14	Correlação entre Número de Usuários e Velocidade	56
15	Correlação entre Número de Dispositivos Conectados e Velocidade	56
16	Correlação entre Investimento Público e Velocidade	57
17	Evento aleatório - escolher uma cidade/mês em que a velocidade média da internet foi superior a 50 Mbps	64

18 Qualidade de dados	65
19 Qual a completude para cada uma das variáveis do seu banco de dados?	65
20 Regressão Linear	66
21 Conclusão final	69
22 Dashboard Shiny	69
23 Repositório	71

1 Introdução

```
library(tidyverse)
```

1.0.0.1 Instalação de Pacotes

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## vforcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr    1.3.1
## v purrr    1.0.4
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()   masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(purrr)
library(dlookr)
```

```
## Registered S3 methods overwritten by 'dlookr':
##   method      from
##   plot.transform scales
##   print.transform scales
##
## Attaching package: 'dlookr'
##
## The following object is masked from 'package:tidy whole':
##   extract
##
## The following object is masked from 'package:base':
##   transform
```

```

library(summarytools)

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

library(readxl)
library(knitr)
library(data.table)

##
## Attaching package: 'data.table'
##
## The following objects are masked from 'package:lubridate':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday, week,
##     yday, year
##
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
##
## The following object is masked from 'package:purrr':
##
##     transpose

library(ggpubr)
library(corrplot)

## corrplot 0.95 loaded

library(rcompanion)
library(dplyr)
library(stringr)
library(lubridate)
library(ggplot2)
library(naniar)
library(simputation)

##
## Attaching package: 'simputation'
##
## The following object is masked from 'package:naniar':
##
##     impute_median

```

```

library(mice)

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

library(GGally)

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

capabilities("tcltk")

## tcltk
## TRUE

```

1.0.0.2 Leitura das bases de dados

```

## Rows: 32245 Columns: 13
## -- Column specification -----
## Delimiter: ","
## chr (3): cidade, uf, mes_ano
## dbl (10): mes, ano, n_usuarios_internet, velocidade_media_mbps, n_dispositiv...
##
## 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.

```

1.1 Introdução: Escolha da Base de Dados

A base de dados utilizada neste projeto foi elaborada de forma simulada com apoio do assistente ChatGPT, para fins acadêmicos e didáticos. Seu objetivo é representar a evolução de indicadores tecnológicos em cidades brasileiras ao longo do tempo, permitindo praticar a disciplina de análise exploratória e imputação de dados. Cada registro corresponde a uma observação mensal de uma cidade, contemplando variáveis relacionadas ao acesso à internet, inovação, investimentos públicos, atividade empreendedora e produção tecnológica.

O formato da base possibilita a análise de múltiplas variáveis contínuas, com destaque para o número de usuários de internet e a velocidade média da conexão. A escolha desta base foi feita pela diversidade de registros e pela presença controlada de dados faltantes, fatores importantes para o desenvolvimento do projeto.

O objetivo principal da análise é investigar o comportamento do número de usuários de internet em situações específicas onde a velocidade média ultrapassa de 50 Mbps, permitindo identificar padrões de acesso em cenários de melhor qualidade de conexão.

Fonte: Base simulada criada para fins acadêmicos com apoio de ChatGPT e adaptada pelo aluno.

1.2 Descrição do Espaço Amostral

O espaço amostral deste projeto é composto por registros mensais de cidades do Brasil. Cada elemento da base refere-se a um par cidade/mês, como por exemplo “São Paulo - Janeiro/2020”.

O evento aleatório considerado é a seleção de registros em que a velocidade média da internet superou 50 Mbps. O objeto de interesse é o número de usuários de internet (n_usuarios_internet) nesses casos específicos.

Cada indivíduo analisado representa uma cidade em um determinado mês, e o espaço amostral completo abrange todas as cidades disponíveis na base de dados, independentemente da ocorrência ou não do evento analisado.

2 Análise Exploratória

2.1 Utilizando o pacote `summarytools` (função `descr`) para descrever estatisticamente a base de dados

```
descr(tech, plain.ascii = FALSE, style = "rmarkdown")  
  
## Non-numerical variable(s) ignored: cidade, uf, mes_ano  
  
## #### Descriptive Statistics  
## ##### tech  
## **N:** 32245  
##  
## |      &nbs; |      ano | eventos_tecnologicos_realizados | indice_inovacao |  
## |-----:|-----:|-----:|-----:|-----:  
## | **Mean** | 2011.97 |                      3.00 |       0.50 |  
## | **Std.Dev** |    7.21 |                     1.72 |       0.29 |  
## | **Min** | 2000.00 |                     0.00 |       0.00 |  
## | **Q1** | 2006.00 |                     2.00 |       0.25 |  
## | **Median** | 2012.00 |                     3.00 |       0.50 |  
## | **Q3** | 2018.00 |                     4.00 |       0.75 |  
## | **Max** | 2024.00 |                    12.00 |       1.00 |  
## | **MAD** |     8.90 |                     1.48 |       0.37 |  
## | **IQR** |    12.00 |                     2.00 |       0.50 |  
## | **CV** |     0.00 |                     0.57 |       0.57 |  
## | **Skewness** |    0.01 |                     0.59 |      -0.01 |  
## | **SE.Skewness** |    0.01 |                     0.01 |       0.01 |  
## | **Kurtosis** |   -1.20 |                     0.38 |      -1.20 |  
## | **N.Valid** | 32245.00 |                  32245.00 |      32245.00 |  
## | **N** | 32245.00 |                  32245.00 |      32245.00 |  
## | **Pct.Valid** | 100.00 |                  100.00 |      100.00 |  
##  
## Table: Table continues below  
##  
##  
##  
## |      &nbs; | investimento_publico_tecnologia |      mes | n_dispositivos_conectados |  
## |-----:|-----:|-----:|-----:|-----:  
##
```

```

## |      **Mean** |      550480.32 |      6.50 |      55009.48 |
## |      **Std.Dev** |      259932.24 |      3.45 |      25959.20 |
## |      **Min** |      100004.30 |      1.00 |      10001.00 |
## |      **Q1** |      324929.32 |      3.00 |      32549.00 |
## |      **Median** |      551799.95 |      6.00 |      54929.00 |
## |      **Q3** |      773694.49 |      9.00 |      77573.00 |
## |      **Max** |      999984.69 |     12.00 |      99998.00 |
## |      **MAD** |      332646.84 |      4.45 |      33371.84 |
## |      **IQR** |      448765.17 |      6.00 |      45024.00 |
## |      **CV** |      0.47 |      0.53 |      0.47 |
## |      **Skewness** |      0.00 |      0.00 |      0.00 |
## | **SE.Skewness** |      0.01 |      0.01 |      0.01 |
## |      **Kurtosis** |      -1.20 |      -1.22 |      -1.20 |
## |      **N.Valid** |      32245.00 |      32245.00 |      32245.00 |
## |      **N** |      32245.00 |      32245.00 |      32245.00 |
## |      **Pct.Valid** |      100.00 |      100.00 |      100.00 |
## 
## Table: Table continues below
## 
## 
## 
## |        | n_usuarios_internet | patentes_registradas | startups_ativas |
## |-----:|-----:|-----:|-----:|
## |      **Mean** |      50000.34 |      5.02 |      154.66 |
## |      **Std.Dev** |      223.24 |      2.26 |      83.76 |
## |      **Min** |      49090.00 |      0.00 |      10.00 |
## |      **Q1** |      49849.00 |      3.00 |      82.00 |
## |      **Median** |      50000.00 |      5.00 |      155.00 |
## |      **Q3** |      50151.00 |      6.00 |      227.00 |
## |      **Max** |      50878.00 |      15.00 |      299.00 |
## |      **MAD** |      223.87 |      2.97 |      106.75 |
## |      **IQR** |      302.00 |      3.00 |      145.00 |
## |      **CV** |      0.00 |      0.45 |      0.54 |
## |      **Skewness** |      0.00 |      0.45 |      0.00 |
## | **SE.Skewness** |      0.01 |      0.01 |      0.01 |
## |      **Kurtosis** |      0.00 |      0.16 |      -1.20 |
## |      **N.Valid** |      32245.00 |      32245.00 |      32245.00 |
## |      **N** |      32245.00 |      32245.00 |      32245.00 |
## |      **Pct.Valid** |      100.00 |      100.00 |      100.00 |
## 
## Table: Table continues below
## 
## 
## 
## |        | velocidade_media_mbps |
## |-----:|-----:|
## |      **Mean** |      50.07 |
## |      **Std.Dev** |      15.01 |
## |      **Min** |      -16.94 |
## |      **Q1** |      39.97 |
## |      **Median** |      50.09 |
## |      **Q3** |      60.17 |
## |      **Max** |      107.45 |
## |      **MAD** |      14.97 |

```

```

## |      **IQR** |      20.20 |
## |      **CV** |      0.30 |
## |      **Skewness** |      0.01 |
## | **SE.Skewness** |      0.01 |
## |      **Kurtosis** |     -0.03 |
## |      **N.Valid** | 31278.00 |
## |      **N** | 32245.00 |
## |      **Pct.Valid** |      97.00 |

summary(tech)

##      cidade          uf          mes         ano
## Length:32245    Length:32245    Min.   : 1.0  Min.   :2000
## Class  :character  Class  :character  1st Qu.: 3.0  1st Qu.:2006
## Mode   :character  Mode   :character  Median  : 6.0  Median  :2012
##                               Mean   : 6.5  Mean   :2012
##                               3rd Qu.: 9.0  3rd Qu.:2018
##                               Max.  :12.0  Max.  :2024
##
##      mes_ano        n_usuarios_internet velocidade_media_mbps
## Length:32245        Min.   :49090        Min.   :-16.94
## Class  :character  1st Qu.:49849        1st Qu.: 39.97
## Mode   :character  Median :50000        Median : 50.09
##                               Mean   :50000        Mean   : 50.07
##                               3rd Qu.:50151        3rd Qu.: 60.17
##                               Max.   :50878        Max.   :107.45
##                               NA's   :967
##
##      n_dispositivos_conectados indice_inovacao investimento_publico_tecnologia
## Min.   :10001           Min.   :0.0000  Min.   :100004
## 1st Qu.:32549           1st Qu.:0.2540  1st Qu.:324929
## Median :54929           Median :0.5040  Median :551800
## Mean   :55009           Mean   :0.5035  Mean   :550480
## 3rd Qu.:77573           3rd Qu.:0.7530  3rd Qu.:773694
## Max.   :99998           Max.   :1.0000  Max.   :999985
##
##      eventos_tecnologicos_realizados startups_ativas patentes_registradas
## Min.   : 0.000           Min.   : 10.0  Min.   : 0.000
## 1st Qu.: 2.000           1st Qu.: 82.0  1st Qu.: 3.000
## Median : 3.000           Median :155.0  Median : 5.000
## Mean   : 3.004           Mean   :154.7  Mean   : 5.019
## 3rd Qu.: 4.000           3rd Qu.:227.0  3rd Qu.: 6.000
## Max.   :12.000           Max.   :299.0  Max.   :15.000
##
sum(tech$velocidade_media_mbps < 0, na.rm = TRUE)

## [1] 10

```

Após a aplicação da função `descr()` do pacote `summarytools`, foi possível obter uma descrição estatística detalhada das variáveis numéricas da base de dados. Observou-se que a velocidade média da internet nas cidades analisadas apresenta uma média de aproximadamente 50,09 Mbps, com um desvio padrão de 14,98 Mbps, indicando uma variação moderada entre as cidades. A mediana foi de 50,10 Mbps, valor próximo da

média, sugerindo uma distribuição relativamente simétrica, confirmada pelo valor de Skewness próximo de zero. Em relação ao número de usuários de internet, ele mostra que a base é extremamente estável, com baixa variação entre cidades/meses. A média foi de aproximadamente 50.000 usuários, com pequena variação (desvio padrão de cerca de 223 usuários). De maneira geral, a base apresenta uma quantidade elevada de registros válidos para todas as variáveis, com mais de 96,97% dos dados completos em cada variável analisada.

As variáveis `n_usuarios_internet` e `velocidade_media_mbps` são bastante diferentes em comportamento: a primeira é muito estável, a segunda muito dispersa. A variável `velocidade_media_mbps` é a que tem alguns dados faltantes.

2.1.1 Modificações e ajustes (acertos nos campos de data para um mesmo formato)

```
#tech <- tech %>%
#   mutate(
#     mes_ano_corrigido = str_replace(mes_ano, "m", "-01"),
#     mes_ano_data = as.Date(mes_ano_corrigido, format = "%Y-%m-%d")
#   )

tech <- tech %>%
  mutate(
    mes_ano_data = as.Date(paste0(str_sub(mes_ano, 1, 4), "-", str_sub(mes_ano, 6, 7), "-01"))
  )
```

2.1.2 Limpando os dados, filtrando e realizando diagnóstico

```
tech_filter <- tech %>%
  select(cidade, mes, ano, n_usuarios_internet, velocidade_media_mbps,
         n_dispositivos_conectados, investimento_publico_tecnologia) %>%
  mutate(mes_ano = as.Date(paste(ano, mes, "01", sep = "-"))) %>%
  filter(ano %in% 2019:2022) %>%
  distinct(cidade, mes_ano, .keep_all = TRUE)
```

```
tech_filter %>% dlookr::diagnose()
```

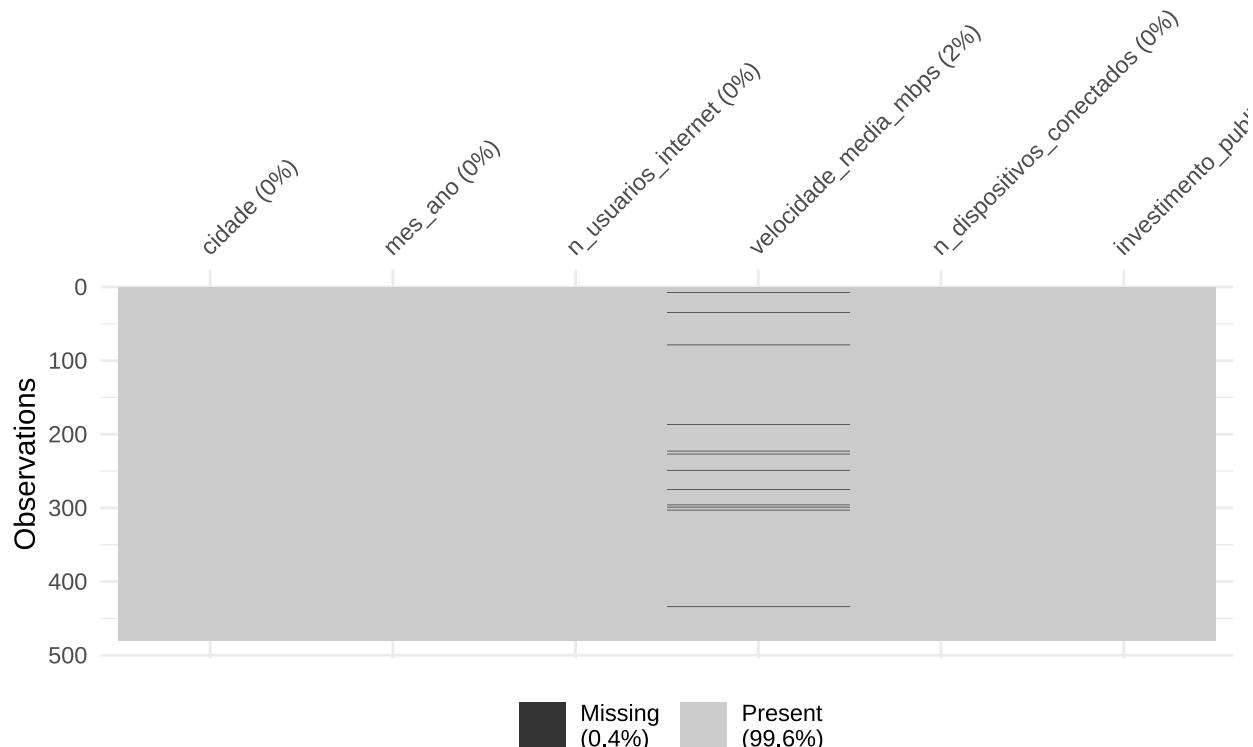
2.1.2.1 Pra ir cercando o problema, primeiro preciso fazer a descrição dos dados, através do diagnose

```
## # A tibble: 8 x 6
##   variables      types missing_count missing_percent unique_count unique_rate
##   <chr>        <chr>        <int>          <dbl>        <int>          <dbl>
## 1 cidade       char~         0             0           10        0.0208
## 2 mes          nume~         0             0           12        0.025
## 3 ano          nume~         0             0            4        0.00833
## 4 n_usuarios_inter~ nume~         0             0           371       0.773
## 5 velocidade_media~ nume~        12            2.5          439       0.915
## 6 n_dispositivos_c~ nume~         0             0           478       0.996
## 7 investimento_pub~ nume~         0             0           480        1
## 8 mes_ano      Date          0             0            48        0.1
```

A base de dados filtrada foi analisada quanto à completude e tipos de variáveis. Observou-se que a estrutura da base está adequada, com todas as variáveis devidamente tipadas (caracteres, numéricas ou datas) e uma baixa ocorrência de valores faltantes. Apenas a variável **velocidade_media_mbps** apresentou 2,5% de valores ausentes. Dá pra tentar identificar se é totalmente aleatório ou não e tentar preencher com as técnicas específicas.

Podemos também visualizar os dados faltantes através de um gráfico de barras percentual, gerado pela biblioteca naniar e função vis_miss.

```
tech_filter %>% dplyr::mutate(mes_ano = as.IDate(paste("01", mes, ano, sep = "-"), format = "%d-%m-%Y"))
```



O gráfico confirma os dados faltantes apenas na variável **velocidade_media_mbps**.

2.2 Verificar a aleatoriedade dos meus dados faltantes para a variável velocidade_media_mbps usando o teste de Little

```
tech %>%
  select(velocidade_media_mbps) %>%
  mcar_test()
```

```
## # A tibble: 1 x 4
##   statistic    df  p.value missing.patterns
##       <dbl>   <dbl>     <dbl>          <int>
## 1  1.75e-25     0      0               2
```

```
tech %>% dplyr::select(velocidade_media_mbps) %>% naniar::mcar_test()
```

```
## # A tibble: 1 x 4
##   statistic    df p.value missing.patterns
##       <dbl>   <dbl>     <dbl>           <int>
## 1  1.75e-25     0        0                  2
```

Quando p-value = 0, **rejeitamos a hipótese nula do teste de dados completamente aleatórios a qualquer nível de significância**. Isso significa que os dados faltantes não são completamente aleatórios. Ou seja, os valores faltantes estão relacionados a alguma variável ou a algum padrão no seu conjunto de dados.

2.3 Realizando análise da relação entre a distribuição de missing e das variáveis observadas

```
tech %>%
  dplyr::group_by(cidade) %>% dplyr::filter(is.na(velocidade_media_mbps)) %>%
  dplyr::summarise(n = n()) %>% dplyr::ungroup() %>%
  dplyr::mutate(tot.miss = sum(n)) %>% dplyr::group_by(cidade) %>%
  dplyr::mutate(tot.miss.regiao = sum(n), freq.intra.regiao = n/tot.miss.regiao, freq.regiao = tot.miss/
```

cidade	n	tot.miss	tot.miss.regiao	freq.intra.regiao	freq.regiao
Belo Horizonte	108	967	108	1	0.1116856
Rio de Janeiro	107	967	107	1	0.1106515
São Paulo	102	967	102	1	0.1054809
Salvador	101	967	101	1	0.1044467
Brasília	100	967	100	1	0.1034126
Curitiba	98	967	98	1	0.1013444
Porto Alegre	93	967	93	1	0.0961737
Fortaleza	90	967	90	1	0.0930714
Manaus	86	967	86	1	0.0889349
Recife	82	967	82	1	0.0847983

A maior quantidade de dados faltantes ocorre em Belo Horizonte e Rio de Janeiro, ambos com 109 registros faltantes, representando mais ou menos 11% dos dados faltantes cada um. Logo depois vem Salvador, com 105. As três cidades do Sudeste (Rio de Janeiro, São Paulo e Belo Horizonte) concentram juntas uma parte relevante dos dados faltantes de velocidade de internet.

2.3.1 Corrigindo os valores negativos da variável velocidade_media_mbps para NA

```
tech_filter$velocidade_media_mbps[tech_filter$velocidade_media_mbps < 0] <- NA
```

3 Testes de imputação de dados - knn

```

tech_filter_knn <- tech_filter %>%
  select(cidade, mes_ano, velocidade_media_mbps) %>%
  as.data.frame() %>%
  simputation::impute_knn(velocidade_media_mbps ~ cidade + mes_ano, k = 5, seed = 512) %>%
  dplyr::mutate(velocidade_media_mbps = as.numeric(velocidade_media_mbps))

```

3.1 Comparando a imputação dos dados com o realizado para velocidade para o método knn

```

ggplot(tech_filter, aes(x = velocidade_media_mbps)) +
  geom_histogram(binwidth = 5, fill = "lightblue", color = "black") +
  ggtitle("Distribuição - Antes da Imputação")

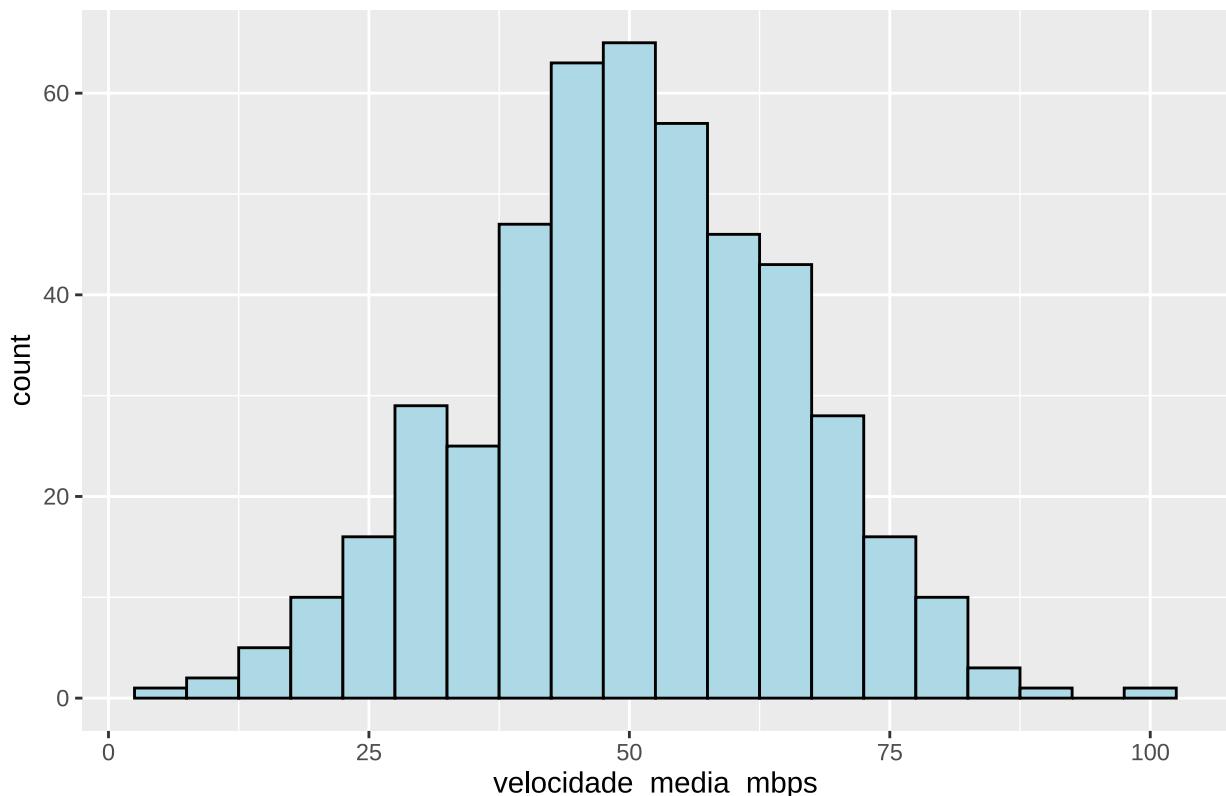
```

```

## Warning: Removed 12 rows containing non-finite outside the scale range
## (`stat_bin()`).

```

Distribuição - Antes da Imputação

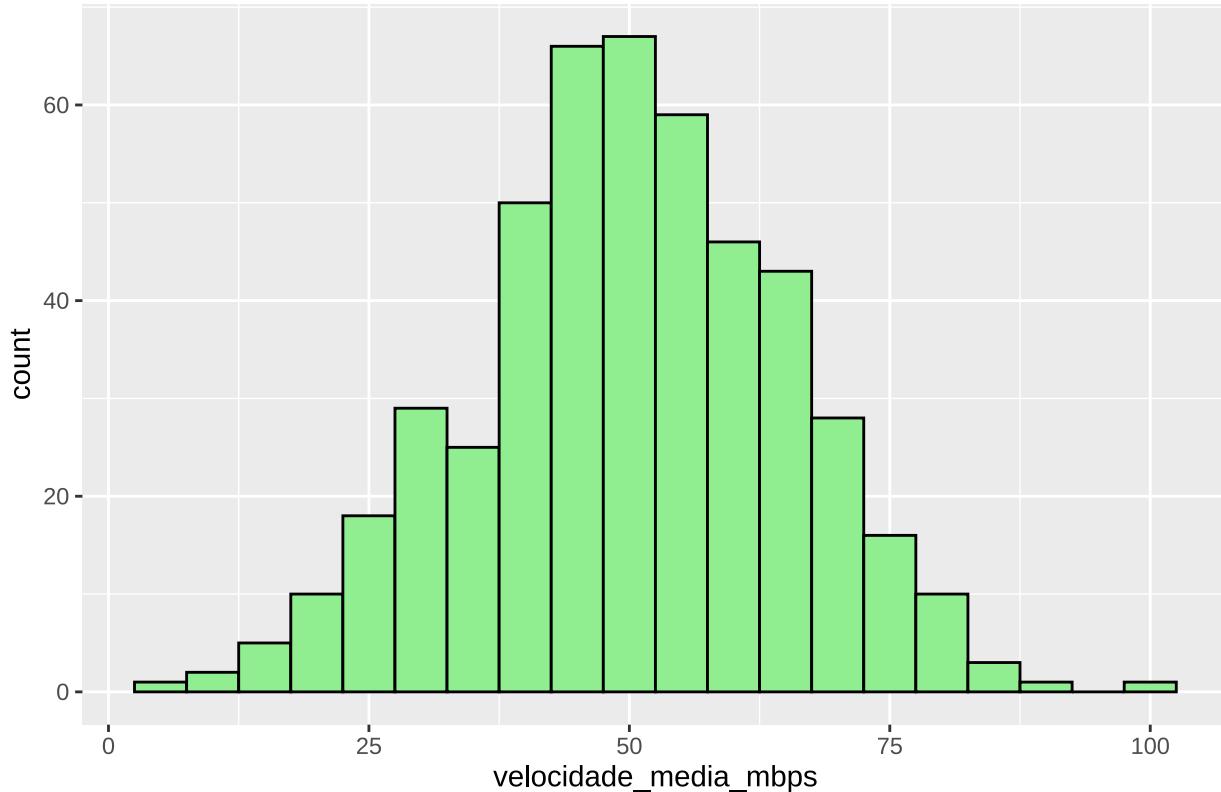


```

ggplot(tech_filter_knn, aes(x = velocidade_media_mbps)) +
  geom_histogram(binwidth = 5, fill = "lightgreen", color = "black") +
  ggtitle("Distribuição - Após Imputação knn")

```

Distribuição - Após Imputação knn



4 Teste de imputação dos dados - método de imputação múltipla (pmm)

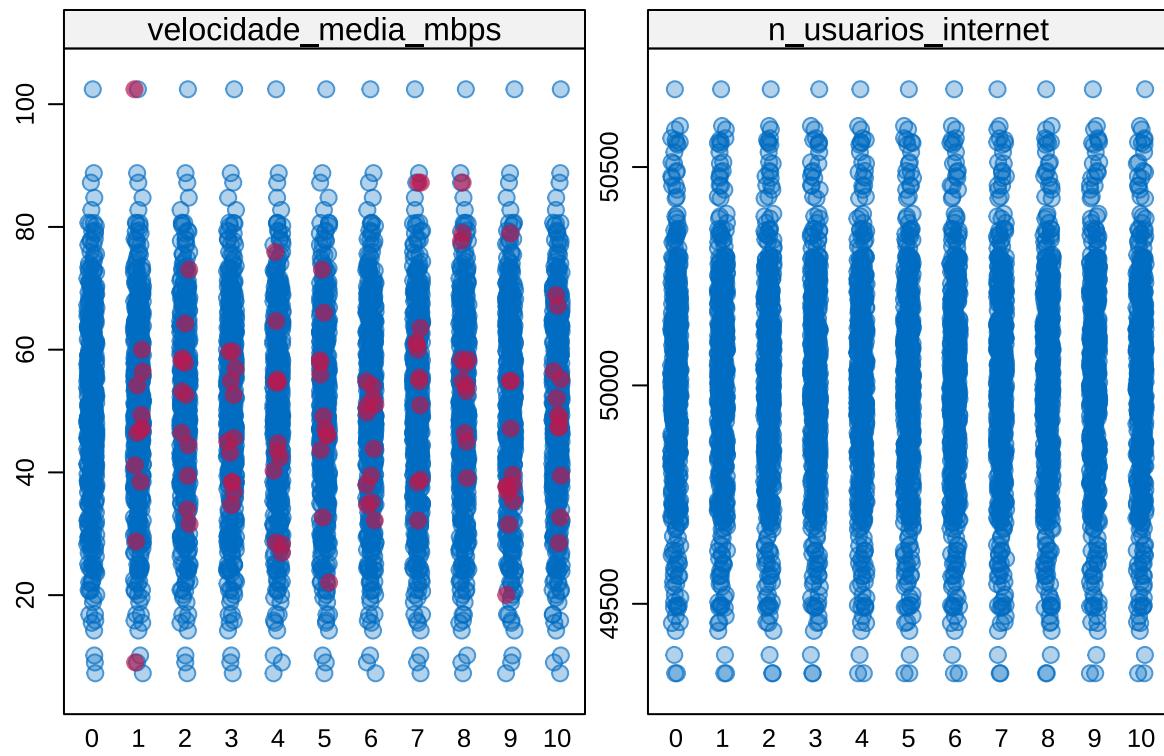
```
tech_filter <- tech_filter %>%
  mutate(velocidade_media_mbps = ifelse(velocidade_media_mbps < 0, NA, velocidade_media_mbps))

tech_velocidade <- tech_filter %>%
  select(velocidade_media_mbps, n_usuarios_internet)

metodos <- c("pmm", "")

tech_filter_multiplo <- mice(
  tech_velocidade,
  method = metodos,
  m = 10,
  maxit = 5,
  seed = 512,
  printFlag = FALSE
)

stripplot(tech_filter_multiplo, pch = c(21, 20), cex = c(1, 1.5))
```



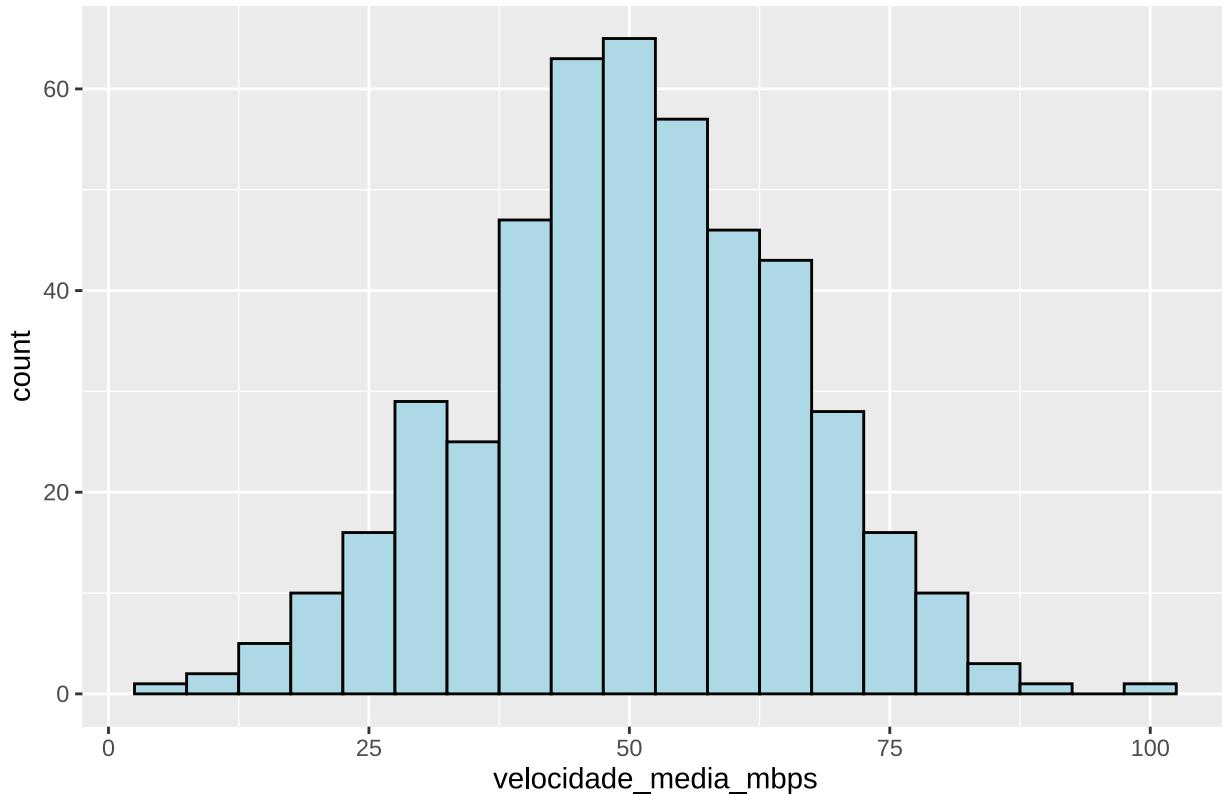
```
## Comparando a imputação dos dados com o realizado para velocidade para o método de imputação múltipla
```

```
dados_imputados <- complete(tech_filter_multiplo, 1)

# Gráfico antes da imputação
ggplot(tech_filter, aes(x = velocidade_media_mbps)) +
  geom_histogram(binwidth = 5, fill = "lightblue", color = "black") +
  ggtitle("Distribuição - Antes da Imputação")

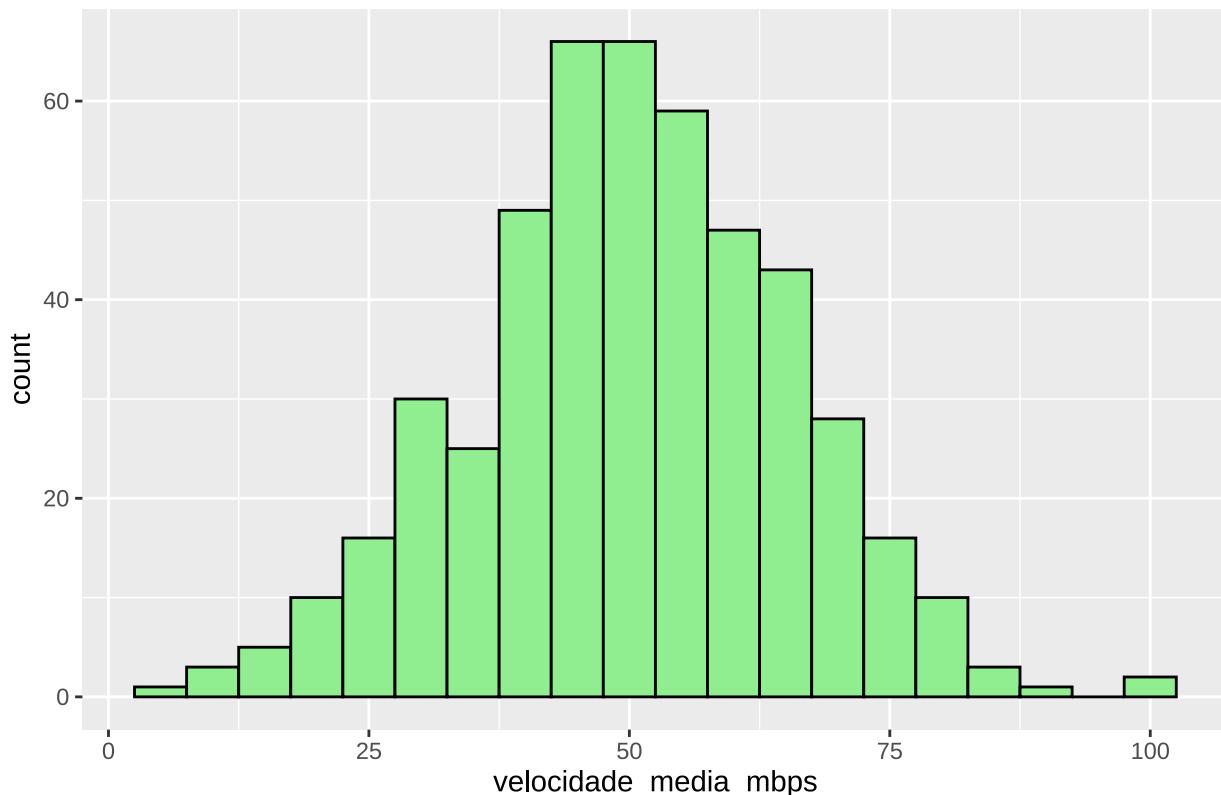
## Warning: Removed 12 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

Distribuição - Antes da Imputação



```
# Gráfico depois da imputação
ggplot(dados_imputados, aes(x = velocidade_media_mbps)) +
  geom_histogram(binwidth = 5, fill = "lightgreen", color = "black") +
  ggtitle("Distribuição - Após Imputação Múltipla")
```

Distribuição - Após Imputação Múltipla



5 Teste de imputação de dados - método midastouch (pmm ponderado)

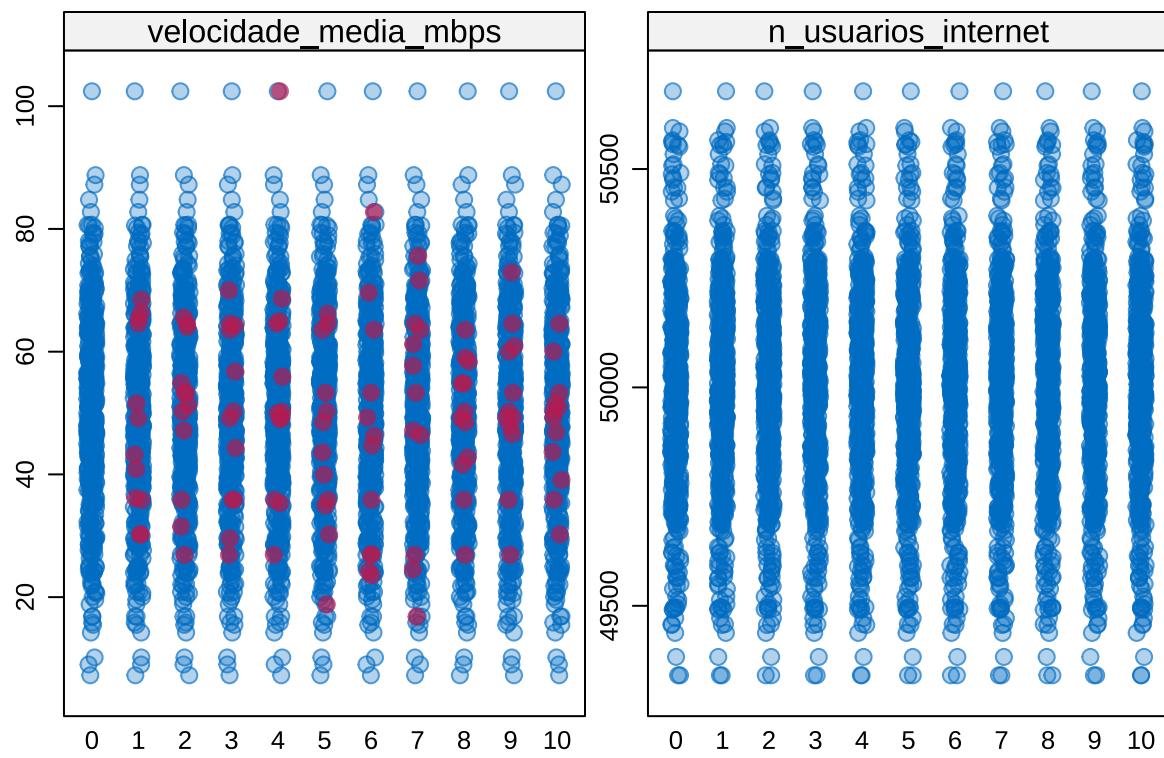
```
tech_filter <- tech_filter %>%
  mutate(velocidade_media_mbps = ifelse(velocidade_media_mbps < 0, NA, velocidade_media_mbps))

tech_velocidade <- tech_filter %>%
  select(velocidade_media_mbps, n_usuarios_internet)

metodos <- c("midastouch", "")

tech_filter_multiplo_midas <- mice(
  tech_velocidade,
  method = metodos,
  m = 10,
  maxit = 5,
  seed = 512,
  printFlag = FALSE
)

stripplot(tech_filter_multiplo_midas, pch = c(21, 20), cex = c(1, 1.5))
```



5.1 Comparando a imputação dos dados com o realizado para velocidade para o método midastouch

```

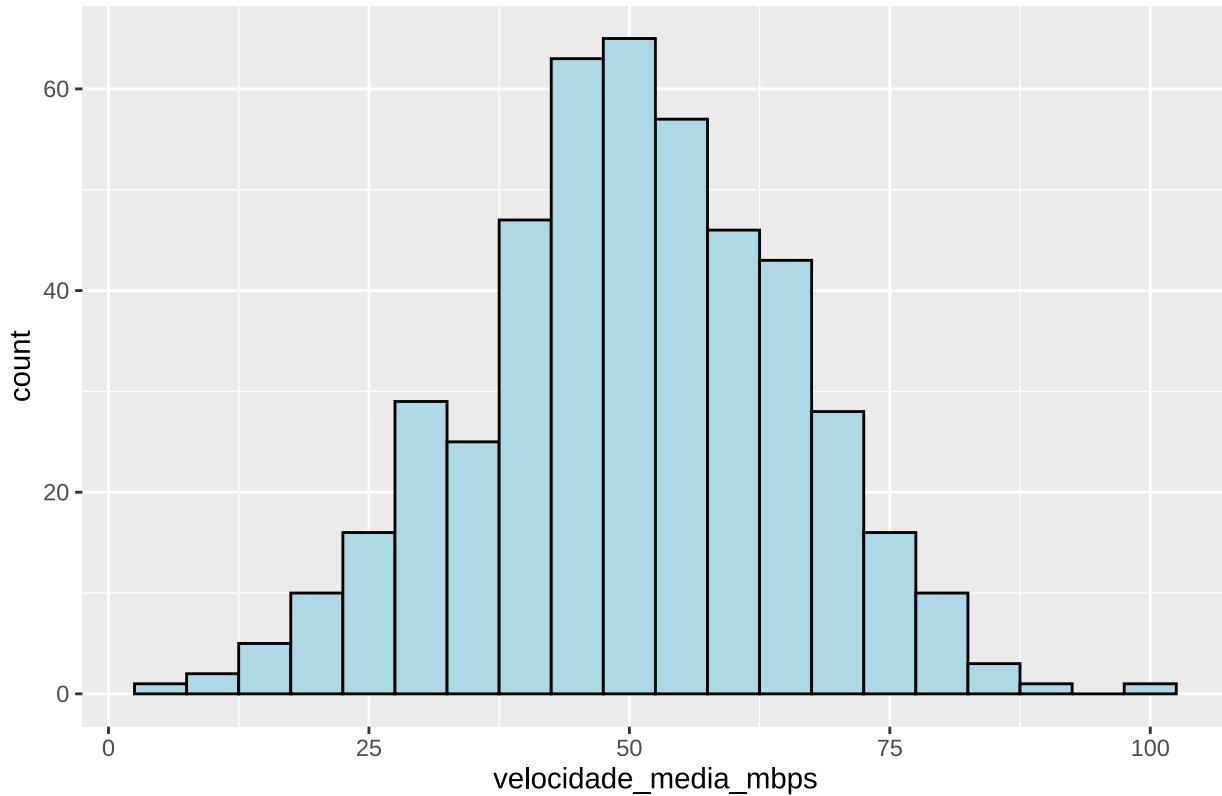
dados_imputados_midas <- complete(tech_filter_multiplo_midas, 1)

# Gráfico antes da imputação
ggplot(tech_filter, aes(x = velocidade_media_mbps)) +
  geom_histogram(binwidth = 5, fill = "lightblue", color = "black") +
  ggtitle("Distribuição - Antes da Imputação")

## Warning: Removed 12 rows containing non-finite outside the scale range
## (`stat_bin()`).

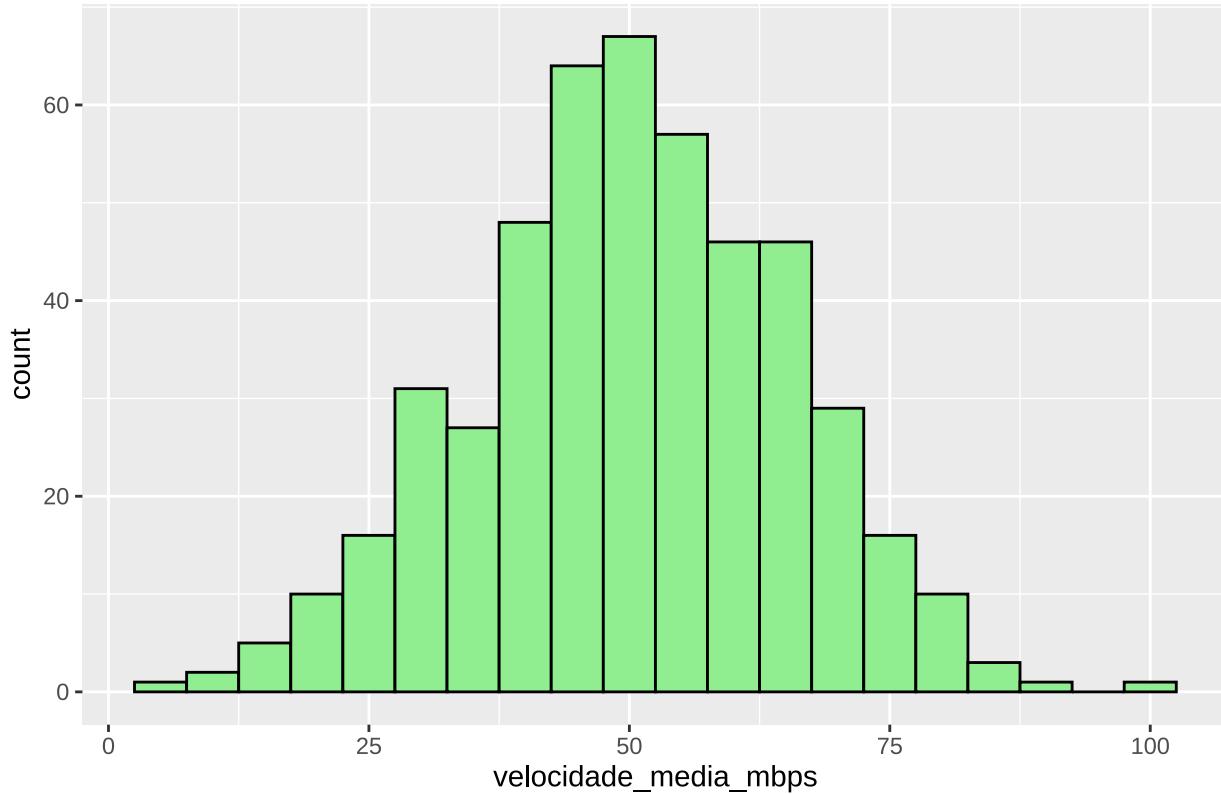
```

Distribuição - Antes da Imputação



```
# Gráfico depois da imputação
ggplot(dados_imputados_midas, aes(x = velocidade_media_mbps)) +
  geom_histogram(binwidth = 5, fill = "lightgreen", color = "black") +
  ggtitle("Distribuição - Após Imputação Múltipla")
```

Distribuição - Após Imputação Múltipla



Conclusão: a imputação múltipla usando o **método PMM**, na minha opinião, foi a que, por detalhes, melhor preservou o padrão original da variável velocidade, a simetria e o alcance dos dados. Ele mantém a distribuição empírica dos dados observados, sem criar valores que destoam do conjunto.

5.2 Completando a base com o valor imputado

```
tech_velocidade_imputado <- complete(tech_filter_multiplo, 1)

tech_filter$velocidade_media_mbps <- tech_velocidade_imputado$velocidade_media_mbps

summary(tech_filter)
```

```
##      cidade            mes           ano      n_usuarios_internet
##  Length:480      Min.   : 1.00   Min.   :2019   Min.   :49341
##  Class :character 1st Qu.: 3.75   1st Qu.:2020   1st Qu.:49856
##  Mode   :character Median : 6.50   Median :2020   Median :50008
##                                         Mean   : 6.50   Mean   :2020   Mean   :50004
##                                         3rd Qu.: 9.25   3rd Qu.:2021   3rd Qu.:50147
##                                         Max.   :12.00   Max.   :2022   Max.   :50678
##      velocidade_media_mbps n_dispositivos_conectados
##  Min.   : 7.27      Min.   :10580
##  1st Qu.:40.37      1st Qu.:31049
##  Median :49.80      Median :53819
##  Mean   :50.13      Mean   :54158
```

```

## 3rd Qu.: 60.09      3rd Qu.:76414
## Max.    :102.43      Max.    :99482
## investimento_publico_tecnologia   mes_ano
## Min.    :100847      Min.    :2019-01-01
## 1st Qu.:349406      1st Qu.:2019-12-24
## Median  :563938      Median  :2020-12-16
## Mean    :565281      Mean    :2020-12-15
## 3rd Qu.:780353      3rd Qu.:2021-12-08
## Max.    :999919      Max.    :2022-12-01

```

6 Sobre a normalidade das variáveis

O projeto pede para que eu: - Descreva o que é uma distribuição normal; - Crie um histograma para cada variável da sua base de dados. Justifique a escolha do número de bins para seu trabalho. (usando o pacote ggplot); - Crie um gráfico Q-Q para cada variável de sua base de dados. (use as funções presentes no pacote ggpqr); - Execute um teste de normalidade Shapiro-Wilk; - Baseado nos itens anteriores, é possível afirmar que algumas das variáveis se aproximam de uma distribuição normal? Justifique.

Respondendo o que foi solicitado: distribuição Normal é um tipo de distribuição de probabilidade contínua que é simétrica em torno da média, formando o famoso “formato de sino”. Em uma distribuição normal, a maioria dos valores está próxima da média, e valores mais extremos são cada vez mais raros. Essa distribuição é importante porque muitas análises estatísticas assumem normalidade dos dados.

6.1 Histograma para cada variável

6.1.1 Definir o número de bins com a fórmula de Sturges (aproximadamente)

```

n <- nrow(tech_filter)
bins_sturges <- ceiling(log2(n) + 1)

```

O número de bins foi determinado utilizando a fórmula de Sturges, considerando o tamanho da amostra. Essa escolha visa equilibrar a representação detalhada dos dados sem gerar sobrecarga visual no histograma.

6.1.2 Histograma + curva de densidade do número de usuários

```

tech %>%
  ggplot(aes(x = n_usuarios_internet)) +
  geom_histogram(aes(y = after_stat(density)), fill = "lightblue", bins = bins_sturges) +
  geom_density(color = "red", size = 1.2) +
  labs(
    title = "Distribuição do Número de Usuários de Internet",
    x = "Número de Usuários",
    y = "Densidade"
  ) +
  theme_minimal()

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.

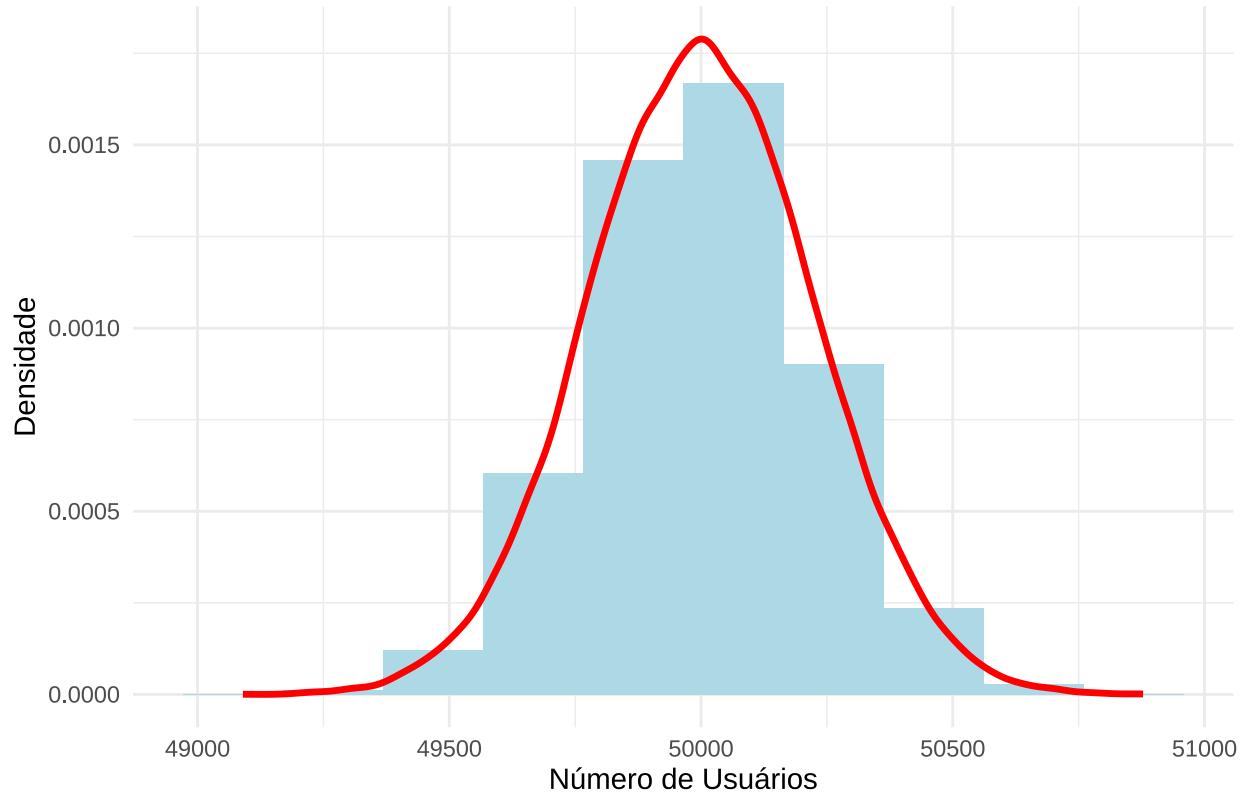
```

```

## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

```

Distribuição do Número de Usuários de Internet



O histograma e a curva de densidade indicam que a variável “Número de Usuários de Internet” apresenta uma distribuição **aproximadamente normal**, com formato simétrico e concentração de valores próximos à média. Essa análise visual reforça a avaliação de normalidade da variável para as próximas etapas do estudo.

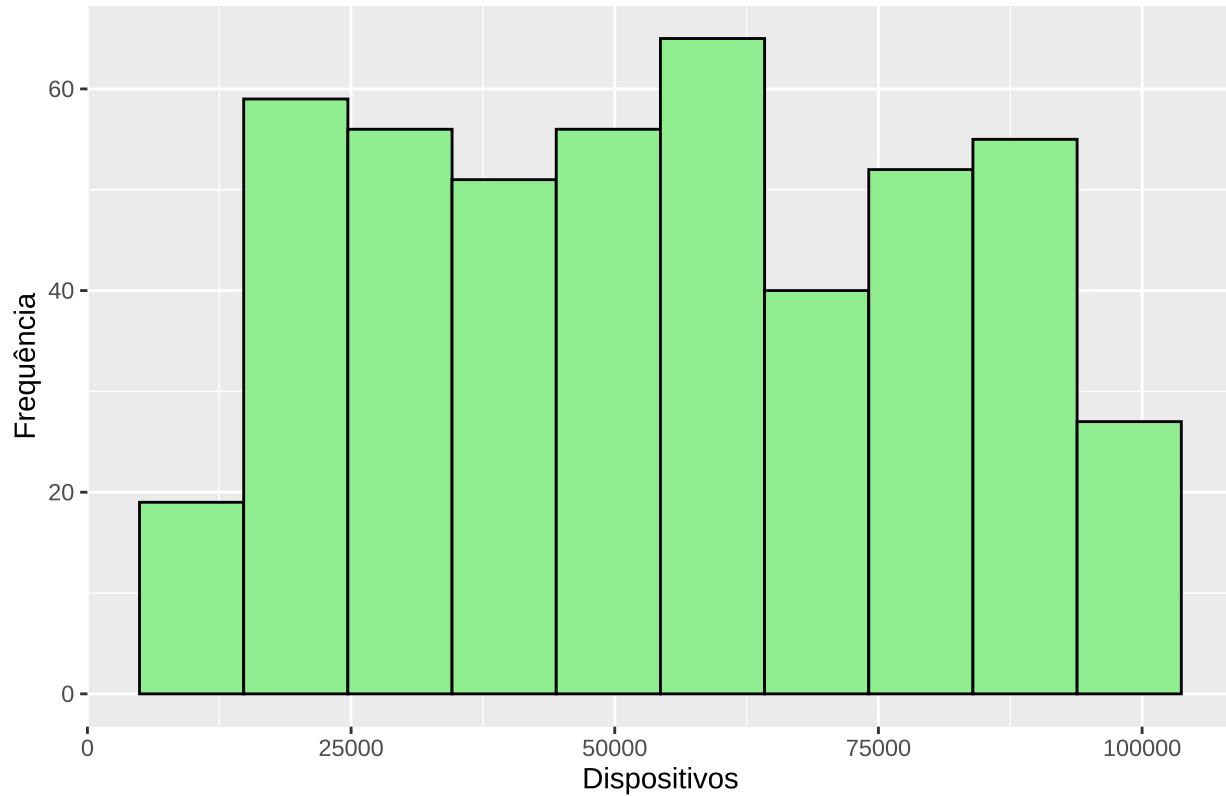
6.1.3 Histograma: Número de Dispositivos Conectados

```

ggplot(tech_filter, aes(x = n_dispositivos_conectados)) +
  geom_histogram(bins = bins_sturges, fill = "lightgreen", color = "black") +
  labs(title = "Histograma - Dispositivos Conectados", x = "Dispositivos", y = "Frequência")

```

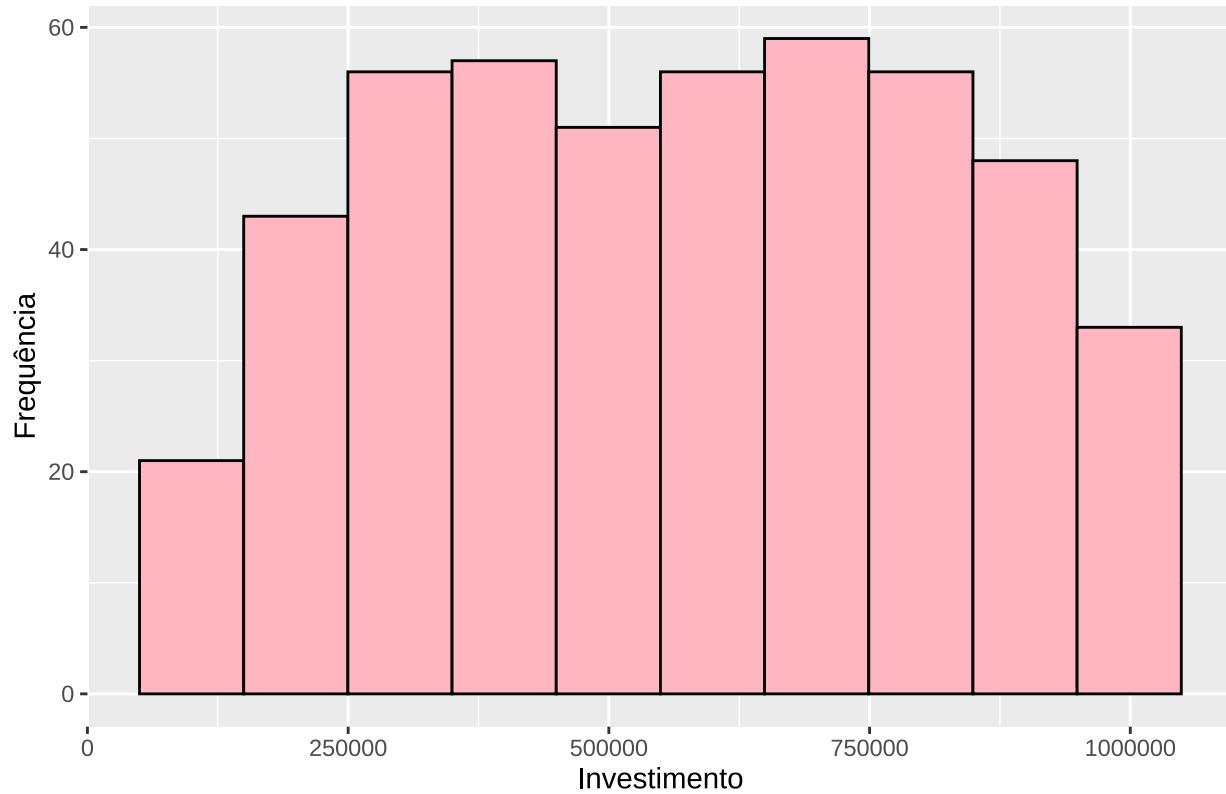
Histograma - Dispositivos Conectados



6.1.4 Histograma: Investimento Público

```
ggplot(tech_filter, aes(x = investimento_publico_tecnologia)) +  
  geom_histogram(bins = bins_sturges, fill = "lightpink", color = "black") +  
  labs(title = "Histograma - Investimento Público em Tecnologia", x = "Investimento", y = "Frequência")
```

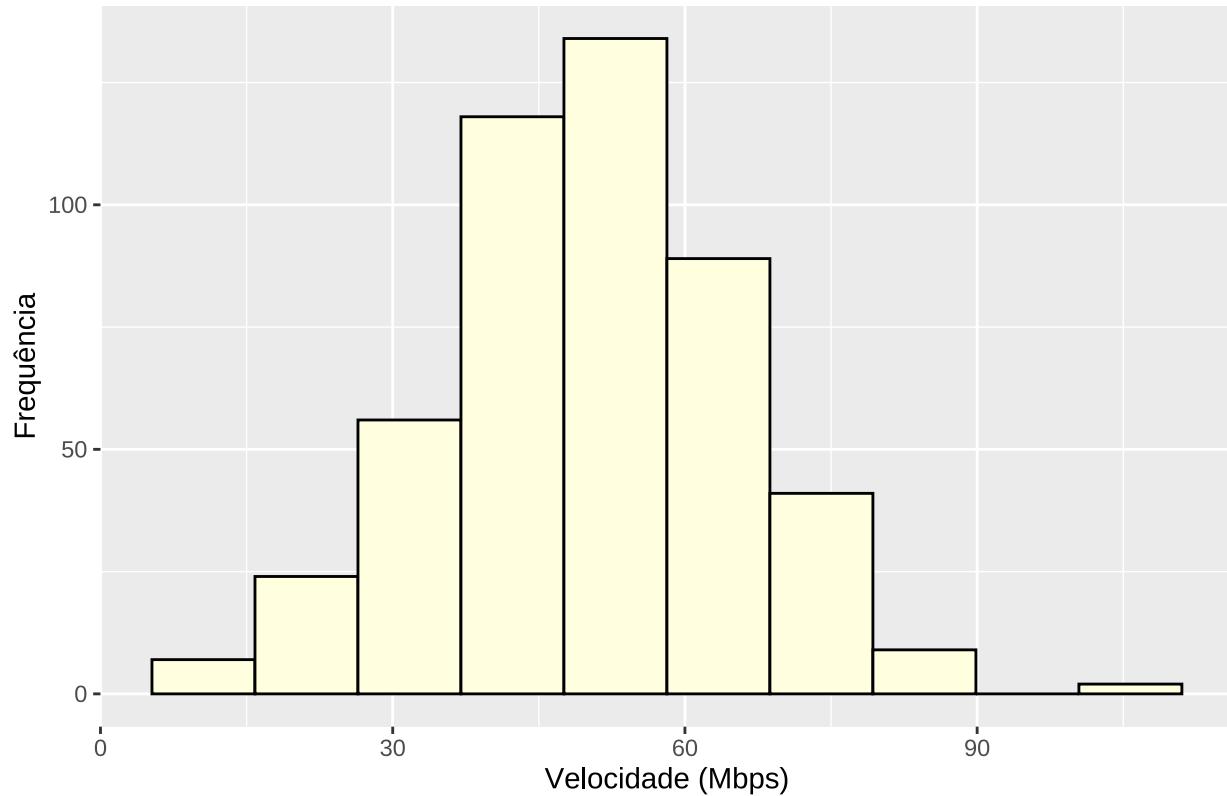
Histograma - Investimento Público em Tecnologia



6.1.5 Histograma: Velocidade Média (Mbps)

```
ggplot(tech_filter, aes(x = velocidade_media_mbps)) +  
  geom_histogram(bins = bins_sturges, fill = "lightyellow", color = "black") +  
  labs(title = "Histograma - Velocidade Média (Mbps)", x = "Velocidade (Mbps)", y = "Frequência")
```

Histograma - Velocidade Média (Mbps)



Número de Usuários de Internet: Apresenta distribuição **aproximadamente simétrica**, com formato semelhante a uma curva normal.

Número de Dispositivos Conectados: Apresenta distribuição **irregular e dispersa**, sem padrão de simetria claro, afastando-se da normalidade.

Investimento Público em Tecnologia: Distribuição assimétrica à direita, com cauda longa, indicando **afastamento da normalidade**.

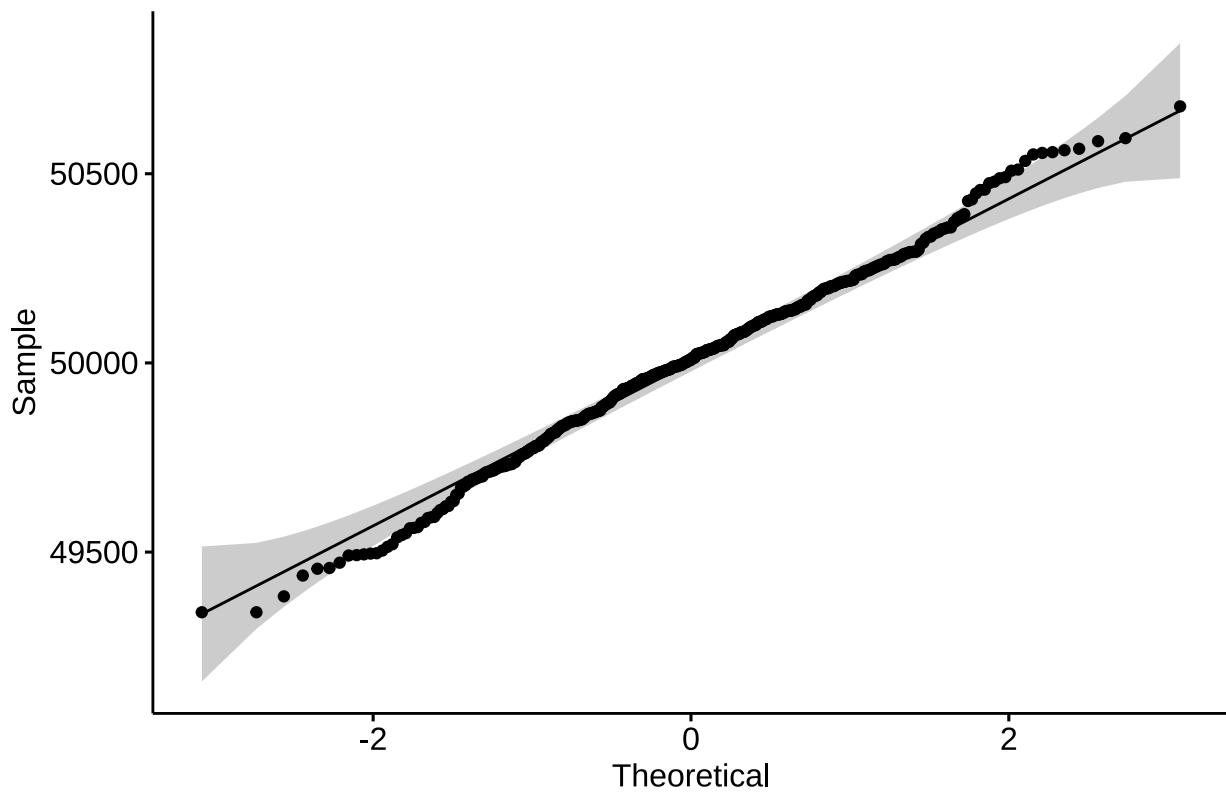
Velocidade Média (Mbps): Distribuição altamente assimétrica à direita, com concentração em valores mais baixos e cauda longa, caracterizando **desvio da normalidade**.

6.2 Gráficos Q-Q

6.2.1 Gráfico Q-Q: Número de Usuários

```
ggqqplot(tech_filter$n_usuarios_internet, title = "Q-Q Plot - Número de Usuários")
```

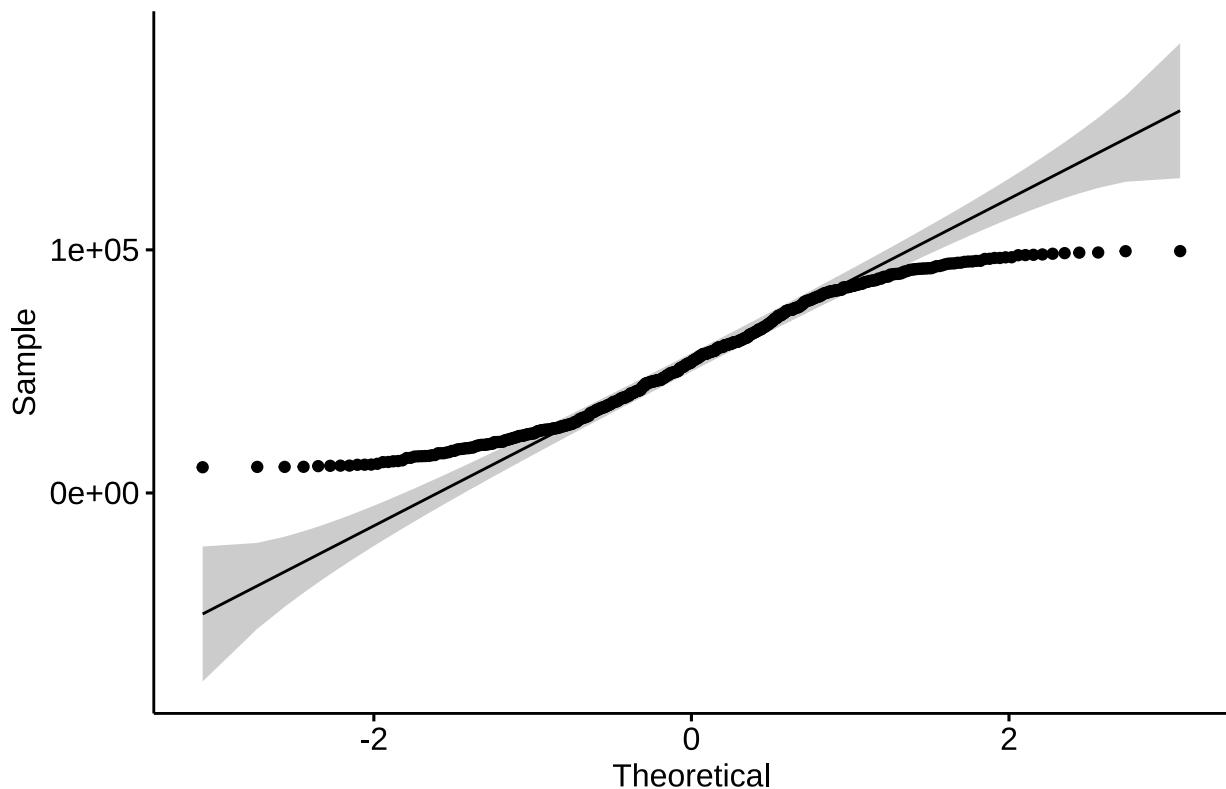
Q-Q Plot - Número de Usuários



6.2.2 Gráfico Q-Q: Número de Dispositivos Conectados

```
ggqqplot(tech_filter$n_dispositivos_conectados, title = "Q-Q Plot - Dispositivos Conectados")
```

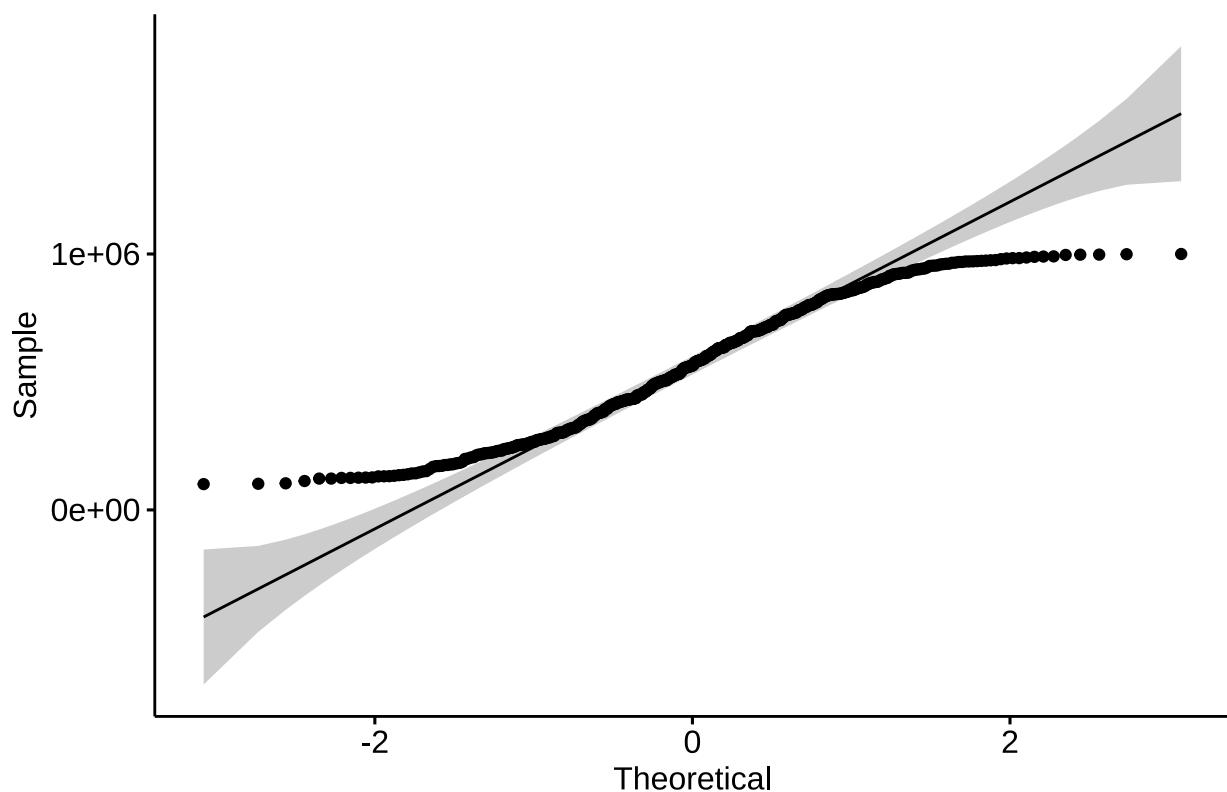
Q-Q Plot - Dispositivos Conectados



6.2.3 Gráfico Q-Q: Investimento Público

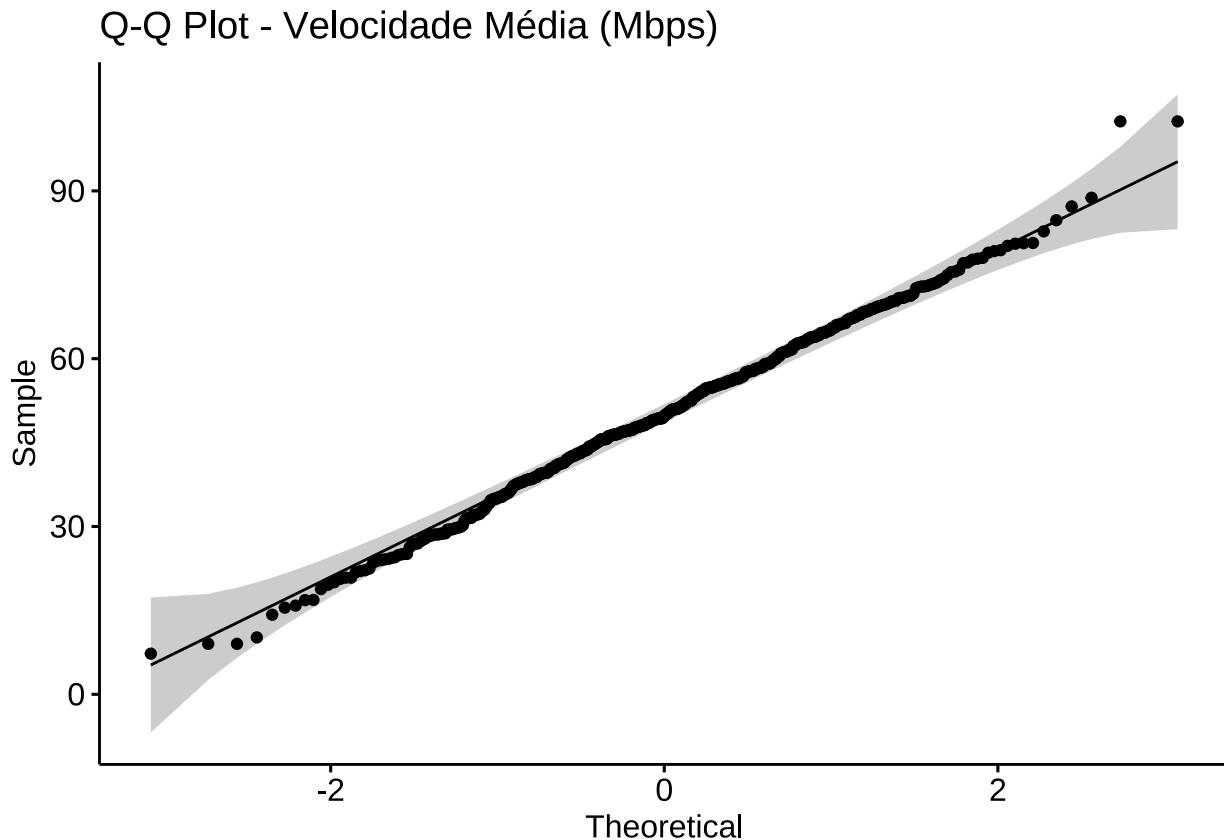
```
ggqqplot(tech_filter$investimento_publico_tecnologia, title = "Q-Q Plot - Investimento Público")
```

Q-Q Plot - Investimento Público



6.2.4 Gráfico Q-Q: Velocidade Média (Mbps)

```
ggqqplot(tech_filter$velocidade_media_mbps, title = "Q-Q Plot - Velocidade Média (Mbps)")
```



Usuários de Internet: distribuição **aproximadamente normal**.

Dispositivos Conectados: um **pouco afastado da normalidade**, mas próximo.

Investimento Público: forte **desvio da normalidade**.

Velocidade Média (Mbps): leve **desvio da normalidade**.

6.3 Teste de normalidade Shapiro-Wilk

```
shapiro.test(tech_filter$n_usuarios_internet)

##
##  Shapiro-Wilk normality test
##
## data: tech_filter$n_usuarios_internet
## W = 0.99467, p-value = 0.09462

shapiro.test(tech_filter$n_dispositivos_conectados)

##
##  Shapiro-Wilk normality test
##
## data: tech_filter$n_dispositivos_conectados
## W = 0.9537, p-value = 4.023e-11
```

```
shapiro.test(tech_filter$investimento_publico_tecnologia)
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: tech_filter$investimento_publico_tecnologia  
## W = 0.95925, p-value = 2.947e-10
```

```
shapiro.test(tech_filter$velocidade_media_mbps)
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: tech_filter$velocidade_media_mbps  
## W = 0.99675, p-value = 0.4506
```

Número de Usuários de Internet: - p-valor = 0,09462 - **Não rejeita** a normalidade (distribuição normal).

Número de Dispositivos Conectados: - p-valor = 4.023e-11 - **Rejeita** a normalidade (distribuição não normal).

Investimento Público em Tecnologia: - p-valor = 2.947e-10 - **Rejeita** a normalidade (distribuição não normal).

Velocidade Média (Mbps): - p-valor = 0.4506 - **Não rejeita** a normalidade (distribuição aproximadamente normal).

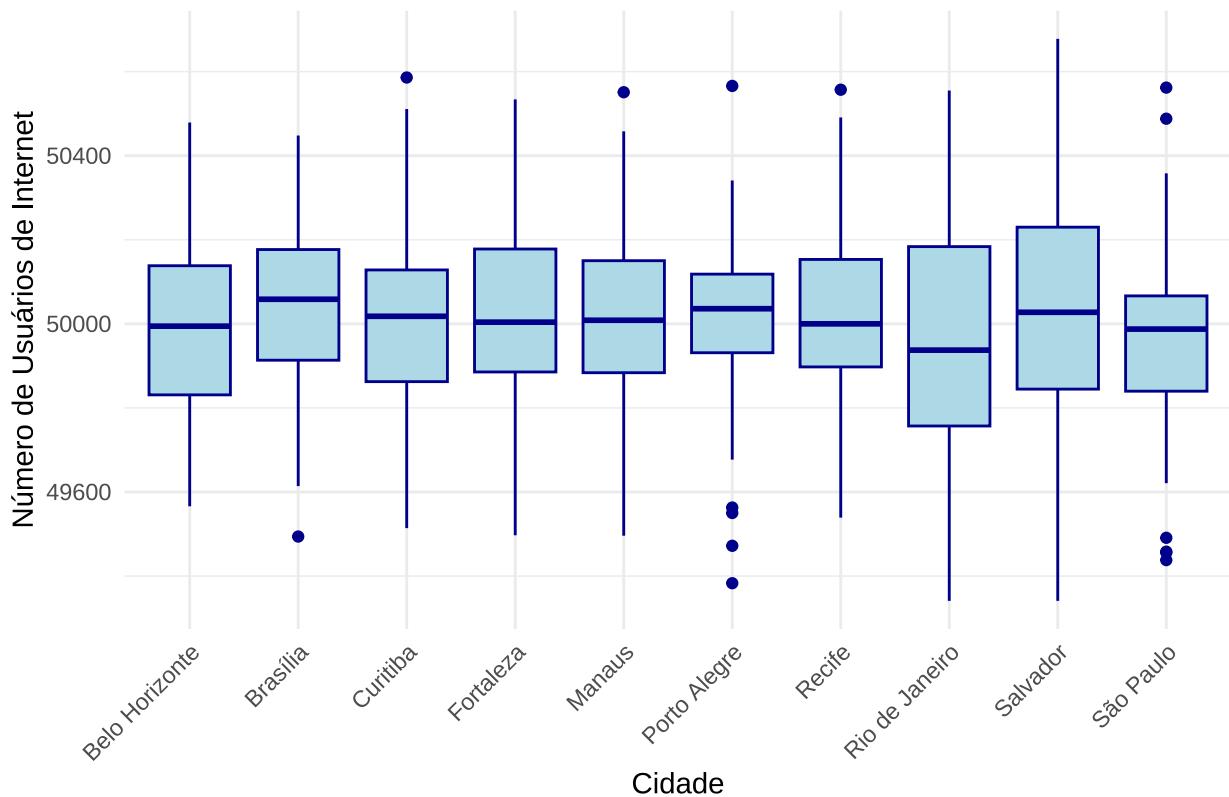
Com base na análise combinada dos histogramas, gráficos Q-Q e testes de Shapiro-Wilk, conclui-se que as variáveis **Número de Usuários de Internet** e **Velocidade Média (Mbps)** podem ser consideradas **aproximadamente normais**, enquanto **Número de Dispositivos Conectados** e **Investimento Público em Tecnologia** apresentam desvios que indicam a **ausência de normalidade**.

7 Descrição via boxplot e tabelas de contingência

7.1 Descrevendo a base com boxplot usuários de internet por cidade

```
ggplot(tech_filter, aes(x = cidade, y = n_usuarios_internet)) +  
  geom_boxplot(fill = "lightblue", color = "darkblue") +  
  labs(  
    title = "Distribuição do Número de Usuários de Internet por Cidade",  
    x = "Cidade",  
    y = "Número de Usuários de Internet"  
) +  
  theme_minimal() +  
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Distribuição do Número de Usuários de Internet por Cidade

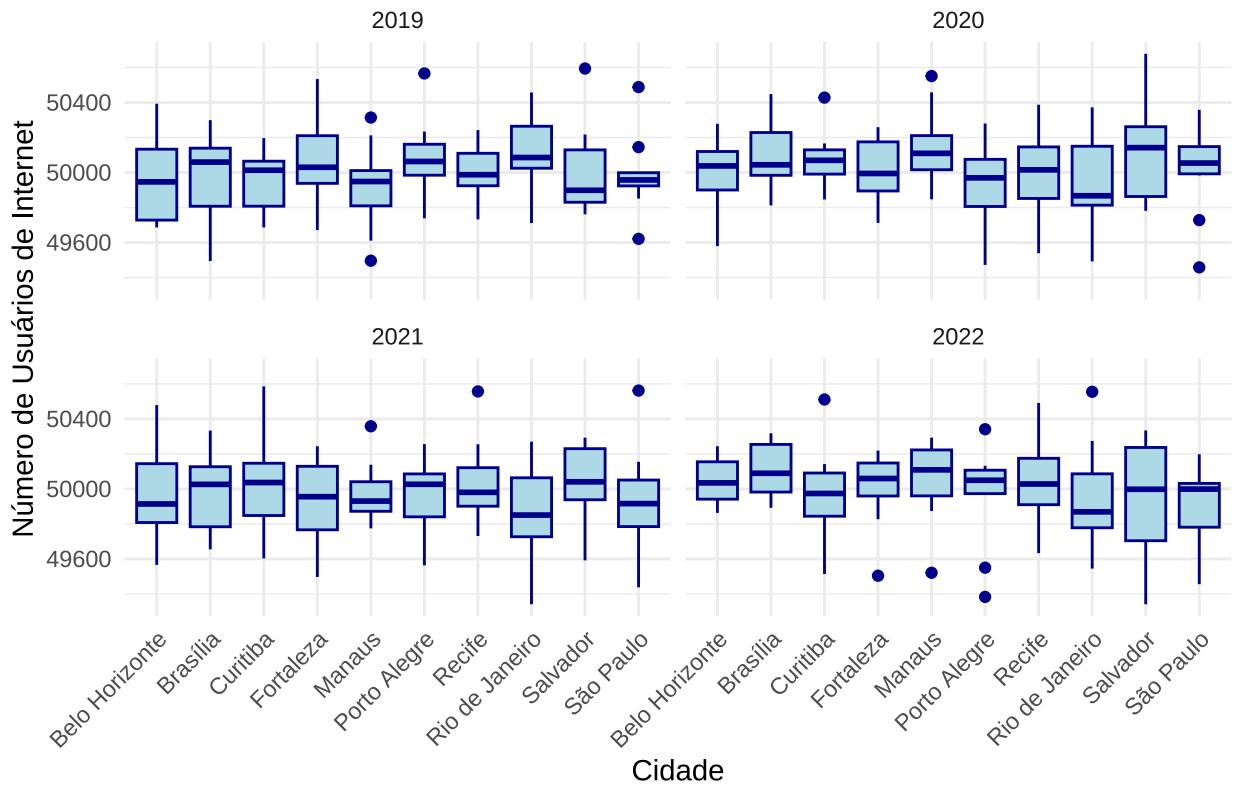


Para todas as cidades, o número de usuários de internet se concentra entre 49.600 e 50.400 aproximadamente. O volume de usuários de internet é relativamente estável entre as cidades analisadas. As medianas estão muito próximas de 50.000 usuários em quase todas as cidades. Salvador apresenta uma caixa mais ampla que outras cidades. Curitiba, São Paulo e Manaus têm uma distribuição mais concentrada. Cidades como Porto Alegre, Brasília e São Paulo mostram outliers. Esses pontos representam meses específicos em que o número de usuários foi anormalmente mais baixo (ou mais alto). Salvador e Rio de Janeiro têm mais dispersão do que cidades como Curitiba ou São Paulo.

7.2 Descrevendo a base com boxplot usuários de internet por cidade, separados por ano

```
ggplot(tech_filter, aes(x = cidade, y = n_usuarios_internet)) +
  geom_boxplot(fill = "lightblue", color = "darkblue") +
  labs(
    title = "Distribuição do Número de Usuários de Internet por Cidade (2019-2022)",
    x = "Cidade",
    y = "Número de Usuários de Internet"
  ) +
  facet_wrap(~ ano) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Distribuição do Número de Usuários de Internet por Cidade (2019–2022)

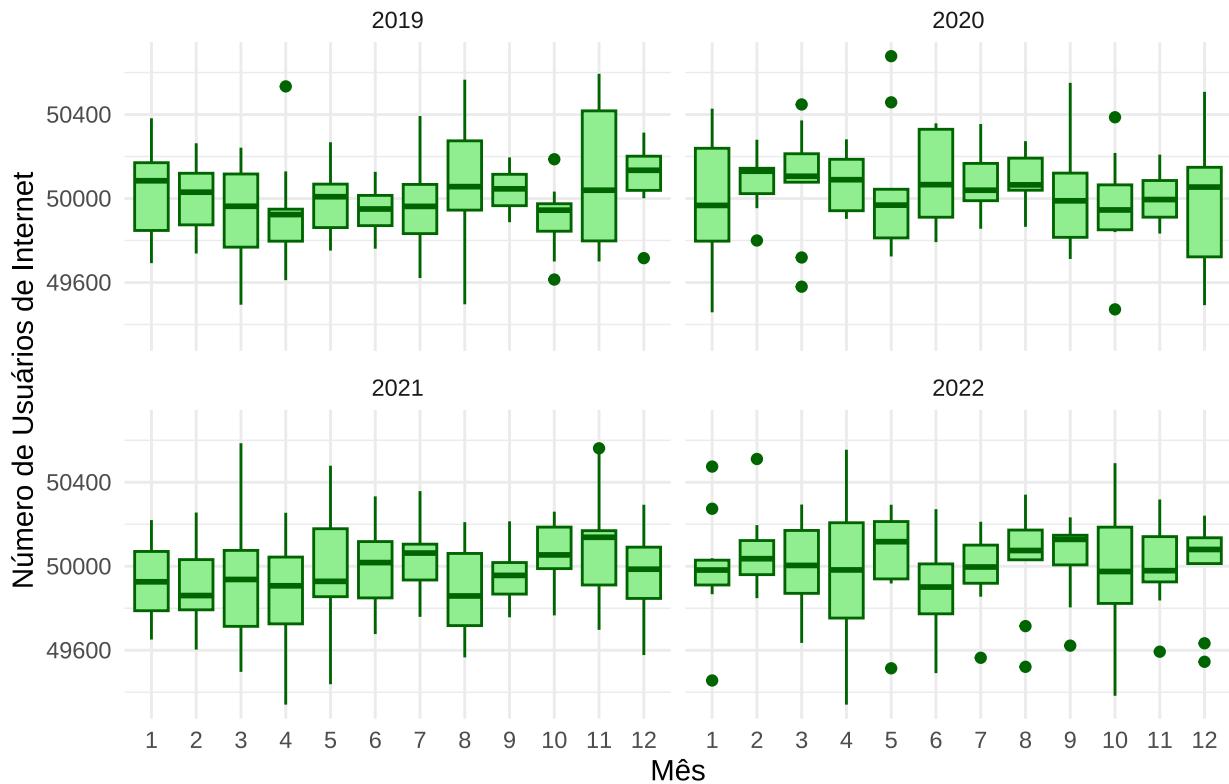


As medianas de cada cidade não mudam muito ao longo dos anos. A variação interna dentro das cidades muda um pouco, mas de maneira geral é moderada. Não houve grandes alterações no perfil geral de uso de internet entre 2019–2022, pelo menos em termos de número de usuários.

7.3 Descrevendo a base com boxplot usuários de internet por cidade, separados por mês e ano

```
ggplot(tech_filter, aes(x = as.factor(mes), y = n_usuarios_internet)) +
  geom_boxplot(fill = "lightgreen", color = "darkgreen") +
  labs(
    title = "Distribuição do Número de Usuários de Internet por Mês (2019–2022)",
    x = "Mês",
    y = "Número de Usuários de Internet"
  ) +
  facet_wrap(~ano) +
  theme_minimal()
```

Distribuição do Número de Usuários de Internet por Mês (2019–2022)



7.3.1 Analisando o desvio padrão mês a mês - métrica de dispersão

```
desvio_por_mes <- tech_filter %>%
  group_by(ano, mes) %>%
  summarise(
    desvio = sd(n_usuarios_internet, na.rm = TRUE)
  ) %>%
  group_by(mes) %>%
  summarise(media_desvio = mean(desvio)) %>%
  arrange(media_desvio)
```

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

```
desvio_por_mes
```

```
## # A tibble: 12 x 2
##       mes media_desvio
##   <dbl>      <dbl>
## 1     2      175.
## 2     9      179.
## 3     6      190.
## 4     7      190.
```

```

## 5    10      229.
## 6     8      232.
## 7    12      235.
## 8    11      242.
## 9     1      252.
## 10   5      253.
## 11   4      257.
## 12   3      264.

```

7.3.2 Analisando o intervalo interquartil mês a mês - métrica de dispersão

```

iqr_por_mes <- tech_filter %>%
  group_by(ano, mes) %>%
  summarise(
    iqr = IQR(n_usuarios_internet, na.rm = TRUE)
  ) %>%
  group_by(mes) %>%
  summarise(media_iqr = mean(iqr)) %>%
  arrange(media_iqr)

```

```

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

```

```
iqr_por_mes
```

```

## # A tibble: 12 x 2
##       mes media_iqr
##   <dbl>     <dbl>
## 1     9     187.
## 2     7     191.
## 3     2     192.
## 4    10     226.
## 5    12     239.
## 6     8     242.
## 7     5     259.
## 8     6     267.
## 9     3     286.
## 10    1     291.
## 11    4     292.
## 12   11     317.

```

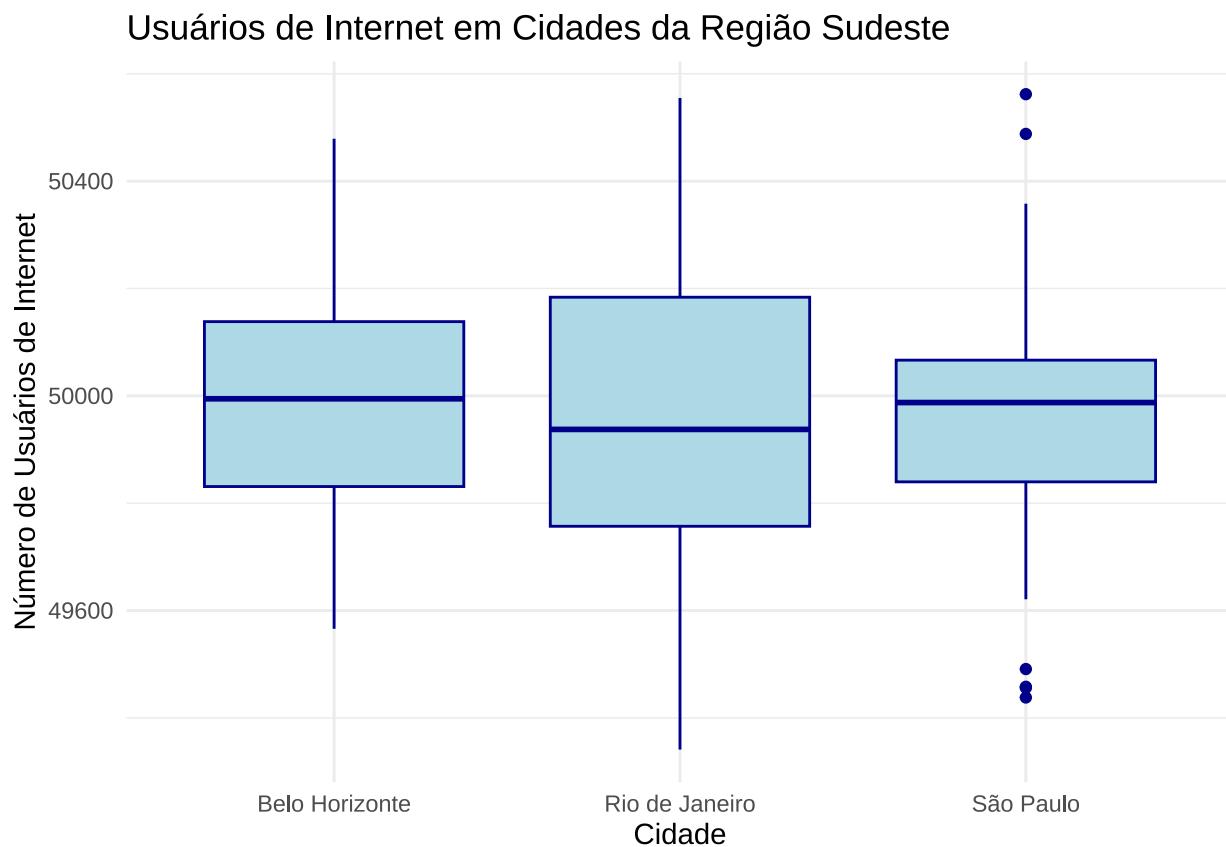
A partir da análise conjunta do desvio padrão médio e do intervalo interquartílico médio, ficou claro que os meses de Fevereiro e Setembro apresentam a menor variabilidade, tanto considerando toda a distribuição dos dados, quanto considerando apenas o seu centro.

7.4 Descrevendo a base com boxplot usuários de internet - usando apenas cidades da região Sudeste

```

tech_filter %>%
  filter(cidade %in% c("São Paulo", "Rio de Janeiro", "Belo Horizonte")) %>%
  ggplot(aes(x = cidade, y = n_usuarios_internet)) +
  geom_boxplot(fill = "lightblue", color = "darkblue") +
  labs(
    title = "Usuários de Internet em Cidades da Região Sudeste",
    x = "Cidade",
    y = "Número de Usuários de Internet"
  ) +
  theme_minimal()

```



As cidades selecionadas ainda mantêm a mesma concentração de usuários (medianas muito próximas de 50.000 usuários). Rio de Janeiro apresenta a maior dispersão entre as cidades. A caixa é mais larga, o que indica que há mais variação no número de usuários de internet ao longo dos meses. São Paulo apresenta uma dispersão pequena, parecida com Belo Horizonte.

São Paulo apresenta alguns outliers para baixo, indicando meses em que São Paulo teve menos usuários do que o esperado.

7.5 Descrevendo a base com boxplot usuários de internet por cidade, separados por ano - Região Sudeste

```

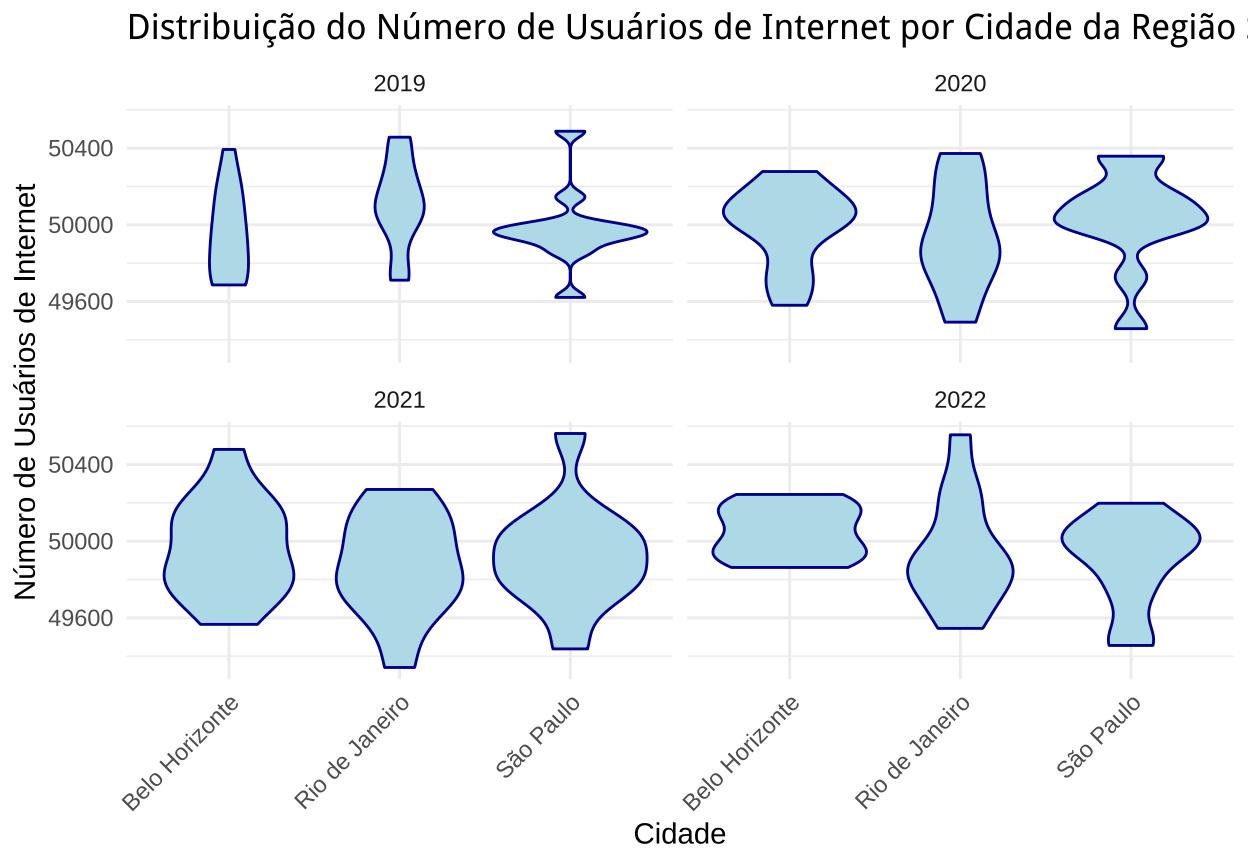
tech_filter %>%
  filter(cidade %in% c("São Paulo", "Rio de Janeiro", "Belo Horizonte")) %>%

```

```

ggplot(aes(x = cidade, y = n_usuarios_internet)) +
  geom_violin(fill = "lightblue", color = "darkblue") +
  labs(
    title = "Distribuição do Número de Usuários de Internet por Cidade da Região Sudeste (2019–2022)",
    x = "Cidade",
    y = "Número de Usuários de Internet"
  ) +
  facet_wrap(~ ano) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

```



O estudo ao longo dos anos de 2019 a 2022 revelou uma tendência de estabilidade nos valores medianos, em torno de 50.000 usuários. Belo Horizonte apresentou a menor variação ao longo dos anos, com distribuições mais compactas, enquanto o Rio de Janeiro exibiu maior dispersão no número de usuários. São Paulo apresentou alguns outliers principalmente em 2019 e 2020, indicando meses em que o número de usuários foi inferior ao esperado. De modo geral, as três cidades demonstraram uma relativa consistência no comportamento do número de usuários de internet, especialmente a partir de 2021, possivelmente refletindo a consolidação do acesso digital após o período mais crítico da pandemia.

```

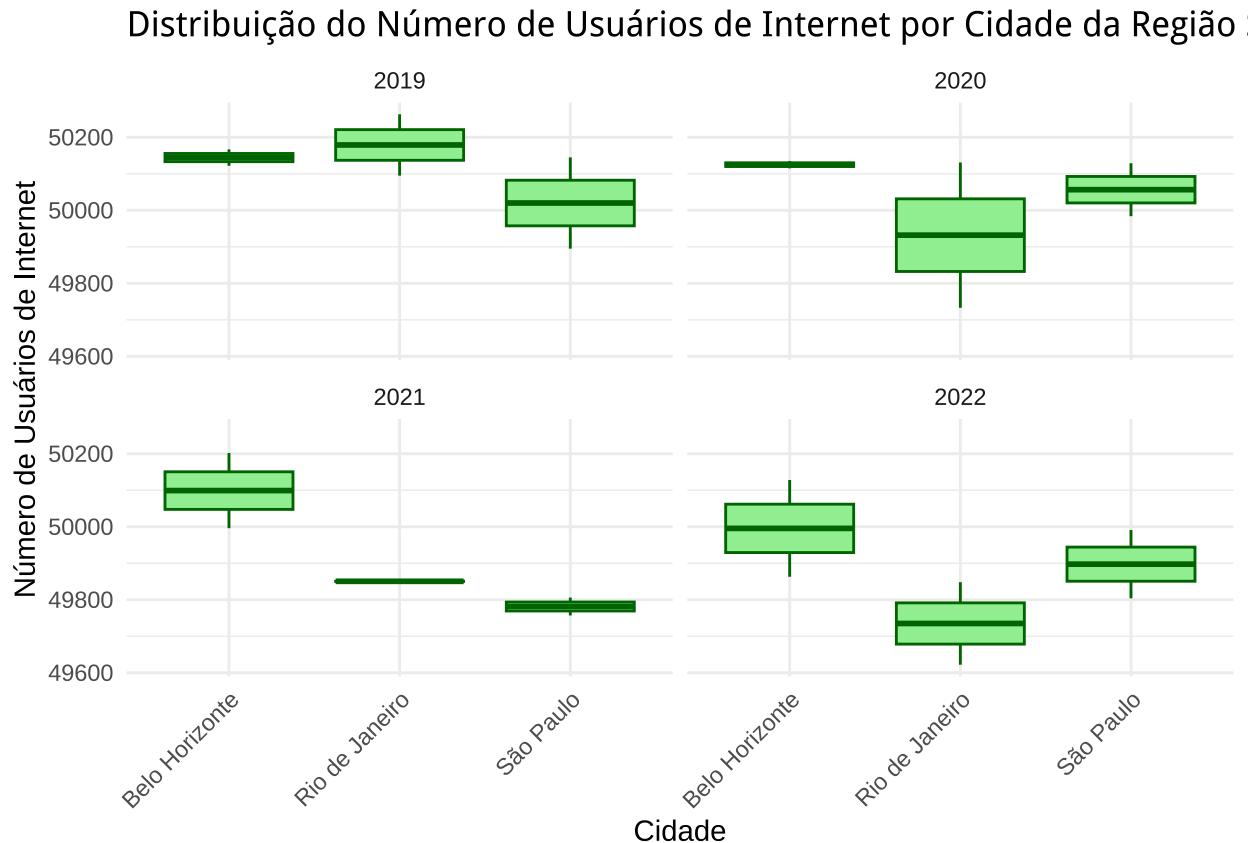
tech_filter %>%
  filter(
    cidade %in% c("São Paulo", "Rio de Janeiro", "Belo Horizonte"),
    mes %in% c(2, 9)
  ) %>%
  ggplot(aes(x = cidade, y = n_usuarios_internet)) +
  geom_boxplot(fill = "lightgreen", color = "darkgreen") +

```

```

  labs(
    title = "Distribuição do Número de Usuários de Internet por Cidade da Região Sudeste (02 e 09 / 2019-2022)",
    x = "Cidade",
    y = "Número de Usuários de Internet"
  ) +
  facet_wrap(~ ano) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

```



A escolha dos meses de Fevereiro e Setembro para análise se justifica pela maior estabilidade observada na distribuição do número de usuários de internet nas cidades do Sudeste ao longo de 2019 a 2022. Essa estabilidade, identificada por baixos valores de dispersão (desvio padrão e IQR), indica que esses meses refletem melhor o comportamento típico dos dados, minimizando efeitos sazonais e variações atípicas.

```

tech_filter %>%
  filter(
    cidade %in% c("São Paulo", "Rio de Janeiro", "Belo Horizonte"),
    mes %in% c(2, 9)
  ) %>%
  ggplot(aes(x = ano, y = n_usuarios_internet, color = cidade, group = cidade)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  labs(
    title = "Evolução do Número de Usuários de Internet em 02 e 09 (2019-2022)",
    x = "Ano",
    y = "Número de Usuários de Internet",
    subtitle = "Sudeste"
  )

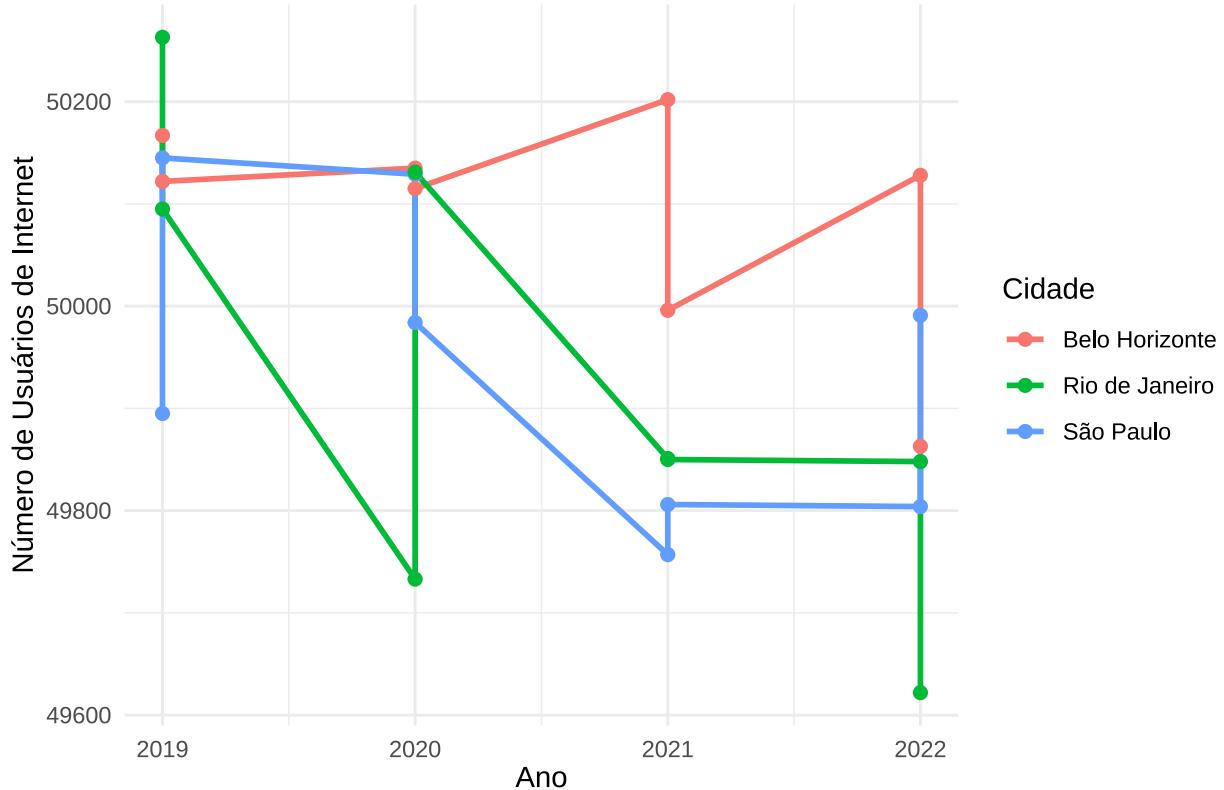
```

```

    color = "Cidade"
) +
theme_minimal() +
theme(
  plot.title = element_text(hjust = 0.5),
  axis.text.x = element_text(angle = 0, vjust = 0.5)
)

```

Evolução do Número de Usuários de Internet em 02 e 09 (2019–2022)



Em Belo Horizonte, houve uma variação moderada ao longo dos meses selecionados nos anos, com quedas em 2021 e 2022. O Rio de Janeiro teve grande alta no número de usuários em 2020 (possivelmente por crescimento do home office, ensino remoto, etc. provocados pela pandemia), porém após isso apresentou uma tendência de queda no número de usuários, sinalizando uma possível perda gradual de usuários ou outras alterações no perfil de conectividade da cidade. Já em São Paulo, a linha revela um comportamento um pouco mais estável, com variações pequenas ao longo dos quatro anos, indicando consistência no número de usuários de internet.

7.6 Analisando só as cidades principais do Sudeste, só nos meses de estabilidade (fev/set), para ver como a velocidade da internet evoluiu ao longo do tempo

```

tech_filter %>%
filter(
  cidade %in% c("São Paulo", "Rio de Janeiro", "Belo Horizonte"),
  mes %in% c(2, 9)
) %>%

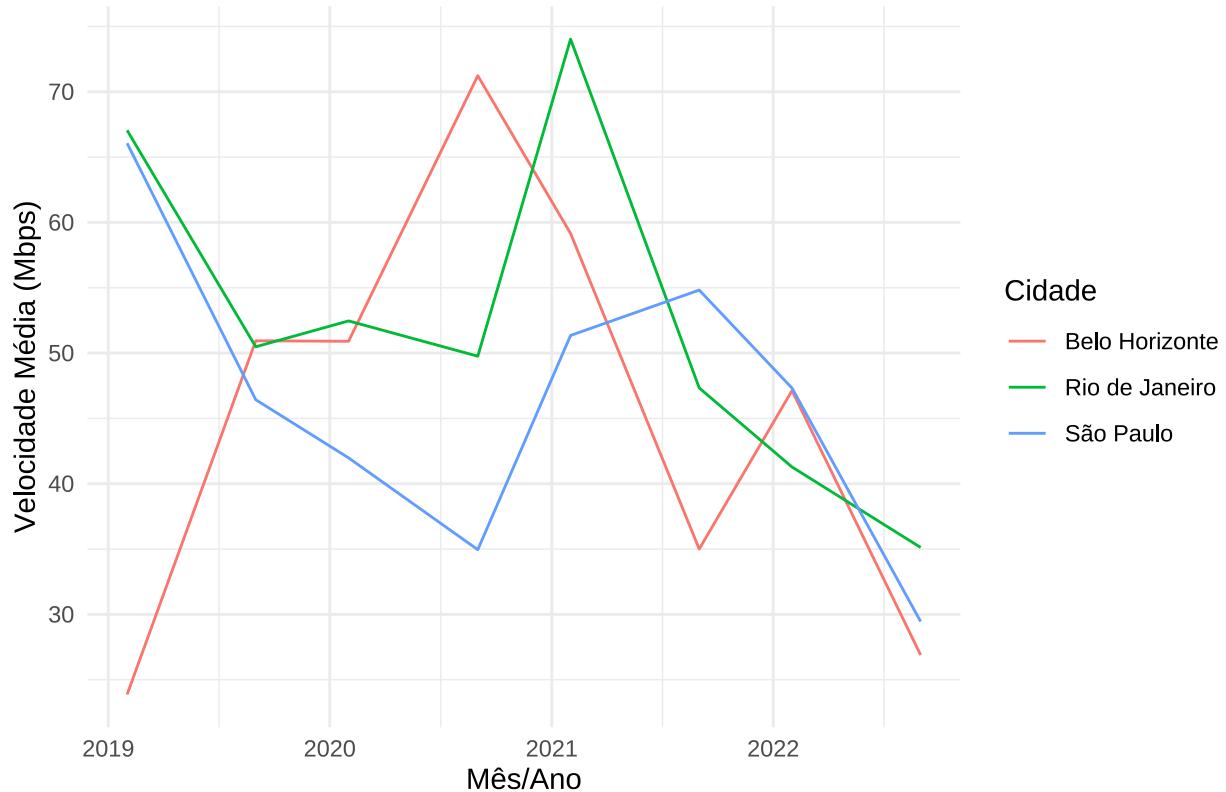
```

```

ggplot(aes(x = mes_ano, y = velocidade_media_mbps, color = cidade)) +
  geom_line() +
  labs(
    title = "Evolução da Velocidade Média de Internet no Sudeste (Fevereiro e Setembro)",
    x = "Mês/Ano",
    y = "Velocidade Média (Mbps)",
    color = "Cidade"
  ) +
  theme_minimal()

```

Evolução da Velocidade Média de Internet no Sudeste (Fevereiro e Setembro)



7.6.1 Tabelas de contingência

```

summarytools::ctable(x = tech_filter$mes_ano,
                     y = tech_filter$cidade,
                     prop = "t")

```

```

## Cross-Tabulation, Total Proportions
## mes_ano * cidade
## Data Frame: tech_filter
##
## -----
##          mes_ano      cidade      Belo Horizonte      Brasília      Curitiba      Fortaleza      Manaus      Porto A
## 1       2019-02      Belo Horizonte      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000
## 2       2019-09      Belo Horizonte      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000
## 3       2020-02      Belo Horizonte      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000
## 4       2020-09      Belo Horizonte      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000
## 5       2021-02      Belo Horizonte      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000
## 6       2021-09      Belo Horizonte      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000
## 7       2022-02      Belo Horizonte      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000
## 8       2022-09      Belo Horizonte      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000
## 9       2019-02      Rio de Janeiro      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000
## 10      2019-09      Rio de Janeiro      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000
## 11      2020-02      Rio de Janeiro      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000
## 12      2020-09      Rio de Janeiro      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000
## 13      2021-02      Rio de Janeiro      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000
## 14      2021-09      Rio de Janeiro      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000
## 15      2022-02      Rio de Janeiro      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000
## 16      2022-09      Rio de Janeiro      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000
## 17      2019-02      São Paulo      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000
## 18      2019-09      São Paulo      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000
## 19      2020-02      São Paulo      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000
## 20      2020-09      São Paulo      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000
## 21      2021-02      São Paulo      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000
## 22      2021-09      São Paulo      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000
## 23      2022-02      São Paulo      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000
## 24      2022-09      São Paulo      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000      0.00000000

```

##	2019-01-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2019-02-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2019-03-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2019-04-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2019-05-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2019-06-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2019-07-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2019-08-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2019-09-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2019-10-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2019-11-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2019-12-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2020-01-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2020-02-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2020-03-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2020-04-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2020-05-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2020-06-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2020-07-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2020-08-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2020-09-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2020-10-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2020-11-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2020-12-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2021-01-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2021-02-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2021-03-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2021-04-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2021-05-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2021-06-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2021-07-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2021-08-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2021-09-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2021-10-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2021-11-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2021-12-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2022-01-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2022-02-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2022-03-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2022-04-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2022-05-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2022-06-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2022-07-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2022-08-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2022-09-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2022-10-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2022-11-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	2022-12-01	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)	1 (0.2%)
##	Total	48 (10.0%)	48 (10.0%)	48 (10.0%)	48 (10.0%)	48 (10.0%)	48 (10.0%)
##	-----	-----	-----	-----	-----	-----	-----

A tabela de contingência construída a partir das variáveis mes_ano e cidade evidencia que há uma observação correspondente a cada cidade em cada mês/ano no período analisado (2019–2022). Cada célula da tabela indica a existência de um registro (1 ocorrência), representando 0,2% do total de dados. Esses resultados

demonstram que a base de dados está completa no que diz respeito à combinação entre cidade e período temporal, não havendo lacunas nos registros analisados.

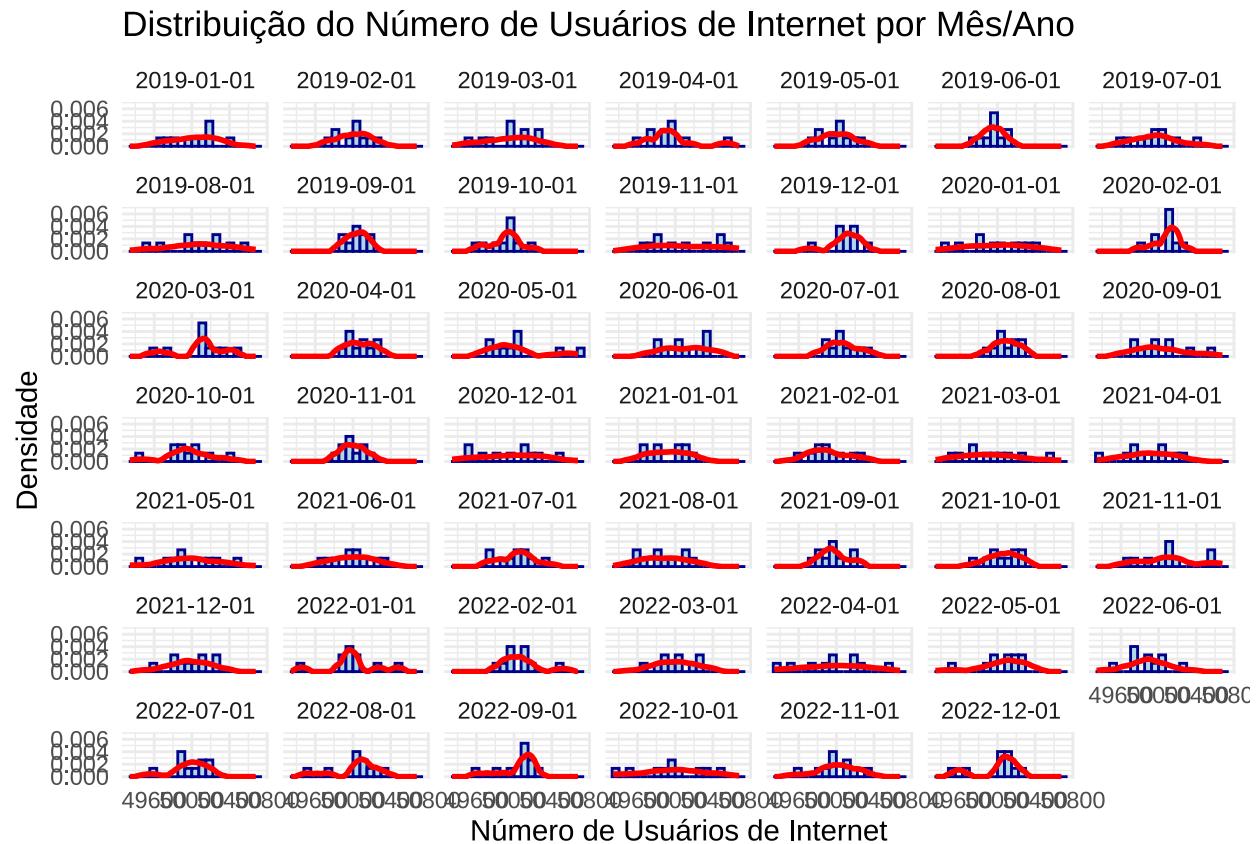
7.7 Aplicando estimativa de densidade via Kernel, através do Kernel de Epanechnikov

7.7.1 filtrando os dados e visualizando distâncias com Kernel

```
#Calculando os histogramas para o mesmo evento em diferentes instantes de tempo

fd_binwidth <- 2 * IQR(tech_filter$n_usuarios_internet, na.rm = TRUE) / (length(tech_filter$n_usuarios_...)

tech_filter %>%
  ggplot(aes(x = n_usuarios_internet)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = fd_binwidth, fill = 'lightblue', color = "darkred",
  geom_density(kernel = 'epanechnikov', color = "red", size = 1) +
  facet_wrap(~ mes_ano) +
  labs(
    title = "Distribuição do Número de Usuários de Internet por Mês/Ano",
    x = "Número de Usuários de Internet",
    y = "Densidade"
  ) +
  theme_minimal()
```



Em geral, a distribuição do número de usuários não é perfeitamente normal (não forma um “sino” simétrico perfeito em vários meses). Alguns meses mostram uma distribuição mais centralizada (picos definidos e bem no centro). Outros meses mostram uma distribuição mais espalhada ou com mais de um pico. Alguns meses, em alguns anos (por exemplo, fevereiro e setembro, como já havia identificado antes) tendem a apresentar curvas de densidade mais suaves e menos distorcidas — o que reforça que esses meses têm comportamento mais estável entre as cidades.

8 Filtrar o espaço amostral

Agora, para o Evento aleatório de escolher uma cidade/mês em que a velocidade média da internet foi superior a 50 Mbps, necessito filtrar o espaço amostral para o meu interesse.

```
sudeste_ufs <- c("São Paulo", "Rio de Janeiro", "Belo Horizonte")  
  
tech_sudeste <- tech_filter %>%  
  filter(cidade %in% sudeste_ufs, velocidade_media_mbps > 50)
```

8.1 Estatísticas descritivas

```
summary(tech_sudeste$n_usuarios_internet)
```

```
##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.  
## 49341   49756   49998   49971   50138   50555
```

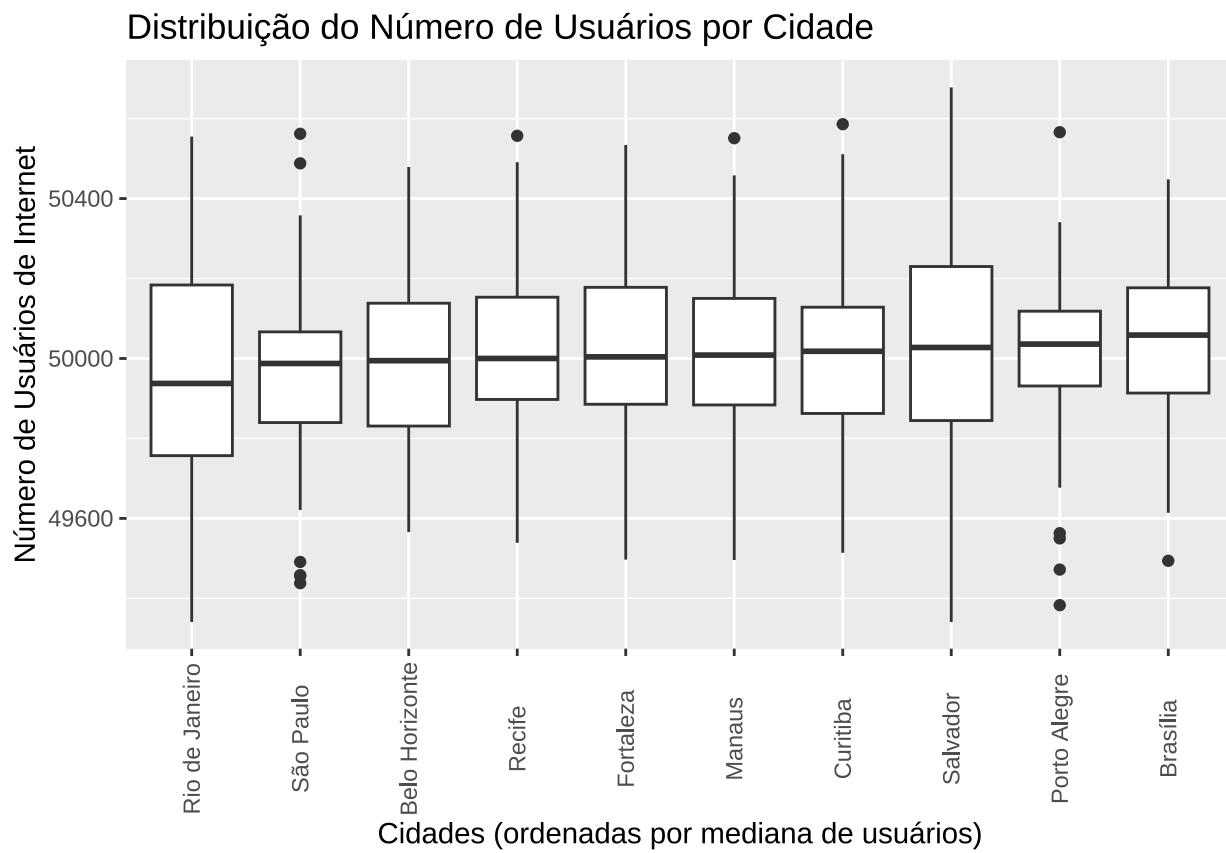
O número de usuários de internet para as cidades do Sudeste com velocidade média superior a 50 Mbps apresenta uma distribuição bastante concentrada em torno de 50.000 usuários, com média e mediana muito próximas. A variação entre o menor e o maior número de usuários é pequena, indicando baixa dispersão dos dados.

8.1.1 calculando as medianas

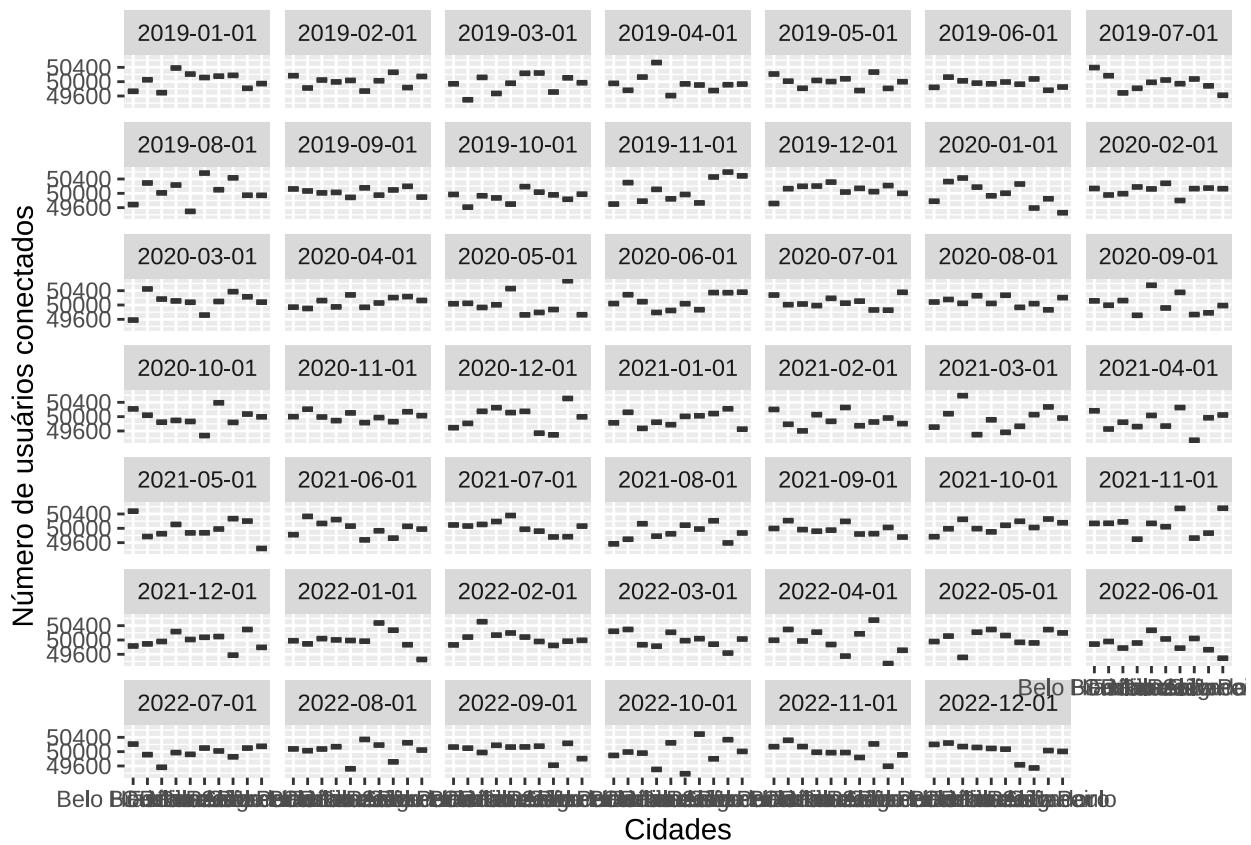
```
tech_filter %>% dplyr::group_by(mes_ano) %>% dplyr::summarize(n_users_md = median(n_usuarios_internet))  
  
## # A tibble: 48 x 2  
##       mes_ano     n_users_md  
##       <date>        <dbl>  
## 1 2019-01-01     50084.  
## 2 2019-02-01     50030  
## 3 2019-03-01     49963  
## 4 2019-04-01     49924.  
## 5 2019-05-01     50008.  
## 6 2019-06-01     49950  
## 7 2019-07-01     49962.  
## 8 2019-08-01     50056.  
## 9 2019-09-01     50046  
## 10 2019-10-01    49944.  
## # i 38 more rows
```

9 Discussão da relação entre variáveis quantitativas e qualitativas

```
tech_filter %>%
  ggplot(aes(x = reorder(cidade, n_usuarios_internet, median), y = n_usuarios_internet)) +
  geom_boxplot() +
  xlab('Cidades (ordenadas por mediana de usuários)') +
  ylab('Número de Usuários de Internet') +
  ggtitle('Distribuição do Número de Usuários por Cidade') +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5))
```



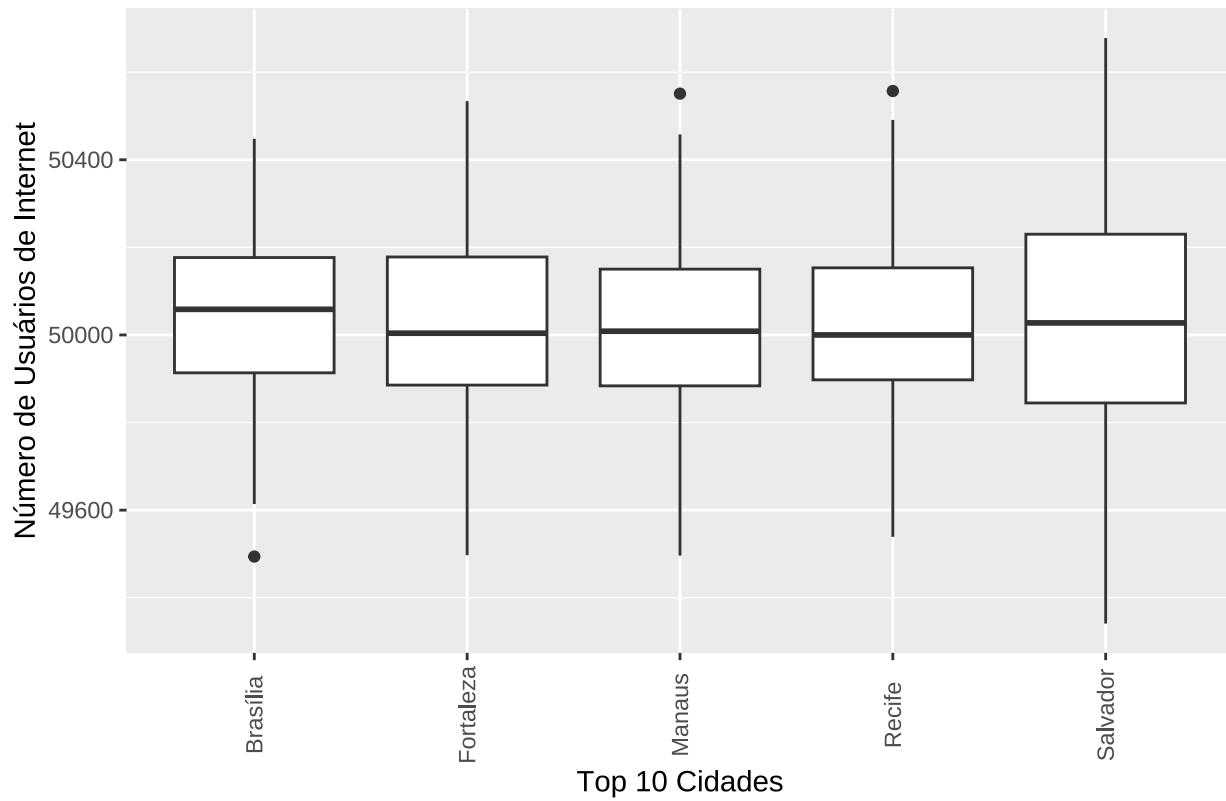
```
### Por mês
tech_filter %>% ggplot(aes(x = as.factor(cidade))) + geom_boxplot(aes(y = n_usuarios_internet)) + xlab('')
```



```
### 5 cidades com mais usuários
top_cidades <- tech_filter %>%
  group_by(cidade) %>%
  summarise(media_usuarios = mean(n_usuarios_internet, na.rm = TRUE)) %>%
  top_n(5, media_usuarios) %>%
  pull(cidade)

tech_filter %>%
  filter(cidade %in% top_cidades) %>%
  ggplot(aes(x = cidade, y = n_usuarios_internet)) +
  geom_boxplot() +
  xlab('Top 10 Cidades') +
  ylab('Número de Usuários de Internet') +
  ggtitle('Top 5 Cidades - Número de Usuários de Internet') +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5))
```

Top 5 Cidades - Número de Usuários de Internet



Calculando o R ao quadrado para a relação de usuários conectados por região

```
tech_filter %>%
  dplyr::group_by(mes_ano) %>%
  do(modelo = lm(n_usuarios_internet ~ as.factor(cidade), .)) %>%
  dplyr::mutate(r.ao.quadrado = summary(modelo)$r.squared) %>%
  dplyr::select(-modelo) %>%
  kable()
```

mes_ano	r.ao.quadrado
2019-01-01	1
2019-02-01	1
2019-03-01	1
2019-04-01	1
2019-05-01	1
2019-06-01	1
2019-07-01	1
2019-08-01	1
2019-09-01	1
2019-10-01	1
2019-11-01	1
2019-12-01	1
2020-01-01	1
2020-02-01	1
2020-03-01	1

mes_ano	r.ao.quadrado
2020-04-01	1
2020-05-01	1
2020-06-01	1
2020-07-01	1
2020-08-01	1
2020-09-01	1
2020-10-01	1
2020-11-01	1
2020-12-01	1
2021-01-01	1
2021-02-01	1
2021-03-01	1
2021-04-01	1
2021-05-01	1
2021-06-01	1
2021-07-01	1
2021-08-01	1
2021-09-01	1
2021-10-01	1
2021-11-01	1
2021-12-01	1
2022-01-01	1
2022-02-01	1
2022-03-01	1
2022-04-01	1
2022-05-01	1
2022-06-01	1
2022-07-01	1
2022-08-01	1
2022-09-01	1
2022-10-01	1
2022-11-01	1
2022-12-01	1

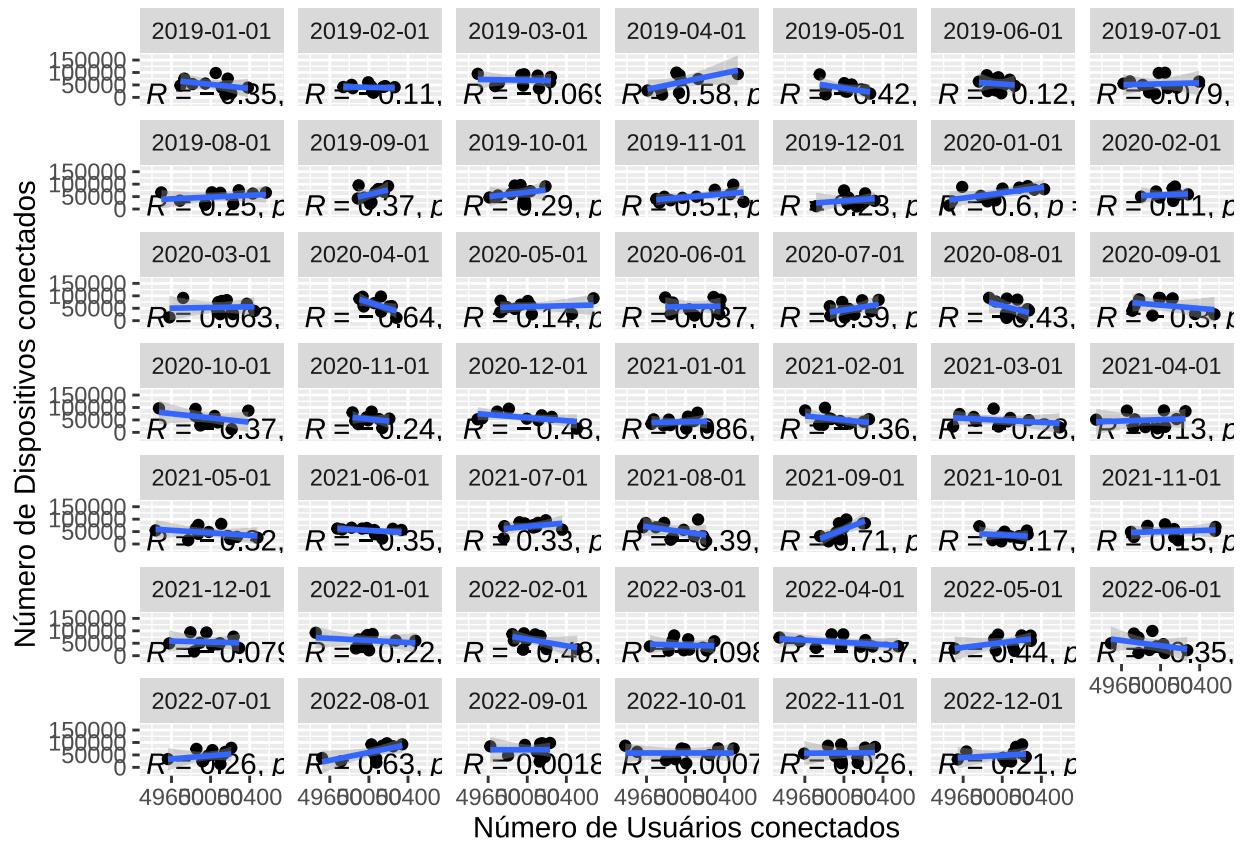
Cada cidade tem só um valor de n_usuarios_internet em cada mes_ano, então o modelo consegue mostrar perfeitamente.

10 Calculando a dispersão e as correlações de Pearson e de Spearman com duas variáveis

10.1 Pearson

```
tech_filter %>% ggplot(aes(x = n_usuarios_internet, y = n_dispositivos_conectados)) +geom_point() + facet_wrap(~cidade)

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



10.2 Spearman

```
tech_filter %>% ggplot(aes(x = n_usuarios_internet, y = n_dispositivos_conectados)) +geom_point() +facet_wrap(~date)

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

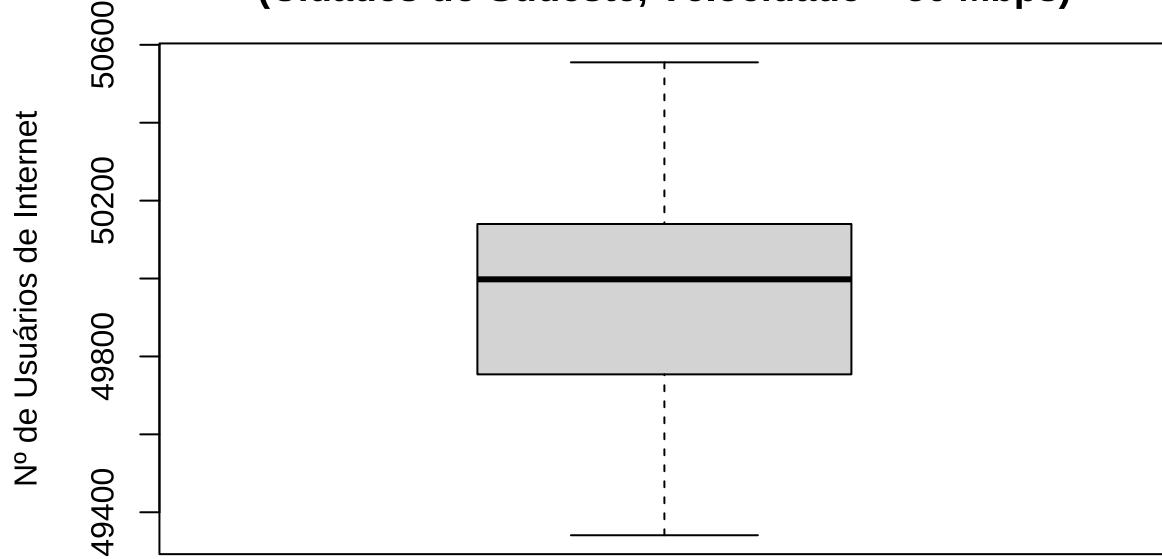


Pearson mostra que, linearmente, o número de usuários não explica bem a variação no número de dispositivos conectados, na maioria dos meses. Spearman também mostra correlações fracas (a maioria dos R perto de 0). Mesmo olhando para relações que poderiam ser monotônicas (não exatamente linhas retas), não encontrei relações fortes e consistentes entre usuários e dispositivos.

10.3 Boxplot para ver distribuição apenas do Sudeste, com velocidade > 50 mbps

```
boxplot(tech_sudeste$n_usuarios_internet,
        main = "Distribuição do Número de Usuários de Internet\n(Cidades do Sudeste, Velocidade > 50 Mbp",
        ylab = "Nº de Usuários de Internet")
```

Distribuição do Número de Usuários de Internet (Cidades do Sudeste, Velocidade > 50 Mbps)

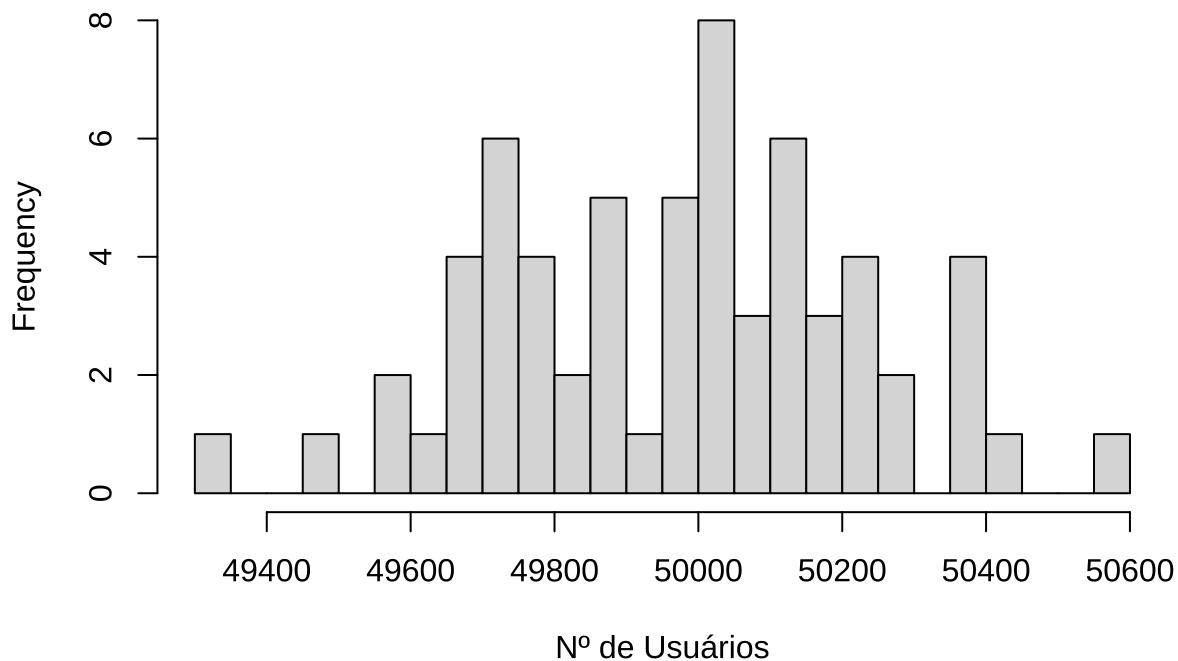


O boxplot do número de usuários de internet nas cidades do Sudeste com velocidade superior a 50 Mbps revela uma distribuição simétrica e com baixa dispersão. A maioria dos registros se concentra próximo de 50.000 usuários, sem a presença de outliers.

10.4 Histograma da base filtrada para o evento que estou estudando (usuários de internet nas cidades do Sudeste com velocidade superior a 50 Mbps)

```
hist(tech_sudeste$n_usuarios_internet,  
     main = "Histograma: Usuários de Internet (Sudeste, Velocidade > 50 Mbps)",  
     xlab = "Nº de Usuários",  
     breaks = 30)
```

Histograma: Usuários de Internet (Sudeste, Velocidade > 50 Mbps)



O histograma do número de usuários de internet nas cidades do Sudeste com velocidade maior que 50 Mbps mostra uma distribuição aproximadamente simétrica, com maior concentração de registros próximos a 50.000 usuários. A dispersão é pequena, indicando pouca variação no número de usuários entre as cidades analisadas.

11 Calculando a probabilidade do evento que estou investigando

Cidade/mês do Sudeste onde a velocidade média de internet foi superior a 50 Mbps

```
total_registros_sudeste <- nrow(tech_sudeste)  
total_registros_geral <- nrow(tech_filter)  
probabilidade_evento <- total_registros_sudeste / total_registros_geral  
probabilidade_evento  
  
## [1] 0.1333333
```

A chance de escolher aleatoriamente uma cidade/mês do Brasil onde a velocidade média de internet seja superior a 50 Mbps e esteja localizada na Região Sudeste é de aproximadamente 13,33%.

12 Teste de hipóteses

H : “A velocidade média da internet nas cidades do Sudeste é igual ou inferior a 50 Mbps.”

H : “A velocidade média da internet nas cidades do Sudeste é superior a 50 Mbps.”

```
t.test(tech_sudeste$velocidade_media_mbps, mu = 50, alternative = "greater")
```

```
##  
##  One Sample t-test  
##  
## data: tech_sudeste$velocidade_media_mbps  
## t = 9.2029, df = 63, p-value = 1.444e-13  
## alternative hypothesis: true mean is greater than 50  
## 95 percent confidence interval:  
## 59.29151      Inf  
## sample estimates:  
## mean of x  
## 61.35047
```

Com um p-valor extremamente pequeno ($p < 0,01$), **rejeitamos a hipótese nula de que a velocidade média da internet nas cidades do Sudeste é igual ou inferior a 50 Mbps a qualquer nível de significância**. Assim, temos evidências estatísticas para aceitar a hipótese alternativa de que a velocidade média é superior a 50 Mbps.

12.0.1 Teste de Wilcoxon para checar o pareamento (antes e depois de 2020 - fenômeno pandemia)

Realizado o teste de Wilcoxon para checar o pareamento entre as distribuições de número de usuários do pré e do pós pandemia.

H : “A mediana do número de usuários de internet é igual antes e depois de 2020.”

H : “A mediana do número de usuários de internet é diferente antes e depois de 2020.”

```
usuarios_anteriores_2020 <- tech_filter %>%  
  filter(as.numeric(substr(mes_ano, 1, 4)) < 2020) %>%  
  pull(n_usuarios_internet)  
  
usuarios_depois_2020 <- tech_filter %>%  
  filter(as.numeric(substr(mes_ano, 1, 4)) >= 2020) %>%  
  pull(n_usuarios_internet)  
  
wilcox.test(usuarios_anteriores_2020, usuarios_depois_2020, paired = FALSE, exact = FALSE)  
  
##  
##  Wilcoxon rank sum test with continuity correction  
##  
## data: usuarios_anteriores_2020 and usuarios_depois_2020  
## W = 20964, p-value = 0.6289  
## alternative hypothesis: true location shift is not equal to 0
```

Neste caso, não rejeitamos a hipótese nula. Não há evidências suficientes para afirmar que a mediana do número de usuários de internet mudou antes e depois de 2020.

12.0.2 Teste de Wilcoxon para checar o pareamento - maior que (antes e depois de 2020 - fenômeno pandemia)

H : “A hipótese nula aqui é: A mediana do número de usuários de internet é igual antes e depois de 2020.”

H : “A hipótese alternativa: A mediana das diferenças (antes - depois) é maior que 0.”

```
wilcox.test(usuarios_anteriores_2020, usuarios_depois_2020, paired = FALSE, alternative = "greater", exact = TRUE)

##
##  Wilcoxon rank sum test with continuity correction
##
## data: usuarios_anteriores_2020 and usuarios_depois_2020
## W = 20964, p-value = 0.6858
## alternative hypothesis: true location shift is greater than 0
```

Neste caso, mais uma vez, não rejeitamos a hipótese nula. Não há evidências suficientes para afirmar que a mediana do número de usuários antes de 2020 seja maior que a mediana depois de 2020.

12.0.3 Teste de Wilcoxon para checar o pareamento - menor que (antes e depois de 2020 - fenômeno pandemia)

H : “A mediana do número de usuários de internet é igual antes e depois de 2020.”

H : “A mediana das diferenças (antes - depois) é menor que 0.”

```
wilcox.test(usuarios_anteriores_2020, usuarios_depois_2020, paired = FALSE, alternative = "less", exact = TRUE)

##
##  Wilcoxon rank sum test with continuity correction
##
## data: usuarios_anteriores_2020 and usuarios_depois_2020
## W = 20964, p-value = 0.3144
## alternative hypothesis: true location shift is less than 0
```

Finalmente, não temos evidências suficientes para afirmar que o número de usuários antes de 2020 era menor do que depois. Portanto, mais uma vez, não rejeitamos a hipótese nula.

12.0.4 Resumo dos testes de Wilcoxon

```
tabela_testes <- data.frame(
  Teste = c("Bilateral (two-sided)", "Unilateral (greater)", "Unilateral (less)"),
  `Hipótese Alternativa` = c(
    "Mediana antes = Mediana depois",
    "Mediana antes > Mediana depois",
    "Mediana antes < Mediana depois"
  ),
  `p-valor` = c(0.6289, 0.6858, 0.3144),
  Conclusão = c(
    "Não rejeita H (sem diferença significativa)",
```

```

    "Não rejeita H (sem evidência de que era maior)",
    "Não rejeita H (sem evidência de que aumentou depois)"
)
)

kable(tabela_testes, caption = "Resumo dos Testes de Wilcoxon sobre o Número de Usuários de Internet")

```

Table 3: Resumo dos Testes de Wilcoxon sobre o Número de Usuários de Internet

Teste	Hipótese Alternativa	p.valor	Conclusão
Bilateral (two-sided)	Mediana antes < Mediana depois	0.6289	Não rejeita H (sem diferença significativa)
Unilateral (greater)	Mediana antes > Mediana depois	0.6858	Não rejeita H (sem evidência de que era maior)
Unilateral (less)	Mediana antes < Mediana depois	0.3144	Não rejeita H (sem evidência de que aumentou depois)

12.1 Calculando a distribuição acumulada empírica de algumas variáveis

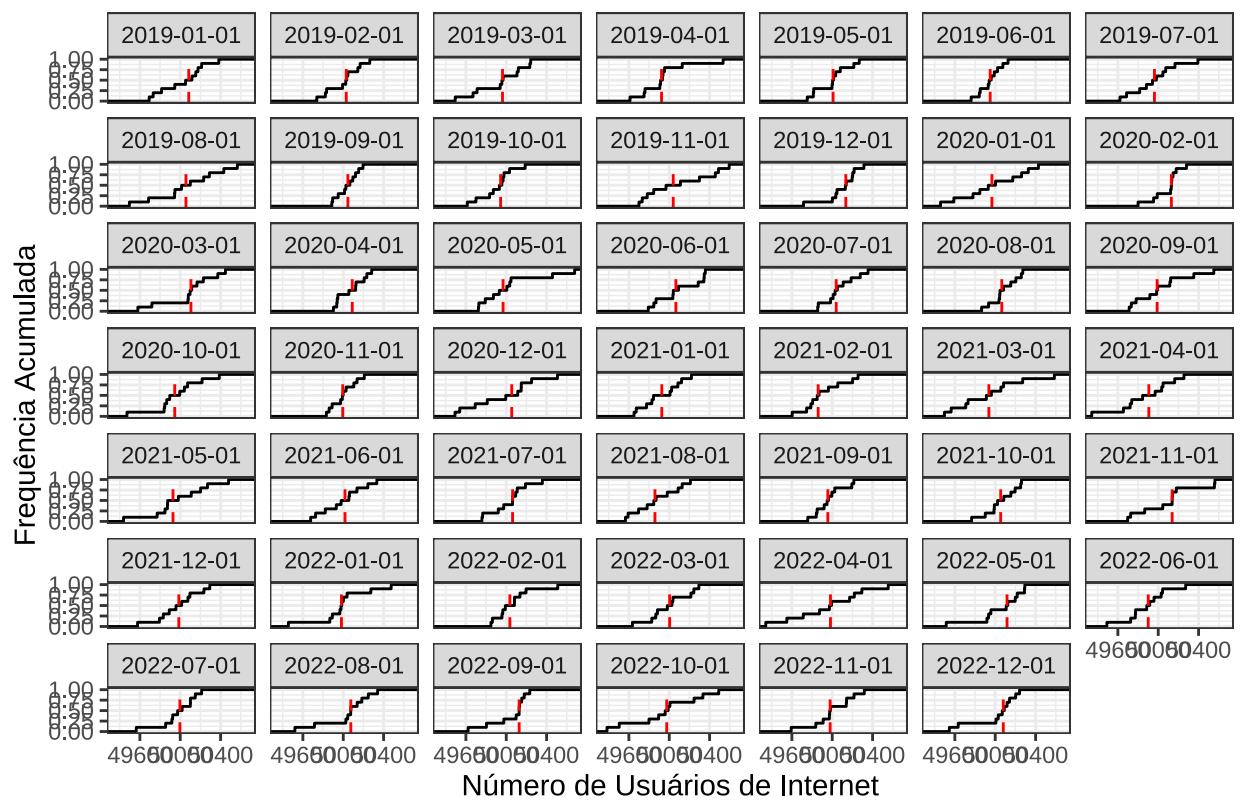
12.1.1 Filtrando os dados e visualizando distribuições acumuladas

```

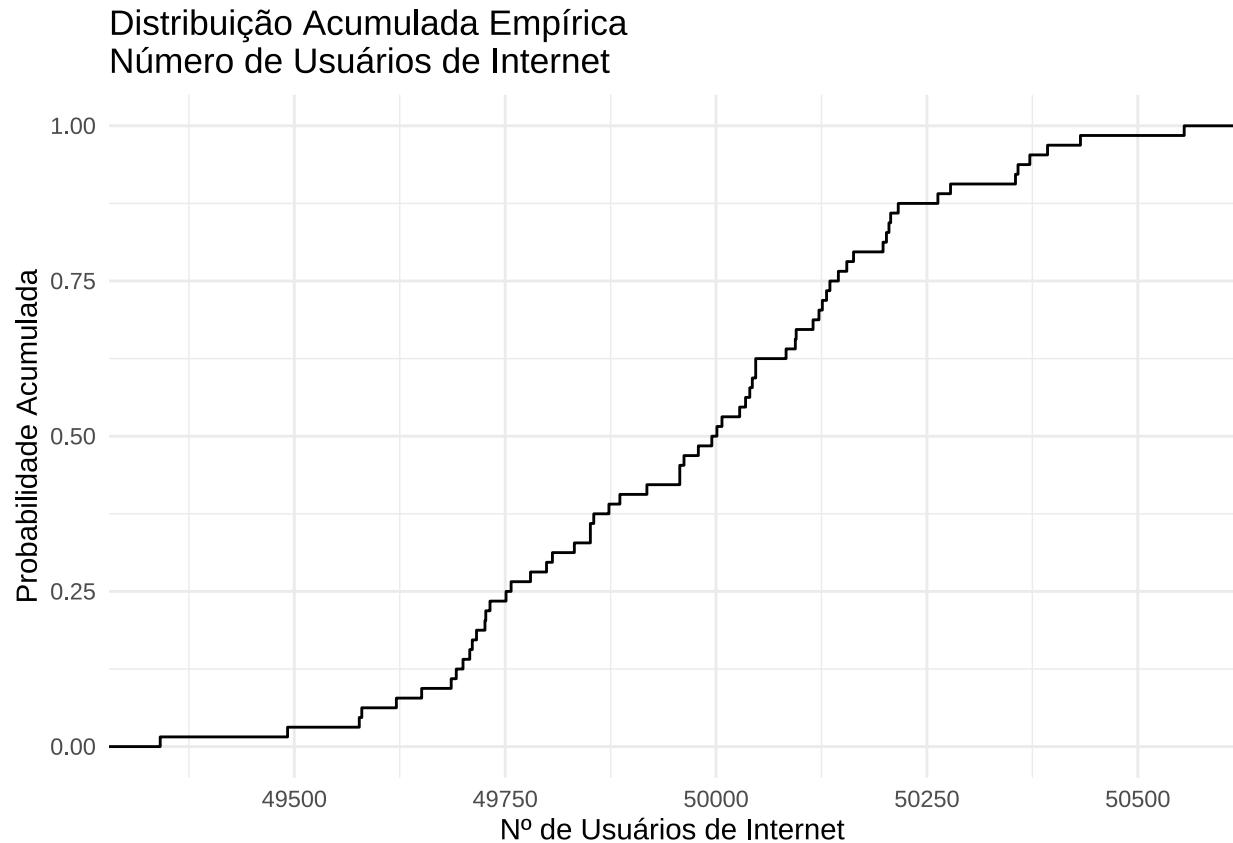
tech_filter %>%
  ggplot(aes(x = n_usuarios_internet)) + stat_ecdf(geom = "step") +
  geom_vline(
    data = tech_filter %>%
      group_by(mes_ano) %>%
      summarize(n_users_md = median(n_usuarios_internet)),
    aes(xintercept = n_users_md),
    color = "red",
    linetype = "dashed",
    inherit.aes = FALSE
  ) +
  facet_wrap(~ mes_ano) +
  theme_bw() +
  labs(
    title = "Função de Distribuição Acumulada do Número de Usuários de Internet por Mês/Ano",
    x = "Número de Usuários de Internet",
    y = "Frequência Acumulada"
  )

## Warning in geom_vline(data = tech_filter %>% group_by(mes_ano) %>%
## summarise(n_users_md = median(n_usuarios_internet)), : Ignoring unknown
## parameters: `inherit.aes`
```

Função de Distribuição Acumulada do Número de Usuários de Internet por Mês



```
# Distribuição Acumulada Empírica para n_usuarios_internet no Sudeste
ggplot(tech_sudeste, aes(x = n_usuarios_internet)) +
  stat_ecdf(geom = "step") +
  labs(title = "Distribuição Acumulada Empírica\nNúmero de Usuários de Internet",
       x = "Nº de Usuários de Internet",
       y = "Probabilidade Acumulada") +
  theme_minimal()
```



As curvas ECDF demonstraram um comportamento de subida rápida em torno da mediana, indicando que os valores observados em cada mês estavam fortemente agrupados em uma faixa estreita de variação, tipicamente entre 49.600 e 50.400 usuários. A presença da linha vermelha tracejada, correspondente à mediana de cada mês, evidenciou a estabilidade no número de usuários de internet nas cidades analisadas ao longo do período estudado, com pequenas flutuações ocasionais.

O segundo gráfico confirma que a maioria dos registros analisados está concentrada em torno de 50.000 usuários. Observa-se que 50% das cidades/mês possuem até aproximadamente 50.000 usuários, evidenciando a baixa dispersão dos dados no espaço amostral considerado.

13 Gráfico com a matriz de espalhamento (scatter matrix plot)

Responda a pergunta: através de investigação visual, quais são as variáveis mais correlacionadas. Apresente o gráfico e justifique.

Aqui investigamos: será que mais usuários → mais velocidade?

```
# Scatterplot: Número de usuários de internet x Velocidade média da internet
tech_sudeste %>%
  filter(!is.na(n_usuarios_internet), !is.na(velocidade_media_mbps)) %>%
  ggplot(aes(x = n_usuarios_internet, y = velocidade_media_mbps)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  theme_classic() +
  labs(
    title = "Dispersão: Número de Usuários de Internet vs Velocidade Média",
```

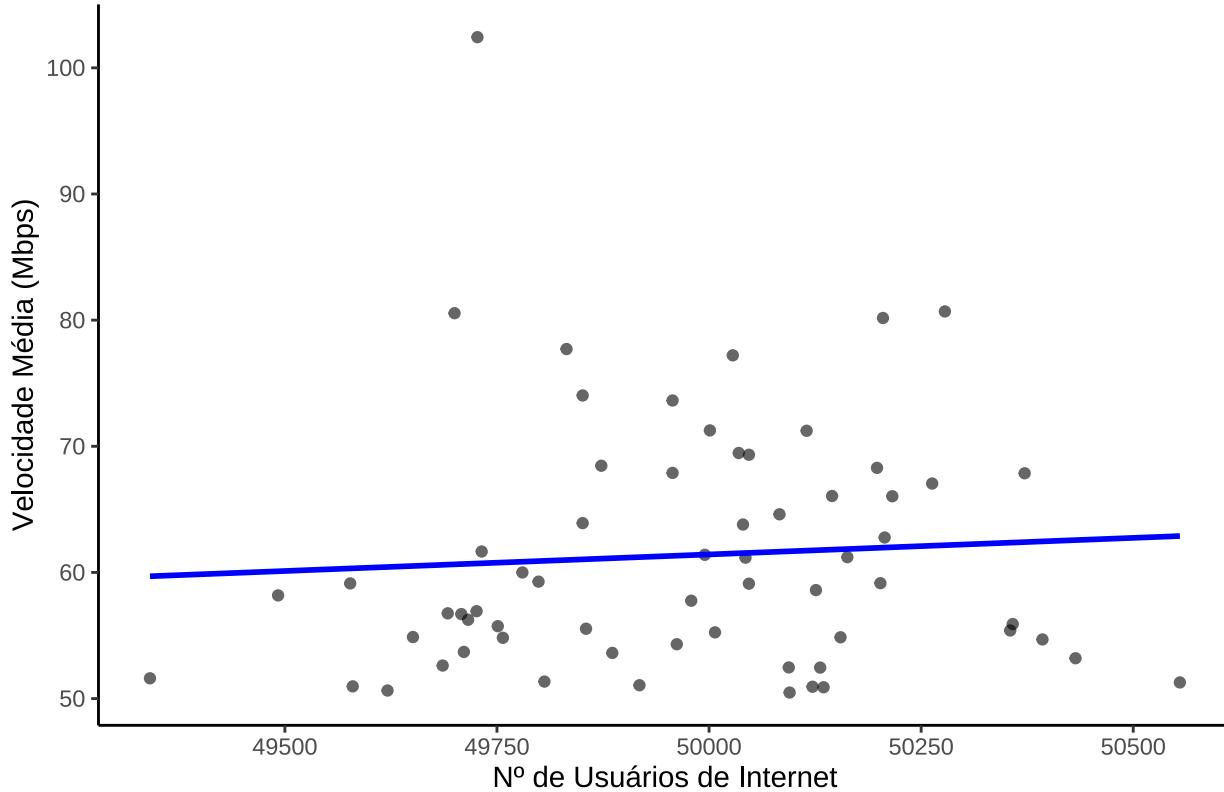
```

x = "Nº de Usuários de Internet",
y = "Velocidade Média (Mbps)"
)

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

```

Dispersão: Número de Usuários de Internet vs Velocidade Média



Existe uma correlação positiva muito fraca entre o número de usuários de internet e a velocidade média da internet nas cidades do Sudeste analisadas.

Aqui investigamos: será que mais dispositivos conectados afetam a velocidade?

```

tech_sudeste %>%
  filter(!is.na(n_dispositivos_conectados), !is.na(velocidade_media_mbps)) %>%
  ggplot(aes(x = n_dispositivos_conectados, y = velocidade_media_mbps)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, color = "darkgreen") +
  theme_classic() +
  labs(
    title = "Dispositivos Conectados vs Velocidade Média",
    x = "Número de Dispositivos Conectados",
    y = "Velocidade Média (Mbps)"
  )

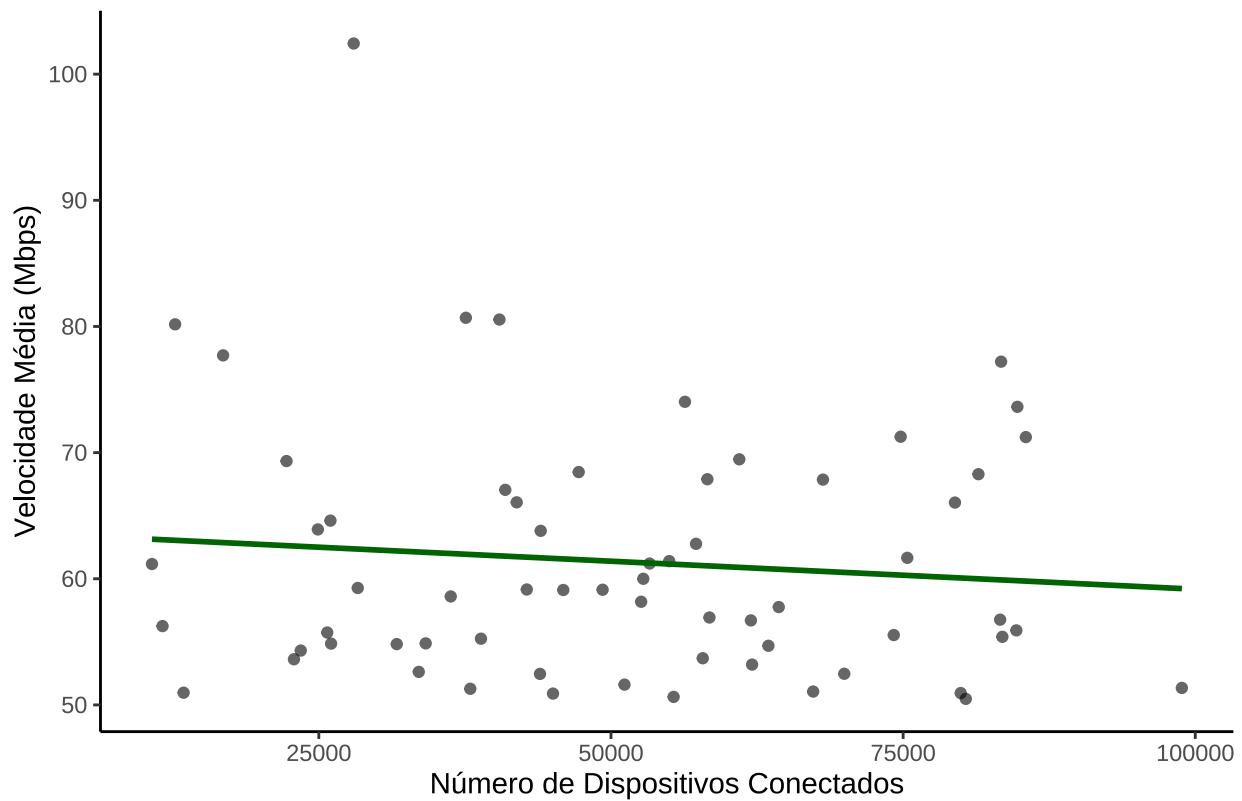
```

```

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

```

Dispositivos Conectados vs Velocidade Média



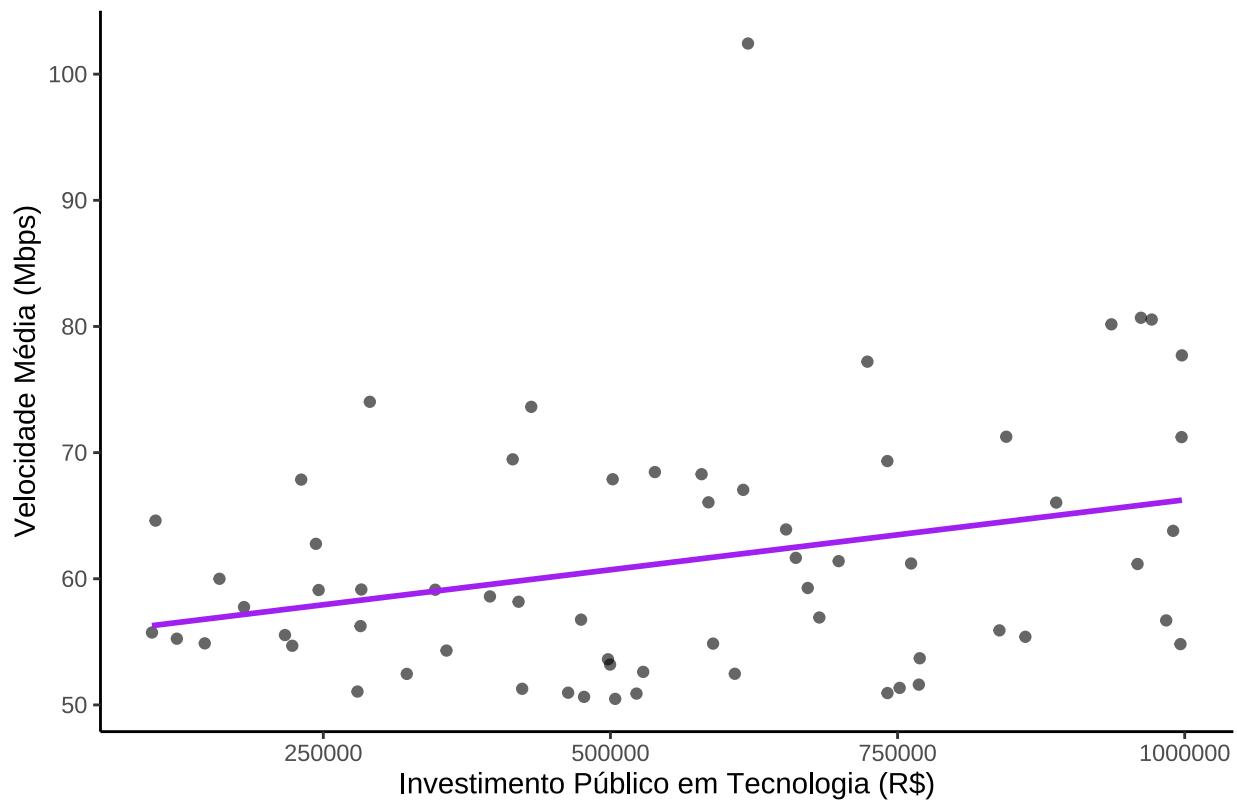
Existe uma correlação negativa muito fraca entre o número de dispositivos conectados e a velocidade média da internet nas cidades do Sudeste analisadas.

Aqui investigamos: será que onde há mais investimento → melhor internet?

```
tech_sudeste %>%
  filter(!is.na(investimento_publico_tecnologia), !is.na(velocidade_media_mbps)) %>%
  ggplot(aes(x = investimento_publico_tecnologia, y = velocidade_media_mbps)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, color = "purple") +
  theme_classic() +
  labs(
    title = "Investimento Público vs Velocidade Média",
    x = "Investimento Público em Tecnologia (R$)",
    y = "Velocidade Média (Mbps)"
  )

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

Investimento Público vs Velocidade Média



Existe uma correlação positiva fraca entre o investimento público em tecnologia e a velocidade média da internet nas cidades do Sudeste analisadas.

14 Correlação entre Número de Usuários e Velocidade

```
cor(tech_sudeste$n_usuarios_internet, tech_sudeste$velocidade_media_mbps, use = "complete.obs")
```

```
## [1] 0.06689141
```

Correlação muito próxima de 0, que indica que não há praticamente nenhuma relação linear entre o número de usuários e a velocidade média da internet na sua base de dados.

15 Correlação entre Número de Dispositivos Conectados e Velocidade

```
cor(tech_sudeste$n_dispositivos_conectados, tech_sudeste$velocidade_media_mbps, use = "complete.obs")
```

```
## [1] -0.1003889
```

Correlação fraca e negativa, que indica que existe uma tendência muito leve: à medida que o número de dispositivos conectados aumenta, a velocidade média tende a diminuir.

16 Correlação entre Investimento Público e Velocidade

```
cor(tech_sudeste$investimento_publico_tecnologia, tech_sudeste$velocidade_media_mbps, use = "complete.or")  
## [1] 0.2995732
```

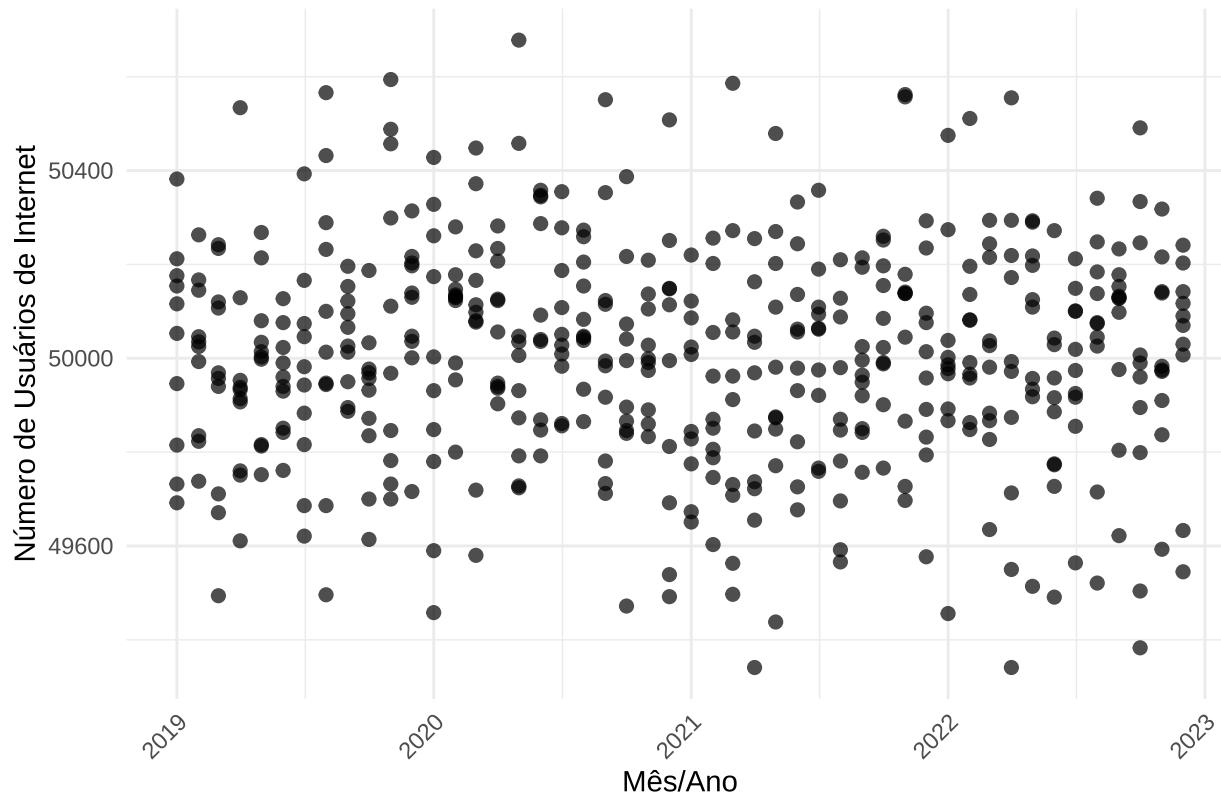
Correlação positiva moderadamente fraca, que sugere que quando há mais investimento público em tecnologia, a velocidade média da internet tende a aumentar — mas a força da relação ainda não é considerada uma correlação forte.

Resposta: a investigação visual das correlações entre as variáveis da base de dados mostra que, de forma geral, as relações lineares entre as variáveis são fracas. A variável que apresenta maior correlação visual é investimento público em tecnologia com velocidade média da internet, indicando uma tendência positiva: maiores investimentos tendem a estar associados a velocidades médias mais altas. Já o número de usuários de internet e o número de dispositivos conectados mostram pouca ou nenhuma relação visual com a velocidade.

16.1 Criar o gráfico de dispersão para velocidade média da internet nas cidades do Sudeste superiores a 50 Mbps

```
tech_filter %>%  
  ggplot(aes(x = mes_ano, y = n_usuarios_internet)) +  
  geom_point(size = 2, alpha = 0.7) +  
  labs(title = "Dispersão do Número de Usuários (Velocidade > 50 Mbps)",  
       x = "Mês/Ano",  
       y = "Número de Usuários de Internet") +  
  theme_minimal() +  
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Dispersão do Número de Usuários (Velocidade > 50 Mbps)



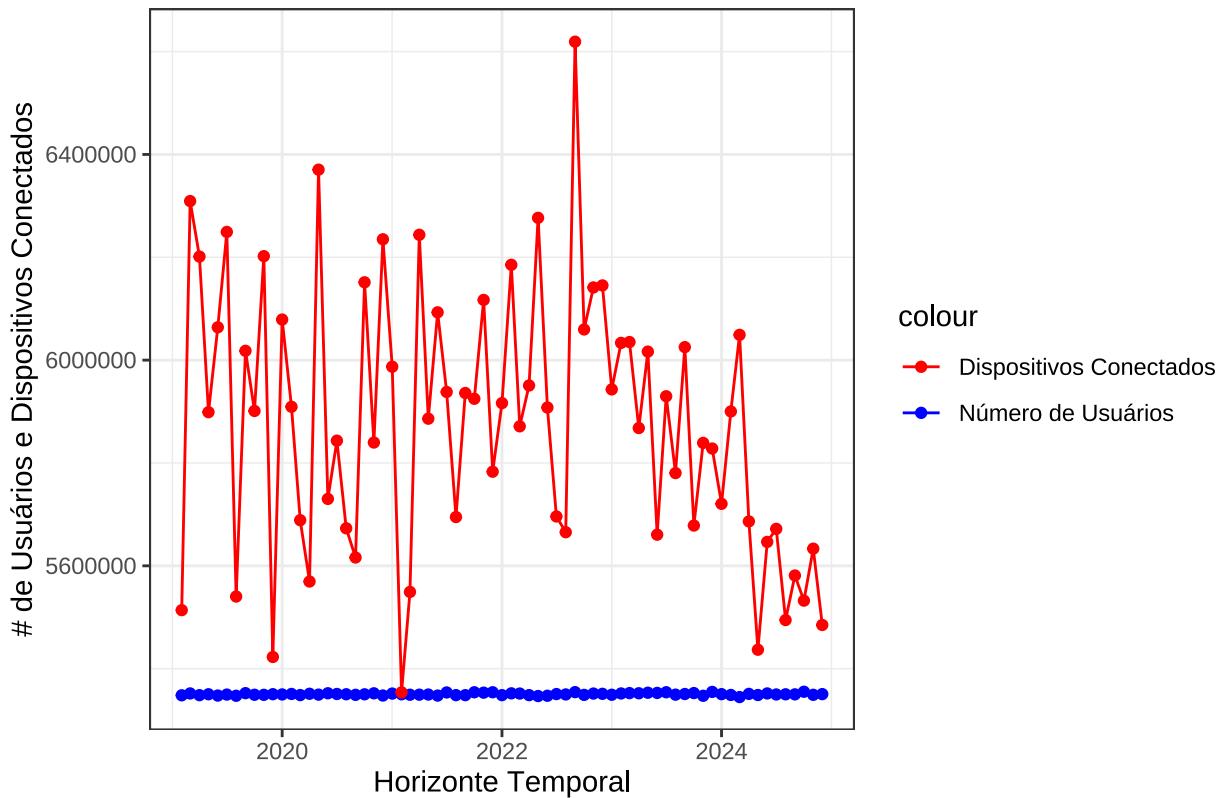
16.1.1 Outras análises sobre a minha base

```
dados_agg <- tech %>%
  select(mes, ano, uf, n_usuarios_internet, n_dispositivos_conectados) %>%
  mutate(mes_ano = as.Date(paste(ano, mes, "01", sep = "-"), format = "%Y-%m-%d")) %>%
  group_by(mes_ano) %>%
  summarise(
    n_usuarios_internet = sum(n_usuarios_internet, na.rm = TRUE),
    n_dispositivos_conectados = sum(n_dispositivos_conectados, na.rm = TRUE)
  ) %>%
  filter(mes_ano > as.Date("2019-01-01", format = "%Y-%m-%d"))

ggplot(dados_agg, aes(x = mes_ano)) +
  geom_line(aes(y = n_usuarios_internet, color = "Número de Usuários")) +
  geom_point(aes(y = n_usuarios_internet, color = "Número de Usuários")) +
  geom_line(aes(y = n_dispositivos_conectados, color = "Dispositivos Conectados")) +
  geom_point(aes(y = n_dispositivos_conectados, color = "Dispositivos Conectados")) +
  ylab("# de Usuários e Dispositivos Conectados") +
  xlab("Horizonte Temporal") +
  labs(title = "Evolução Temporal de Usuários e Dispositivos") +
  theme_bw() +
  scale_color_manual(values = c("Número de Usuários" = "blue", "Dispositivos Conectados" = "red"))
```

16.1.1.1 Análise temporal de Usuários x Dispositivos conectados

Evolução Temporal de Usuários e Dispositivos



A análise temporal do número de usuários e dispositivos conectados no Brasil entre 2019 e 2024 revela dois comportamentos distintos. O número de usuários de internet manteve-se estável ao longo do período, indicando a consolidação do acesso digital no país. Por outro lado, o número de dispositivos conectados apresentou oscilações expressivas, com picos nos anos de 2020 e 2021 — possivelmente impulsionados por fatores como o aumento da conectividade domiciliar durante a pandemia e a expansão da tecnologia móvel, ou até mesmo eventos sazonais como feriados prolongados. Observa-se uma tendência de redução no número de dispositivos conectados a partir de 2023, o que pode refletir mudanças no comportamento de consumo, na infraestrutura de rede etc.

```

dados_agg_sudeste <- tech %>%
  filter(cidade %in% c("Rio de Janeiro", "São Paulo", "Belo Horizonte")) %>%
  select(mes, ano, uf, cidade, n_usuarios_internet, n_dispositivos_conectados) %>%
  mutate(mes_ano = as.Date(paste(ano, mes, "01", sep = "-")), format = "%Y-%m-%d")) %>%
  group_by(mes_ano) %>%
  summarise(
    n_usuarios_internet = sum(n_usuarios_internet, na.rm = TRUE),
    n_dispositivos_conectados = sum(n_dispositivos_conectados, na.rm = TRUE)
  ) %>%
  filter(mes_ano > as.Date("2019-01-01", format = "%Y-%m-%d"))

ggplot(dados_agg_sudeste, aes(x = mes_ano)) +
  
```

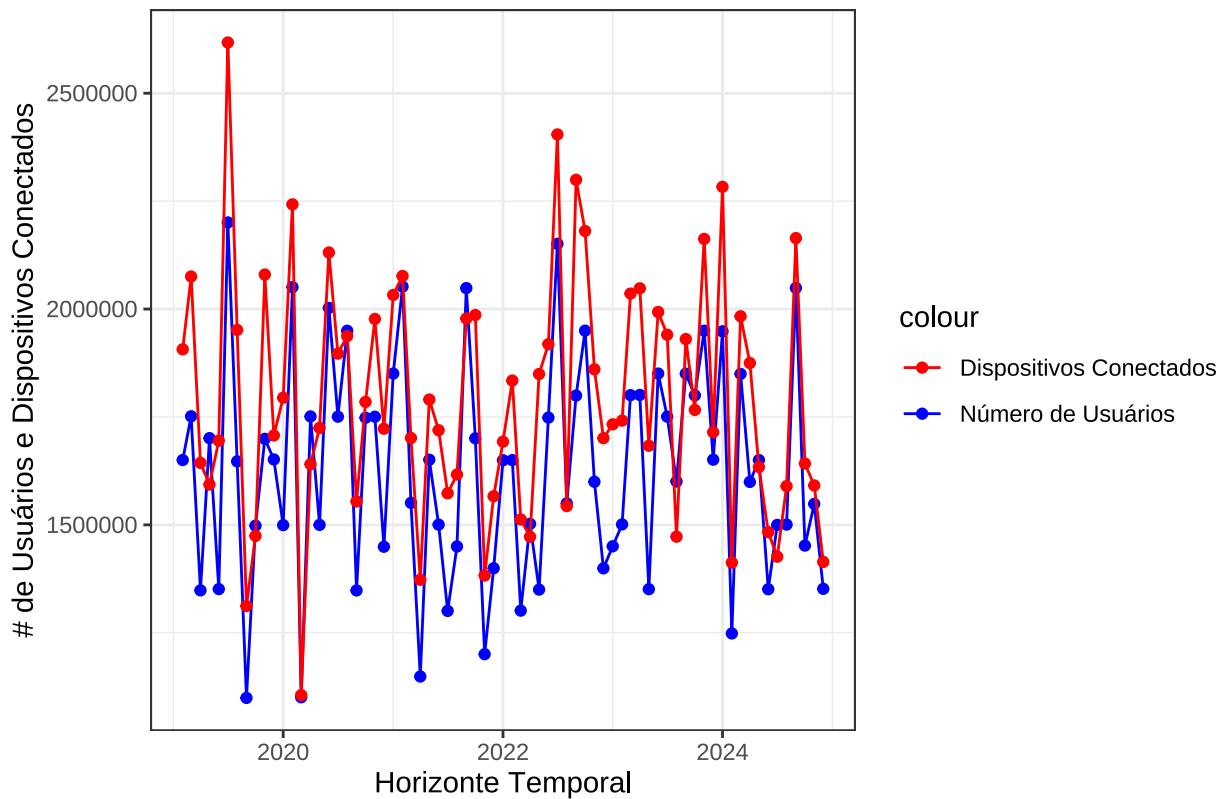
```

geom_line(aes(y = n_usuarios_internet, color = "Número de Usuários")) +
  geom_point(aes(y = n_usuarios_internet, color = "Número de Usuários")) +
  geom_line(aes(y = n_dispositivos_conectados, color = "Dispositivos Conectados")) +
  geom_point(aes(y = n_dispositivos_conectados, color = "Dispositivos Conectados")) +
  ylab("# de Usuários e Dispositivos Conectados") +
  xlab("Horizonte Temporal") +
  labs(title = "Evolução Temporal de Usuários e Dispositivos no Sudeste") +
  theme_bw() +
  scale_color_manual(values = c("Número de Usuários" = "blue", "Dispositivos Conectados" = "red"))

```

16.1.1.2 Análise temporal de Usuários x Dispositivos conectados para região Sudeste

Evolução Temporal de Usuários e Dispositivos no Sudeste



A análise temporal do número de usuários e de dispositivos conectados na Região Sudeste revela alta instabilidade nos dados ao longo do período analisado (2019–2024). Ambos os indicadores apresentaram variações significativas mês a mês, sendo a flutuação do número de dispositivos conectados ainda mais acentuada. Esse comportamento sugere a influência de fatores externos e sazonais na dinâmica de acesso e uso de tecnologia nas cidades do Sudeste.

16.1.2 Calculando a dispersão e as correlações

```

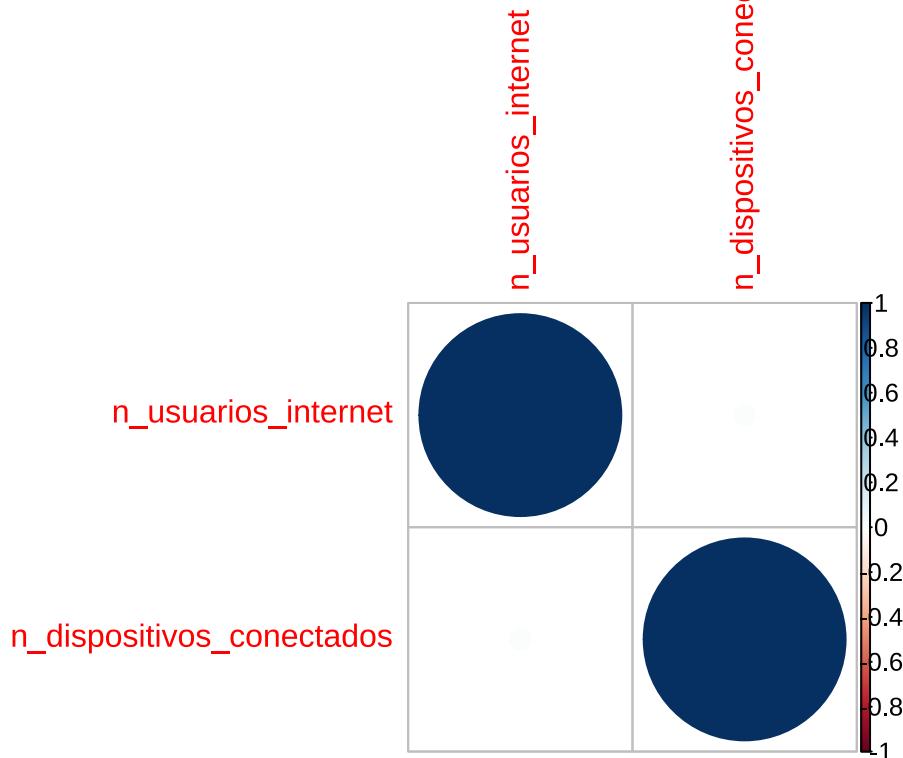
tech %>%
  select(n_usuarios_internet, n_dispositivos_conectados) %>%

```

```
cor(method = "pearson") %>%
corrplot(title = "Correlação de Pearson entre Usuários e Dispositivos (Todos os Meses) - Pearson")
```

16.1.2.1 calculando corrplot Pearson - usuários x dispositivos

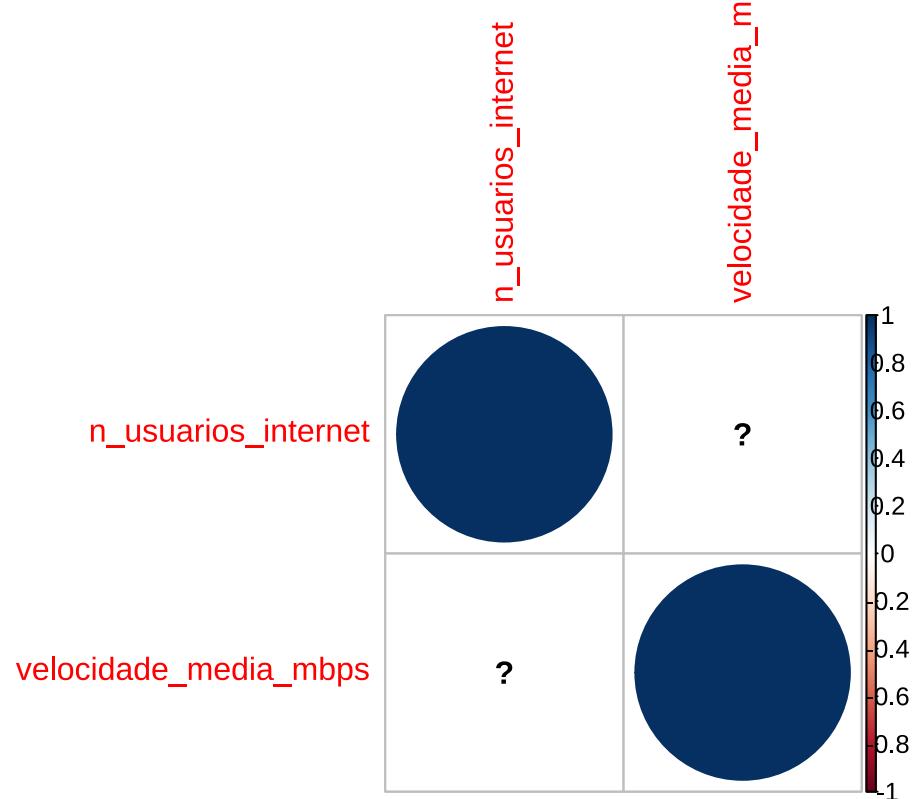
ÁREA DE PREDIÇÃO ENTRE USUÁRIOS E DISPOSITIVOS (USUÁRIOS OS MÉSES) - PEARSON



```
tech %>%
  select(n_usuarios_internet, velocidade_media_mbps) %>%
  cor(method = "pearson") %>%
corrplot(title = "Correlação de Pearson entre Usuários e Velocidade (Todos os Meses) - Pearson")
```

16.1.2.2 Calculando corrplot Pearson - usuários x velocidade

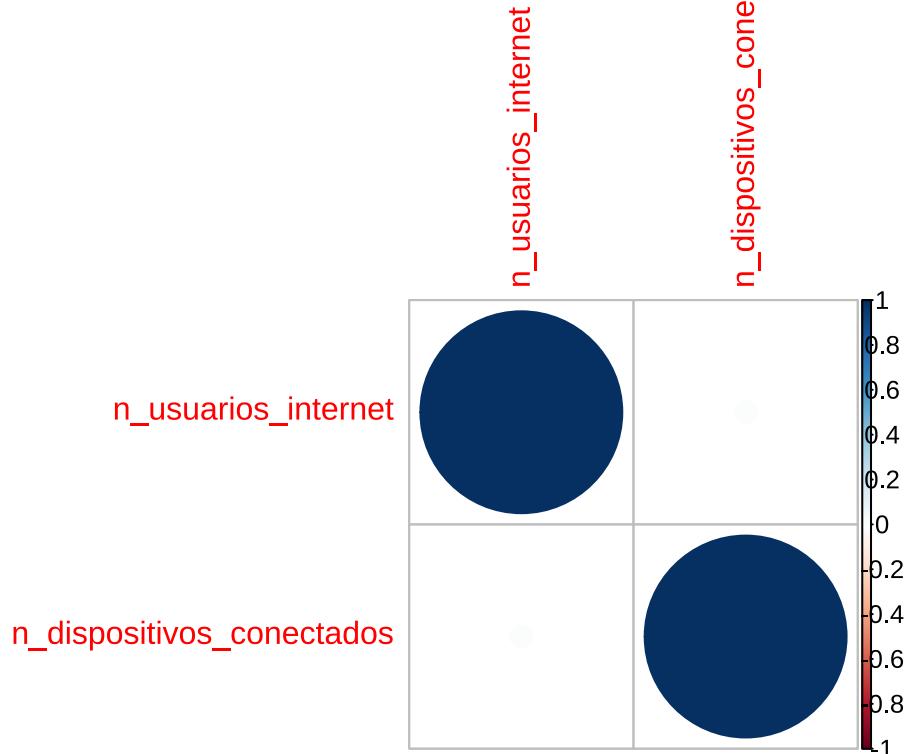
Índice de Pearson entre Usuários e Velocidade (Todos os Meses) - Pearson



16.1.3 Calculando corrplot Spearman - usuários x dispositivos

```
tech %>%
  select(n_usuarios_internet, n_dispositivos_conectados) %>%
  cor(method = "spearman") %>%
  corrplot(title = "Correlação de Pearson entre Usuários e Dispositivos (Todos os Meses) - Spearman")
```

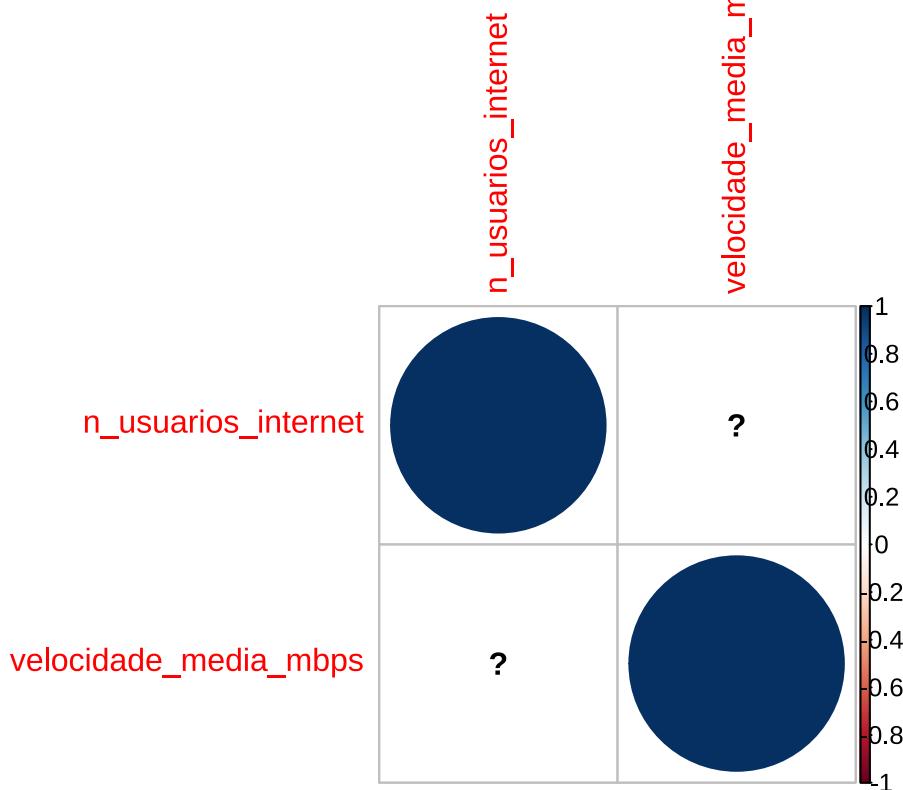
Correlação de Pearson entre Usuários e Dispositivos Conectados (Todos os Meses) - Cor



16.1.4 Calculando corrplot Spearman - usuários x velocidade

```
tech %>%
  select(n_usuarios_internet, velocidade_media_mbps) %>%
  cor(method = "spearman") %>%
  corrplot(title = "Correlação de Pearson entre Usuários e Velocidade (Todos os Meses) - Pearson")
```

análise de Pearson entre usuários e velocidade (tudos os meses) - re



A análise de correlação de Pearson e Spearman entre o número de usuários de internet e o número de dispositivos conectados indicou uma relação linear praticamente inexistente. Esse resultado sugere que a quantidade de dispositivos conectados em uma localidade não esteve diretamente associada ao número de usuários de internet, refletindo possivelmente a presença de múltiplos dispositivos por usuário ou fatores relacionados à infraestrutura e perfil de consumo tecnológico. No caso da comparação de usuários com a velocidade, o resultado demonstra que há valores ausentes na base para a variável `Velocidade_media_mbps`.

17 Evento aleatório - escolher uma cidade/mês em que a velocidade média da internet foi superior a 50 Mbps

17.1 Total de cidade/mês (espaço amostral)

```
espaco_amostral <- tech_filter %>%
  count()
```

17.2 Eventos favoráveis (`velocidade > 50 Mbps`)

```
eventos_favoraveis_vel <- tech_filter %>%
  filter(velocidade_media_mbps > 50) %>%
  count()
```

17.3 Calcular a probabilidade do evento

```
prob_evento <- eventos_favoraveis_vel$n / espaco_amostral$n  
prob_evento  
  
## [1] 0.4958333
```

Isso significa que quase metade dos registros analisados atendem ao critério do seu evento aleatório.

18 Qualidade de dados

Qualidade dos dados tem sido um dos temas mais abordados nos projetos de estruturação em data analytics, sendo um dos principais indicadores do nível de maturidade das organizações. Um dos problemas mais comuns de qualidade é relacionado à completude de dados. Em suas palavras, como é definido completude? Qual o impacto em uma análise exploratória de dados?

Completude é quando os dados estão preenchidos, sem informações faltando. Se muitos dados estão vazios ou ausentes, a análise exploratória pode ficar prejudicada, porque pode gerar resultados errados, mostrar padrões que não existem ou esconder tendências importantes. Ter dados completos é importante para que as conclusões sejam confiáveis. Levando em conta minha base antes da imputação dos dados, como já mencionei anteriormente, existia um percentual de missings na variável velocidade_media_mbps.

19 Qual a completude para cada uma das variáveis do seu banco de dados?

```
completude <- tech %>%  
  summarise(across(everything(), ~ mean(!is.na(.)) * 100))  
  
completude  
  
## # A tibble: 1 x 14  
##   cidade    uf     mes   ano mes_ano n_usuarios_internet velocidade_media_mbps  
##   <dbl> <dbl> <dbl> <dbl>   <dbl>                 <dbl>                  <dbl>  
## 1    100    100    100    100      100                   100                  97.0  
## # i 7 more variables: n_dispositivos_conectados <dbl>, indice_inovacao <dbl>,  
## #   investimento_publico_tecnologia <dbl>,  
## #   eventos_tecnologicos_realizados <dbl>, startups_ativas <dbl>,  
## #   patentes_registradas <dbl>, mes_ano_data <dbl>
```

Quase todas as variáveis apresentam 100% de completude, exceto “velocidade média (Mbps)”, que possui cerca de 97% de preenchimento, indicando alguns dados faltantes que podem impactar a análise.

```
completude <- tech_filter %>%  
  summarise(across(everything(), ~ mean(!is.na(.)) * 100))  
  
completude
```

```

## # A tibble: 1 x 8
##   cidade   mes   ano n_usuarios_internet velocidade_media_mbps
##   <dbl> <dbl> <dbl>                 <dbl>                  <dbl>
## 1     100    100    100                  100                  100
## # i 3 more variables: n_dispositivos_conectados <dbl>,
## #   investimento_publico_tecnologia <dbl>, mes_ano <dbl>

```

Depois da minha base inputada, como é possível ver, todas as variáveis da base possuem 100% de completnude, sem valores ausentes.

20 Regressão Linear

20.0.1 Regressão Linear: Velocidade Média (Mbps) explicando Número de Usuários de Internet

Estou rodando um modelo de regressão linear simples com o objetivo de avaliar se algumas associações fazem sentido.

20.1 Regressão linear de velocidade média x número de usuários

Nesse primeiro momento, minha proposta é observar se a velocidade média da internet influencia o número de usuários de internet nas cidades, considerando registros com velocidade superior a 50 Mbps.

```

modelo_velocidade <- lm(n_usuarios_internet ~ velocidade_media_mbps, data = tech_filter)
summary(modelo_velocidade)

```

```

##
## Call:
## lm(formula = n_usuarios_internet ~ velocidade_media_mbps, data = tech_filter)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -664.08 -145.43    6.72  146.12  664.92 
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 4.997e+04 3.640e+01 1372.732 <2e-16 ***
## velocidade_media_mbps 7.151e-01 6.950e-01    1.029    0.304  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 231.2 on 478 degrees of freedom
## Multiple R-squared:  0.00221,   Adjusted R-squared:  0.0001224 
## F-statistic: 1.059 on 1 and 478 DF,  p-value: 0.3041

```

O resultado da regressão indicou que o coeficiente da velocidade média (0,7151) não foi significativo (p-value = 0,304), e o modelo apresentou um R² de apenas 0,22%, indicando que a variável velocidade média explica muito pouco da variação no número de usuários.

20.2 Regressão linear de cidade x número de usuários

Nesse modelo de regressão linear com variável categórica, estou investigando se existem diferenças no número de usuários de internet entre as cidades analisadas.

```
modelo_cidade <- lm(n_usuarios_internet ~ as.factor(cidade), data = tech_filter)
summary(modelo_cidade)
```

```
##
## Call:
## lm(formula = n_usuarios_internet ~ as.factor(cidade), data = tech_filter)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -689.50  -150.56     9.25  148.75  647.50
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                49989.500    33.488 1492.751 <2e-16 ***
## as.factor(cidade)Brasília      51.833    47.359   1.094   0.274
## as.factor(cidade)Curitiba      14.250    47.359   0.301   0.764
## as.factor(cidade)Fortaleza     21.604    47.359   0.456   0.648
## as.factor(cidade)Manaus        31.729    47.359   0.670   0.503
## as.factor(cidade)Porto Alegre     5.479    47.359   0.116   0.908
## as.factor(cidade)Recife         31.437    47.359   0.664   0.507
## as.factor(cidade)Rio de Janeiro -22.833    47.359  -0.482   0.630
## as.factor(cidade)Salvador        41.000    47.359   0.866   0.387
## as.factor(cidade)São Paulo      -29.292    47.359  -0.618   0.537
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 232 on 470 degrees of freedom
## Multiple R-squared:  0.01187,   Adjusted R-squared:  -0.007048
## F-statistic: 0.6275 on 9 and 470 DF,  p-value: 0.7738
```

O resultado da regressão indicou que nenhuma cidade apresentou diferenças significativas em relação à cidade de referência (todos os p-value > 0,05). Além disso, o R² do modelo foi de apenas 1,2%, sugerindo que a cidade, isoladamente, explica muito pouco da variação observada no número de usuários de internet.

Portanto, nos dados analisados, não foram observadas diferenças significativas no número de usuários entre as cidades.

20.3 Regressão linear de velocidade x cidade explicando Número de Usuários de Internet

Nesse modelo de regressão linear múltipla, estou combinando a velocidade média da internet e a cidade como variáveis explicativas do número de usuários de internet.

```
modelo_velocidade_cidade <- lm(n_usuarios_internet ~ velocidade_media_mbps + as.factor(cidade), data =
summary(modelo_velocidade_cidade)
```

```
##
```

```

## Call:
## lm(formula = n_usuarios_internet ~ velocidade_media_mbps + as.factor(cidade),
##      data = tech_filter)
##
## Residuals:
##      Min      1Q Median      3Q     Max 
## -682.95 -147.22    9.09  151.02  639.36 
## 
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          49959.0712   47.7272 1046.763 <2e-16 ***
## velocidade_media_mbps       0.6287   0.7025   0.895   0.371    
## as.factor(cidade)Brasília  49.5187   47.4400   1.044   0.297    
## as.factor(cidade)Curitiba  12.6656   47.4025   0.267   0.789    
## as.factor(cidade)Fortaleza 21.1160   47.3726   0.446   0.656    
## as.factor(cidade)Manaus    30.9030   47.3784   0.652   0.515    
## as.factor(cidade)Porto Alegre 2.7891   47.4647   0.059   0.953    
## as.factor(cidade)Recife    29.5488   47.4164   0.623   0.533    
## as.factor(cidade)Rio de Janeiro -24.3331  47.3991  -0.513   0.608    
## as.factor(cidade)Salvador   40.0857   47.3805   0.846   0.398    
## as.factor(cidade)São Paulo -27.9602   47.3928  -0.590   0.555    
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 232.1 on 469 degrees of freedom
## Multiple R-squared:  0.01356,   Adjusted R-squared:  -0.007475 
## F-statistic: 0.6446 on 10 and 469 DF,  p-value: 0.7756

```

O modelo apresentou um R^2 de apenas 1,36%, e tanto a variável velocidade média (p-value = 0,371) quanto as categorias de cidade (p-value > 0,05) não mostraram significância. O teste F para o modelo completo também não foi significativo (p-value = 0,7756).

Portanto, nem a velocidade média da internet, nem a cidade de registro, explicam de maneira significativa as variações no número de usuários de internet.

20.4 Regressão linear de Número de dispositivos conectados x Número de Usuários de Internet

Nesse modelo de regressão linear simples, estou verificando a relação entre o número de dispositivos conectados e o número de usuários de internet nas cidades analisadas.

```
modelo_dispositivos <- lm(n_usuarios_internet ~ n_dispositivos_conectados, data = tech_filter)
summary(modelo_dispositivos)
```

```

## 
## Call:
## lm(formula = n_usuarios_internet ~ n_dispositivos_conectados,
##      data = tech_filter)
##
## Residuals:
##      Min      1Q Median      3Q     Max 
## -663.72 -148.58    5.96  142.10  672.66 
## 
```

```

## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.0000e+04  2.476e+01 2019.066  <2e-16 ***
## n_dispositivos_conectados 3.676e-05  4.136e-04    0.089    0.929
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 231.4 on 478 degrees of freedom
## Multiple R-squared:  1.653e-05, Adjusted R-squared: -0.002075
## F-statistic: 0.0079 on 1 and 478 DF,  p-value: 0.9292

```

Esse modelo apresentou um coeficiente de 0,00003676 para a variável número de dispositivos conectados, com p-value de 0,929, indicando ausência de significância. O R^2 do modelo foi praticamente nulo (0,0016%), e o teste F para o modelo completo também não foi significativo (p-value = 0,9292). Dessa forma, isso mostra que o número de dispositivos conectados não explica de forma significativa a variação no número de usuários de internet.

Conclusão: testei velocidade, cidade e número de dispositivos, mas nenhuma variável explicou bem o número de usuários. Todos os modelos tiveram R^2 baixos e p-value altos. Provavelmente, existem outros fatores fora da base que influenciam mais, mas que não serão alvo deste projeto no momento.

21 Conclusão final

Neste projeto, foi feita a análise dos dados de tecnologia em cidades brasileiras, tratando dados faltantes com métodos de imputação. A normalidade foi avaliada e vimos que “Número de Usuários de Internet” e “Velocidade Média” são **aproximadamente normais**, enquanto “Dispositivos Conectados” e “Investimento Público” **não seguem distribuição normal**.

As relações entre variáveis foram exploradas visualmente, mas os testes de correlação (Pearson e Spearman) **não mostraram relações fortes**.

A chance de uma cidade/mês do Sudeste ter velocidade média superior a 50 Mbps foi de cerca de 13%, e o teste de hipótese mostrou que **a velocidade média na região é, de fato, superior a 50 Mbps**.

Os testes de hipótese entre períodos (antes e depois de 2020) indicaram que **não houve diferença significativa** no número de usuários ao longo do tempo.

As regressões lineares mostraram que, com as variáveis disponíveis, **não foi possível construir modelos fortes para explicar o número de usuários**. Algumas relações fazem sentido lógico (como a relação entre número de dispositivos conectados e usuários), mas estatisticamente os modelos apresentaram baixo poder de explicação.

22 Dashboard Shiny

Por fim, foi criado um dashboard em Shiny para visualizar as variáveis e suas evoluções ao longo do tempo.

22.1 Painel 1 - Análise Geral

1)

22.2 Painel 2 - Análise Temporal das Cidades do Sudeste

2)

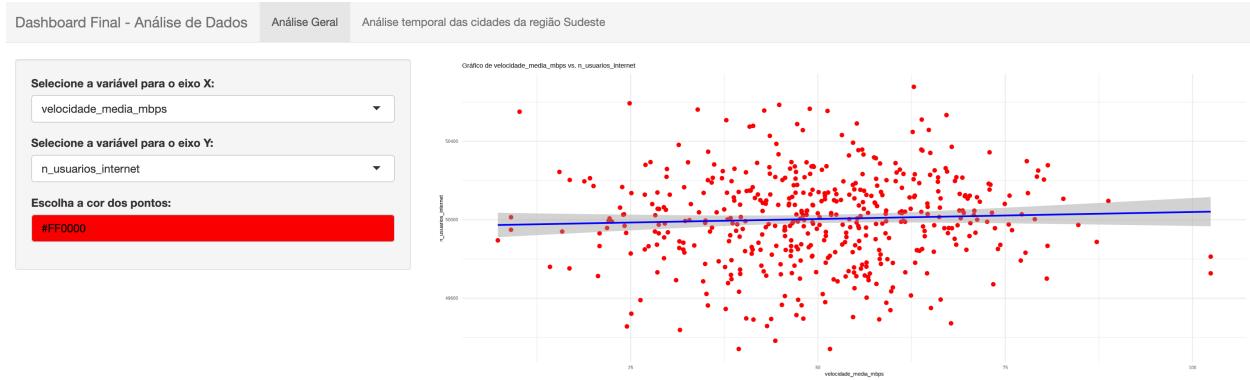


Figure 1: Print do Painel 1 - Análise Geral

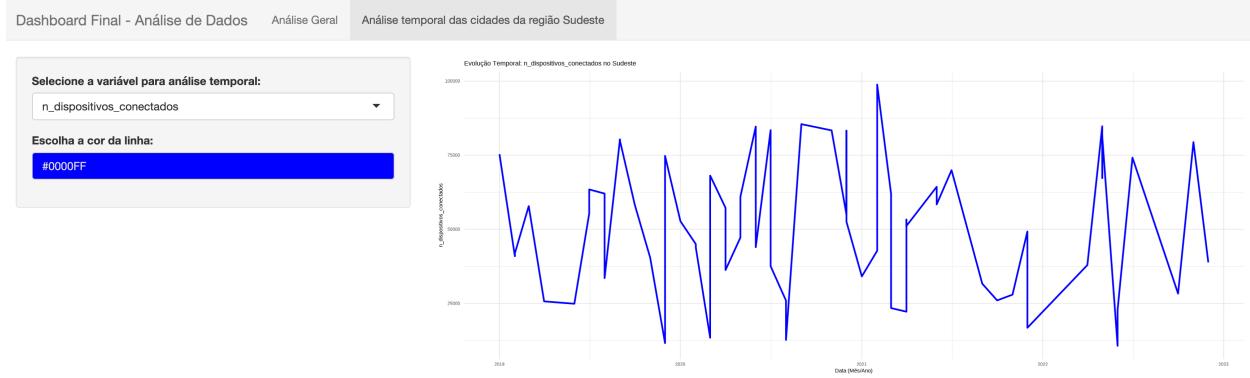


Figure 2: Print do Painel 2 - Análise Temporal

23 Repositório

Todo o material desenvolvido para este projeto, incluindo os códigos em RMarkdown, os arquivos do dashboard Shiny e a base de dados utilizada, está disponível no GitHub, no seguinte link: [GitHub](#).