Universidade Federal do Rio de Janeiro
Engenharia de Computação e Informação
Escola Politécnica
COE241/COS868 - Probabilidade e Estatística

Relatório do Projeto do Curso de Probabilidade e Estatística - Análise de datasets

Nome	Guilherme En Shih Hu
DRE	123224674
Professora	Rosa Maria Meri Leão
Horário	Ter/Qui - 15:00-17:00

Rio de Janeiro, 3 de janeiro de 2025

Conteúdo

1 Introdução **1**

2 Implementação em código **1**

3 Análises das estatísticas gerais **2**
3.1 Histogramas das estatísticas gerais 2
3.2 Função de distribuição empírica 5
3.3 Boxplots 7
3.4 Média, variância e desvio padrão 8
3.5 Interpretação geral dos dados 9

4 Análises das estatísticas por horário **10**
4.1 Boxplots 10
4.2 Média, variância e desvio padrão 12
4.3 Interpretação geral dos dados 15

5 Análises dos horários com maior valor de tráfego **16**
5.1 Passo 1 - Determinação dos datasets 16
5.2 Passo 2 - Histograma para os horários de pico 16
5.3 Passo 3 - QQ Plots 20

6 Análises da correlação entre as taxas de upload e download para os horários com o maior valor de tráfego **21**

7 Conclusão **23**

1 Introdução

Este relatório detalha a implementação do código desenvolvido para gerar os gráficos e calcular as estatísticas definidas no projeto final da disciplina de Probabilidade e Estatística - COE241/COS86. Além disso, apresenta uma análise dos resultados obtidos, destacando as conclusões derivadas dos valores calculados e dos gráficos gerados. O trabalho visa consolidar habilidades práticas, como a criação de códigos voltados à análise e visualização de dados, bem como o desenvolvimento da capacidade de interpretar criticamente os resultados e identificar padrões ou tendências relevantes.

Segue abaixo o link para o repositório com o código que gera os gráficos e as estatísticas a serem analisadas no relatório: https://github.com/guilherme-hu/Probest/blob/master/Trabalho.ipynb

Nas próximas seções, será inicialmente descrito o processo de implementação do código, abordando as principais etapas e escolhas realizadas. Em seguida, cada estudo do trabalho será analisado em seções específicas, nas quais serão apresentados os resultados obtidos, acompanhados de discussões sobre as conclusões possíveis a partir dos gráficos e valores gerados.

2 Implementação em código

A implementação do projeto foi realizada em Jupyter Notebook, uma ferramenta que permite a exibição dos gráficos e valores diretamente após as seções de código, facilitando a análise e a interpretação dos resultados. O objetivo principal do projeto é analisar as taxas de upload e download de dispositivos Chromecast e Smart TV, utilizando técnicas de estatística e visualização de dados para identificar padrões e compreender a relação entre essas métricas.

O código foi organizado de acordo com as etapas propostas no trabalho, garantindo uma estrutura e coerente. Inicialmente, são importadas as bibliotecas necessárias, como NumPy, Pandas, Matplotlib, Seaborn, Statsmodels e Scipy, que oferecem suporte às análises e visualizações. Em seguida, os datasets dos dispositivos Chromecast e Smart TV são carregados a partir de arquivos CSV e transformados para a escala logarítmica de base 10, normalizando as distribuições das taxas de upload e download.

Na etapa de estatísticas gerais, são apresentadas visualizações como histogramas, funções de distribuição empírica (ECDF) e box plots, além do cálculo de estatísticas descritivas, como médias, variâncias e desvios padrão, desconsiderando-se os horários em que foram gerados. Essas análises fornecem uma visão inicial das distribuições dos dados e das principais características das taxas de upload e download.

Posteriormente, na etapa de estatísticas por horário, os dados são agrupados por hora para calcular estatísticas horárias, como médias, variâncias e desvios padrão, organizadas de forma a identificar padrões temporais. Box plots e gráficos de barras são gerados para visualizar a variação das taxas de upload e download ao longo do dia.

Na seção que aborda os horários de maior tráfego, são identificados os picos de upload e download para ambos os dispositivos. Histogramas são utilizados para visualizar as distribuições das taxas nesses horários, e QQ plots são criados para comparar as taxas entre os dispositivos durante os períodos de pico.

Por fim, são calculados os coeficientes de correlação de Pearson entre as taxas de upload e download para os dois dispositivos, complementados por scatter plots que ilustram a relação entre essas métricas. Para garantir comparabilidade, os conjuntos de dados são ajustados via interpolação linear, igualando o número de pontos. Com esses dados, é possível indicar se existe alguma correlação entre as taxas de download e upload dos dispositivos.

3 Análises das estatísticas gerais

Nos estudos das estatísticas gerais, os dados dos datasets de cada dispositivo (Smart TV e Chromecast) são avaliados sem levar em conta o horário em que foram gerados. Conforme proposto pelo projeto, foram calculados e analisados os seguintes elementos para as taxas de upload e download: histograma, função de distribuição empírica (ECDF), box plot, média, variância e desvio padrão.

Nas subseções que seguem, serão comentados os resultados obtidos para cada gráfico e estatística calculada, destacando as principais observações e interpretações dos dados analisados.

3.1 Histogramas das estatísticas gerais

Abaixo, nas figuras (1), (2), (3) e (4), estão os histogramas com as ocorrências por logaritmo de taxa de upload ou download para cada um dos dispositivos analisados pelo projeto, Smart TV e Chromecast.

O histograma da figura (1) mostra os dados referentes às ocorrências de determinadas faixas de valores de taxa de upload, em log de bps, no dispositivo da Smart TV. Observa-se uma alta concentração de ocorrências próximas a zero, com mais de 1,75 milhões de registros, indicando que a maioria dos uploads tem taxas muito baixas ou próximas de zero. À medida que a taxa de upload aumenta, a frequência de ocorrências diminui significativamente, exibindo uma distribuição mais dispersa entre os valores logarítmicos de 1 a 6. Este padrão sugere que, apesar da predominância de taxas de upload baixas, há uma variação considerável entre os registros com taxas de upload mais altas.

Por outro lado, o histograma da figura (2) exibe os dados relacionados às taxas de download na Smart TV. Embora perceba-se um padrão semelhante às taxas de upload no quesito de haver uma concentração extremamente alta de ocorrências no valor logarítmico zero, revelando o significativo número de dados coletados com taxas muito baixas ou perto de zero, é possível verificar que seu valor é ainda maior que o número visto no histograma do upload para a Smart TV. Além disso, conforme a taxa de download aumenta, as ocorrências se tornam mais dispersas e menos frequentes, o que sugere que, apesar de uma predominância de taxas de download baixas, a Smart TV experimenta uma variedade considerável de taxas ao longo do tempo. Em comparação às taxas de upload, observa-se que não há muitas ocorrências entre os valores logarítimicos de 4 e 6, no entanto, as taxas alcançam valores de logaritmo maiores que 6, indicando que as taxas de download são, em geral, mais altas do que as de upload.

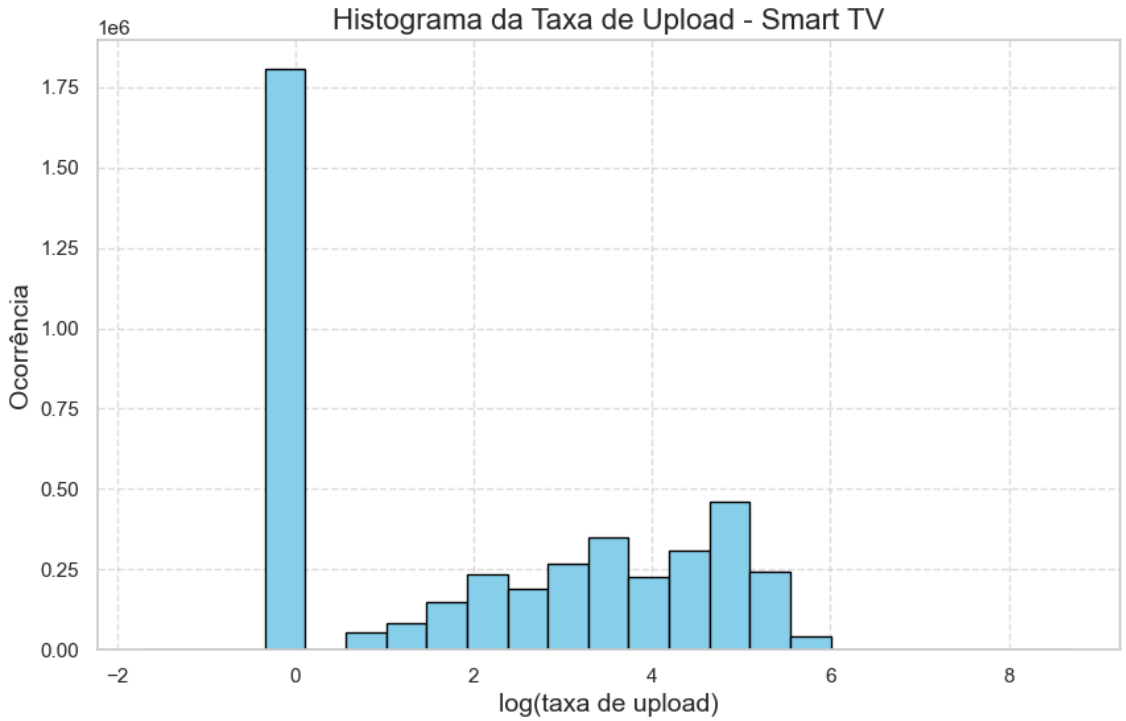


Figura 1: Histograma da Taxa de Upload para Smart TV

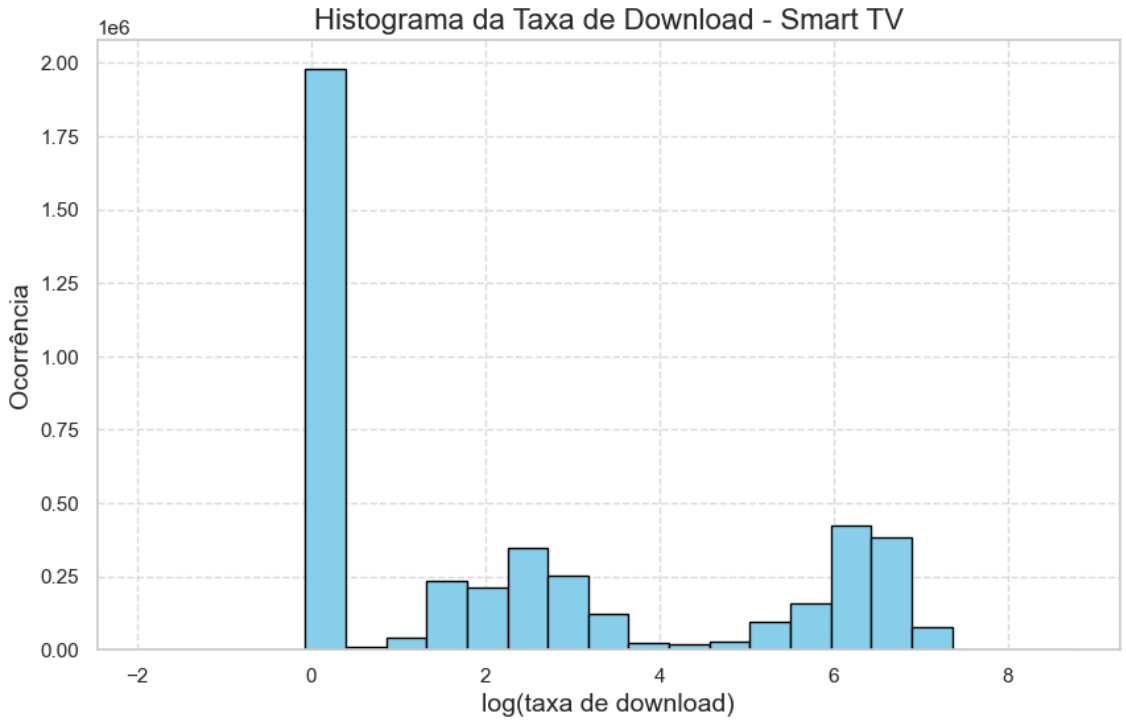


Figura 2: Histograma da Taxa de Download para Smart TV

O histograma da figura (3) ilustra as faixas de valores de ocorrência das taxas de upload para dispositivos Chromecast. Observa-se que a maioria dos dados se concentra entre os valores logarítmicos de 2 a 4, com um pico em torno de 3, indicando que a maioria dos dispositivos Chromecast possui taxas de upload nessa faixa logarítmica. Há algumas ocorrências em valores abaixo da faixa dos 2.5 e mais altos que 4, porém são pouco frequentes.

Já o histograma da figura (4) apresenta os dados das taxas de download para o Chromecast. É possível notar que a maioria dos downloads ocorre em taxas compreendidas entre os valores logarítmicos de 2 a 5, com um pico acentuado próximo a 5. Diferente do histograma referente às taxas de upload, que têm seus dados centralizados entre as faixas de 2 e 4, o gráfico para a taxa de download possui duas regiões com grande presença de dados: entre o 2 e o 3, e entre o 4 e o 5, com os espaços entre as faixas 3 e 4 e as faixas 5 a 7 possuindo menos valores. Isso indica que as taxas de upload para o Chromecast são mais centralizadas e homogêneas, sugerindo uma consistência nas velocidades de upload, enquanto que a dispersão observada nas taxas de download mostram uma maior variabilidade, podendo atingir uma ampla gama de velocidades, inclusive alumas bem maiores do que as taxas de upload poderiam alcançar (entre as faixas de 6 e 7 que apresentam uma boa quantidade de ocorrências).

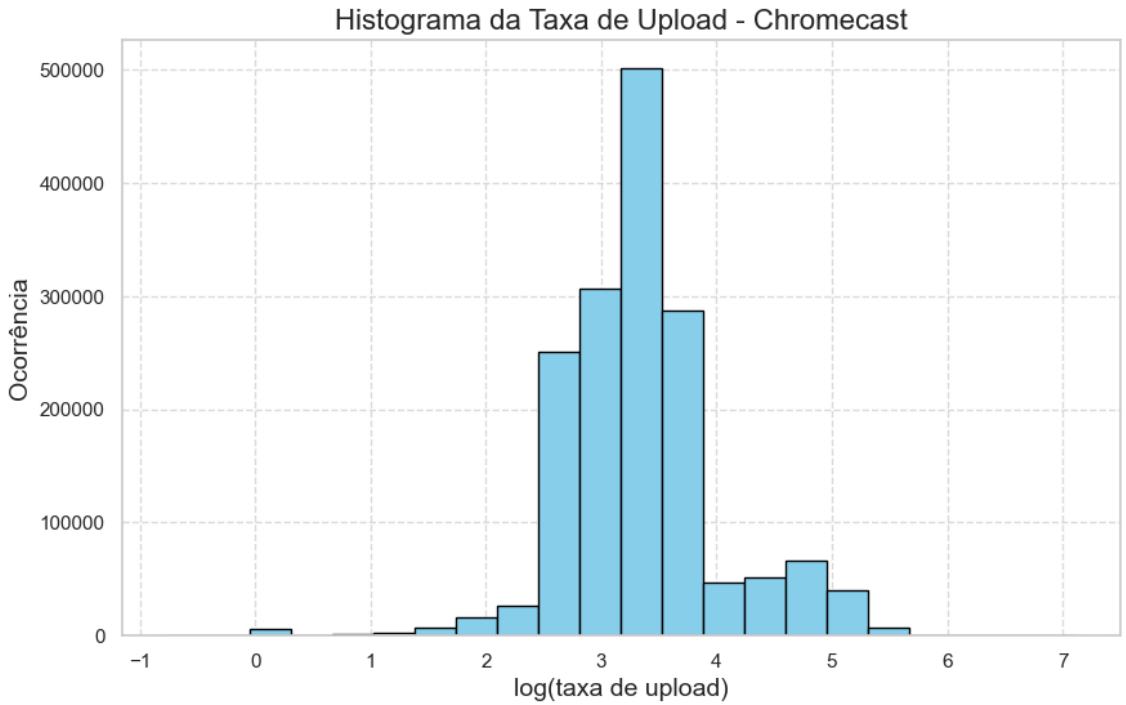


Figura 3: Histograma da Taxa de Upload para Chromecast

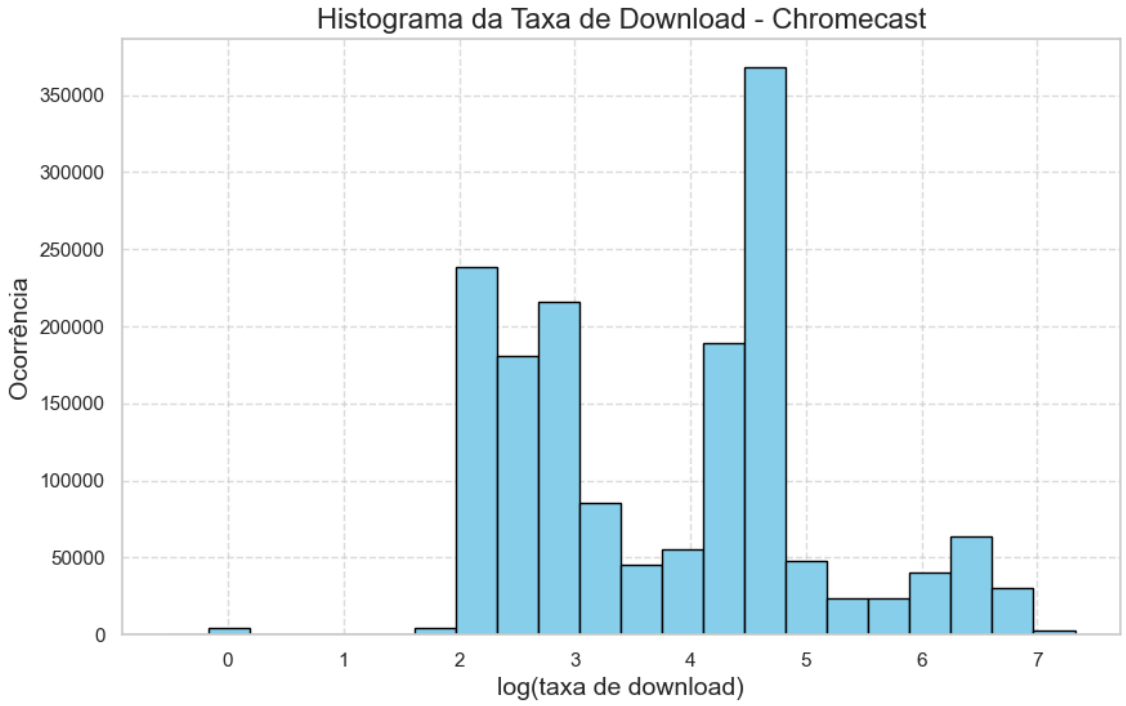


Figura 4: Histograma da Taxa de Download para Chromecast

3.2 Função de distribuição empírica

Abaixo, nas figuras (5), (6), (7) e (8), estão os gráficos ECDFs com as probabilidades de se encontrar um valor dentro do dataset com valor igual ou abaixo do valor no eixo X de log de taxa de upload ou download, para cada um dos dispositivos analisados pelo projeto: Smart TV e Chromecast.

Analisando os gráficos das figuras (5) e (6), que apresentam as funções de distribuição empírica da Smart TV, observa-se um fato já constatado na seção 3.1: há uma grande quantidade de ocorrências de valores próximos ou iguais a zero, com uma probabilidade do dado possuir esse valor sendo em torno de 0.4, e com a taxa de download sendo um pouco maior que a de upload. Ademais, também é possível prever um pouco do comportamento da frequência de cada log de taxa de upload ou download: para a taxa de upload, como a curva de sua ECDF sobe com uma inclinação quase constante até a faixa do valor logarítimico de 6, quase 60% dos valores possuem valores de taxa distante de zero e uma distribuição mais dispersa, sem nenhuma dessas faixa possuir poucos valores, como se observa no histograma (1); por outro lado, para o download, verifica-se 2 partes do gráfico com inclinação mais acentuada (por volta dos valores logarítimicos de 2 e de 6), o que expressa a maior ocorrência de dados com esses valores nesses valores, e 3 partições com pouca inclinação (entre zero e 1.5, perto de 4 e 5 e maior que 7, o que indica poucos valores de ocorrência nessas regiões, como também se verifica no histograma (2).

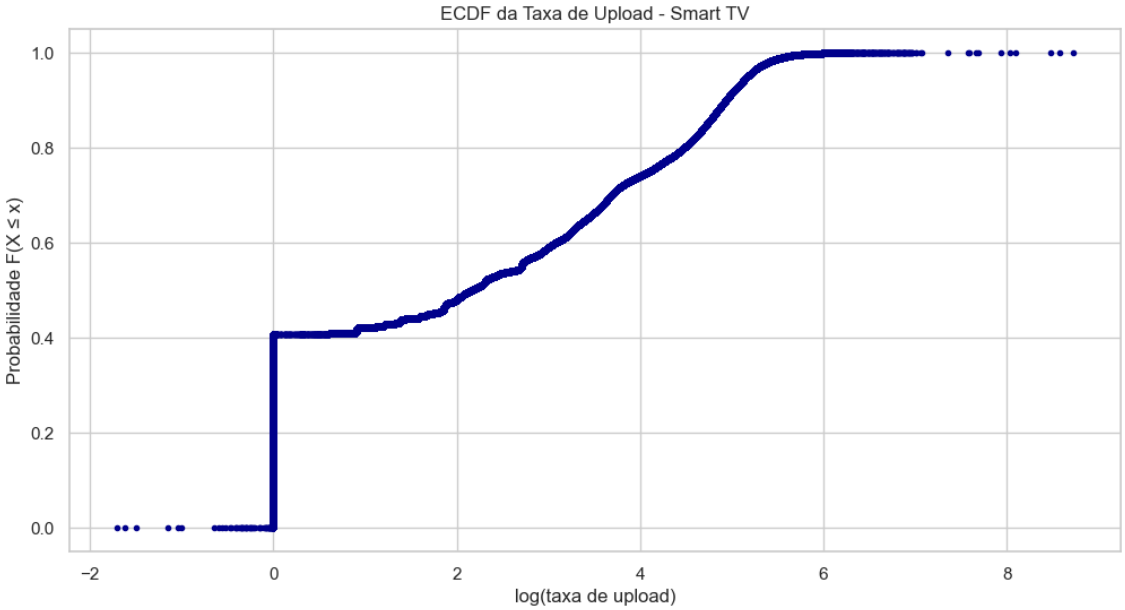


Figura 5: ECDF da Taxa de Upload para Smart TV

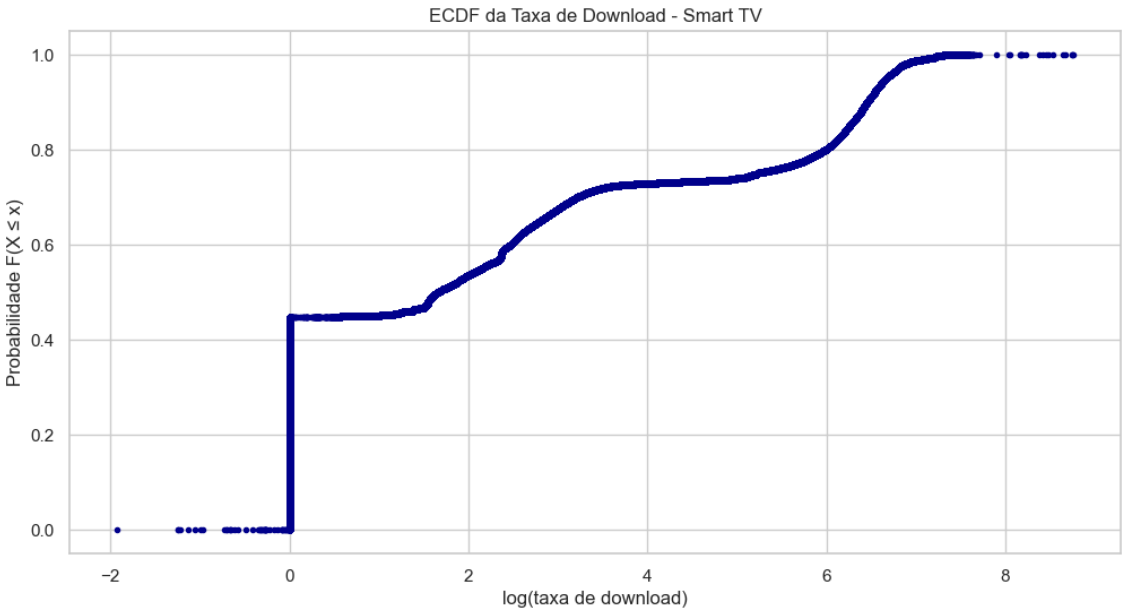


Figura 6: ECDF da Taxa de Download para Smart TV

Da mesma forma, os gráficos das figuras (7) e (8), também exibem fatos já evidenciados pelos histogramas (3) e (4). Para a taxa de upload do Chromecast, percebe-se que há uma curva bastante acentuada dentro da faixa dos valores logarítmicos 2 e 4, o que revela o caráter centralizado dos dados dentro desse intervalo. A taxa de download, por sua vez, apresenta curvas mais inclinadas dentro da faixa do 2 e 3, e entre o 4 e 5, aspectos notados pela existência de duas regiões com muitos dados no gráfico (8).

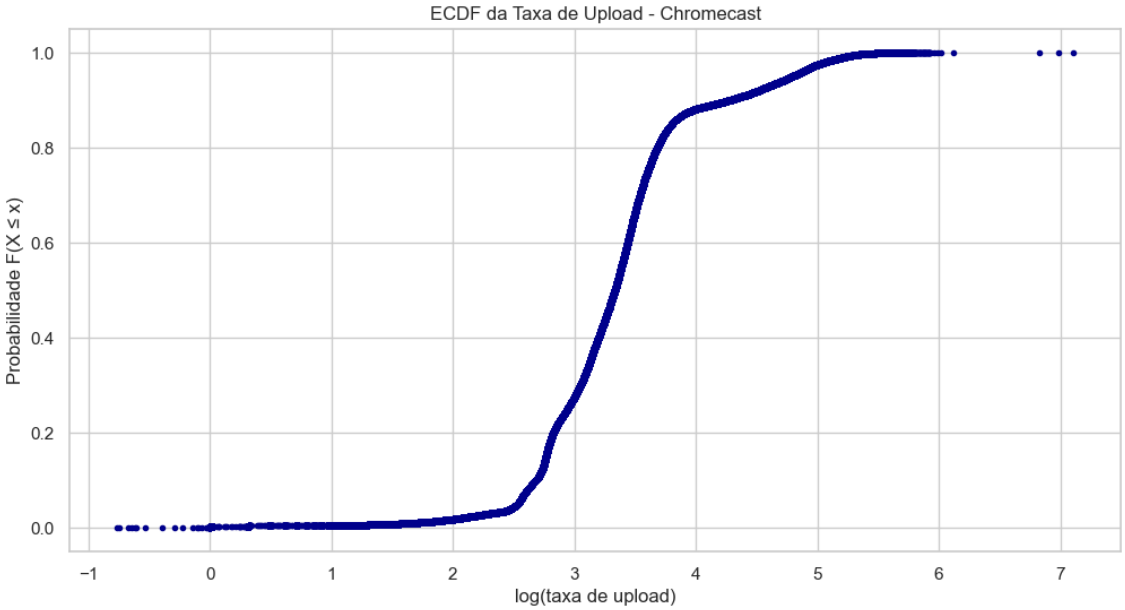


Figura 7: ECDF da Taxa de Upload para Chromecast

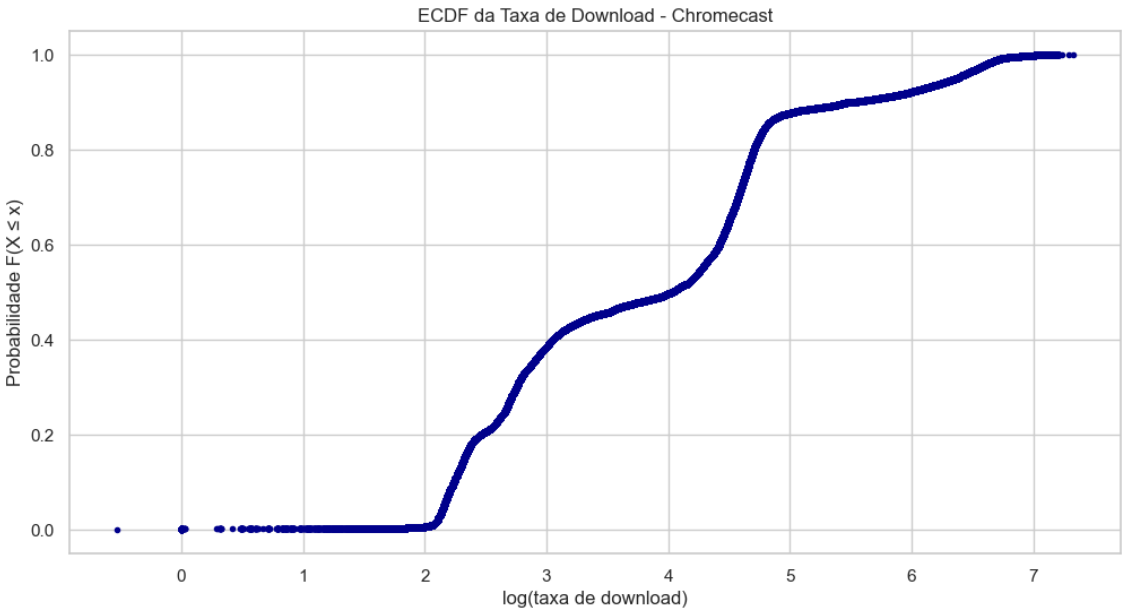


Figura 8: ECDF da Taxa de Download para Chromecast

3.3 Boxplots

Com relação aos box plots, foi gerado um único gráfico contendo os quatro box plots (taxa de upload e taxa de download para Smart TV e para Chromecast), permitindo uma comparação direta entre as métricas dos dois dispositivos. Esse gráfico é mostrado na figura (??).

Ao analisar o boxplot, observa-se que, para a Smart TV, tanto as taxas de

upload quanto as de download possuem uma distribuição mais ampla, com maior variabilidade em comparação ao Chromecast. Em particular, as taxas de download da Smart TV apresentam valores máximos consideravelmente elevados, enquanto as taxas de upload também mostram uma dispersão significativa, mas com valores inferiores. Para o Chromecast, as taxas de upload possuem uma dispersão pequena, concentrando-se em valores próximos à mediana, enquanto as taxas de download apresentam maior variabilidade, mas ainda inferiores às da Smart TV. Além disso, há a presença de outliers nas taxas de upload e download do Chromecast, indicando algumas observações isoladas. Esses padrões sugerem que a Smart TV opera em uma faixa mais ampla de taxas, enquanto o Chromecast tende a ter um desempenho mais estável, porém limitado.

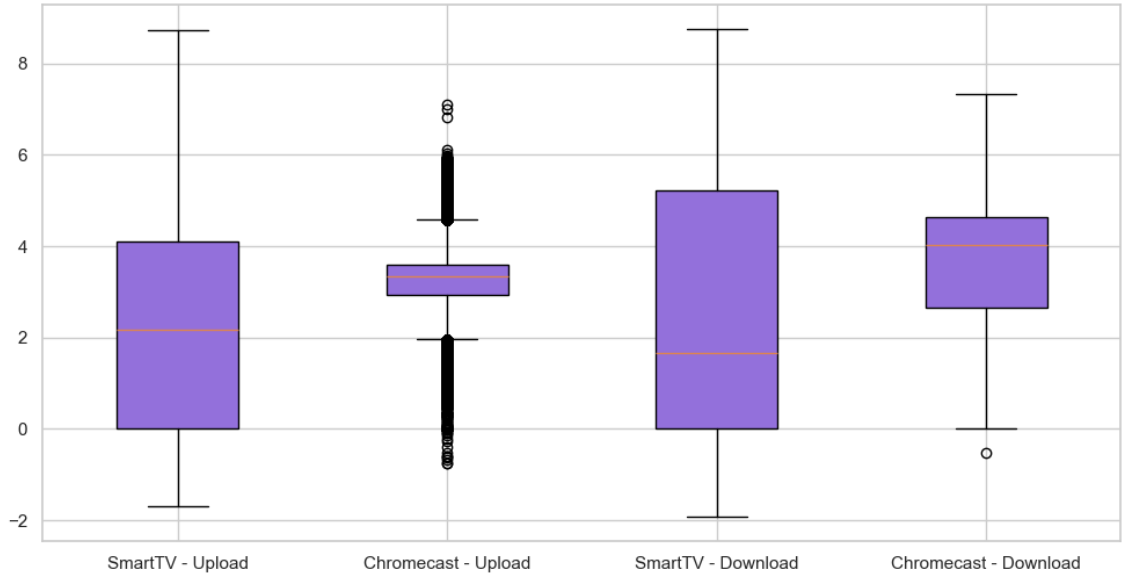


Figura 9: Boxplots

3.4 Média, variância e desvio padrão

Abaixo, na tabela 1, estão os valores calculados de cada estatistica descritiva definida pelo trabalho - média, variância e desvio padrão - para as taxas de upload e download de dos dispositivos de Smart TV e Chromecast. Vale notar que esses valores estão em log na base 10.

	Smart TV		Chromecast	
	Upload	Download	Upload	Download
Média	2.156590	2.350173	3.349672	3.799335
Variância	4.113083	6.723921	0.461600	1.665980
Desvio padrão	2.028074	2.593052	0.679412	1.290728

Tabela 1: Tabela com os valores de média, variância e desvio padrão, em log de 10

A análise dos gráficos feitos anterioremente em conjunto com os valores de média, variância e desvio padrão evidencia importantes diferenças entre os dispositivos

Smart TV e Chromecast, bem como entre as taxas de upload e download. No caso da Smart TV, a maior variabilidade observada no boxplot é corroborada pelos altos valores de variância e desvio padrão, especialmente para a taxa de download (variância de 6,72 e desvio padrão de 2,59). Isso reflete uma ampla dispersão das taxas, incluindo valores máximos significativamente elevados no gráfico, aspecto já observado na seção 3.1. Já a taxa de upload da Smart TV apresenta menor variabilidade em relação ao download, mas ainda mantém valores de variância e desvio padrão (4,11 e 2,03, respectivamente) bem superiores aos do Chromecast.

No Chromecast, os valores de variância e desvio padrão confirmam a menor dispersão observada nos boxplots, com destaque para a taxa de upload, que apresenta uma variância de apenas 0,46 e um desvio padrão de 0,67, indicando alta concentração em torno da média (fato também observado visualmente no histograma 3) . As taxas de download do Chromecast, embora mais dispersas que as de upload (variância de 1,66 e desvio padrão de 1,29), ainda apresentam menor variabilidade do que as taxas observadas na Smart TV. A comparação geral entre os dispositivos revela que a Smart TV opera em uma faixa mais ampla de valores, especialmente para download, enquanto o Chromecast apresenta maior estabilidade e valores médios mais elevados para ambas as taxas.

3.5 Interpretação geral dos dados

A análise conjunta dos boxplots, histogramas, ECDFs e estatísticas descritivas revela diferenças marcantes entre os dispositivos Smart TV e Chromecast, bem como entre as taxas de upload e download. Para a Smart TV, os boxplots mostram uma maior variabilidade nas taxas de download, corroborada pelos altos valores de variância (6,72) e desvio padrão (2,59). O histograma reforça esse padrão ao exibir uma dispersão significativa, com algumas ocorrências em taxas logarítmicas muito altas, enquanto a ECDF confirma que a maior parte dos valores de download se concentra em taxas mais baixas, mas com uma cauda mais longa indicando valores elevados (chegando até o valor logarítmico de 8). As taxas de upload da Smart TV, por sua vez, também apresentam variabilidade considerável (variância de 4,11 e desvio padrão de 2,03), conforme observado nos boxplots e no histograma, mas são mais concentradas em torno da mediana em comparação ao download.

Já para o Chromecast, os boxplots e as estatísticas descritivas destacam uma menor variabilidade geral em relação à Smart TV, especialmente para as taxas de upload, que possuem uma variância de apenas 0,46 e desvio padrão de 0,67. Os histogramas e ECDFs confirmam essa alta concentração das taxas de upload próximas à mediana, com uma distribuição bastante compacta. As taxas de download do Chromecast, embora mais dispersas que as de upload (variância de 1,66 e desvio padrão de 1,29), ainda apresentam uma menor variabilidade em relação à Smart TV, o que é evidente nas distribuições mais concentradas observadas nos histogramas e nas curvas ECDF mais íngremes. Em suma, enquanto o Chromecast se caracteriza por taxas mais consistentes e concentradas, a Smart TV opera com maior variabilidade, especialmente em taxas de download.

4 Análises das estatísticas por horário

Nas análises dessa seção, os dados são avaliados considerando o horário em que foram gerados, independente do dia. Conforme proposto pelo projeto, foram calculados e analisados os seguintes elementos para as taxas de upload e download para cada um dos dispositivos: box plot, média, variância e desvio padrão.

4.1 Boxplots

Abaixo, nas figuras (10), (11), (12) e (13), estão os gráficos de boxplot para cada horário, para cada taxa de upload ou download e para cada um dos dispositivos analisados pelo projeto: Smart TV e Chromecast.

O gráfico da figura (10) apresenta a distribuição das taxas de upload da Smart TV ao longo das 24 horas do dia, utilizando boxplots para cada hora. Observa-se uma maior variabilidade das taxas de upload nas primeiras horas do dia (0h a 1h), com valores mais altos nos quartis superiores e presença de alguns outliers. Entre as 2h e 7h, as taxas de upload mostram uma redução significativa na mediana e menor dispersão, indicando uma atividade mais uniforme, concentrada em taxas mais baixas nesse período. Nesse intervalo também se verifica uma grande quantidade de outliers, que refletem casos isolados de altas taxas de upload nesse momento em que as taxas costumam estar mais baixas. A partir das 8h, a mediana das taxas de upload começa a aumentar, juntamente com a amplitude interquartil, e mantém-se relativamente estável durante a maior parte do dia (8h às 21h), sugerindo maior atividade de upload da Smart TV nesse intervalo. Em geral, o gráfico revela uma tendência de maior variabilidade nas taxas de upload nos extremos do dia (madrugada da noite e manhã), enquanto períodos intermediários apresentam comportamento mais uniforme, com mediana e dispersão mais altas em relação às horas da madrugada.

Por outro lado, o gráfico da imagem (11) trata das taxas de download da Smart TV para cada hora do dia. De forma semelhante a como ocorria no gráfico (10), que também é sobre a Smart TV, nota-se o padrão de baixas taxas ocorrendo nos períodos da madrugada e um pouco da manhã (entre 1h às 8hs), com bastante presença de outliers, e uma crescente atividade a partir das 10hs, com a mediana e a amplitude interquartil aumentando até às 21hs. Esse resultado indica um baixo uso da Smart TV durante as horas da madrugada e um pouco da manhã (entre às 1hs e 8hs), e sugere um maior uso durante o dia (metade da manhã até a noite), com seu comportamento mais uniforme com medianas e dispersão maiores se comparados com as horas da madrugada. Outra fato que pode ser constado a partir da comparação dos gráficos das figuras (10) e (11) é que as taxas de download costumam ser maiores que as de upload, como já analisado na seção 3, visto que as amplitudes interquatil alcançam valores mais altos no eixo Y.

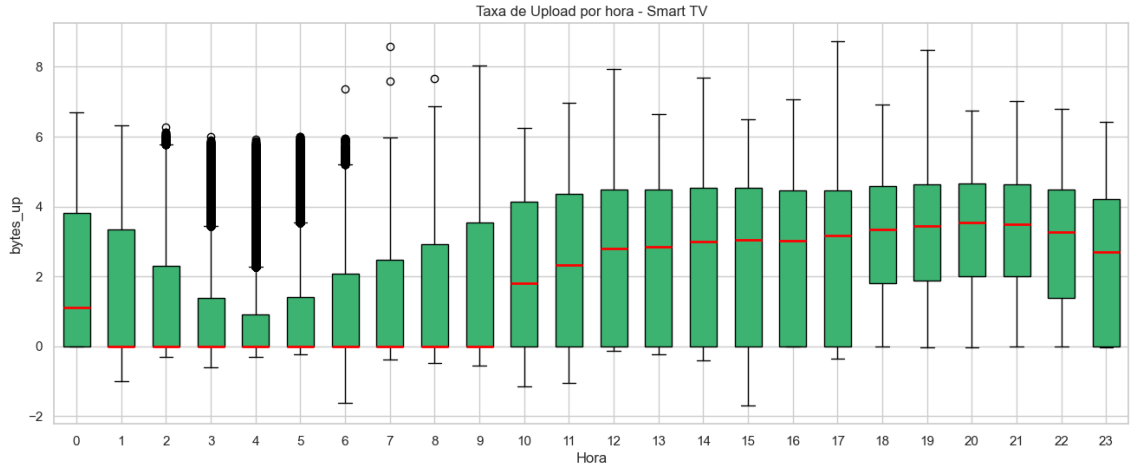


Figura 10: Taxa de Upload para Smart TV por hora

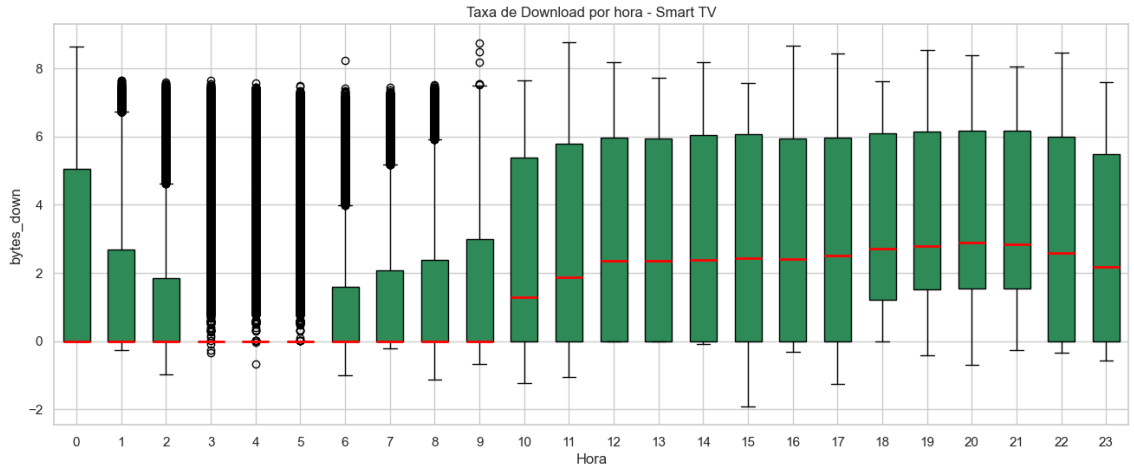


Figura 11: Taxa de Download para Smart TV por hora

Analisando em particular o gráfico da figura (12), agora tratando sobre as taxas de upload por horário para o Chromecast, a mediana da taxa de upload se mantém relativamente constante ao longo das 24 horas, em torno de 3 bytes_up. A dispersão dos dados (representada pelo IQR) também é bastante consistente ao longo do dia. Além disso, existem vários outliers em todas as horas, indicando que há momentos em que a taxa de upload é significativamente diferente da maioria dos dados.

Comparando com o gráfico das taxas de download do Chromecast, da figura (13), nota-se que a mediana das taxas de download é ligeiramente maior, em torno de 4 bytes_down, e também se mantém constante ao longo do dia, com um IQR estável e poucas variações significativas. Os outliers são menos frequentes no gráfico de download, mas ainda presentes, especificamente na hora 23. Essa análise destaca um comportamento relativamente estável para ambas as métricas (upload e download) do Chromecast ao longo do dia, com medianas constantes e dispersão consistente, mas com uma leve variação nos outliers e nas medianas entre as taxas de upload e download.

Com isso, observa-se uma grande diferença em relação aos gráficos da Smart TV: enquanto nos gráficos da Smart TV é possível notar uma explícita variabilidade nas taxas, havendo pouco uso durante a madrugada e mais uso durante a tarde e noite, os boxplots do Chromecast possuem dados com medianas e dispersões bastante consistentes e quase constantes durante o dia todo.

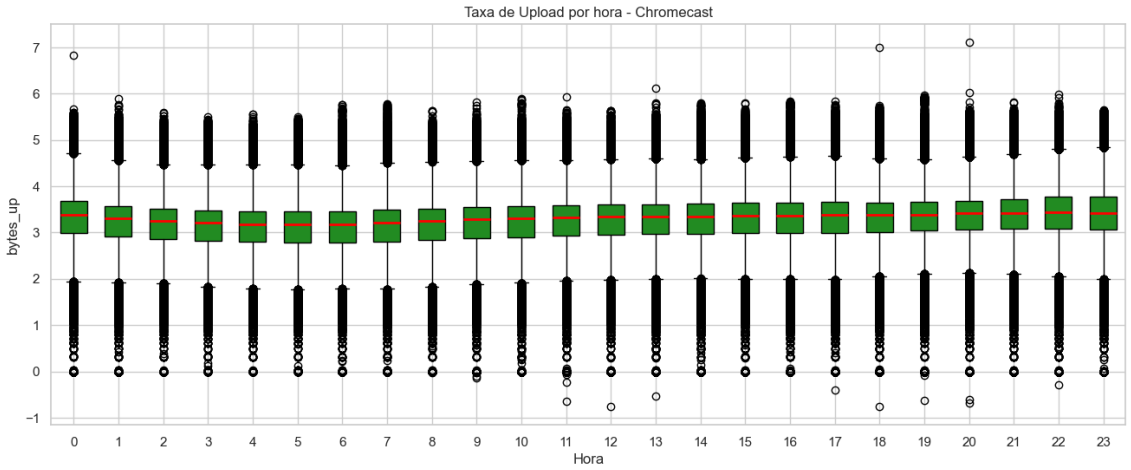


Figura 12: Taxa de Upload para Chromecast por hora

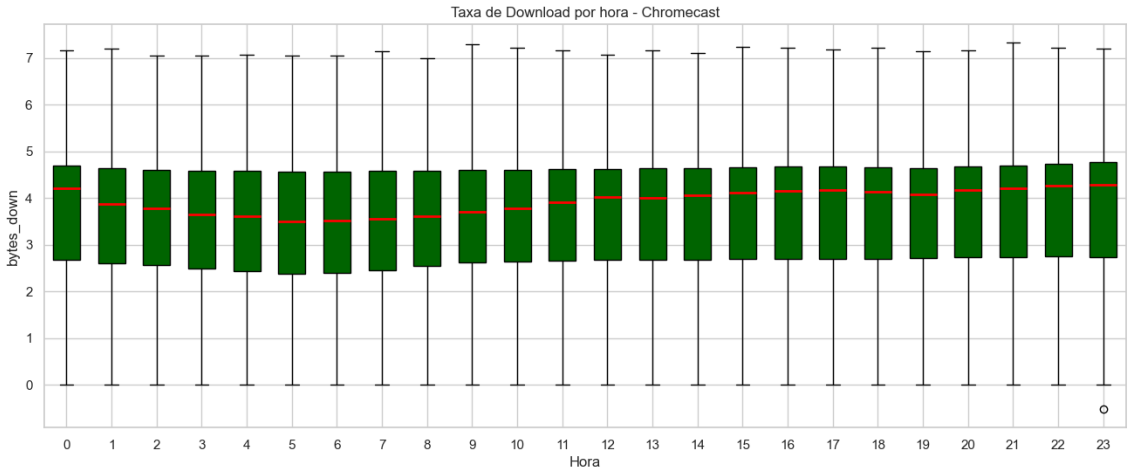


Figura 13: Taxa de Download para Chromecast por hora

4.2 Média, variância e desvio padrão

Para as estatísticas descritivas de média, variância e desvio padrão, foram feitos 4 gráficos representando no eixo X a hora e no eixo Y os valores das três estatísticas para cada taxa coletada, para cada tipo de dispositivo. Esses gráficos estão presentes nas figuras (14), (15), (16) e (17).

Analisando os gráficos relacionados à Smart TV, ou seja, as estatísticas descritivas das taxas de upload e download nas figuras (14) e (15) respectivamente, é possível notar como o formato das linhas de cada estatística reforçam as informações concluídas na subseção 4.1 com os boxplots.

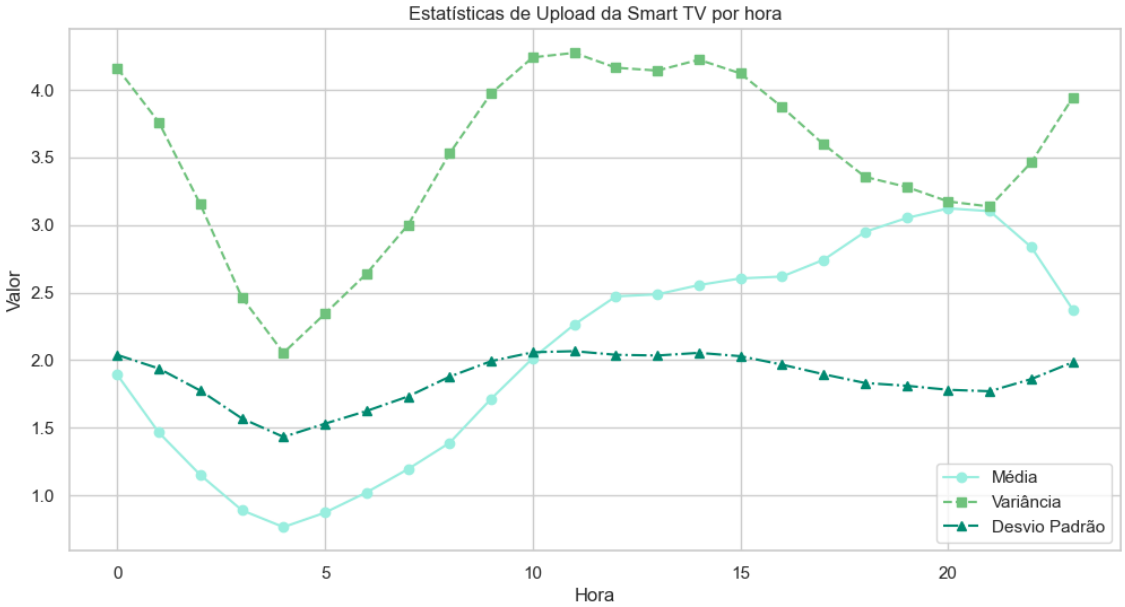


Figura 14: Estatísticas descritivas da Taxa de Upload da Smart TV por hora

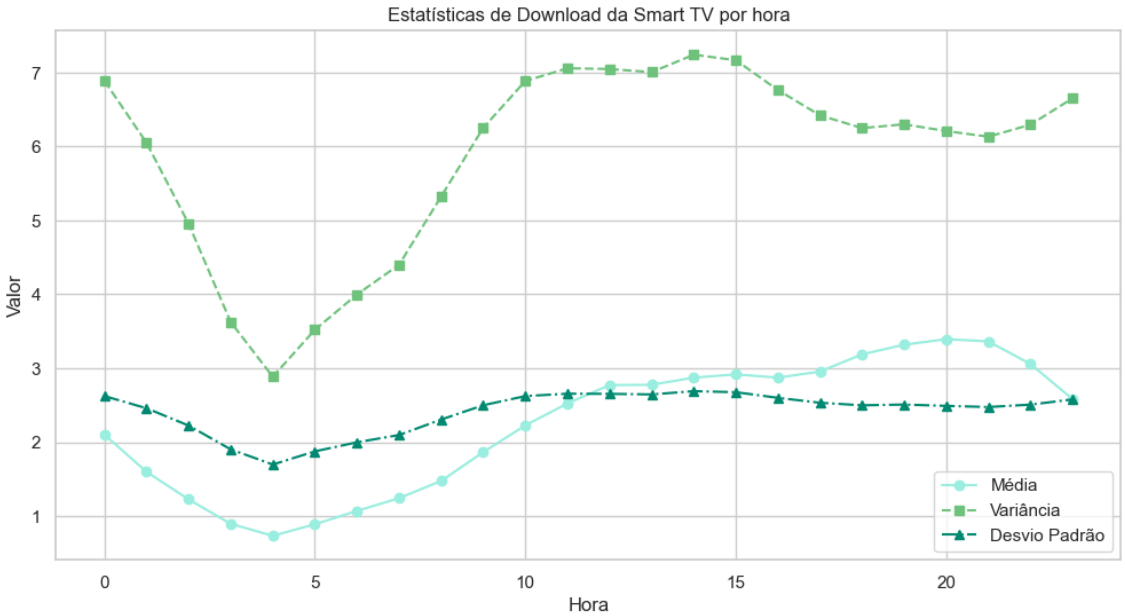


Figura 15: Estatísticas descritivas da Taxa de Download da Smart TV por hora

Primeiramente, observa-se que dentro da faixa das 0hs às 10hs, existe um "poço"no valor da média em ambos os gráficos, o que demonstra o baixo uso da Smart TV durante esses horários. Esse valor de média começa a aumentar por volta das 8hs, e se mantém estável entre 12hs e 17hs, aumentando mais entre 17hs e 20hs, demonstrando como esses horários são os que apresentam maior uso por parte do usuário. Após esses horários, na parte da noite e madrugada do dia, os valores caem novamente. Dessa forma, é possível correlacionar o valor da média calculada nesses estatísticas com o valor da mediana dos boxplots, que aumentam em uma proporção semelhante e confirmam os horários de maior e menor uso como já visto na subseção 4.1 .

Em uma segunda análise, é possível correlacionar os valores da variância e do desvio padrão (que apresenta um comportamento muito semelhante ao da variância) com os boxplots das Figuras (14) e (15). Assim como a média, a variância exibe um "poço"entre 0h e 10h, com o valor mínimo ocorrendo às 4h. Após esse horário, a variância aumenta até as 10h, onde se mantém estabilizada até às 15h, reduzindo em seguida até às 21h. Esse comportamento reflete a dispersão observada nos boxplots (linhas verticais pretas em cada coluna), que apresenta um formato semelhante ao reduzir de 0hs às 4hs, depois aumentar das 4hs até às 11hs, e reduzir entre 17hs e 21hs.

Da mesma maneira, verifica-se que os gráficos relacionados à média, variância e desvio padrão do dispositivo Chromecast, presentes nas figuras (16) e (17), também estão condizentes com seus respectivos gráficos de boxplots, das figuras (12) e (13). Percebe-se que a média varia pouco, com a linha dos gráficos tendo poucas inclinações abruptas e sempre estando entre 3 e 3.5 (valores em log) para o upload e por volta de 3.5 e 4 para o download. Dessa forma, pode-se novamente correlacionar os valores da média com os da mediana, já que ambos apresentam valores próximos se comparados os gráficos em questão

Simultaneamente, a variância e o desvio padrão para as taxas de download dos dispositivos Chromecast se mostram maiores que as de upload, com a primeira variando perto da faixa de 1.5 e 2.0, e a segunda estando por volta de 0.5, o que se comprova ao se analisar a dispersão dos dados nos boxplots das figuras (12) e (13), no qual as dispersões das taxas de download são bem maiores que as de upload. Isso explica por que o gráfico (12) tem tantos outliers: como sua dispersão é baixa, há mais chances de outliers aparecerem do que comparado às taxas de download.

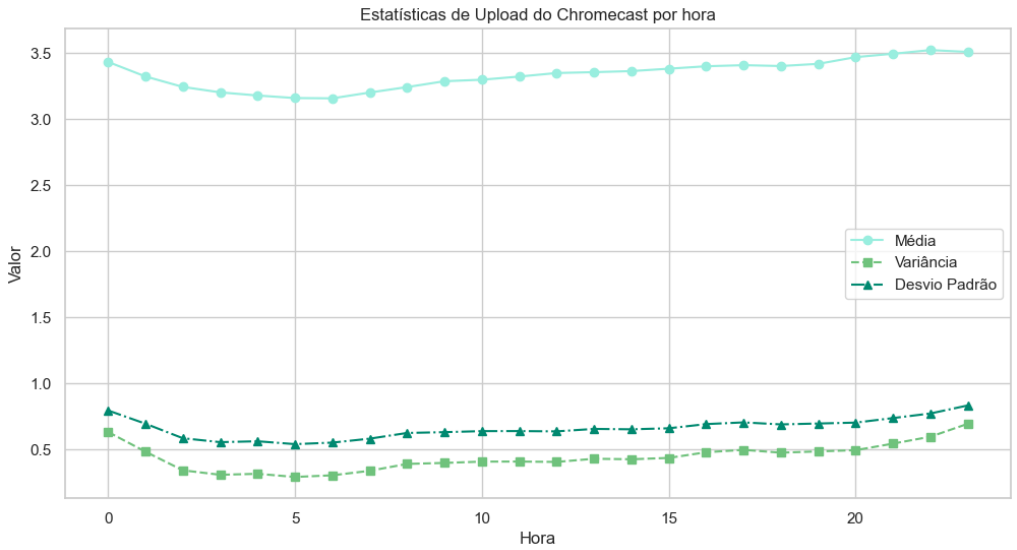


Figura 16: Estatísticas descritivas da Taxa de Upload do Chromecast por hora

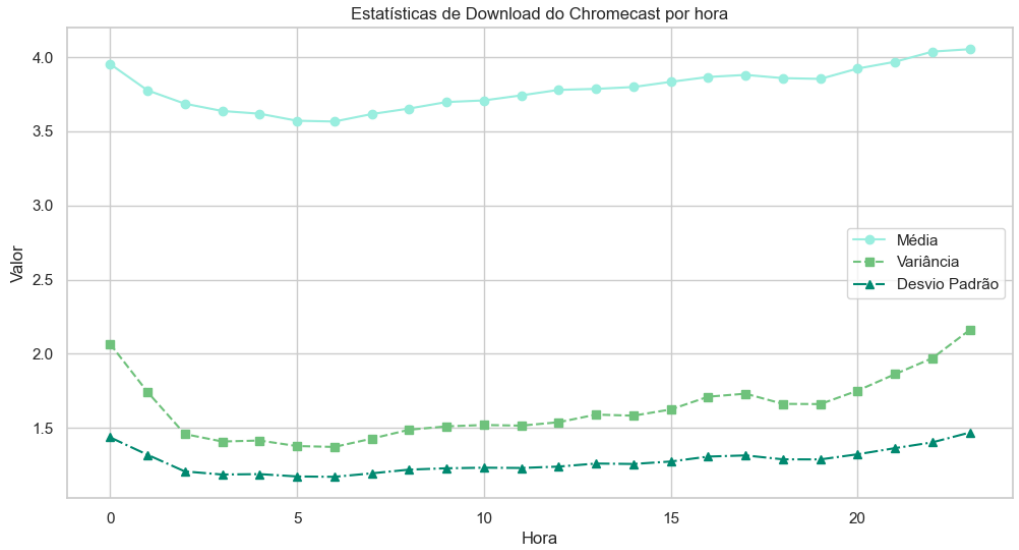


Figura 17: Estatísticas descritivas da Taxa de Download do Chromecast por hora

4.3 Interpretação geral dos dados

Com base nas análises realizadas, é evidente que os dois dispositivos apresentam padrões de uso distintos ao longo do dia, refletidos tanto nas medidas descritivas quanto nos gráficos de boxplot. A Smart TV demonstra um comportamento mais dinâmico, com variações significativas nas taxas de upload e download dependendo do horário. Durante a madrugada e início da manhã, as taxas apresentam valores baixos, tanto na mediana quanto na média, e uma menor dispersão. Já a partir do final da manhã, e especialmente à tarde e à noite, observamos um aumento expressivo nas taxas, com maior dispersão, indicando períodos de uso intenso. Esses padrões sugerem que o uso da Smart TV está relacionado a horários típicos de consumo de mídia, como filmes e vídeos, que ocorrem principalmente durante o dia e a noite.

Por outro lado, o Chromecast apresenta um padrão mais estável ao longo das 24 horas do dia. As taxas de upload e download mantêm-se relativamente constantes, com pequenas variações tanto na mediana quanto na média, e uma dispersão mais uniforme em comparação à Smart TV. Essa estabilidade indica que o Chromecast é utilizado de maneira mais consistente, sem uma forte dependência de horários específicos, o que pode estar relacionado a uma utilização automatizada ou mais frequente ao longo do dia, como streaming contínuo ou integração com outros dispositivos.

As diferenças observadas entre os dois dispositivos também podem ser explicadas pela sua natureza de uso. A Smart TV, por ser um dispositivo de consumo direto de mídia, tende a apresentar picos de uso em horários de maior demanda por entretenimento. Já o Chromecast, sendo um dispositivo de transmissão, reflete padrões mais constantes e menos dependentes de horários. Em termos de dispersão, as taxas de download para ambos os dispositivos são geralmente maiores do que as de upload, o que é consistente com o comportamento esperado para dispositivos focados no consumo de mídia.

5 Análises dos horários com maior valor de tráfego

Nessa seção, o objetivo é analisar os horários com maior valor da média para as taxas de upload e de download para cada tipo de dispositivo. Para esse estudo, a análise foi dividida em 3 passos, de modo a organizar melhor a forma como os dados serão manipulados: primeiramente, serão determinados os 4 datasets contendo os dados dos horários com maior valor de tráfego para cada taxa e dispositivo; em seguida, é feito um histograma para cada um desses 4 datasets; e, por fim, são feitos dois QQ Plots, um para comparar os dados dos datasets relacionados ao upload (datasets 1 e 3) e o outro para comparar os dados associados às taxas de download (datasets 2 e 4).

5.1 Passo 1 - Determinação dos datasets

Para a determinação dos datasets a serem analisados nessa seção, basta escolher os dados que ocorrem nos horários com maior valor de média para cada gráfico da seção 4.2. Com isso, chegou-se nos seguintes horários de pico para cada gráfico:

Horários de Pico

Smart TV
Dataset 1 - Horário de pico de upload da Smart TV: **20hs**
Dataset 2 - Horário de pico de download da Smart TV: **20hs**

Chromecast
Dataset 3 - Horário de pico de upload do Chromecast: **22hs**
Dataset 4 - Horário de pico de download do Chromecast: **23hs**

5.2 Passo 2 - Histograma para os horários de pico

Abaixo, nas figuras (18), (19), (20) e (21), estão disponíveis os histogramas com os dados dos datasets 1, 2, 3 e 4, definidos na seção anterior. Esses histogramas consideram apenas os dados que ocorrem no horário de maior média, e registram o número de ocorrências para cada log na base 10 da taxa de upload ou download.

De um modo geral, pode se observar que os padrões que ocorrem nos gráficos abaixo se assemelham muito aos histogramas gerais que consideram todos os datasets, feitos na subseção 3.1. Para a Smart TV, as taxas de upload e download apresentam alta concentração de ocorrências próximas a zero, indicando predominância de valores baixos. No entanto, as taxas de download mostram valores máximos mais elevados e maior dispersão, sugerindo maior variação em comparação às taxas de upload. Já para o Chromecast, os histogramas indicam comportamentos distintos entre upload e download. As taxas de upload são mais centralizadas, com a maioria das ocorrências entre os valores logarítmicos de 2 e 4, refletindo uma distribuição mais homogênea. Por outro lado, as taxas de download apresentam maior variabilidade, com picos em torno de 2-3 e 4-5, além de valores significativamente altos (entre 6 e 7), sugerindo uma ampla gama de velocidades.

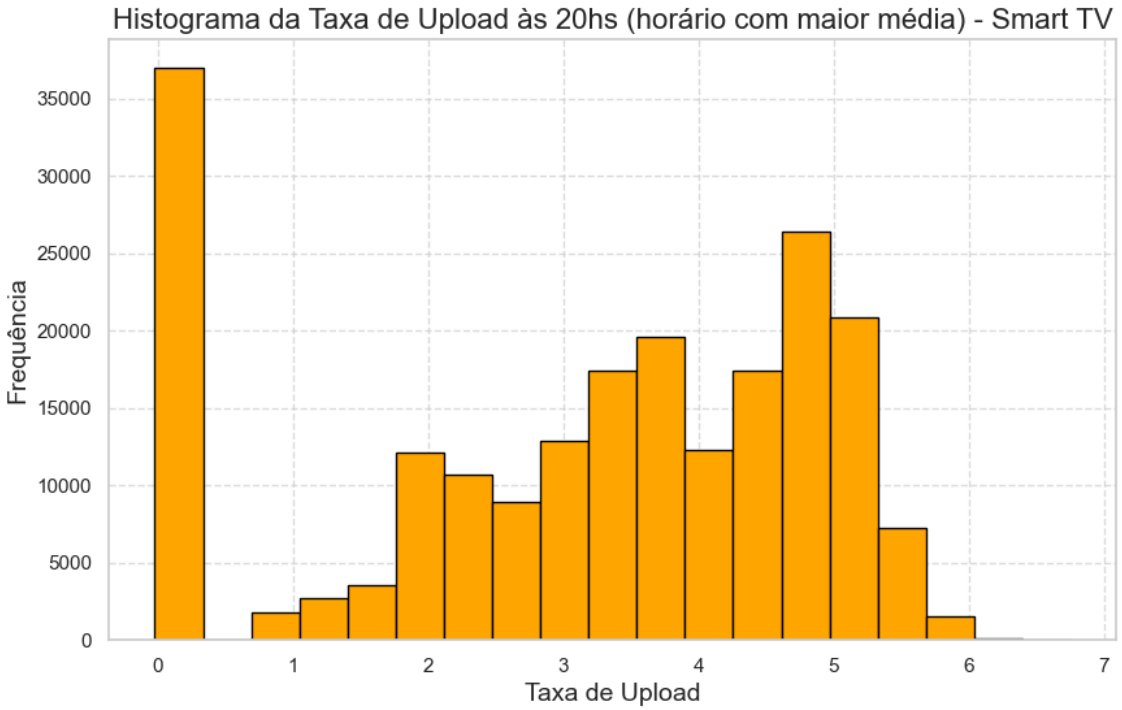


Figura 18: Histograma da taxa de upload às 20hs da Smart TV

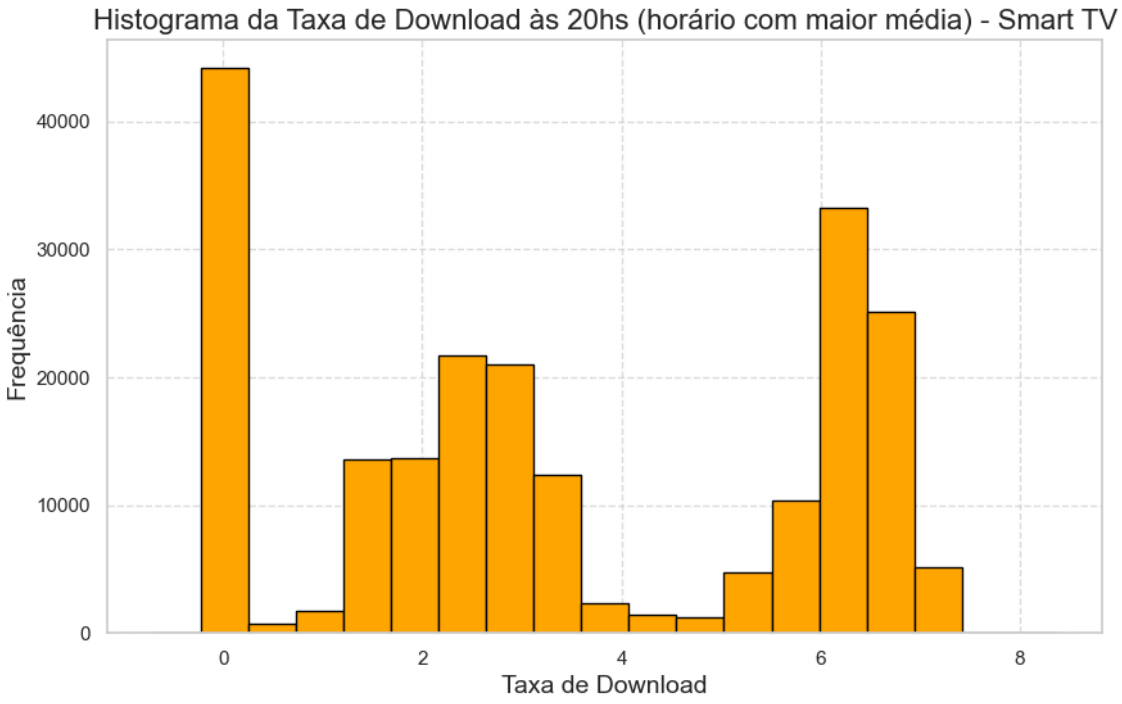


Figura 19: Histograma da taxa de download às 20hs da Smart TV

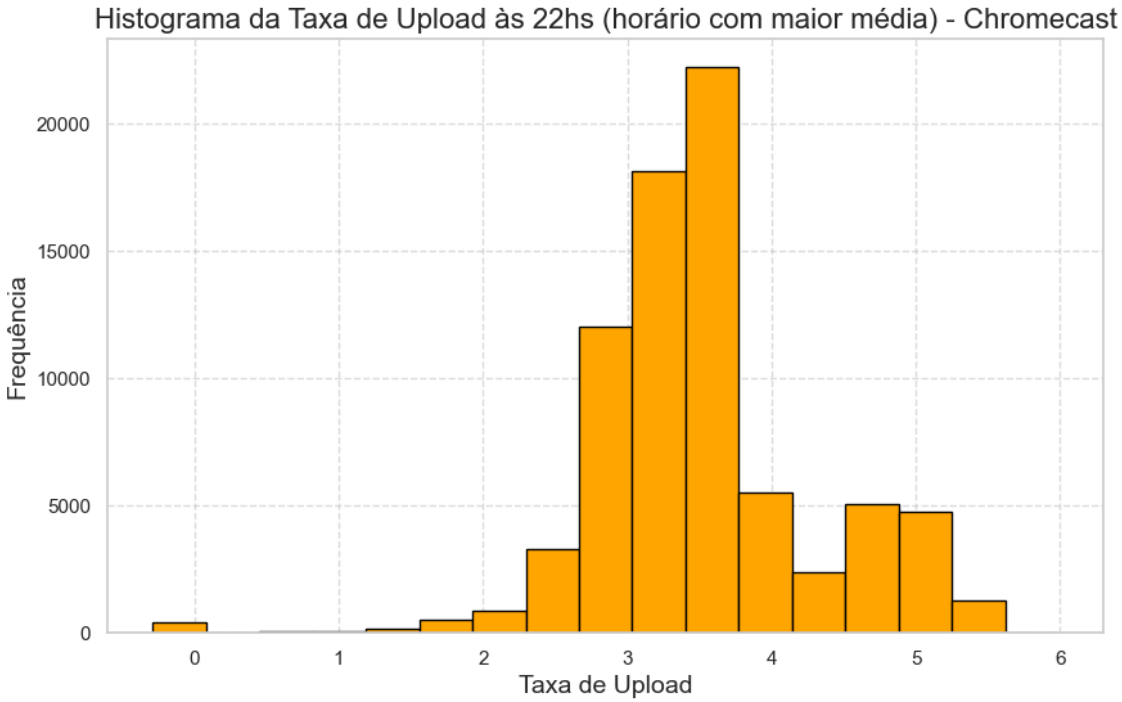


Figura 20: Histograma da taxa de upload às 22hs do Chromecast

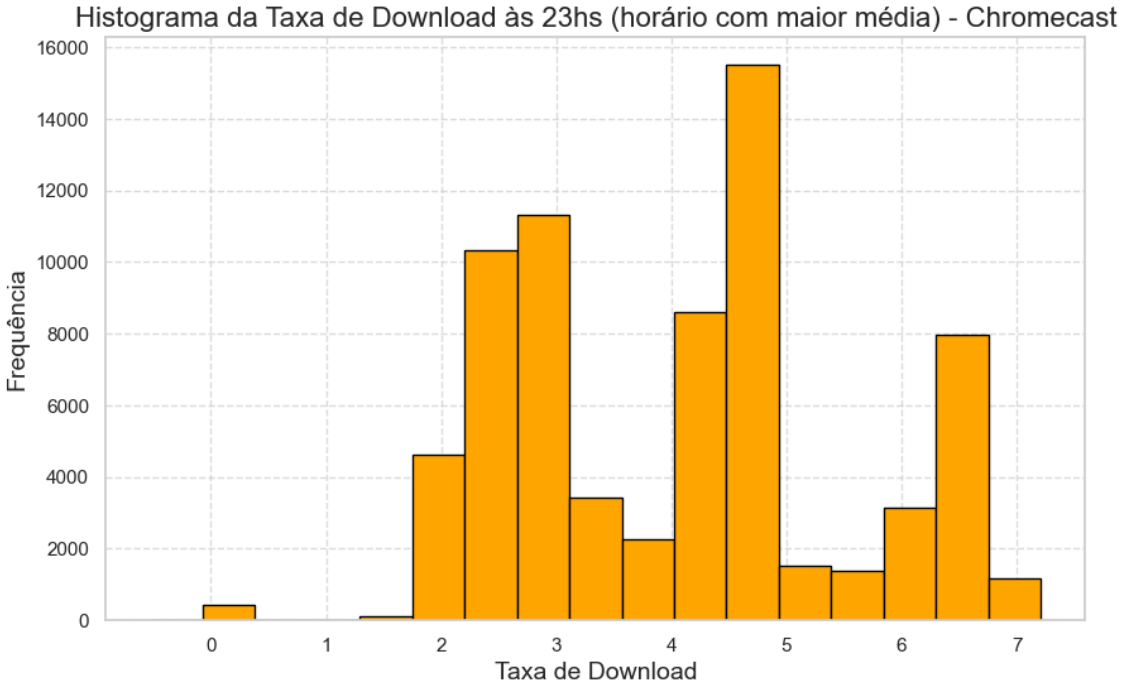


Figura 21: Histograma da taxa de download às 23hs do Chromecast

Tratando das similaridades entre os datasets 1, 2, 3 e 4, as comparações são as mesmas feitas na seção 3.1. Os quatro datasets apresentam padrões distintos, mas também compartilham algumas semelhanças, refletindo os diferentes comportamentos de upload e download para Smart TV e Chromecast.

- Dataset 1 (Upload da Smart TV às 20hs): O histograma exibe uma alta concentração de ocorrências próximas a zero, indicando que a maioria das taxas de upload para a Smart TV é muito baixa. Se comparado com o histograma (10), há mais ocorrências de dados com valores distantes de zero;

- Dataset 2 (Download da Smart TV às 20hs): Embora também apresente concentração próxima a zero, o número de ocorrências é ainda maior do que no upload da Smart TV. Além disso, as taxas de download alcançam valores logarítmicos mais altos, indicando que, em geral, a Smart TV experimenta velocidades de download superiores às de upload, com maior dispersão nos valores. Vale notar que existe uma região em torno da faixa dos valores logarítmicos de 4 e 5 que apresentam poucos dados;

- Dataset 3 (Upload do Chromecast às 22hs): Diferentemente da Smart TV, as taxas de upload para o Chromecast se concentram principalmente entre os valores logarítmicos de 2 a 4, com um pico em torno de 3. Embora a faixa de foco dos dados seja igual no histograma (12), seu pico de ocorrências está entre 3 e 4;

- Dataset 4 (Download do Chromecast às 23hs): O histograma do download do Chromecast para 23hs é mais disperso que o do upload das 22hs, apresentando três picos principais: um entre 2 e 3, um entre entre 4 e 5 e outro 6 e 7. Uma diferença em relação ao histograma (12) é que a faixa 6 e 7 é maior nesse histograma;

Analisando especificamente as semelhanças entre os datasets, a primeira comparação que pode ser feita é a semelhança que o 1 e o 2 possuem entre si por serem ambos da Smart TV, e o 3 com o 4 por serem do Chromecast. Em relação aos datasets da Smart TV, é possível verificar a grande quantidade de dados que são próximos a zero, sendo o pico de ocorrência desses gráficos. Já os datasets do Chromecast apresentam uma distribuição de dadosmais centralizadas nas faixas intermediárias (valores logarítmicos entre 2 e 4 para upload e 4 a 5, com destaques para as faixas 2 e 3 e 6 e 7, para download). Outra semelhança que pode se notar é como as taxas de download sempre alcançam valores logarítmicos maiores que as taxas de upload, como é possível ver comparando os gráficos (19) com (18) e o (21) com o (20)

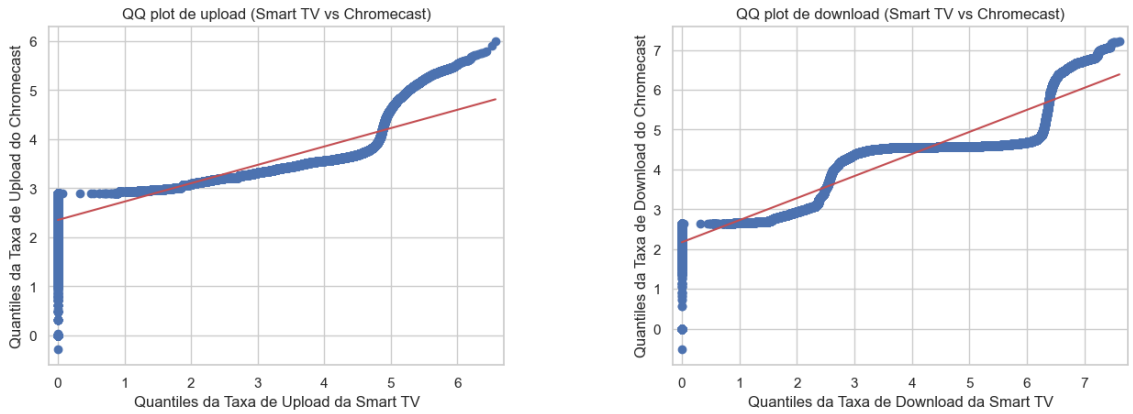
Sobre as diferenças, destacam-se como os datasets 2 e 4, referentes às taxas de download, apresentam regiões com menos dados se comparados com seus respectivos datasets de upload do mesmo dispositivo: o dataset 2 possuindo tal região por volta da faixa de 4 e 5, e o 4 possuindo duas regiões de menos dados em torno de 3 e 4 e perto de 5 e 6 (mas compensando com os mais dados entre o 2 e 3, e o 6 e 7).

5.3 Passo 3 - QQ Plots

No último passo dos estudos sobre os dados dos horários de maior tráfego dos datasets, foram feitos dois QQ Plots, um para comparar os dados dos datasets relacionados ao upload (datasets 1 e 3) e o outro para comparar os dados associados às taxas de download (datasets 2 e 4). Pelo fato desses datasets não necessariamente terem o mesmo número de elementos, foi necessário aplicar o método de interpolação descrito no projeto do trabalho para redimensionar o conjunto de dados maior para corresponder à distribuição de quantis do menor. Os QQ Plots estão disponíveis abaixo, nas imagens (22a) e (22b).

No QQ Plot da figura (22a), que compara as taxas de upload dos dispositivos, observa-se que a maior parte dos dados se alinha bem com a linha de referência diagonal, indicando que as distribuições possuem características semelhantes para valores medianos e intermediários. No entanto, nos extremos da distribuição (valores muito baixos ou muito altos), observam-se desvios significativos da linha diagonal, sugerindo diferenças na variabilidade entre os dois dispositivos. A Smart TV parece apresentar maior dispersão em valores extremos de upload (por conta de sua grande quantidade de dados próximos a zero), enquanto o Chromecast tem uma distribuição mais compacta.

Já para o QQ Plot da figura (22b), referente às taxas de download, o gráfico revela um alinhamento razoável nas partes centrais da distribuição, mas diferenças marcantes aparecem nas caudas. Nos valores mais baixos de download, a Smart TV apresenta maior ocorrência de dados próximos a zero, o que não é observado no Chromecast. Por outro lado, em valores altos de download, as taxas da Smart TV ultrapassam significativamente as do Chromecast, indicando a presença de valores extremos elevados que refletem sua maior variabilidade e amplitude de operação.



(a) QQ Plot da taxa de upload

(b) QQ Plot da taxa de download

Figura 22: Comparação entre as QQ Plots de taxa de upload e taxa de download

6 Análises da correlação entre as taxas de upload e download para os horários com o maior valor de tráfego

Nessa seção, o objetivo de análise é verificar se existe alguma correlação entre a taxa de upload e a de download de um mesmo dispositivo. Para isso, são feitos dois gráficos Scatter Plot a partir dos datasets definidos na subseção 5.1 e, em seguida, é calculado o coeficiente de correlação amostral, um para cada dispositivo. Pelo fato desses datasets não necessariamente terem o mesmo número de elementos, foi necessário aplicar um método de interpolação para equiparar os datasets, mas como a descrição do trabalho não especifica qual deve ser usada nessa seção, foi utilizado no código a interpolação da biblioteca numpy. Os Scatter Plots estão disponíveis abaixo, nas imagens (23) e (24).

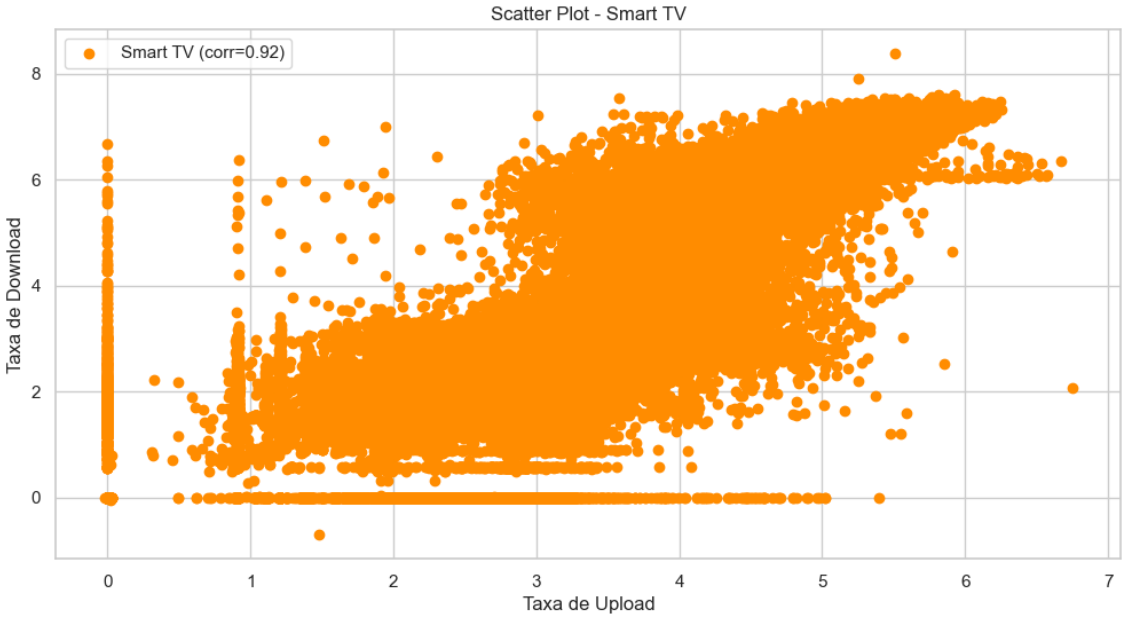


Figura 23: Scatter Plot da Smart TV

A partir da análise do gráfico de dispersão da figura (23), que relaciona as taxas de upload e download para a Smart TV, observa-se uma correlação positiva significativa entre essas variáveis, evidenciada pelo coeficiente de correlação (corr = 0,92) calculado. Esse alto valor de correlação sugere que, à medida que a taxa de upload aumenta, a taxa de download tende a aumentar também, indicando uma relação linear forte entre as duas variáveis.

A dispersão dos pontos no gráfico apresenta uma tendência geral ascendente, corroborando a existência dessa relação positiva. No entanto, é importante notar que existem alguns pontos mais dispersos e valores próximos a zero, indicando possíveis casos de tráfego reduzido ou comportamento anômalo do dispositivo. Esses pontos podem ocorrer em horários de menor uso ou por limitações específicas de rede.

Analisando-se os histogramas feitos na subseção 5.2, é possível notar a semelhança entre os gráficos (18) e (19), e como seus dados aumentam a uma proporção parecida. Existe uma divergência que ocorre pelo fato de haver baixas taxas de download na faixa dos valores logarítmicos de 4 e 5 para as taxas de download, mas, de um modo geral, a proporcionalidade dos dados possui semelhança.

Portanto, com base nos dados observados, conclui-se que há uma forte correlação linear entre as taxas de upload e download do dispositivo analisado nos horários de maior tráfego, o que pode refletir um comportamento consistente de utilização da rede pela Smart TV.

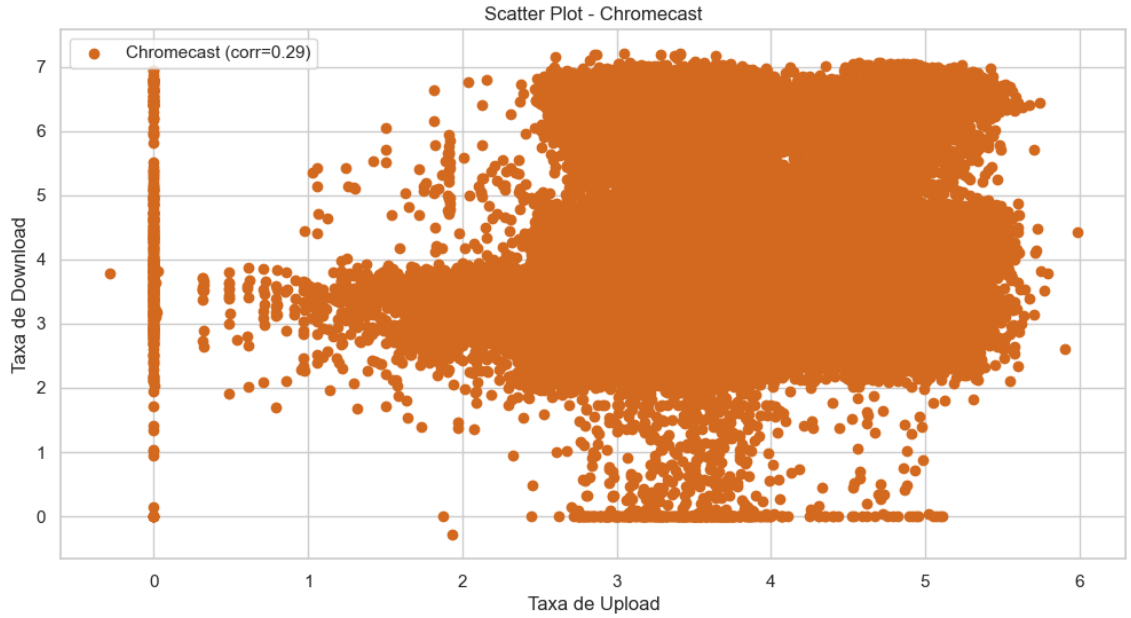


Figura 24: Scatter Plot do Chromecast

Já o Scatter Plot da figura (24), que relaciona as taxas de upload e download para o dispositivo Chromecast, revela um coeficiente de correlação baixo (corr = 0,29). Esse valor indica uma correlação positiva fraca entre as duas variáveis, sugerindo que não há uma relação linear forte entre as taxas de upload e download.

Observando os dados no gráfico, percebe-se uma grande dispersão dos pontos, sem uma tendência clara de alinhamento que indique uma relação significativa entre as variáveis. Existe uma concentração notável de pontos ao longo do eixo y para valores baixos de upload (próximos de zero), o que pode indicar que o dispositivo realiza mais operações de download do que de upload em determinados momentos.

Além disso, analisando-se pelos histogramas (20) e (21) da subseção 5.2, é possível notar que os dois gráficos possuem comportamentos bem distintos. Ao passo que o dataset 3 possui apenas um grande pico de dados entre a faixa dos valores logarítmicos do 2 e 4, o dataset 4 apresenta seu maior pico entre o 4 e o 5, e mais outras duas regiões com foco de ocorrências nas faixas do 2 ao 3, e do 6 ao 7. Isso dificulta uma correlação linear, visto que seus comportamentos e proporcionalidades divergem bastante, tendo o dataset 3 apenas um pico, e o dataset 4 tendo regiões de destaque.

Com base nesses resultados, conclui-se que, para o dispositivo Chromecast, a relação entre as taxas de upload e download é fraca, indicando que essas variáveis não estão diretamente relacionadas de maneira consistente nos horários analisados. Esse comportamento pode ser explicado pelo padrão de uso típico do Chromecast, que geralmente realiza mais downloads (como streaming de conteúdo) e menos uploads. Essa análise pode ser útil para identificar possíveis melhorias na infraestrutura de rede ou otimizar o desempenho do dispositivo em cenários específicos.

7 Conclusão

A partir de todas as análises e estudos feitos sobre os datasets do trabalho, é possível garantir com certeza que os resultados obtidos podem auxiliar o provedor de serviço de Internet a entender os dados que passam pela rede de um usuário. Isso porque as conclusões que podem ser deduzidas a partir dos estudos dos dados dizem bastante sobre o comportamento do usuário em relação ao uso dos dispositivos, e isso pode auxiliar a empresa a direcionar seus esforços a aprimorar seus serviços com base nesse padrão.

São conclusões relevantes, por exemplo, as análises que mostraram que as Smart TVs apresentam uma grande variabilidade nas taxas de tráfego, com picos de uso concentrados principalmente entre 17h e 21h. Durante esses horários, as taxas de download tendem a ser significativamente maiores do que as de upload, evidenciando a natureza do dispositivo como ferramenta de consumo de mídia. Juntamente disso, foi visto como durante os horários da madrugada o dispositivo é pouco utilizado, e como existem muitos dados coletados que indicam taxas de upload e download com valor igual ou muito perto de zero.

Por outro lado, os dispositivos Chromecast se destacam por um comportamento mais estável ao longo do dia, com taxas de upload e download consistentes e menos dependentes de horários específicos. Essa característica sugere que esses dispositivos estão associados a um uso mais contínuo e automatizado, com menor variação nas demandas de rede.

Uma análise da correlação entre as taxas de upload e download revelou diferenças marcantes entre os dois dispositivos. Nas Smart TVs, há uma forte relação linear, com o aumento de uma taxa acompanhando diretamente o aumento da outra. Isso indica que os dispositivos tendem a utilizar ambas as direções de tráfego de maneira conjunta, especialmente durante os horários de maior uso. Em contraste, os Chromecasts apresentam uma correlação fraca, evidenciando que suas taxas de upload e download não estão diretamente relacionadas, reforçando o foco do dispositivo em operações de streaming e consumo de mídia.

Essas descobertas trazem implicações importantes para empresas que desejam otimizar sua infraestrutura ou aprimorar a experiência do usuário. Identificar horários de pico, como 20h para as Smart TVs e 22h-23h para os Chromecasts, permite direcionar recursos de rede para suportar demandas intensas de tráfego nesses períodos. Além disso, compreender a predominância de downloads em Smart TVs e a estabilidade das taxas nos Chromecasts ajuda a ajustar a alocação de banda e otimizar o desempenho dos serviços oferecidos.