Dados Videos

Vinnicius Pereira

2025-04-12

```
library(readxl)
library(ggplot2)
library(summarytools)
library(GGally)
```

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

```
library(ggpubr)
library(mice)
```

```
##
## Anexando pacote: 'mice'
```

```
## O seguinte objeto é mascarado por 'package:stats':
##
##     filter
```

```
## Os seguintes objetos são mascarados por 'package:base':
##
##     cbind, rbind
```

```
library(shiny)
library(colourpicker)
```

```
##
## Anexando pacote: 'colourpicker'
```

```
## O seguinte objeto é mascarado por 'package:shiny':
##
##     runExample
```

```
library(flexdashboard)
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ lubridate 1.9.4     ✓ tibble    3.2.1
## ✓ purrr     1.0.4     ✓ tidyr     1.3.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks mice::filter(), stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ✖ tibble::view()  masks summarytools::view()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(knitr)
library(rmarkdown)
library(kableExtra)
```

```
##
## Anexando pacote: 'kableExtra'
##
## O seguinte objeto é mascarado por 'package:dplyr':
##
##     group_rows
```

```
library(stargazer)
```

```
##
## Please cite as:
##
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
library(shiny)
library(mice)
```

Introdução

A base de dados utilizada neste projeto se chama `dados_videos.xlsx` e contém informações sobre vídeos curtos e suas métricas de engajamento. A escolha se deu por apresentar variáveis numéricas suficientes para a análise e por conter dados faltantes, o que é ideal para validar técnicas de imputação. Os resultados esperados incluem uma descrição estatística da base, visualização de padrões de correlação entre variáveis e avaliação da normalidade dos dados, além de aplicar técnicas de imputação e desenvolver um dashboard interativo com Shiny.

2. Escolha da Base de Dados

Importando os dados - Dados Vídeos

```
dados_videos <- read_excel("dados_videos.xlsx")
View(dados_videos)
```

Verificando as informações iniciais e uma síntese dos dados inicialmente

```
head(dados_videos)
```

```
## # A tibble: 6 × 11
##   video_id status_reinvidicacao duracao_video transcricao_video
##      <dbl> <chr>                        <dbl> <chr>
## 1        1 claim                           59 someone shared with me that drone…
## 2        2 claim                           32 someone shared with me that there…
## 3        3 claim                           31 someone shared with me that ameri…
## 4        4 claim                           25 someone shared with me that the m…
## 5        5 claim                           19 someone shared with me that the n…
## 6        6 claim                           35 someone shared with me that gross…
## # ℹ 7 more variables: status_verificacao <chr>, status_video <chr>,
## #   qtd_visualizacoes <dbl>, qtd_curtidas <dbl>, qtd_compartilhamento <dbl>,
## #   qtd_downloads <dbl>, qtd_comentarios <dbl>
```

```
summary(dados_videos)
```

```
##     video_id    status_reinvidicacao duracao_video   transcricao_video
##  Min.   :   1   Length:4121          Min.   : 5.00   Length:4121
##  1st Qu.:1031   Class :character     1st Qu.:18.00   Class :character
##  Median :2061   Mode  :character     Median :32.00   Mode  :character
##  Mean   :2061                        Mean   :32.25
##  3rd Qu.:3091                        3rd Qu.:47.00
##  Max.   :4121                        Max.   :60.00
##
##  status_verificacao status_video       qtd_visualizacoes  qtd_curtidas
##  Length:4121        Length:4121        Min.   :    35    Min.   :     0
##  Class :character   Class :character   1st Qu.:  4837    1st Qu.:   749
##  Mode  :character   Mode  :character   Median :  9897    Median :  3232
##                                        Mean   :252978    Mean   : 84460
##                                        3rd Qu.:492709    3rd Qu.:122786
##                                        Max.   :999127    Max.   :654588
##                                        NA's   :60        NA's   :60
##  qtd_compartilhamento qtd_downloads   qtd_comentarios
##  Min.   :     0       Min.   :    0   Min.   :   0.0
##  1st Qu.:   105       1st Qu.:    6   1st Qu.:   1.0
##  Median :   695       Median :   45   Median :   9.0
##  Mean   : 16832       Mean   : 1033   Mean   : 343.6
##  3rd Qu.: 18837       3rd Qu.: 1135   3rd Qu.: 289.0
##  Max.   :240154       Max.   :14308   Max.   :8481.0
##  NA's   :60           NA's   :60      NA's   :60
```

3. Utilizando a função DESCR() - Estatísticas Descritivas

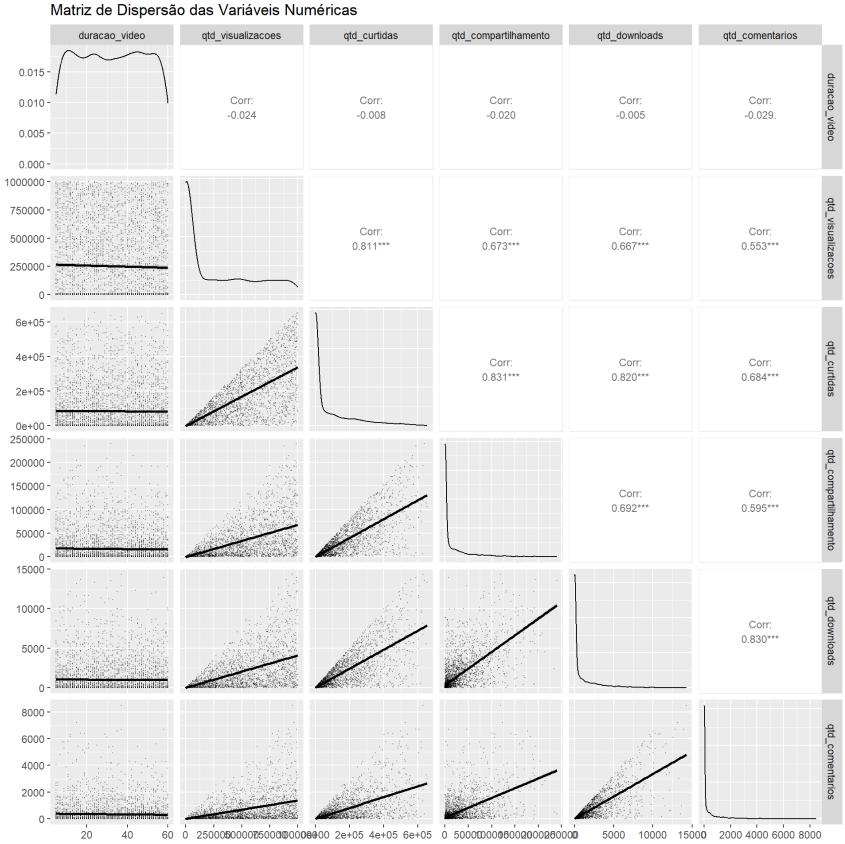
```
dados_videos %>%
  dplyr::select(qtd_curtidas, qtd_visualizacoes, qtd_compartilhamento, qtd_downloads, qtd_comentarios) %>%
  summarytools::descr(round.digits = 2, scientific = FALSE, style = "simple", transpose = TRUE) %>%
  kableExtra::kbl(caption = "Estatísticas Descritivas das variáveis objetivo") %>%
  kableExtra::kable_material(c("striped", "hover"))
```

```
## Error in table(names(candidates))[["tested"]]: índice fora dos limites
```

Estatísticas Descritivas das variáveis objetivo

	Mean	Std.Dev	Min	Q1	Median	Q3	Max	MAD	IQR	
qtd_comentarios	343.6183	800.395	0	1	9	289	8481	13.3434	288	2.32
qtd_compartilhamento	16832.2632	32523.935	0	105	695	18837	240154	1018.5462	18732	1.93
qtd_curtidas	84460.3282	135064.483	0	749	3232	122786	654588	4722.0810	122037	1.59
qtd_downloads	1033.0138	1976.879	0	6	45	1135	14308	65.2344	1129	1.91
qtd_visualizacoes	252977.5592	321991.205	35	4837	9897	492709	999127	14434.5936	487872	1.27

4. Matriz de Dispersão



5. Análise de Normalidade das Variáveis

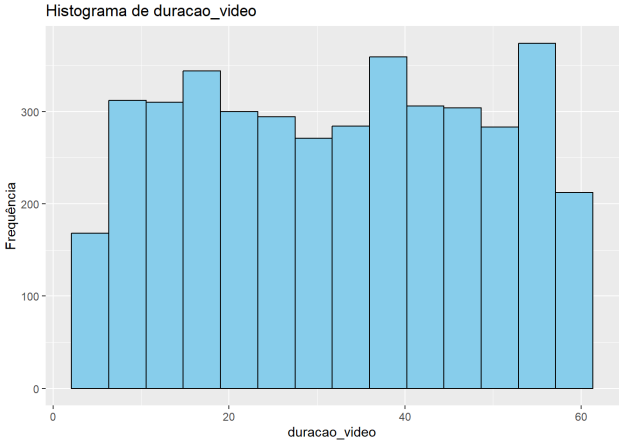
5A. O que é uma Distribuição Normal?

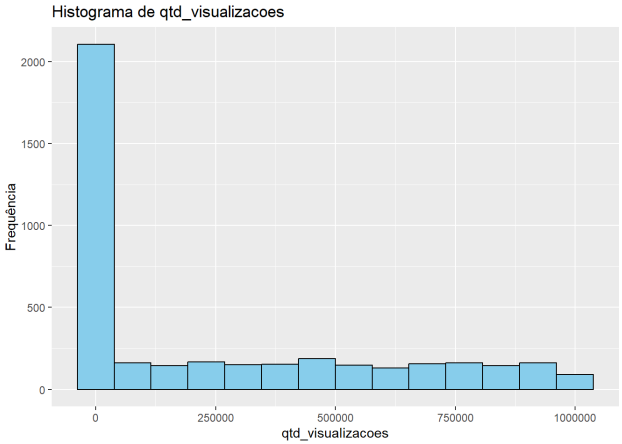
A distribuição normal, também conhecida como distribuição Gaussiana, é uma distribuição de probabilidade contínua caracterizada por uma curva simétrica em forma de sino. Ela é definida por sua média (μ) e desvio padrão (σ), sendo que: - A média, mediana e moda são iguais; - Aproximadamente 68% dos dados estão a um desvio padrão da média, 95% a dois desvios e 99,7% a três; - É amplamente utilizada em estatística devido ao Teorema Central do Limite, que afirma que, sob certas condições, a média de várias amostras independentes de uma população terá distribuição normal.

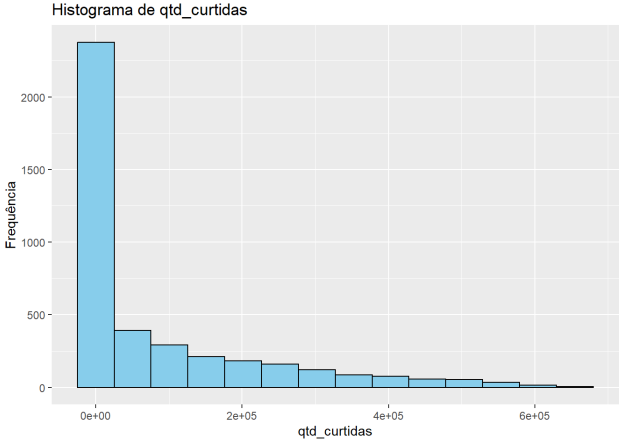
5B. Histogramas por Variável

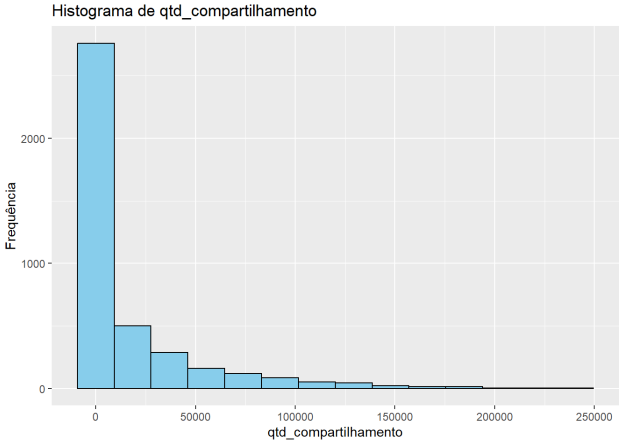
O número de *bins* foi escolhido com base na regra de Sturges: → k = 1 + 3,322 log(n) Onde k é o número de classes e n é o número total de observações. num_bins = 14.

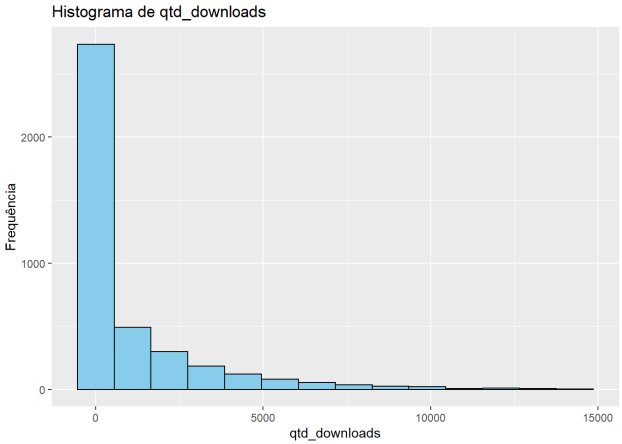
```
variaveis <- c("duracao_video", "qtd_visualizacoes", "qtd_curtidas",
               "qtd_compartilhamento", "qtd_downloads", "qtd_comentarios")

num_bins <- ceiling(log2(nrow(dados_videos)) + 1)

for (var in variaveis) {
  print(
    ggplot(dados_videos, aes_string(x = var)) +
      geom_histogram(bins = num_bins, fill = "skyblue", color = "black") +
      labs(title = paste("Histograma de", var), x = var, y = "Frequência")
  )
}
```

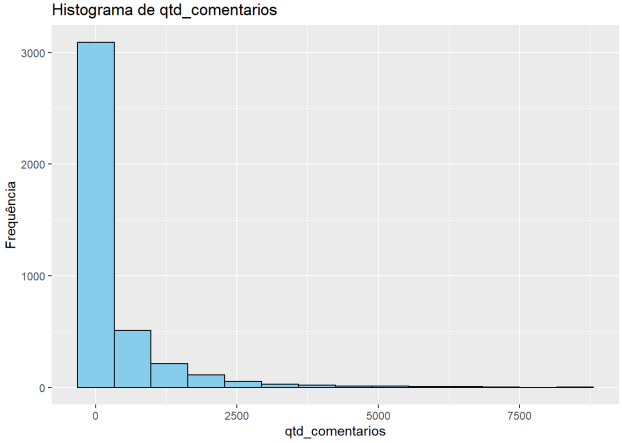

Histograma de duracao_video
Frequência
0
100
200
300
0
20
40
60
duracao_video


Histograma de qtd_visualizacoes
Frequência
0
500
1000
1500
2000
0
250000
500000
750000
1000000
qtd_visualizacoes


Histograma de qtd_curtidas
Frequência
0
500
1000
1500
2000
0e+00
2e+05
4e+05
6e+05
qtd_curtidas

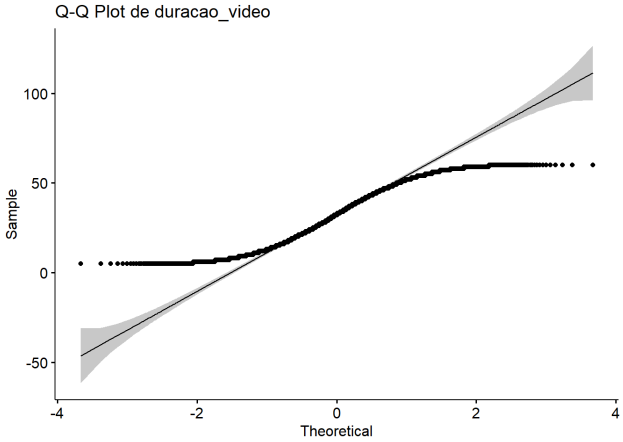

Histograma de qtd_compartilhamento
Frequência
2000
1000
0
0
50000
100000
150000
200000
250000
qtd_compartilhamento

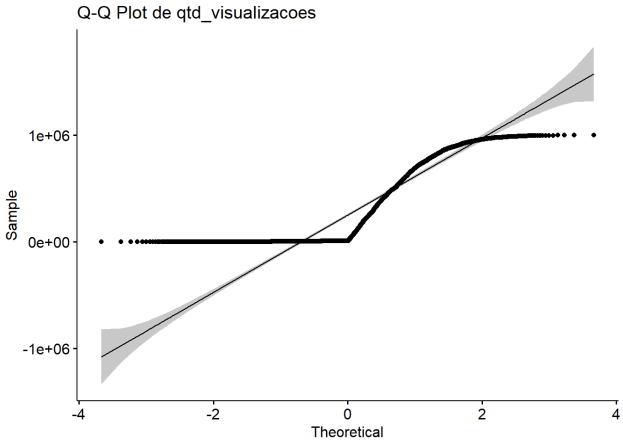

Histograma de qtd_downloads
Frequência
2000
1000
0
0
5000
10000
15000
qtd_downloads

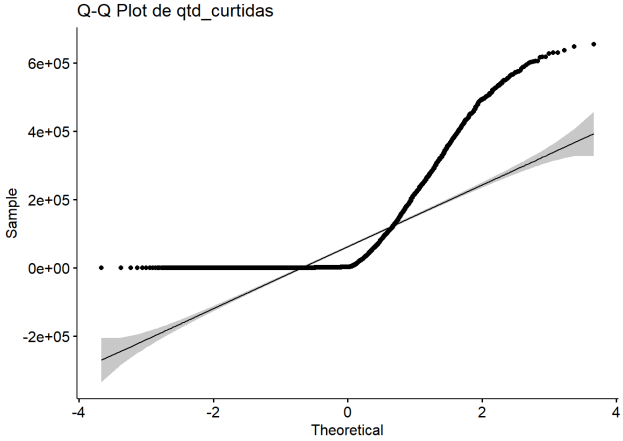

Histograma de qtd_comentarios
Frequência
3000
2000
1000
0
0
2500
5000
7500
qtd_comentarios

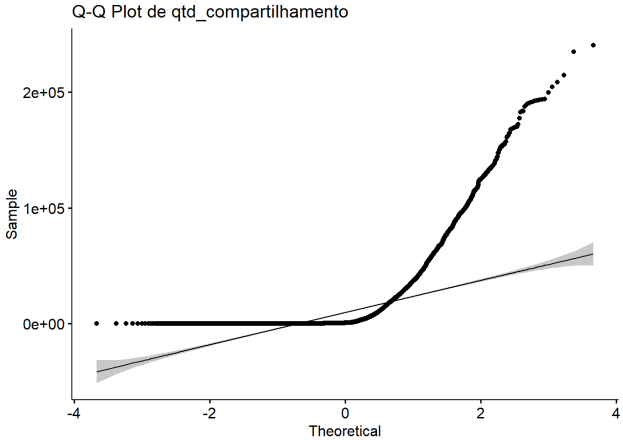
5C. Q-Q Plot por Variável

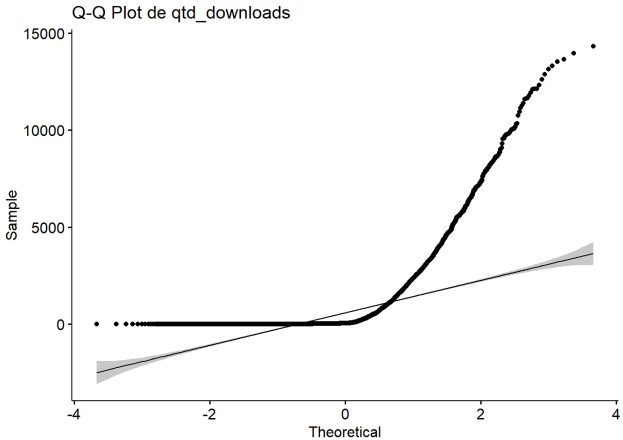
```
for (var in variaveis) {
  print(
    ggqqplot(dados_videos[[var]], title = paste("Q-Q Plot de", var))
  )
}
```

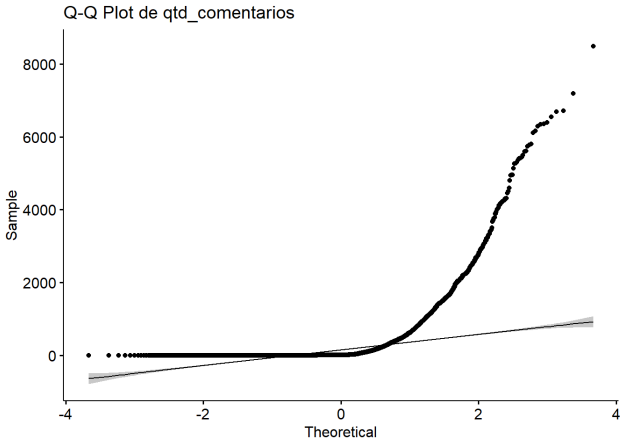

Q-Q Plot de duracao_video
Sample
Theoretical
100
50
0
-50
-4
-2
0
2
4


Q-Q Plot de qtd_visualizacoes
Sample
Theoretical
1e+06
0e+00
-1e+06
-4
-2
0
2
4


Q-Q Plot de qtd_curtidas
Sample
Theoretical
6e+05
4e+05
2e+05
0e+00
-2e+05
-4
-2
0
2
4







5D e 5E. Teste de Normalidade (Shapiro-Wilk)

Desta forma, nenhuma das variáveis numéricas da base de dados segue uma distribuição normal.

Segundo o teste de Shapiro-Wilk, rejeita-se a hipótese nula de normalidade sempre que o p-valor é inferior a 0.05. Neste caso, todos os p-valores são muito inferiores a esse limite, indicando que os dados não são normalmente distribuídos.

```
shapiro_resultados <- sapply(dados_videos[variaveis], function(x) shapiro.test(x)$p.value)
shapiro_resultados
```

```
##        duracao_video    qtd_visualizacoes         qtd_curtidas
##         5.083467e-35         3.188442e-60         6.863126e-66
## qtd_compartilhamento       qtd_downloads      qtd_comentarios
##         2.786075e-71         4.570776e-71         8.155866e-76
```

6. Completude dos Dados

Completude: Em pesquisa ao dicionário Oxford, seria referido a algo como qualidade, estado ou propriedade do que é completo, perfeito, acabado.

Mas para a visão dos dados, seria algo referente a ter todas as informações necessárias presentes no conjunto, no caso, ter os dados inteiros, sem colunas ou linhas vazias, campos nulos e afins. E isso é totalmente um dos pilares quando falamos em qualidade dos dados, porque sempre buscamos a confiabilidade, consistência e que sejam dados totalmente confiáveis. E o impacto na análise exploratória de dados (EDA) é praticamente o que significa a completude, a busca para que possamos fazer com que os dados sejam de máxima credibilidade, assertivos, consistentes e essa ação de tratar os dados, é para buscarmos a qualidade de completude ao negócio, pesquisa ou afins.

Podemos abordar diversas práticas de governança, metodologias e afins, mas organizações avançadas sempre estão monitorando a completude de seus dados, justamente por ser facilmente mensurável e diretamente ligado à confiança nas análises.

Por isso, a completude de dados não é apenas uma questão técnica — é também uma questão de credibilidade. Sem dados completos, a análise deixa de ser levada em consideração no negócio e passa a ser um empecilho, já que os dados são extremamente importantes para apoiar decisões estratégicas, e caso não tenhamos visão clara disso tudo, é como andar de carro, subindo uma serra cheia de neblina, um perigo enorme.

7. Completude dos Dados

```
completude <- sapply(dados_videos, function(x) sum(!is.na(x)) / length(x))
completude
```

```
##             video_id status_reinvidicacao        duracao_video
##            1.0000000            0.9854404            1.0000000
##    transcricao_video   status_verificacao         status_video
##            0.9854404            1.0000000            1.0000000
##    qtd_visualizacoes         qtd_curtidas qtd_compartilhamento
##            0.9854404            0.9854404            0.9854404
##       qtd_downloads      qtd_comentarios
##            0.9854404            0.9854404
```

8. Imputação de Dados com MICE

```
imp <- mice(dados_videos, m = 5, method = "pmm", seed = 123)
```

```
##
##  iter imp variable
##   1   1  qtd_visualizacoes  qtd_curtidas  qtd_compartilhamento  qtd_downloads  qtd_comentarios
##   1   2  qtd_visualizacoes  qtd_curtidas  qtd_compartilhamento  qtd_downloads  qtd_comentarios
##   1   3  qtd_visualizacoes  qtd_curtidas  qtd_compartilhamento  qtd_downloads  qtd_comentarios
##   1   4  qtd_visualizacoes  qtd_curtidas  qtd_compartilhamento  qtd_downloads  qtd_comentarios
##   1   5  qtd_visualizacoes  qtd_curtidas  qtd_compartilhamento  qtd_downloads  qtd_comentarios
##   2   1  qtd_visualizacoes  qtd_curtidas  qtd_compartilhamento  qtd_downloads  qtd_comentarios
##   2   2  qtd_visualizacoes  qtd_curtidas  qtd_compartilhamento  qtd_downloads  qtd_comentarios
##   2   3  qtd_visualizacoes  qtd_curtidas  qtd_compartilhamento  qtd_downloads  qtd_comentarios
##   2   4  qtd_visualizacoes  qtd_curtidas  qtd_compartilhamento  qtd_downloads  qtd_comentarios
##   2   5  qtd_visualizacoes  qtd_curtidas  qtd_compartilhamento  qtd_downloads  qtd_comentarios
##   3   1  qtd_visualizacoes  qtd_curtidas  qtd_compartilhamento  qtd_downloads  qtd_comentarios
##   3   2  qtd_visualizacoes  qtd_curtidas  qtd_compartilhamento  qtd_downloads  qtd_comentarios
##   3   3  qtd_visualizacoes  qtd_curtidas  qtd_compartilhamento  qtd_downloads  qtd_comentarios
##   3   4  qtd_visualizacoes  qtd_curtidas  qtd_compartilhamento  qtd_downloads  qtd_comentarios
##   3   5  qtd_visualizacoes  qtd_curtidas  qtd_compartilhamento  qtd_downloads  qtd_comentarios
##   4   1  qtd_visualizacoes  qtd_curtidas  qtd_compartilhamento  qtd_downloads  qtd_comentarios
##   4   2  qtd_visualizacoes  qtd_curtidas  qtd_compartilhamento  qtd_downloads  qtd_comentarios
##   4   3  qtd_visualizacoes  qtd_curtidas  qtd_compartilhamento  qtd_downloads  qtd_comentarios
##   4   4  qtd_visualizacoes  qtd_curtidas  qtd_compartilhamento  qtd_downloads  qtd_comentarios
##   4   5  qtd_visualizacoes  qtd_curtidas  qtd_compartilhamento  qtd_downloads  qtd_comentarios
##   5   1  qtd_visualizacoes  qtd_curtidas  qtd_compartilhamento  qtd_downloads  qtd_comentarios
##   5   2  qtd_visualizacoes  qtd_curtidas  qtd_compartilhamento  qtd_downloads  qtd_comentarios
##   5   3  qtd_visualizacoes  qtd_curtidas  qtd_compartilhamento  qtd_downloads  qtd_comentarios
##   5   4  qtd_visualizacoes  qtd_curtidas  qtd_compartilhamento  qtd_downloads  qtd_comentarios
##   5   5  qtd_visualizacoes  qtd_curtidas  qtd_compartilhamento  qtd_downloads  qtd_comentarios
```

```
dados_imputados <- complete(imp, 1)
head(dados_imputados)
```

```
##   video_id status_reinvidicacao duracao_video
## 1        1               claim            59
## 2        2               claim            32
## 3        3               claim            31
## 4        4               claim            25
## 5        5               claim            19
## 6        6               claim            35
##
transcricao_video
## 1                                      someone shared with me that drone deliveries are already happening and will bec
ome common by 2025
## 2                            someone shared with me that there are more microorganisms in one teaspoon of soil than pe
ople on the planet
## 3 someone shared with me that american industrialist andrew carnegie had a net worth of $475 million usd, worth over $300
billion usd today
## 4       someone shared with me that the metro of st. petersburg, with an average depth of hundred meters, is the deepest
metro in the world
## 5         someone shared with me that the number of businesses allowing employees to bring pets to the workplace has gro
wn by 6% worldwide
## 6           someone shared with me that gross domestic product (gdp) is the best financial indicator of a country's overa
ll trade potential
##   status_verificacao status_video qtd_visualizacoes qtd_curtidas
## 1       not verified under review            343296        19425
## 2       not verified       active            140877        77355
## 3       not verified       active            902185        97690
## 4       not verified       active            437506       239954
## 5       not verified       active             56167        34987
## 6       not verified under review            336647       175546
##   qtd_compartilhamento qtd_downloads qtd_comentarios
## 1                  241            1               0
## 2                19034         1161             684
## 3                 2858          833             329
## 4                34812         1234             584
## 5                 4110          547             152
## 6                62303         4293            1857
```

```
colSums(is.na(dados_imputados))
```

```
##             video_id status_reinvidicacao        duracao_video
##                    0                   60                    0
##    transcricao_video   status_verificacao         status_video
##                   60                    0                    0
##    qtd_visualizacoes         qtd_curtidas qtd_compartilhamento
##                    0                    0                    0
##       qtd_downloads      qtd_comentarios
##                    0                    0
```

Verificando quantidade de ajustes foi realizado pelo Pacote MICE - modelo 2.

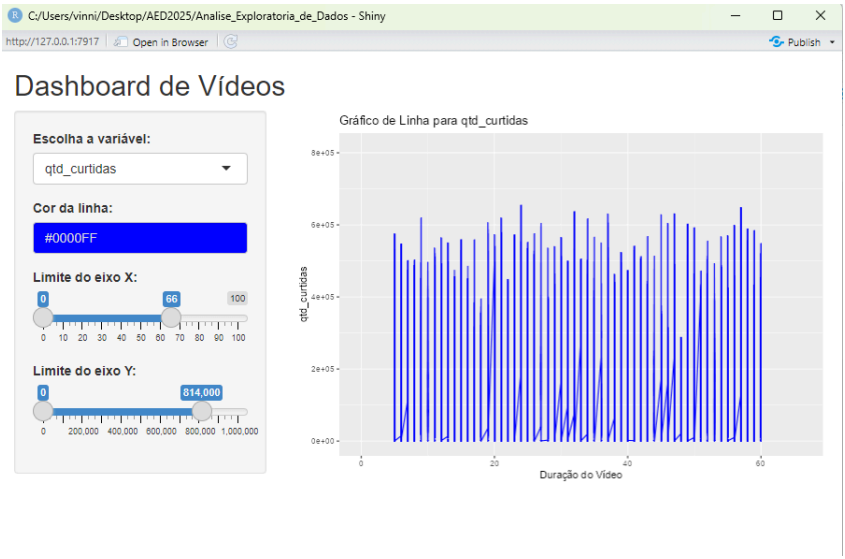
```
colSums(is.na(dados_imputados))
```

```
##             video_id status_reinvidicacao        duracao_video
##                    0                   60                    0
##    transcricao_video   status_verificacao         status_video
##                   60                    0                    0
##    qtd_visualizacoes         qtd_curtidas qtd_compartilhamento
##                    0                    0                    0
##       qtd_downloads      qtd_comentarios
##                    0                    0
```

9. Dashboard com Shiny

```
# Interface do usuário
ui <- fluidPage(
  titlePanel("Dashboard de Vídeos"),
  sidebarLayout(
    sidebarPanel(
      selectInput("variavel", "Escolha a variável:",
                  choices = c("qtd_visualizacoes", "qtd_curtidas",
                              "qtd_compartilhamento", "qtd_downloads", "qtd_comentarios")),
      colourInput("cor", "Cor da linha:", value = "blue"),
      sliderInput("xlim", "Limite do eixo X:", min = 0, max = 100, value = c(0, 100)),
      sliderInput("ylim", "Limite do eixo Y:", min = 0, max = 1000000, value = c(0, 500000))
    ),
    mainPanel(
      plotOutput("grafico")  # Corrigido: nome em minúsculo
    )
  )
)

# Servidor
server <- function(input, output) {
  output$grafico <- renderPlot({
    ggplot(dados_videos, aes_string(x = "duracao_video", y = input$variavel)) +
      geom_line(color = input$cor) +
      coord_cartesian(xlim = input$xlim, ylim = input$ylim) +
      labs(title = paste("Gráfico de Linha para", input$variavel),
           x = "Duração do Vídeo", y = input$variavel)
  })
}

# Rodar o app
shinyApp(ui = ui, server = server)
```

Foto do Dashboard com Shiny



Referências

- Fonte dos dados: base simulada `dados_videos.xlsx`
- Pacotes utilizados: `ggplot2`, `summarytools`, `GGally`, `ggpubr`, `mice`, `shiny`, `colourpicker`
- #10 Código disponibilizado em: https://github.com/VinniciusL/EDA2025 (https://github.com/VinniciusL/EDA2025)
- Curso: Análise Exploratória de Dados com R