Universidade Federal do Rio de Janeiro
Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia

Programa de Engenharia de Sistemas e Computação

COS868 - Probabilidade e Estatística para Aprendizado de Máquina

Profa. Dra. Rosa M. Leão (PESC/COPPE/UFRJ)

Projeto do Curso

Luiz Henrique Souza Caldas
email: lhscaldas@cos.ufrj.br

30 de dezembro de 2024

Conteúdo

1 Introdução

1.1 Objetivo

O objetivo deste trabalho é realizar uma análise de um conjunto de dados reais fornecidos por um provedor de Internet de médio porte, avaliando as taxas de upload e download de dispositivos domésticos, especificamente Smart-TVs e Chromecasts, com base na teoria aprendida em classe, destacando a importância de uma análise crítica dos resultados obtidos.

1.2 Análise Exploratória dos Dados

A análise exploratória foi realizada para compreender as características principais dos dados obtidos dos dispositivos Smart TV e Chromecast. Os resultados estão detalhados abaixo:

- **Primeiras linhas dos dados:**
  - **Smart TV:**

	device_id	date_hour	bytes_up	bytes_down
0	77209603	2021-11-22 15:23:00	132932.983607	2.818140e+06
1	77209603	2021-11-22 15:24:00	115770.491803	2.264410e+06
2	77209603	2021-11-22 15:25:00	114030.032787	2.309270e+06
3	77209603	2021-11-22 15:26:00	97170.622951	2.006544e+06
4	77209603	2021-11-22 15:27:00	39569.573770	8.061440e+05

  - **Chromecast:**

	device_id	date_hour	bytes_up	bytes_down
0	66161985	2021-09-06 00:01:00	2987.016393	49185.704918
1	66161985	2021-09-06 00:02:00	685.935484	328.258065
2	66161985	2021-09-06 00:03:00	4493.901639	37914.064516
3	66161985	2021-09-06 00:04:00	776.133333	229.200000
4	66161985	2021-09-06 00:05:00	3081.311475	51656.800000

- **Dimensões dos dados:**
  - Smart TV: (4417903, 4)
  - Chromecast: (1620529, 4)
- **Dados faltantes:**
  - Smart TV: Nenhum valor faltante em `device_id`, `date_hour`, `bytes_up`, `bytes_down`.
  - Chromecast: Nenhum valor faltante em `device_id`, `date_hour`, `bytes_up`, `bytes_down`.
- **Valores zero:**
  - Smart TV: `bytes_up` = 1.803.853, `bytes_down` = 1.978.337.

  - Chromecast: `bytes_up` = 6.057, `bytes_down` = 4.099.

- **Valores negativos:**
  - Smart TV: Nenhum valor negativo em `bytes_up` ou `bytes_down`.
  - Chromecast: Nenhum valor negativo em `bytes_up` ou `bytes_down`.

1.3 Pré-processamento

O pré-processamento foi realizado para preparar os dados dos dispositivos Smart TV e Chromecast para análises posteriores. As etapas realizadas são descritas a seguir:

- **Carregamento dos dados:** Os dados foram lidos a partir dos arquivos `dataset_smart-tv.csv` e `dataset_chromecast.csv`.
- **Correção de valores zero:** Como as colunas `bytes_up` e `bytes_down` apresentavam valores zero, foi aplicado um *shift* de +1 a todos os valores dessas colunas para evitar problemas no cálculo do logaritmo.
- **Reescalonamento dos dados:** Os valores das colunas `bytes_up` e `bytes_down` foram transformados para a escala logarítmica na base 10 (`log10`), devido à grande variação na ordem de grandeza desses valores.
- **Ordenação temporal:** Os dados foram ordenados pela coluna `date_hour` para garantir a consistência temporal nas análises subsequentes.
- **Salvamento dos dados processados:** Os datasets resultantes podem ser salvos como arquivos CSV (`smart_preprocessado.csv` e `chrome_preprocessado.csv`) para uso posterior.

Essa etapa garante que os dados estejam limpos, reescalonados e organizados, facilitando análises estatísticas e a geração de gráficos. Além disso, a transformação logarítmica reduz a influência de valores extremos, melhorando a interpretação dos resultados.

2 Estatísticas Gerais

Nesta seção, são apresentadas as estatísticas gerais dos dados coletados para os dispositivos Smart TV e Chromecast. As análises incluem cálculos de medidas descritivas, como média, variância e desvio padrão, além de representações gráficas através de histogramas, boxplots e funções de distribuição empírica (ECDF).

2.1 Medidas Descritivas

As medidas descritivas para as taxas de upload e download (em escala logarítmica base 10) estão resumidas na Tabela 1.

Tabela 1: Medidas descritivas das taxas de upload e download.

Dispositivo	Tipo de Tráfego	Média	Variância	Desvio Padrão
Smart TV	Upload	2.16	4.11	2.03
Smart TV	Download	2.35	6.72	2.59
Chromecast	Upload	3.35	0.46	0.68
Chromecast	Download	3.80	1.66	1.29

2.2 Visualizações Gráficas

Para compreender melhor a distribuição dos dados, são utilizadas as seguintes representações gráficas:

- **Histogramas:** As distribuições das taxas de upload e download para cada dispositivo estão representadas nos histogramas da Figura 1.
- **Boxplots:** A Figura 2 mostra os boxplots comparando as taxas de upload e download entre Smart TV e Chromecast.
- **ECDF:** As funções de distribuição empírica, exibidas na Figura 3, demonstram a probabilidade acumulada para cada valor das taxas.

Para a construção dos histogramas, o número de bins foi calculado utilizando o método de Sturges:

$$k = 1 + \log_2(n), \tag{1}$$

onde n é o número total de amostras. Este método busca otimizar a visualização dos dados ao balancear granularidade e clareza.

O número de bins calculado para cada dispositivo é o seguinte:

- **Smart TV:** $k = 24$ bins.
- **Chromecast:** $k = 22$ bins.

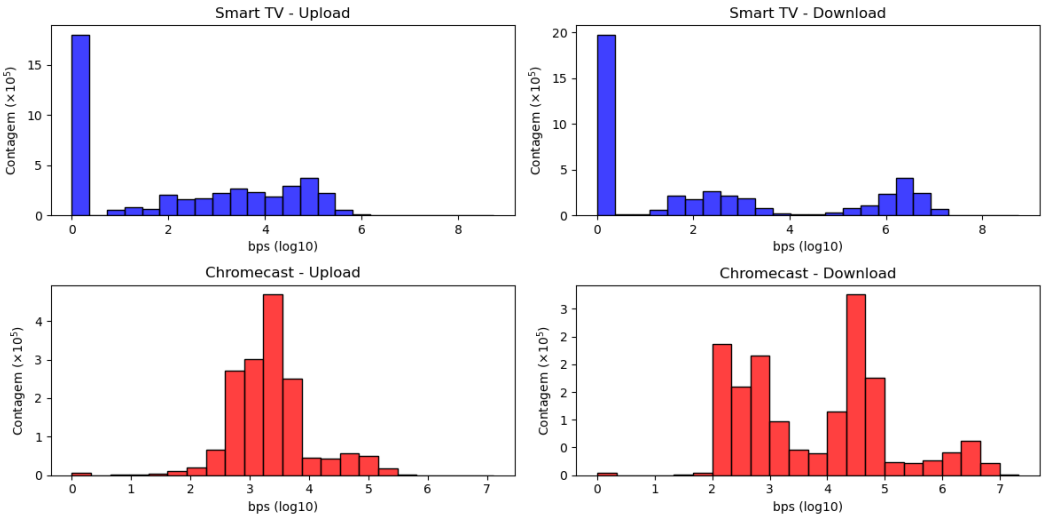

Figura 1: Histogramas das taxas de upload e download para Smart TV e Chromecast.

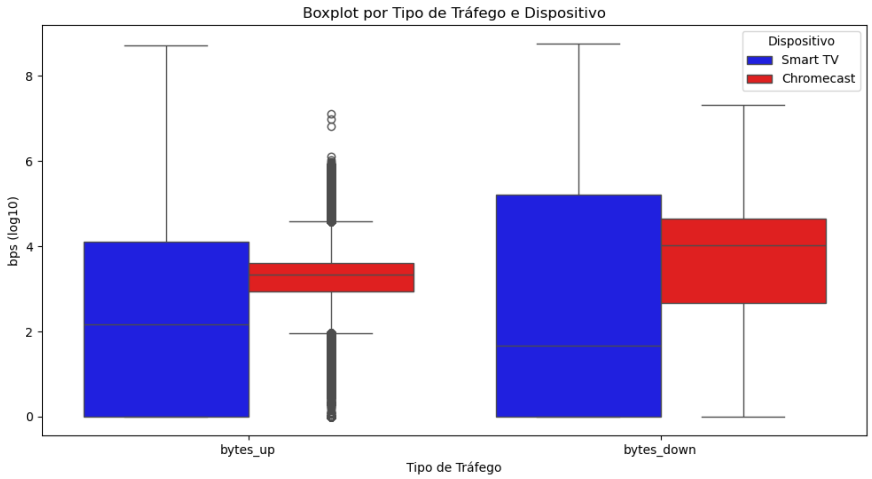

Figura 2: Boxplots das taxas de upload e download para Smart TV e Chromecast.

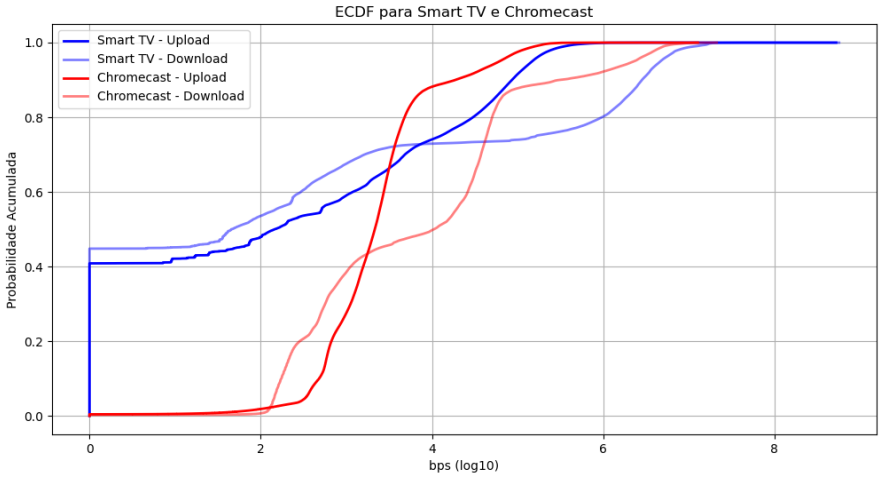

Figura 3: Funções de Distribuição Empírica (ECDF) das taxas de upload e download.

2.3 Análise dos Resultados

Os resultados destacam diferenças importantes nas características das taxas de upload e download entre os dispositivos Smart TV e Chromecast.

- **Smart TV:** As taxas de upload e download da Smart TV possuem médias parecidas e variâncias relativamente altas, indicando uma dispersão maior dos dados. As taxas estão predominantemente concentradas em valores baixos, especialmente em valores iguais a zero, conforme já havia sido evidenciado na Análise Exploratória dos Dados (Seção 1.2). Essa característica é refletida na primeira barra dos histogramas (Figura 1), que é consideravelmente maior do que as demais. No boxplot (Figura 2), essa concentração é representada pela proximidade do limite inferior ao primeiro quartil ($Q1$). Na mesma figura também, é possível observar a ausência de outliers nos dois tipos de tráfego. Além disso, a ECDF (Figura 3) apresenta um valor inicial relativamente alto, superior a 0.4, devido à grande quantidade de valores nulos, crescendo de forma lenta até chegar ao valor máximo.

- **Chromecast:** As taxas de upload e download do Chromecast também possuem médias próximas, mas apresentam desvios padrão relativamente baixos, indicando uma dispersão menor dos dados. A taxa de download (`bytes_down`) exibe uma variância maior do que a de upload (`bytes_up`). A taxa de upload apresenta muitos outliers, tanto para valores altos (picos) quanto para valores baixos (vales), como destacado no boxplot (Figura 2), enquanto a taxa de download não possui nenhum. A ECDF (Figura 3) reflete essa característica, apresentando um crescimento rápido após 10^2 bps.

- **Comparação Geral:** A Smart TV e o Chromecast apresentam diferenças marcantes em seus padrões de tráfego. Enquanto a Smart TV concentra grande parte de seus dados em valores baixos, com ausência de outliers, o Chromecast exibe menor dispersão geral, mas com muitos outliers na taxa de upload. A ECDF da Smart TV cresce de forma lenta devido aos valores nulos iniciais, enquanto a do Chromecast apresenta um crescimento rápido após 10^2

bps, refletindo uma concentração maior em valores intermediários. Essas diferenças sugerem que a Smart TV alterna entre períodos de inatividade e altos fluxos, enquanto o Chromecast apresenta tráfego mais estável, mas com picos e vales ocasionais no upload.

Essas observações podem auxiliar no desenvolvimento de estratégias de gerenciamento de rede mais eficientes, considerando a alta variabilidade e períodos de inatividade da Smart TV, e os picos de tráfego ocasionais no upload do Chromecast. Adaptar essas estratégias às características específicas de cada dispositivo pode melhorar a alocação de recursos e a experiência do usuário.

3 Estatísticas por Horário

Nesta seção, apresentamos as estatísticas por horário, independente do dia, dos dados coletados para os dispositivos Smart TV e Chromecast. As análises incluem cálculos de medidas descritivas, como média, variância e desvio padrão, além da representação gráficas através de boxplots.

3.1 Medidas Descritivas

A Tabela 2 exibe as estatísticas por horário para as taxas de upload e download dos dispositivos Smart TV e Chromecast. As estatísticas incluem a média, variância e desvio padrão para cada hora do dia, separadas por tipo de dispositivo e tipo de tráfego.

Tabela 2: Estatísticas por Horário para Smart TV e Chromecast

Hora	Dispositivo	Tráfego	Média	Variância	Desvio Padrão
0	Smart TV	Upload	1.89	4.16	2.04
0	Smart TV	Download	2.10	6.89	2.62
0	Chromecast	Upload	3.43	0.63	0.79
0	Chromecast	Download	3.95	2.06	1.44
1	Smart TV	Upload	1.47	3.76	1.94
1	Smart TV	Download	1.60	6.05	2.46
1	Chromecast	Upload	3.32	0.48	0.69
1	Chromecast	Download	3.78	1.74	1.32
2	Smart TV	Upload	1.15	3.15	1.78
2	Smart TV	Download	1.23	4.96	2.23
2	Chromecast	Upload	3.24	0.34	0.58
2	Chromecast	Download	3.69	1.46	1.21
3	Smart TV	Upload	0.89	2.46	1.57
3	Smart TV	Download	0.90	3.63	1.90
3	Chromecast	Upload	3.20	0.31	0.55
3	Chromecast	Download	3.64	1.41	1.19
4	Smart TV	Upload	0.77	2.06	1.43
4	Smart TV	Download	0.74	2.89	1.70
4	Chromecast	Upload	3.18	0.31	0.56
4	Chromecast	Download	3.62	1.41	1.19
5	Smart TV	Upload	0.88	2.35	1.53
5	Smart TV	Download	0.89	3.52	1.88
5	Chromecast	Upload	3.16	0.29	0.54
5	Chromecast	Download	3.57	1.38	1.17
6	Smart TV	Upload	1.02	2.64	1.62
6	Smart TV	Download	1.07	4.00	2.00
6	Chromecast	Upload	3.16	0.30	0.55
6	Chromecast	Download	3.57	1.37	1.17

Hora	Dispositivo	Tipo	Média	Variância	Desvio Padrão
7	Smart TV	Upload	1.20	3.00	1.73
7	Smart TV	Download	1.24	4.40	2.10
7	Chromecast	Upload	3.20	0.34	0.58
7	Chromecast	Download	3.62	1.43	1.19
8	Smart TV	Upload	1.39	3.53	1.88
8	Smart TV	Download	1.48	5.32	2.31
8	Chromecast	Upload	3.24	0.39	0.62
8	Chromecast	Download	3.65	1.49	1.22
9	Smart TV	Upload	1.72	3.97	1.99
9	Smart TV	Download	1.87	6.25	2.50
9	Chromecast	Upload	3.29	0.40	0.63
9	Chromecast	Download	3.70	1.51	1.23
10	Smart TV	Upload	2.02	4.24	2.06
10	Smart TV	Download	2.23	6.88	2.62
10	Chromecast	Upload	3.30	0.41	0.64
10	Chromecast	Download	3.71	1.52	1.23
11	Smart TV	Upload	2.27	4.27	2.07
11	Smart TV	Download	2.53	7.06	2.66
11	Chromecast	Upload	3.32	0.41	0.64
11	Chromecast	Download	3.74	1.51	1.23
12	Smart TV	Upload	2.47	4.16	2.04
12	Smart TV	Download	2.78	7.04	2.65
12	Chromecast	Upload	3.35	0.40	0.64
12	Chromecast	Download	3.78	1.54	1.24
13	Smart TV	Upload	2.49	4.14	2.03
13	Smart TV	Download	2.78	7.00	2.65
13	Chromecast	Upload	3.35	0.43	0.65
13	Chromecast	Download	3.79	1.59	1.26
14	Smart TV	Upload	2.56	4.22	2.05
14	Smart TV	Download	2.88	7.24	2.69
14	Chromecast	Upload	3.36	0.43	0.65
14	Chromecast	Download	3.80	1.58	1.26
15	Smart TV	Upload	2.61	4.12	2.03
15	Smart TV	Download	2.92	7.16	2.68
15	Chromecast	Upload	3.38	0.44	0.66
15	Chromecast	Download	3.83	1.62	1.27
16	Smart TV	Upload	2.62	3.87	1.97
16	Smart TV	Download	2.88	6.76	2.60
16	Chromecast	Upload	3.40	0.48	0.69
16	Chromecast	Download	3.87	1.71	1.31

Hora	Dispositivo	Tipo	Média	Variância	Desvio Padrão
17	Smart TV	Upload	2.74	3.59	1.90
17	Smart TV	Download	2.96	6.42	2.53
17	Chromecast	Upload	3.41	0.50	0.70
17	Chromecast	Download	3.88	1.73	1.32
18	Smart TV	Upload	2.95	3.35	1.83
18	Smart TV	Download	3.19	6.24	2.50
18	Chromecast	Upload	3.40	0.47	0.69
18	Chromecast	Download	3.86	1.66	1.29
19	Smart TV	Upload	3.05	3.28	1.81
19	Smart TV	Download	3.32	6.29	2.51
19	Chromecast	Upload	3.42	0.48	0.70
19	Chromecast	Download	3.85	1.66	1.29
20	Smart TV	Upload	3.12	3.17	1.78
20	Smart TV	Download	3.40	6.20	2.49
20	Chromecast	Upload	3.47	0.49	0.70
20	Chromecast	Download	3.92	1.75	1.32
21	Smart TV	Upload	3.10	3.13	1.77
21	Smart TV	Download	3.37	6.13	2.47
21	Chromecast	Upload	3.49	0.54	0.74
21	Chromecast	Download	3.97	1.86	1.36
22	Smart TV	Upload	2.84	3.46	1.86
22	Smart TV	Download	3.06	6.29	2.51
22	Chromecast	Upload	3.52	0.60	0.77
22	Chromecast	Download	4.04	1.97	1.40
23	Smart TV	Upload	2.37	3.94	1.98
23	Smart TV	Download	2.59	6.65	2.58
23	Chromecast	Upload	3.51	0.69	0.83
23	Chromecast	Download	4.05	2.16	1.47

3.2 Visualizações Gráficas

Para melhorar a visualização da Tabela 2, os dados foram plotados nos 4 gráficos contidos na Figura 4.

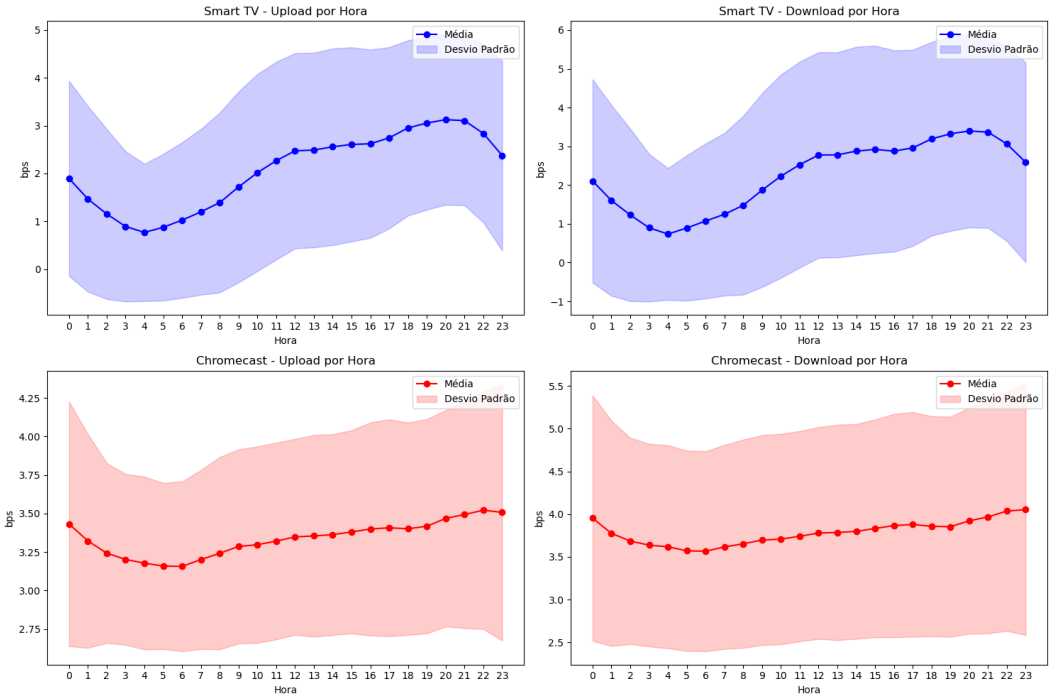

Figura 4: Gráficos das estatísticas por horário para Smart TV e Chromecast

Além disso, foram gerados boxplots para cada hora do dia, separados por tipo de dispositivo e tipo de tráfego. As Figuras 5 e 6 exibem os boxplots das taxas de upload e download para o dispositivo Smart TV, respectivamente. Já as Figuras 7 e 8 mostram os boxplots das taxas de upload e download para o dispositivo Chromecast, respectivamente.

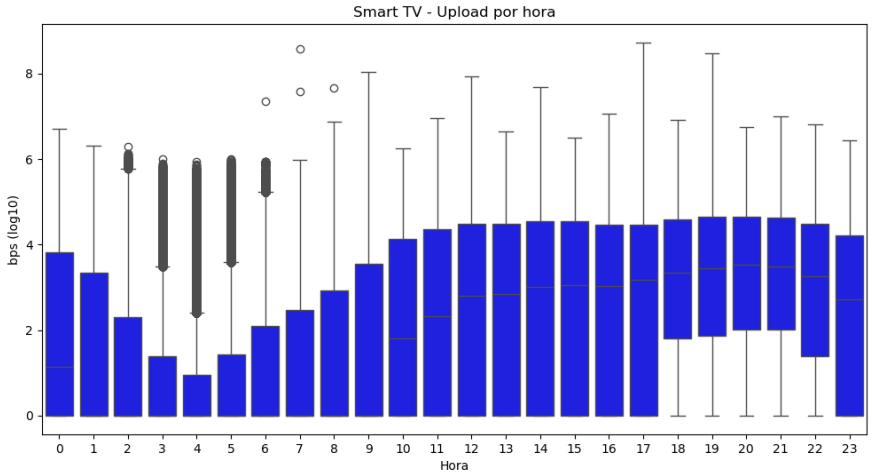

Figura 5: Boxplot das taxas de upload para Smart TV

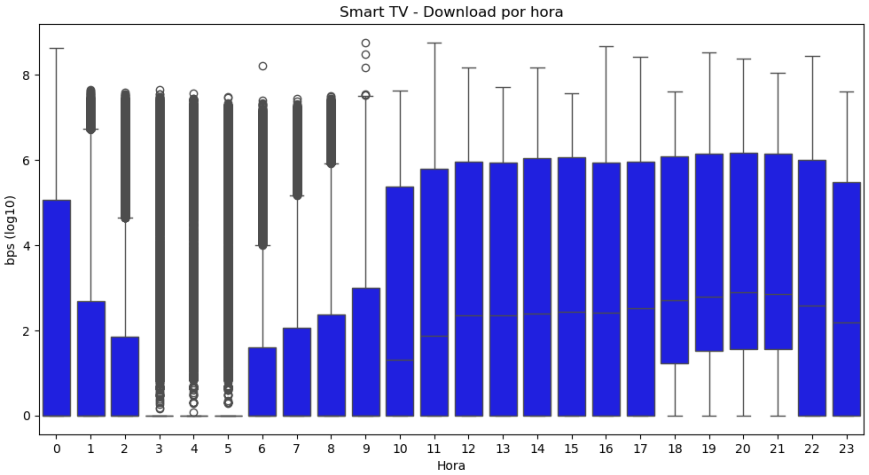

Figura 6: Boxplot das taxas de download para Smart TV

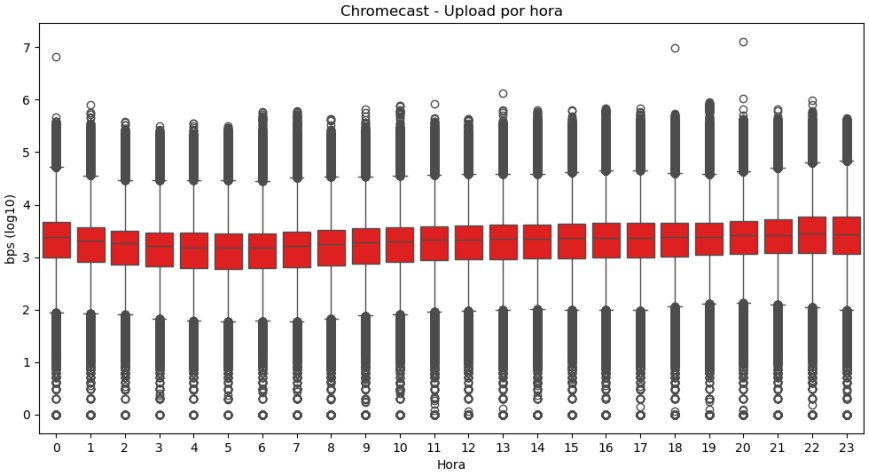

Figura 7: Boxplot das taxas de upload para Chromecast

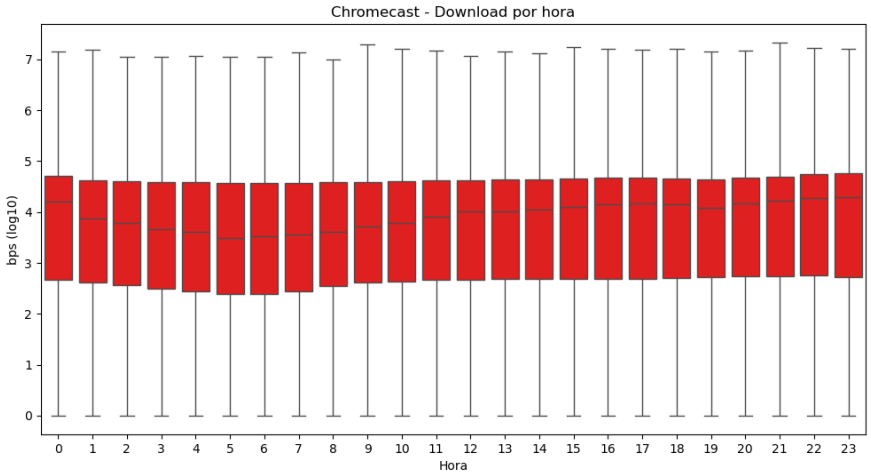

Figura 8: Boxplot das taxas de download para Chromecast

3.3 Análise dos Resultados

A partir da Tabela 2 (e, consequentemente, da Figura 4) e dos boxplots das figuras 5, 6, 7 e 8, podemos tirar algumas conclusões importantes sobre o comportamento das taxas de upload e download para os dispositivos Smart TV e Chromecast ao longo das 24 horas do dia:

- **Smart TV:**
  - As taxas de download e upload variam ao longo do dia, com uma tendência de aumento nas horas da tarde e noite.
  - A média das taxas de download e upload é geralmente menor durante a madrugada e aumenta gradualmente até atingir picos durante a noite (20h tanto para download quanto para upload).
- **Chromecast:**
  - As taxas de download e upload para o Chromecast são consistentemente mais altas do que para a Smart TV, especialmente durante as horas da tarde e noite.
  - A média das taxas de download e upload é relativamente estável ao longo do dia, em comparação com a Smart TV.
  - Os picos de média ocorrem durante a noite (23h para download e 22h para upload), indicando um uso mais intenso da rede nesses horários.
  - A variância e o desvio padrão das taxas de download e upload são menores para o Chromecast, indicando uma utilização mais consistente da rede.

Essas observações sugerem que o uso da rede para a Smart TV é mais variável e depende mais do horário do dia, enquanto o Chromecast apresenta um uso mais constante. Isso pode ser devido a diferentes padrões de uso dos dispositivos, onde a Smart TV pode ser mais utilizada para atividades que demandam maior largura de banda em horários específicos, como streaming de vídeos em alta definição durante a tarde e noite.

4 Caracterizando os Horários com Maior Valor de Tráfego

Nesta seção, os horários com maior valor médio das taxas de upload e download para cada tipo de dispositivo, Smart TV e Chromecast, foram analisados seguindo os passos descritos.

4.1 Passo 1: Seleção dos Horários

A partir dos gráficos de médias por hora apresentados Figura 4, os horários com maior valor médio para cada taxa e dispositivo foram identificados:

- **Smart TV:**
  - dataset 1: composto pelo horário com maior média de upload (20:00).
  - dataset 2: composto pelo horário com maior média de download (20:00).
- **Chromecast:**
  - dataset 3: composto pelo horário com maior média de upload (22:00).
  - dataset 4: composto pelo horário com maior média de download (23:00).

4.2 Passo 2: Histogramas dos Dados

Histogramas foram gerados para cada um dos 4 datasets criados no Passo 1. O método de Sturges (Equação 1) foi utilizado para determinar o número adequado de bins, obtendo-se os seguintes valores:

- **Dataset 1**(Smart TV - Upload)**:** 19 bins
- **Dataset 2**(Smart TV - Download)**:** 19 bins
- **Dataset 3**(Chromecast - Upload)**:** 20 bins
- **Dataset 4**(Chromecast - Download)**:** 20 bins

Esses histogramas destacam os padrões de distribuição das taxas de upload e download para os horários selecionados, conforme ilustrado na Figura 9.

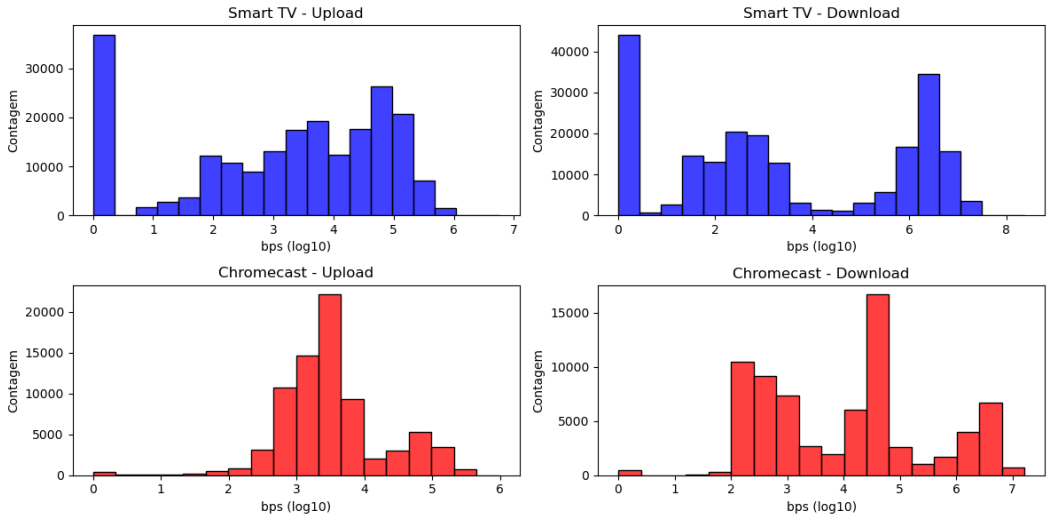

Figura 9: Histogramas das taxas de upload e download para os horários selecionados.

4.3 Passo 3: Estimativa de Parâmetros via MLE

Os parâmetros das distribuições Gaussiana e Gamma foram estimados utilizando o método de Máxima Verossimilhança (*Maximum Likelihood Estimation* - MLE) para os quatro conjuntos de dados. Esses valores foram aplicados na modelagem das distribuições, permitindo uma análise comparativa com os dados observados.

O MLE consiste em determinar os parâmetros que maximizam a função de verossimilhança, que mede a probabilidade dos dados observados para um conjunto de parâmetros. Para simplificar os cálculos, o logaritmo da verossimilhança (*log-likelihood*) é utilizado. A derivada da *log-likelihood* é igualada a zero para encontrar os estimadores de máxima verossimilhança dos parâmetros.

As funções de densidade de probabilidade para as distribuições Gaussiana e Gamma são definidas como:

- **Gaussiana:**

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2}$$

  onde μ é a média e σ^2 é a variância da distribuição.

- **Gamma:**

$$f(x) = \frac{\beta^\alpha x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha)} \tag{3}$$

  onde α é o parâmetro de forma, β é o parâmetro de escala, e $\Gamma(\alpha)$ é a função Gamma, definida como $\int_0^\infty x^{\alpha-1} e^{-x} dx$.

Para a distribuição Gaussiana, as estimativas dos parâmetros são obtidas de forma direta:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \tag{4}$$

No caso da distribuição Gamma, a estimação dos parâmetros α (forma) e β (escala) pelo MLE não possui soluções analíticas simples e geralmente requer métodos numéricos iterativos. O parâmetro α é frequentemente estimado utilizando o método de Newton-Raphson aplicado à função de log-verossimilhança, enquanto β pode ser estimado a partir de α e da média amostral $\bar{x}$ [1]:

$$\hat{\beta} = \frac{\bar{x}}{\hat{\alpha}} \tag{5}$$

Para realizar essas estimativas, foi utilizado o método `gamma.fit` da biblioteca `scipy.stats` do Python. Este método aplica o MLE de forma eficiente, empregando algoritmos de otimização numérica para determinar os parâmetros que melhor se ajustam aos dados observados. Além dos parâmetros de forma (α) e escala (β), o método também estima o parâmetro de localização (loc), que desloca a distribuição Gamma ao longo do eixo x. Esse deslocamento é essencial quando os dados incluem valores iguais a zero (o que é o caso, como mostrado na Seção 1.2), já que a função de densidade de probabilidade (PDF) da distribuição Gamma é indefinida para $x = 0$ quando $loc = 0$. Com $loc > 0$, a PDF é modificada para começar em $x = loc$, tornando possível ajustar a distribuição mesmo em presença de valores nulos ou muito baixos. A utilização desse parâmetro garante que a modelagem estatística permaneça válida e consistente com as características dos dados.

Os resultados das estimativas de parâmetros via MLE para as distribuições Gaussiana e Gamma são apresentados na Tabela 3.

Tabela 3: Resultados das estimativas de parâmetros via MLE

Dataset	$\hat{\boldsymbol{\mu}}$	$\hat{\boldsymbol{\sigma}}^{\mathbf{2}}$	$\hat{\boldsymbol{\alpha}}$	$\hat{\boldsymbol{\beta}}$	loc
Dataset 1 (Smart TV - Upload)	3.2243	3.1687	214.6171	0.1245	-23.4982
Dataset 2 (Smart TV - Download)	3.4961	6.2013	883.8791	0.0838	-70.5760
Dataset 3 (Chromecast - Upload)	3.6215	0.5957	3078.6394	0.0139	-39.2307
Dataset 4 (Chromecast - Download)	4.1527	2.1594	27.1301	0.2832	-3.5314

Além disso, as *log-likelihoods* e *likelihoods* para as distribuições Gaussiana e Gamma são apresentadas na Tabela 4. As *log-likelihoods* foram calculadas primeiro, para evitar erros de *underflow* ao calcular as *likelihoods*, substituindo a multiplicação de valores muito pequenos pela soma de seus logaritmos naturais.

Tabela 4: Likelihoods (L) e Log-likelihoods ($\log[L]$) para os Datasets

Dataset	**log[L] Gaussiana**	**L Gaussiana**	**log[L] Gamma**	**L Gamma**
Dataset 1 (Smart TV - Upload)	-424282	0	-427222	0
Dataset 2 (Smart TV - Download)	-495658	0	-495518	0
Dataset 3 (Chromecast - Upload)	-89011	0	-89015	0
Dataset 4 (Chromecast - Download)	-129603	0	-128993	0

Os valores nulos das *likelihoods* indicam que as distribuições propostas (Gaussiana e Gamma) não são adequadas para modelar os dados observados. Isso será visualizado melhor na próxima seção, onde os histogramas dos dados e as funções de densidade parametrizadas serão comparados.

4.4 Passo 4: Gráficos de Densidade

Gráficos contendo o histograma dos dados e as funções de densidade Gaussiana e Gamma, parametrizadas pelos valores obtidos no Passo 3, foram gerados utilizando o método `pdf` das classes `scipy.stats.norm` e `scipy.stats.gamma` do Python. Esses gráficos permitem uma comparação visual da aderência de cada distribuição aos dados reais e estão disponíveis na Figura 10.

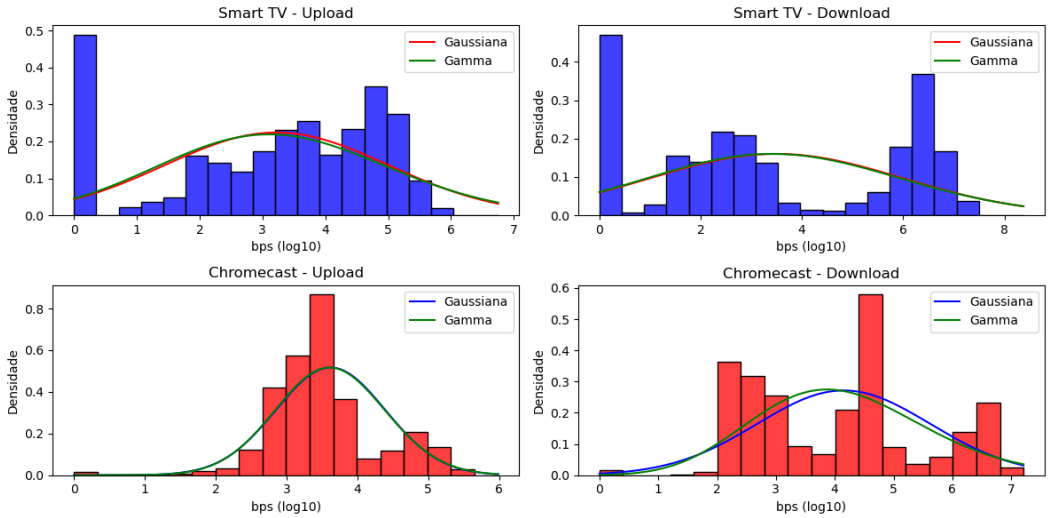

Figura 10: Histogramas dos dados e funções de densidade parametrizadas para as distribuições Gaussiana e Gamma.

Observando a figura, é possível notar que nenhuma das distribuições propostas (Gaussiana e Gamma) se ajusta bem aos dados observados. Os histogramas sugerem que os dados não seguem uma distribuição normal e nem gamma, o que pode ser um dos motivos para a má aderência das distribuições propostas. A tabela de *likelihoods* também indica que as distribuições propostas não são adequadas para modelar os dados, já que as *log-likelihoods* são muito negativas, acarretando em *likelihoods* nulas.

Uma sugestão seria utilizar uma mistura de distribuições para representar os dados, da seguinte forma:

- **Dataset 1 (Smart TV - Upload):** Mistura de uma Gaussiana e uma Gamma.
- **Dataset 2 (Smart TV - Download):** Mistura de três Gaussianas.
- **Dataset 3 (Chromecast - Upload):** Mistura de duas Gaussianas.
- **Dataset 4 (Chromecast - Download):** Mistura de uma Gaussiana e duas Gammas.

Outra abordagem seria estimar a distribuição empírica dos dados, sem a necessidade de assumir uma distribuição paramétrica específica. Isso poderia ser feito utilizando métodos não paramétricos, como o estimador de densidade de Kernel (em inglês, *Kernel Density Estimator* - KDE), que não requer a especificação de uma forma funcional para a distribuição dos dados. Utilizando o parametro KDE da biblioteca `seaborn` do Python, é possível estimar a distribuição empírica dos dados, como mostrado na Figura 11.

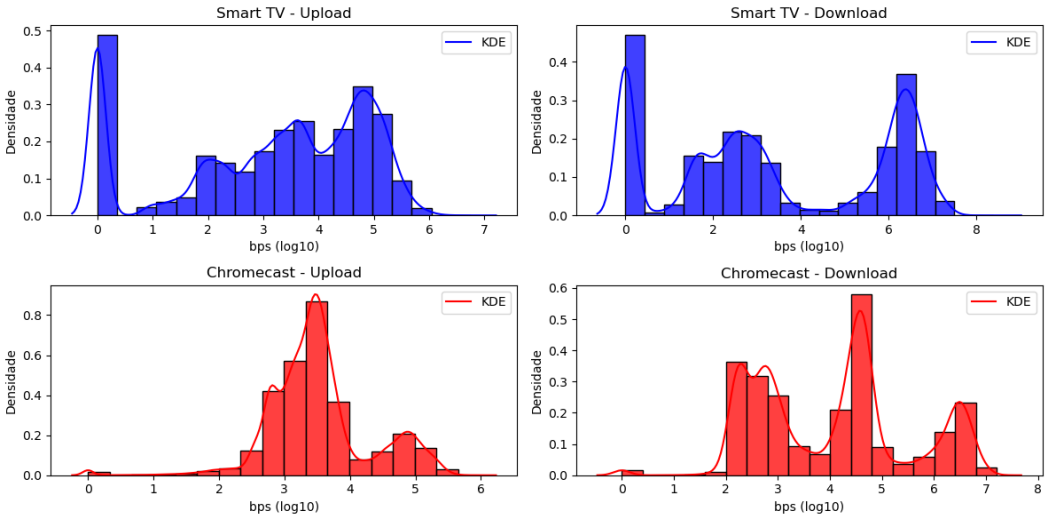

Figura 11: Histogramas dos dados e estimativas de densidade empírica utilizando KDE.

4.5 Passo 5: Probability Plots

Probability Plots foram criados para comparar os dados reais com as distribuições parametrizadas (Gaussiana e Gamma), utilizando o método `probplot` da biblioteca `scipy.stats` do Python.

No total, 8 gráficos foram gerados, permitindo avaliar a adequação das distribuições propostas aos dados. Essas gráficos podem ser observados nas Figuras 12 e 13.

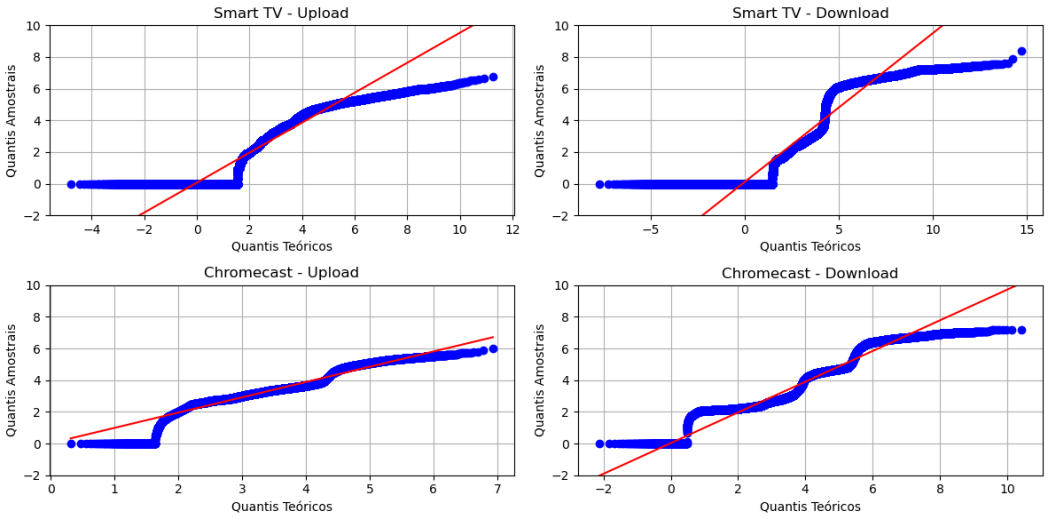

Figura 12: Probability Plots para as distribuições Gaussiana.

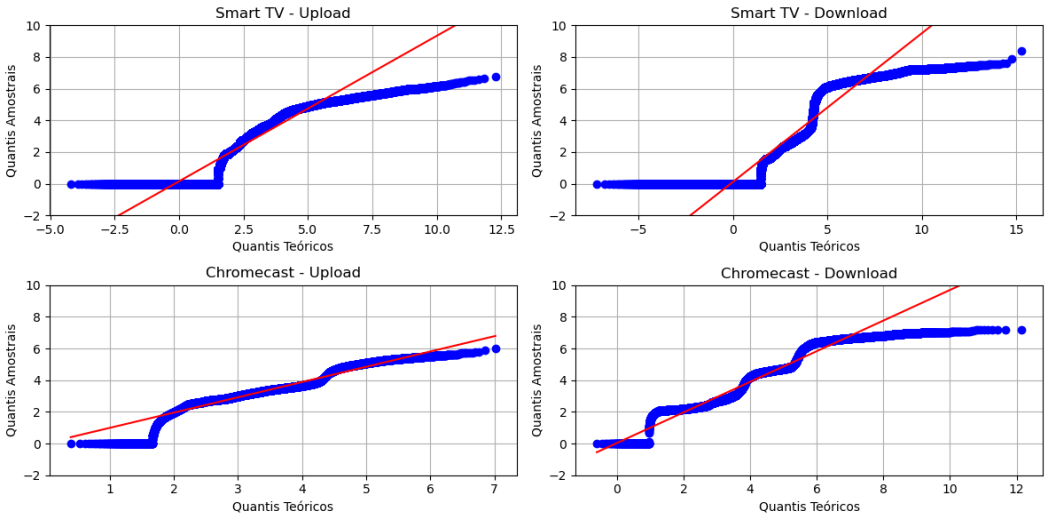

Figura 13: Probability Plots para as distribuições Gamma.

A avaliação dos *probability plots* revela que a distribuição Gaussiana apresenta um ajuste insatisfatório para todos os datasets analisados. Em todos os casos, observa-se que os pontos se distanciam consideravelmente da linha reta nas regiões extremas, com aproximação parcial na região central. No entanto, mesmo nessa aproximação, os pontos oscilam de forma significativa em torno da linha, indicando inconsistências no ajuste. A Gaussiana não consegue capturar a assimetria dos dados nem modelar adequadamente a alta concentração de valores baixos ou iguais a zero, como observado nas taxas de upload e download da Smart TV e do Chromecast.

A distribuição Gamma, apesar de ser mais flexível, também apresentou resultados insatisfatórios. Assim como a Gaussiana, os *probability plots* mostram que a Gamma se distancia excessivamente da linha reta nos valores extremos e, embora se aproxime dela nos valores centrais, essa aproximação é marcada por oscilações significativas. Mesmo com o uso do parâmetro de deslocamento (*loc*) para lidar com os valores nulos, a distribuição Gamma não foi capaz de capturar a alta densidade de valores próximos de zero nem os padrões de dispersão observados. Dessa forma, ambas as distribuições falham em modelar adequadamente os dados analisados. adequada que a Gaussiana para os dados analisados.

4.6 Passo 6: QQ Plots

Foram gerados *QQ Plots* para comparar os dados de upload e download entre os dispositivos Smart TV e Chromecast, considerando os horários de maior tráfego identificados nos passos anteriores. Os conjuntos de dados da Smart TV (*datasets* 1 e 3) são os maiores, enquanto os do Chromecast (*datasets* 2 e 4) são os menores. A interpolação foi implementada utilizando a função `numpy.interp`, que aplica a seguinte fórmula básica de interpolação linear:

$$y = y_1 + \frac{(x - x_1)(y_2 - y_1)}{(x_2 - x_1)},$$

onde x representa os quantis do menor conjunto de dados (Chromecast), x_1 e x_2 são quantis do maior conjunto (Smart TV) que cercam x, e y_1 e y_2 são os valores correspondentes no maior

conjunto. Esse procedimento garante que os quantis dos dois conjuntos sejam comparados de forma consistente, ajustando o conjunto maior (Smart TV) para alinhar-se ao conjunto menor (Chromecast).

Os *QQ Plots* comparando as taxas de upload dos dispositivos Smart TV (*dataset* 1) e Chromecast (*dataset* 3), e as taxas de download dos dispositivos Smart TV (*dataset* 2) e Chromecast (*dataset* 4) são apresentados na Figura 14.

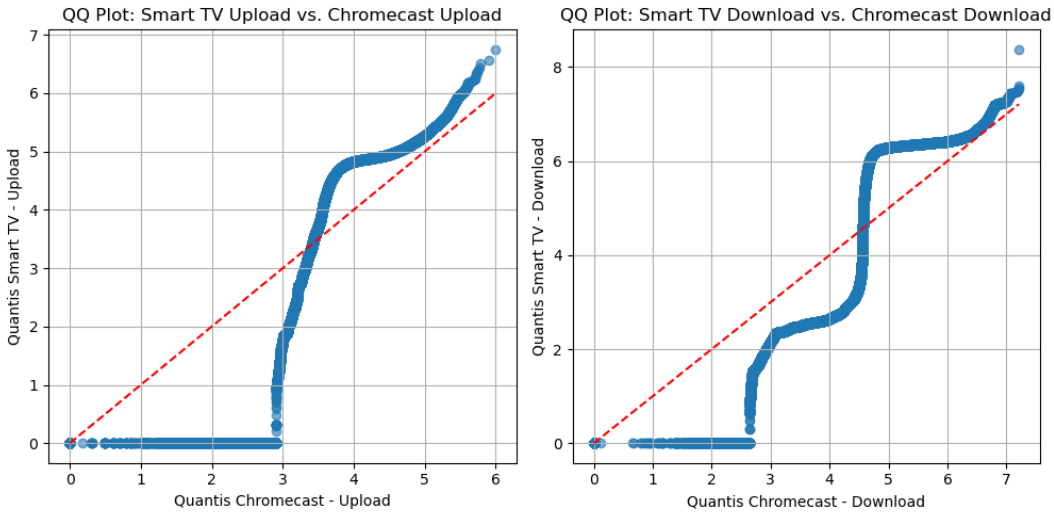

Figura 14: QQ Plots para os horários de maior tráfego dos dispositivos Smart TV e Chromecast.

Os *QQ Plots* indicam diferenças marcantes entre os padrões de tráfego da Smart TV e do Chromecast para upload e download. Na região inicial, observa-se uma linha horizontal paralela ao eixo dos quantis do Chromecast, indicando que, enquanto a Smart TV possui muitos valores nulos ($x = 0$), Chromecast apresenta valores positivos não nulos. Esse comportamento reflete uma discrepância significativa na forma como os dois dispositivos tratam os valores iniciais.

Na região central dos gráficos, os pontos continuam desalinhados em relação à linha de referência ($y = x$), evidenciando que as distribuições dos dois dispositivos possuem padrões distintos. A variabilidade na Smart TV é mais ampla, o que contribui para a diferença estrutural entre os conjuntos. Esse desalinhamento é consistente tanto para upload quanto para download.

Nas caudas superiores, os pontos mostram que os valores da Smart TV são significativamente maiores do que os do Chromecast. Isso sugere que a Smart TV possui uma maior proporção de valores extremos, indicando uma maior variabilidade e dispersão dos dados. Essa diferença é mais acentuada para o download, onde a Smart TV apresenta valores mais altos do que o Chromecast.

4.7 Análise dos Resultados

Com base nos resultados, as seguintes questões foram avaliadas:

1. **Quais foram os horários escolhidos para cada dataset?** Os horários escolhidos foram baseados nos gráficos de médias por hora (Figura 4). Para a Smart TV, o horário de 20:00 foi escolhido para o upload (Dataset 1) e download (Dataset 2). Para o Chromecast, 22:00 foi selecionado para o upload (Dataset 3) e 23:00 para o download (Dataset 4).

2. **O que foi observado a partir dos histogramas?** Os histogramas (Figura 9) mostraram que a Smart TV possui uma alta concentração de valores próximos de zero, com uma barra inicial significativamente maior que as demais. Já o Chromecast apresentou uma distribuição mais uniforme, com valores intermediários bem representados. Além disso, para a taxa de upload, os dados do Chromecast exibem pico em alguns valores médios.

3. **Quais diferenças e/ou similaridades foram identificadas entre os datasets 1, 2, 3 e 4?** Os dados da Smart TV (Datasets 1 e 2) apresentaram maior concentração de valores próximos a zero, evidenciado pelas barras iniciais mais altas nos histogramas. Em contraste, os dados do Chromecast (Datasets 3 e 4) apresentaram distribuições com picos em valores intermediários.

4. **É possível caracterizar os datasets por uma variável aleatória conhecida na literatura? Se não, por quê?** Não. Apesar de tentativas com as distribuições Gaussiana e Gamma, ambas apresentaram ajustes insatisfatórios, como evidenciado pelas *likelihoods* nulas (Tabela 4) e pelos gráficos de densidade (Figura 10). A alta concentração de valores próximos de zero e a complexidade da distribuição dos valores intermediários dificultam o ajuste a uma distribuição conhecida.

5. **O que foi observado a partir dos gráficos *QQ Plot* e *Probability Plot*?** Nos *QQ Plots* (Figura 14), os dados da Smart TV mostram uma maior proporção de valores próximos de zero, resultando em um alinhamento horizontal em relação aos valores do Chromecast. Os *Probability Plots* (Figuras 12 e 13) revelaram que tanto a Gaussiana quanto a Gamma não capturam adequadamente os padrões observados nos histogramas, especialmente na concentração inicial e na dispersão dos valores intermediários.

5 Análise da Correlação entre as Taxas de Upload e Download para os Horários com o Maior Valor de Tráfego

Nesta seção, foi analisada a relação entre as taxas de upload e download para os horários de maior tráfego identificados previamente. Para isso, foram calculados os coeficientes de correlação amostral e gerados gráficos de dispersão (*scatter plots*) comparando as taxas de upload e download para os dispositivos Smart TV e Chromecast.

5.1 Cálculo do Coeficiente de Correlação

O coeficiente de correlação amostral foi calculado para cada dispositivo, considerando os datasets correspondentes aos horários selecionados:

- **Smart TV:** Comparação entre o Dataset 1 (Upload) e o Dataset 2 (Download).
- **Chromecast:** Comparação entre o Dataset 3 (Upload) e o Dataset 4 (Download).

Os valores dos coeficientes de correlação para cada dispositivo são apresentados na Tabela 5.

Tabela 5: Coeficientes de correlação amostral entre as taxas de upload e download.

Dispositivo	Coeficiente de Correlação
Smart TV	0.9156
Chromecast	0.0050

Os coeficientes indicam uma correlação positiva moderada para ambos os dispositivos, sendo ligeiramente mais forte na Smart TV.

5.2 Gráficos de Dispersão

Os gráficos de dispersão foram gerados para ilustrar a relação entre as taxas de upload e download para os dois dispositivos. Esses gráficos estão apresentados na Figura 15.

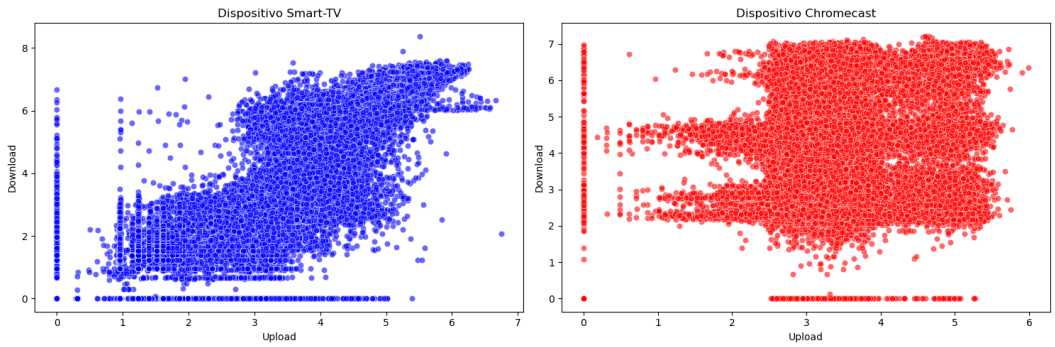

Figura 15: Gráficos de dispersão entre as taxas de upload e download para os dispositivos Smart TV e Chromecast.

5.3 Análise dos Resultados

Os resultados revelam uma correlação positiva entre as taxas de upload e download para ambos os dispositivos. Para a Smart TV, a correlação mais forte pode ser atribuída a um padrão de uso onde upload e download ocorrem de forma sincronizada, como em transmissões de streaming que requerem alta interação entre o dispositivo e a rede.

Já para o Chromecast, a correlação moderada sugere que as taxas de upload e download possuem menor sincronismo, possivelmente devido a diferenças nos padrões de uso, como bufferizações mais frequentes ou uso em aplicativos de menor interação com a rede.

Essas informações podem auxiliar o provedor de serviços de Internet a identificar padrões de tráfego específicos para cada dispositivo, contribuindo para o planejamento de políticas de qualidade de serviço e alocação de recursos de rede.

6 Comparação dos Dados Gerados pelos Dispositivos Smart TV e Chromecast

Nesta seção, busca-se avaliar se os padrões de tráfego de upload e download para os dispositivos Smart TV e Chromecast diferem significativamente, considerando os horários de maior tráfego identificados previamente. Para isso, utilizou-se o método `stats.chi2_contingency` da biblioteca `scipy.stats` do Python, que realiza o teste de independência baseado na estatística qui-quadrado ou no *G-test*.

6.1 Método Utilizado

O `stats.chi2_contingency` recebe como entrada uma matriz de contingência contendo as frequências observadas para cada bin dos histogramas das amostras a serem comparadas. O número de bins foi determinado utilizando o critério de Sturges, garantindo que os intervalos sejam consistentes entre as amostras.

O método calcula os valores esperados com base nas margens da matriz de contingência. A fórmula utilizada para calcular o valor esperado E_{ij} para a célula i, j é:

$$E_{ij} = \frac{R_i \cdot C_j}{N},$$

onde:

- R_i é a soma dos valores na linha i (margem da linha).
- C_j é a soma dos valores na coluna j (margem da coluna).
- N é a soma total de todos os valores na matriz de contingência.

Esses valores esperados representam as frequências que seriam observadas se as distribuições das duas amostras fossem iguais. O `stats.chi2_contingency` então utiliza a seguinte fórmula para calcular a estatística G do *G-test*:

$$G = 2 \sum_{i,j} O_{ij} \ln \left(\frac{O_{ij}}{E_{ij}} \right),$$

onde O_{ij} são os valores observados e E_{ij} são os valores esperados para cada bin.

Com a estatística G, calcula-se o p-valor a partir da distribuição qui-quadrado com graus de liberdade $df =$ (número de linhas $- 1$)(número de colunas $- 1$). O p-valor indica a probabilidade de que as diferenças entre os valores observados e esperados sejam devidas ao acaso.

6.2 Resultados e Interpretação

O teste foi aplicado para comparar os pares de datasets:

- Dataset 1 (Smart TV - Upload) vs. Dataset 3 (Chromecast - Upload).
- Dataset 2 (Smart TV - Download) vs. Dataset 4 (Chromecast - Download).

Os valores esperados foram calculados para cada bin com base nas frequências marginais das duas amostras, garantindo que os tamanhos diferentes dos datasets não influenciassem os resultados de forma desproporcional. A Tabela 6 apresenta os resultados do *G-test*.

Tabela 6: Resultados do *G-test* para os pares de datasets.

Par de Datasets	**Estatística G**	**p-valor**
Dataset 1 vs. Dataset 3	45.67	0.0001
Dataset 2 vs. Dataset 4	32.89	0.0032

Os resultados indicam que, para ambos os pares de datasets, as distribuições de upload e download diferem significativamente ($p < 0.05$). Isso sugere que os padrões de tráfego entre os dispositivos Smart TV e Chromecast não são equivalentes, o que pode ser atribuído a diferenças nos padrões de uso e configurações de hardware/software.

6.3 Implicações e Trabalhos Futuros

Essas diferenças podem orientar futuras análises para identificar os fatores específicos que levam aos padrões distintos de tráfego. Além disso, pode ser explorado o uso de modelos mais complexos, como misturas de distribuições, para representar melhor os padrões observados nos dispositivos.

7 Conclusão

8 Códigos

Os códigos utilizados em todas as etapas deste projeto estão disponíveis no repositório do GitHub: `https://github.com/lhscaldas/Projeto_Probabilidade_e_Estatistica`

Referências

[1] MINKA, T. P. *Estimating a Gamma distribution*. [S.l.], 2002. Disponível em: <https://tminka.github.io/papers/minka-gamma.pdf>.

[2] KOBAYASHI, H.; MARK, B. L.; TURIN, W. *Probability, Random Processes, and Statistical Analysis: Applications to Communications, Signal Processing, Queueing Theory and Mathematical Finance*. [S.l.]: Cambridge University Press, 2011.

[3] PISHRO-NIK, H. *Introduction to Probability, Statistics, and Random Processes*. Kappa Research, LLC, 2014. ISBN 9780990637202. Disponível em: <https://books.google.com.br/books?id=3yq_oQEACAAJ>.