Estatística e Modelos Probabilísticos

Trabalho Final da Disciplina

Engenharia de Computação e Informação UFRJ

Carolina Santiago de Medeiros 122053305

Link do Repositório: `https://github.com/carol-santiago/Estatistica-Projeto-Final`

1 Introdução

Este relatório apresenta os resultados de um projeto de estatística cujo objetivo é analisar dados provenientes de dois dispositivos de streaming muito conhecidos: o Chromecast e a Smart TV. O projeto se baseia na aplicação prática da teoria estatística aprendida em sala de aula, implementando diversas técnicas apresentadas. Os dados representam a taxa de dados enviados (taxa de upload) e a taxa de dados recebidos (taxa de download) em bps, de dispositivos localizados nas residências dos usuários do provedor. Uma peculiaridade importante destacada é a necessidade de reescalonar os dados para logaritmo na base 10, uma vez que as taxas abrangem diversas ordens de grandeza.

A análise estatística foi conduzida com o auxílio da linguagem de programação Python, utilizando uma variedade de bibliotecas para a montagem de gráficos de variados tipos, para explorar e visualizar as nuances dos conjuntos de dados de maneira completa. Este relatório está estruturado de maneira a fornecer uma compreensão abrangente do desempenho individual de cada dispositivo, bem como uma comparação detalhada entre o Chromecast e a Smart TV.

2 Estatísticas Gerais

Nesta seção, será realizada uma análise das características fundamentais dos conjuntos de dados. O objetivo central é desvendar os possíveis padrões e tendências que permeiam esses dados de forma "generalizada", oferecendo uma visão um pouco mais embasada sobre o desempenho dos dispositivos. Através da utilização de diversas ferramentas estatísticas, como histogramas, funções de distribuição e boxplots, procura-se elucidar o máximo de informações sobre as taxas de dados em questão.

A seguir, estão expostos os valores reais para a média, desvio padrão e variância de cada uma das taxas, para ambos os dispositivos, com um arredondamento de 3 casas decimais. Os dados da Tabela 1 foram calculados de forma direta a partir dos Datasets fornecidos, com auxílio da biblioteca Pandas, e serão utilizados como parâmetro para as análises posteriores.

	Chromecast		**Smart TV**	
	Download	**Upload**	**Download**	**Upload**
MÉDIA	3.799	3.350	2.350	2.157
DESVIO PADRÃO	1.291	0.679	2.593	2.028
VARIÂNCIA	1.666	0.462	6.724	4.113

Tabela 1: Dados extraídos dos Datasets.

Ainda considerando ambos dispositivos de maneira conjunta, foi gerado um gráfico com os *Boxplots* para cada situação, sendo estas respectivamente: Taxa de Download (extraída da coluna `"bytes_down"`) para o Chromecast; Taxa de Upload (extraída da coluna `"bytes_up"`); seguido pelo mesmo para o dispositivo Smart TV.

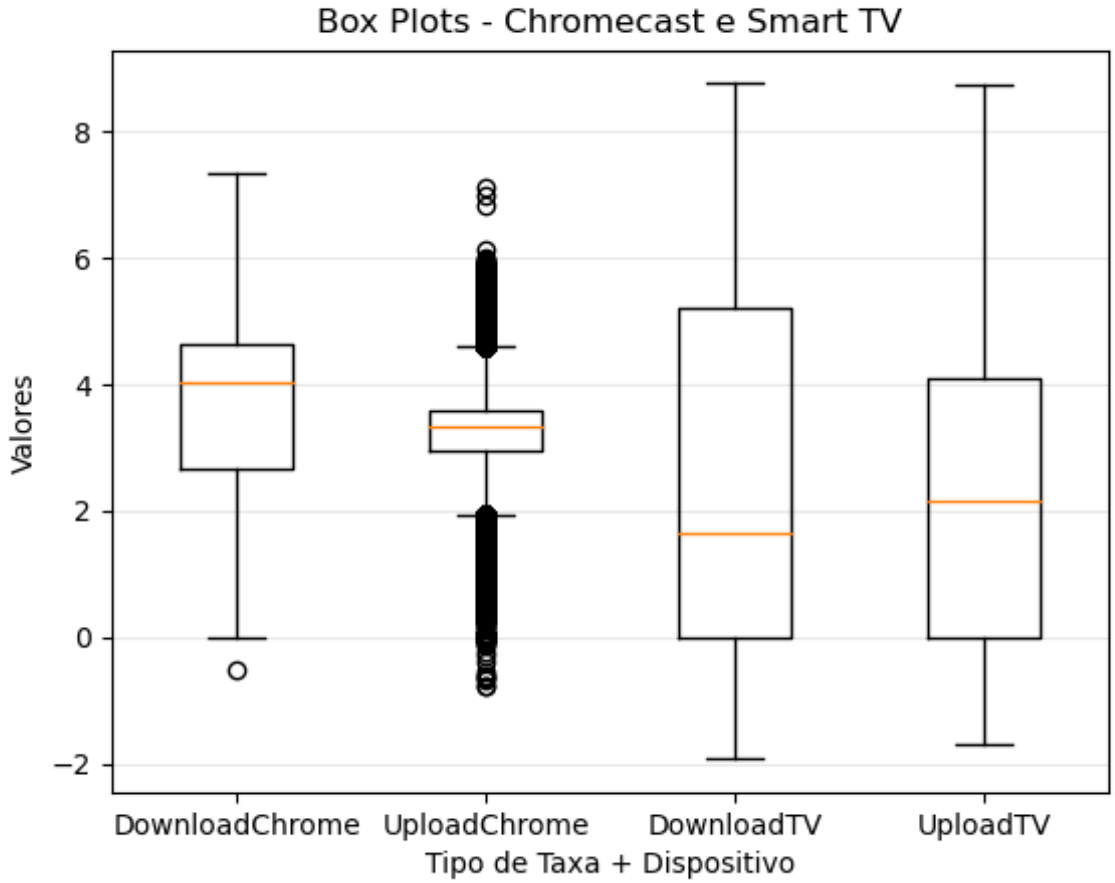

Figura 1: Boxplot Geral

Para o DownloadChromecast, um outlier menor que 0 foi identificado, e a mediana é aproximadamente 4. É importante lembrar que as taxas estão numa escala logarítmica, então um valor negativo de, por exemplo, -1, é o equivalente à 0.1 bps (10**-1). Já no *Boxplot* de UploadChromecast existem muitos outliers, a mediana está entre 3 e 4, e a variância é notadamente mais baixa do que nos outros.

Nos *Boxplots* da Smart TV não foram identificados outliers. A mediana está um pouco abaixo de 2 em DownloadTV, e um pouco acima em UploadTV. Um fato interessante a se notar é que em ambos o limite inferior é negativo e próximo a -2 (0.01 bps); e o valor do primeiro quartil é 0.

2.1 Chromecast

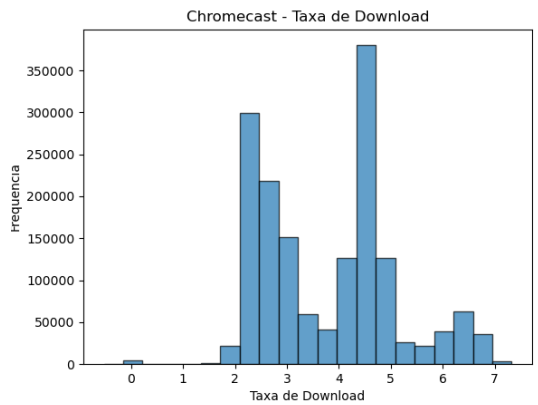

Figura 2: Histograma da Taxa de Download do Chromecast.

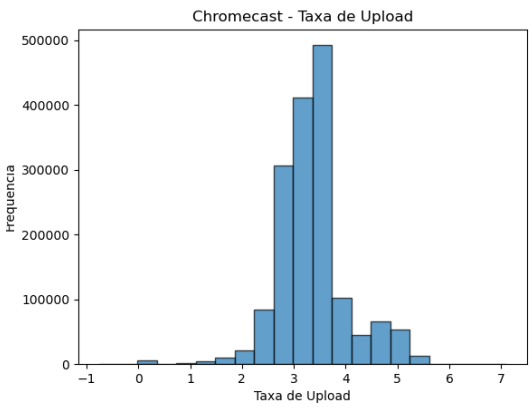

Figura 3: Histograma da Taxa de Upload do Chromecast.

Observando o primeiro histograma (Figura 2), é possível notar que a distribuição da taxa de download exibe dois picos distintos: um entre 2 e 3, com frequência de 300.000, e outro entre 4 e 5, com frequência superior a 350.000. Juntando esses dois fatos, o histograma revelaria uma predominância de taxas próximo a 3.5, que, curiosamente, é próximo da média real da taxa de download do Chromecast: 3.8.

Já no histograma da taxa de upload, é possível notar uma distribuição que se assemelha muito a uma curva gaussiana de baixa variância, com média entre 3 e 4, com uma frequência altíssima em relação aos outros valores. A "simetria"e a concentração das barras indicam uma consistência notável nas taxas de upload. Novamente, a média "entre 3 e 4"condiz com a média real de 3.35; e a variância real de 0.46 agrega à dedução de que esses dados seguem uma distribuição normal.

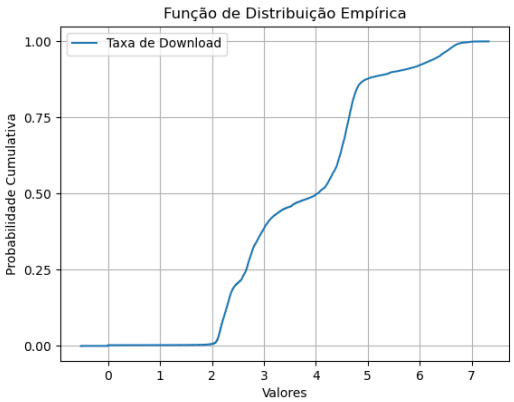

Figura 4: ECDF da Taxa de Download do Chromecast.

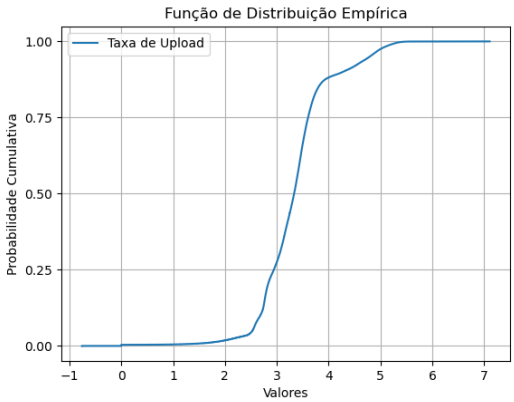

Figura 5: ECDF da Taxa de Upload do Chromecast.

As funções de distribuição cumulativa indicam um crescimento muito parecido com um exponencial suave, para ambas as taxas, alcançando a probabilidade máxima próximo ao valor 7 (para a taxa de download) e um valor pouco acima de 5 (para a taxa de upload). No gráfico de download, observa-se a presença de pequenas estagnações no crescimento da curva, especialmente entre os valores 3 e 4, sugerindo que nesta variável existe uma instabilidade maior. Isto condiz com os dados da Figura 2, que têm uma grande queda nesse intervalo.

2.2 Smart TV

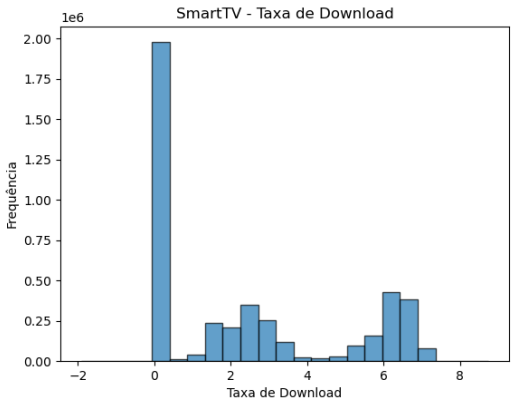

Figura 6: Histograma da Taxa de Download da Smart TV.

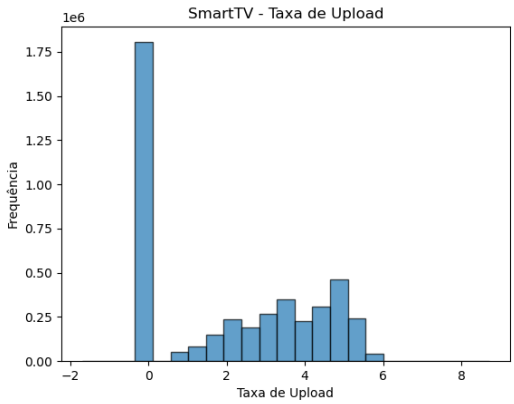

Figura 7: Histograma da Taxa de Upload da Smart TV.

Ambos os histogramas apresentam uma quantidade significativa de zeros, que podem representar dados coletados com o dispositivo desligado, ou em um modo analógico (como TV aberta), que não utiliza a internet. Para a Taxa de Download, as barras parecem ter uma média de frequência em torno de 0.25e6 para a maioria dos valores, porém decrescendo gradualmente quando a taxa tende a 4. O histograma da Taxa de Upload exibe um crescimento semelhante ao logarítmico, porém decresce a partir de, aproximadamente, 5.

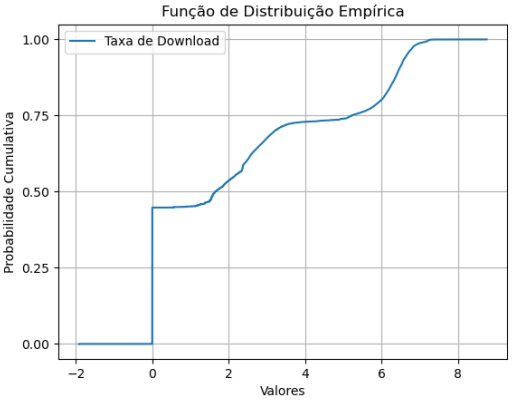

Figura 8: ECDF da Taxa de Download da Smart TV.

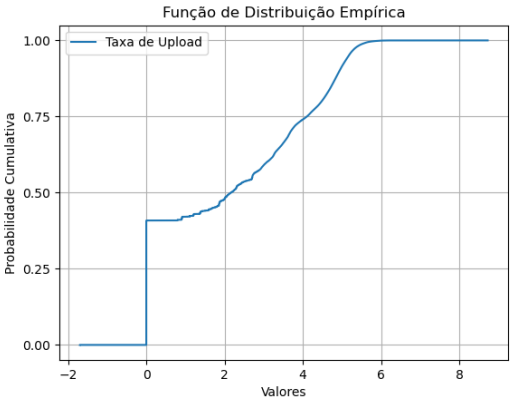

Figura 9: ECDF da Taxa de Upload da Smart TV.

As curvas, para valores acima de 0, são notavelmente semelhantes às do Chromecast. A diferença em destaque é que a probabilidade em zero atinge quase 0.5, o que reforça a alta quantidade de valores nulos nos dados.

3 Estatísticas por Horário

Esta seção explora possíveis correlações dentro dos conjuntos de dados, explorando as estatísticas de taxas de upload e download para os dispositivos Chromecast e Smart TV com relação ao horário da coleta. O objetivo principal é avaliar como essas estatísticas variam em diferentes momentos do dia, destacando padrões que possam influenciar o desempenho desses dispositivos, ou expor informações sobre o seu uso. A análise será conduzida por meio de gráficos informativos, proporcionando uma compreensão detalhada das flutuações nas médias, variâncias e desvios padrão. O objetivo não é de apenas descrever essas estatísticas, mas também interpretar seu significado em relação aos padrões observados anteriormente, fornecendo *insights* cruciais sobre o comportamento temporal das taxas de upload e download.

3.1 Chromecast

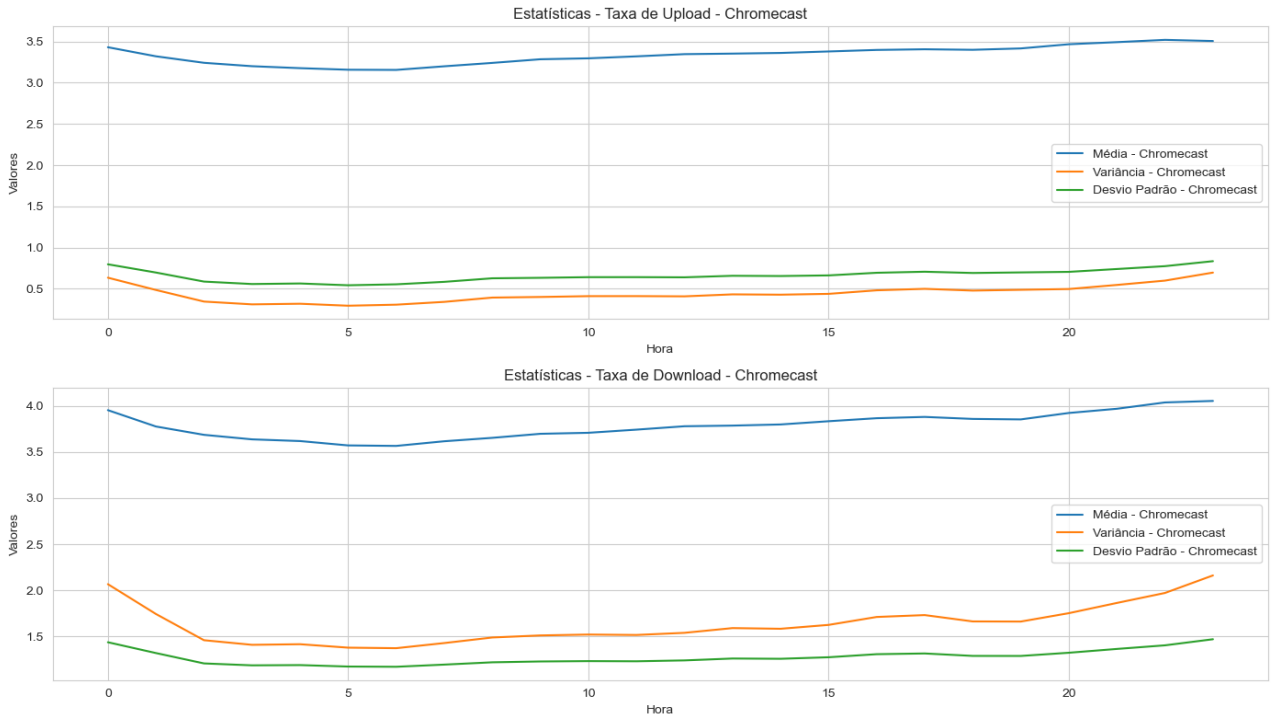

Figura 10: Plots das taxas de upload e download, retratando a média, variância e desvio padrão por hora do dia.

Os gráficos revelam uma uniformidade surpreendente nas estatísticas ao longo do dia, principalmente na Taxa de Upload. A pequena queda nas primeiras horas da manhã e a elevação à noite são perceptíveis, mas tão suaves que pode-se concluir que a consistência geral é mantida. No gráfico referente às características da Taxa de Download, a maior diferença é que a curva de variância está acima da curva de desvio padrão, e possui mais variações que na Taxa de Upload (descidas e subidas mais aparentes, porém ainda muito suaves).

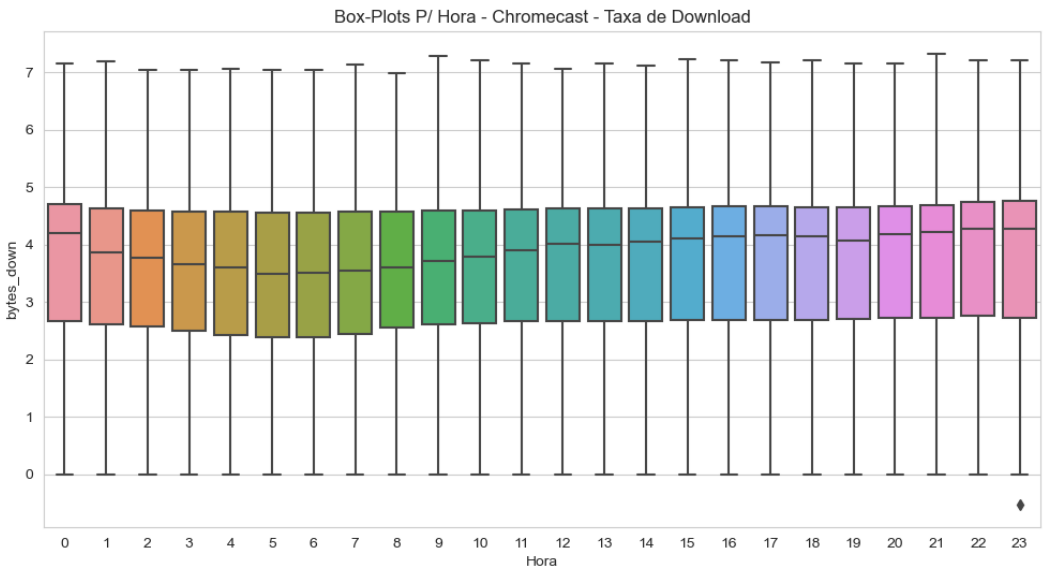

Figura 11: Boxplots da Taxa de Download para cada hora do dia.

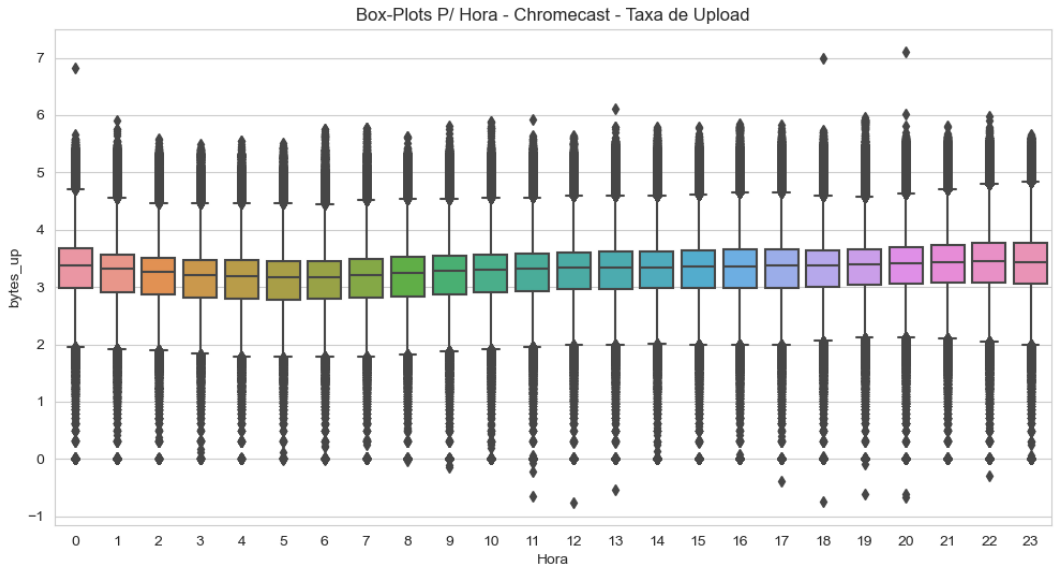

Figura 12: Boxplots da Taxa de Upload para cada hora do dia.

Os *Boxplots* do Chromecast confirmam a uniformidade das taxas destacada na Figura 10, com valores consistentes ao longo do dia. Essa uniformidade pode trazer a tona uma hipótese de que as taxas não estão necessariamente correlacionadas aos padrões de uso diários do dispositivo. De fato, um artigo da *IPSTAR Broadband*, uma operadora de internet australiana, fortalece essa possibilidade:

"Um Chromecast consome dados de duas formas. A primeira é se o utilizar para ver vídeos de serviços de transmissão em contínuo como o Netflix [...] ou o YouTube no seu telefone ou tablet. Durante todo o tempo em que está assistindo, o Chromecast transmite o conteúdo através da sua rede Wi-Fi [...] **A segunda maneira pela qual um Chromecast usa dados é quando está inativo** e nada está a ser transmitido para ele. Pode ter reparado que na tela da sua TV há um papel de parede que muda a cada poucos segundos, juntamente ao clima. O papel de parede que aparece na televisão é baixado dos servidores do Google, às vezes a cada 5 segundos. **Basicamente, é uma transmissão que nunca acaba.**"– Diz a empresa em "Could Your Chromecast Be Using All Your Data Without You Knowing?"

3.2 Smart TV

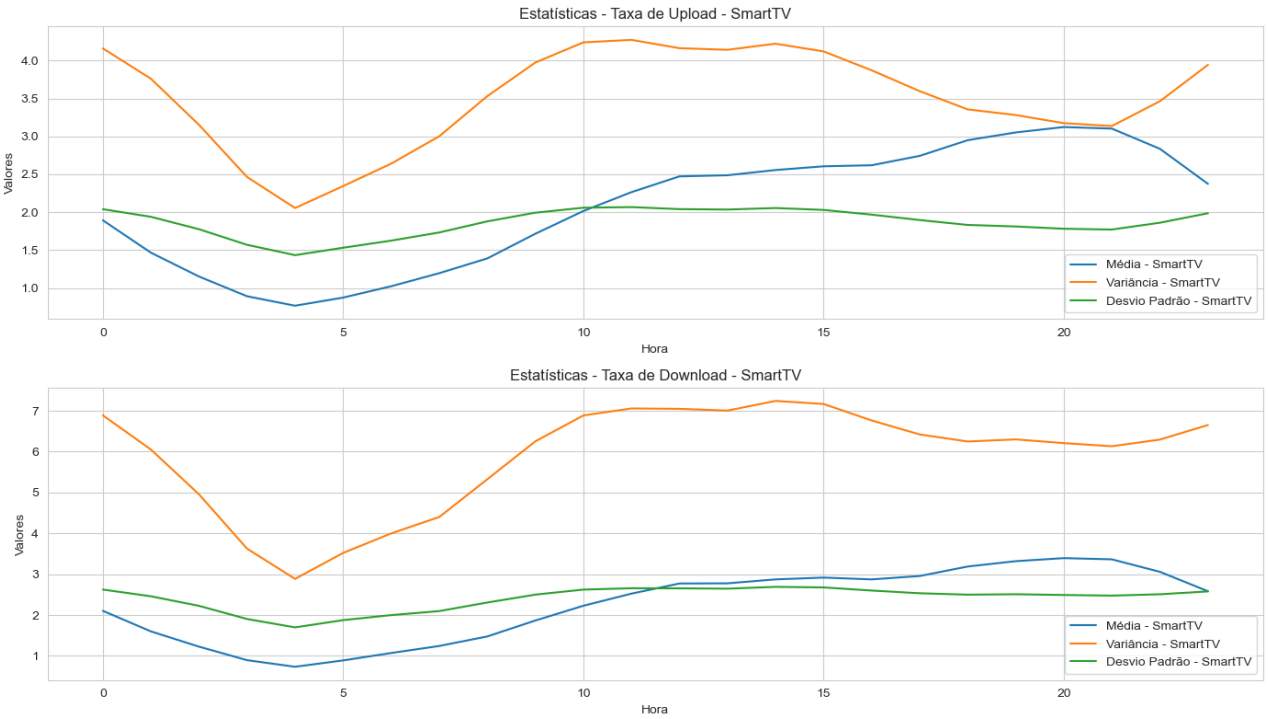

Figura 13: Plots das taxas de upload e download, retratando a média, variância e desvio padrão por hora do dia.

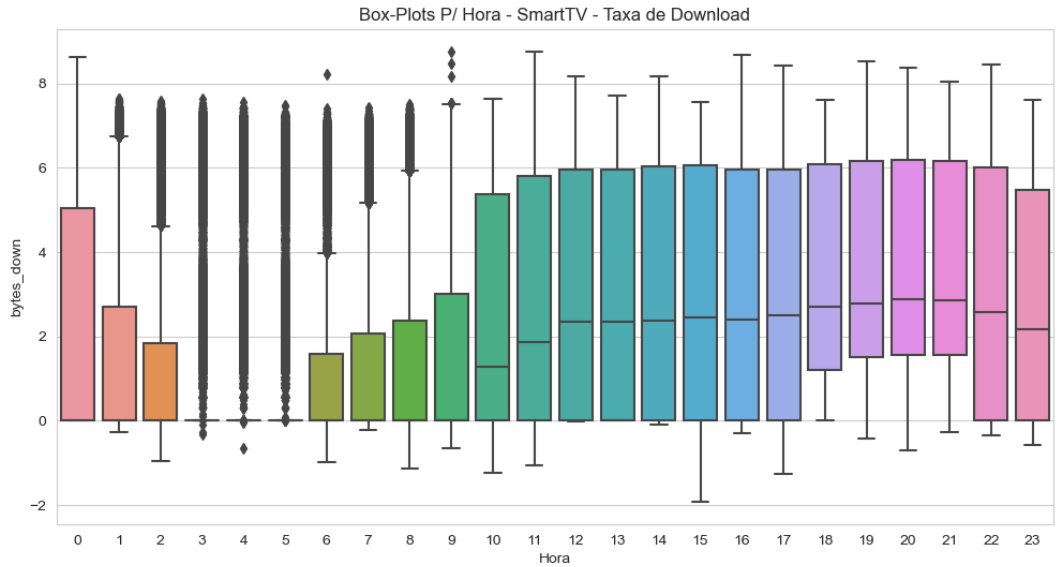

Figura 14: Boxplots da Taxa de Download para cada hora do dia.

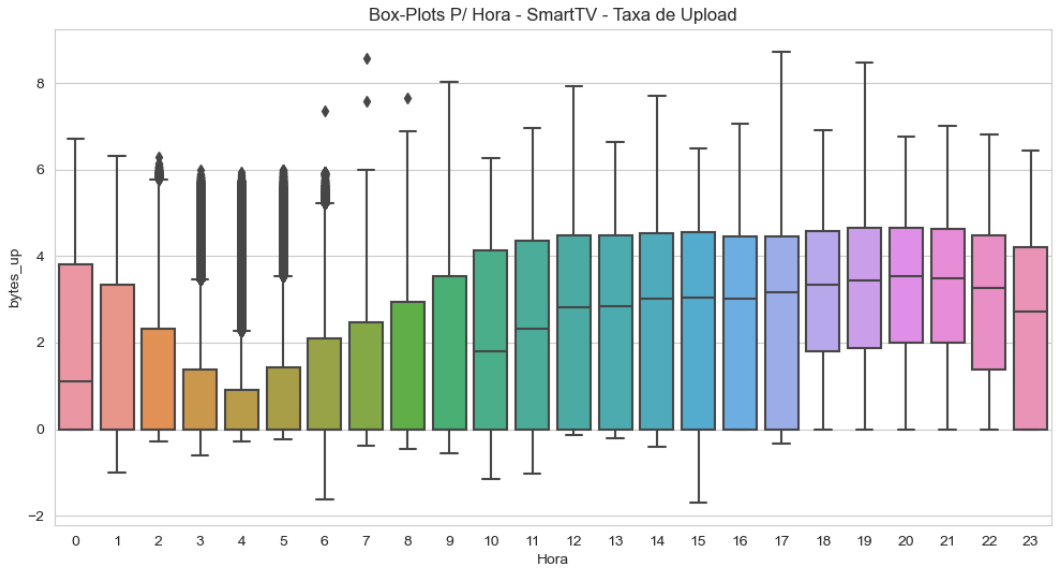

Figura 15: Boxplots da Taxa de Upload para cada hora do dia.

Ao contrário do Chromecast, os gráficos para a Smart TV exibem variações bastantes significativas nas curvas de média, variância e desvio padrão ao longo do dia. A variância consistentemente atinge seu mínimo aproximadamente às 4 horas e segue em seu "máximo"durante o período da tarde, das 11 horas às 17 horas, indicando os momentos de maior e menor estabilidade nas taxas. Os *Boxplots* revelam uma distribuição mais caótica, especialmente nos primeiros horários da manhã. A alta presença de *outliers* nesse horário sugere uma maior variabilidade nas taxas de transferência da Smart TV, possivelmente relacionada a padrões de uso muito distintos dentre os usuários.

A uniformidade nas estatísticas do Chromecast ao longo do dia indica uma certa independência das taxas com o uso real do dispositivo. Em contraste, a variabilidade mais acentuada na Smart TV sugere um comportamento menos previsível, possivelmente relacionado a uma correlação maior das taxas de transferência com os padrões de uso.

De acordo com o *Blog* de tecnologia *CertSimple*, essa hipótese pode ser verdadeira – "Não existe uma resposta definitiva para esta pergunta [A Smart TV usa internet quando desligada?], uma vez que depende da Smart TV em questão e das suas características e definições específicas. No entanto, em geral, **a maioria das Smart TVs não utiliza a Internet quando está desligada**. Isto porque não precisam de estar constantemente ligadas à Internet para funcionarem corretamente e porque, se o fizessem, gastariam muita energia desnecessária."– Diz o artigo aqui.

A partir dessas informações, é possível também assumir que a grande queda nas taxas da Smart TV no período da madrugada se dá pelo fato de que esse é o período do dia com menor uso, e, consequentemente, o período da tarde seria o de maior uso. Essas deduções condizem perfeitamente com o esperado de um usuário médio.

4 Caracterizando os horários com maior valor de tráfego

Para esta seção, os dados foram agrupados da seguinte forma: para cada hora do dia, foi calculada uma média da taxa neste horário. Em seguida, foi escolhido o horário com a maior média (nomeado de `HMT_(dispositivo)_(taxa)` no código do projeto), apresentados na tabela:

HMT	Download	Upload
Chromecast	23h	22h
Smart TV	20h	20h

Tabela 2: Horários com maior valor de tráfego para cada taxa, em ambos os dispositivos.

Com isso, os dados foram reorganizados em 4 datasets, que servirão de base para a elaboração dos gráficos posteriores.

- Dataset 1: Dados da taxa de Upload na Smart TV, no horário com maior média (20h).
- Dataset 2: Dados da taxa de Download na Smart TV, no horário com maior média (20h).
- Dataset 3: Dados da taxa de Upload no Chromecast, no horário com maior média (23h*).
- Dataset 4: Dados da taxa de Upload no Chromecast, no horário com maior média (23h).

* Como o objetivo desta seção é comparar dados de download e upload de um mesmo horário, e no dispositivo Chromecast os horários com maiores médias para as taxas são diferentes, foi utilizado o horário com maior média da Taxa de **Download**.

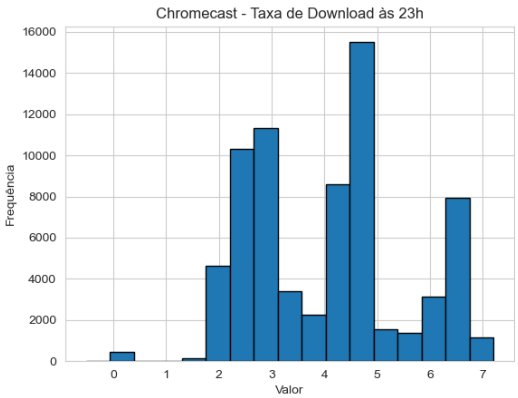

Figura 16: Histograma dos valores da Taxa de Download no horário de maior tráfego.

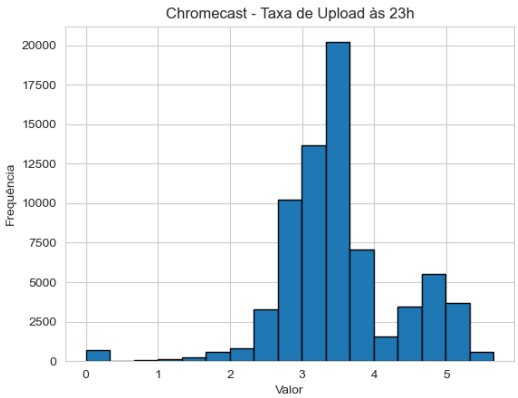

Figura 17: Histograma dos valores da Taxa de Upload no horário de maior tráfego.

Os histogramas indicam que, nos horários de maior tráfego no Chromecast, as distribuições das taxas de download e upload são extremamente semelhantes às distribuições dos histogramas gerais, sugerindo consistência nas características temporais dessas taxas. Três picos distintos são observados nos histogramas de Download, refletindo uma estrutura similar à Figura 2, e uma distribuição "normal"é notada na Taxa de Upload, assim como na Figura 3, com uma média em torno de 3.5 e variança baixa; o que condiz com os dados gerais, coletados e apresentados na Tabela 1.

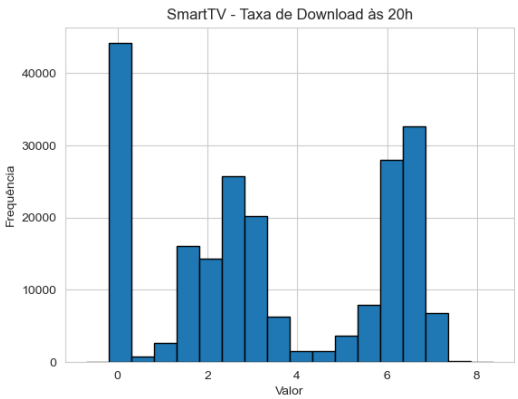

Figura 18: Histograma dos valores da Taxa de Download no horário de maior tráfego.

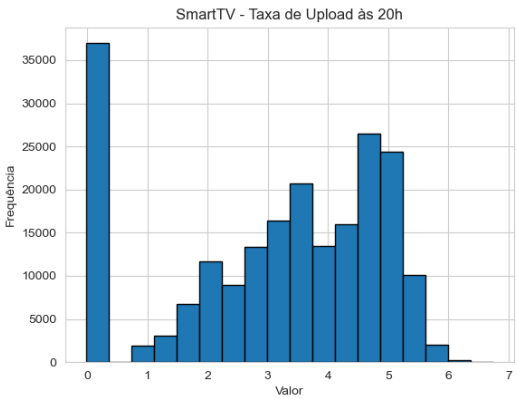

Figura 19: Histograma dos valores da Taxa de Upload no horário de maior tráfego.

Os histogramas para a Smart TV, no horário de pico, também reproduzem as mesmas características gerais das distribuições dos histogramas gerais nas Figuras 4 e 5. A consistência nas formas das barras sugere que, apesar da variação na quantidade de dados por hora, a distribuição das taxas se mantém. A presença significativa de zeros também se repete, alinhando-se com a descrição dos gráficos gerais previamente expostos deste dispositivo.

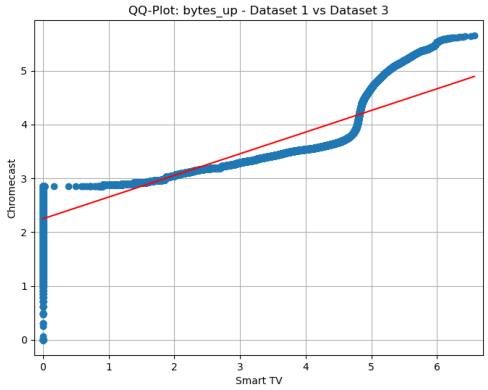

Figura 20: QQPlot da Taxa de Upload.

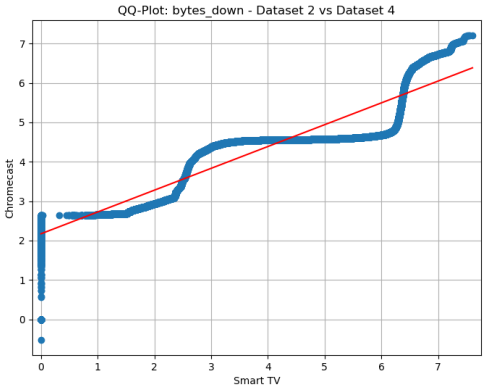

Figura 21: QQPlot da Taxa de Download.

Os QQ-Plots, ou gráficos quantil-quantil, são utilizados para comparar duas distribuições de dados. Eles mostram, graficamente, se os quantis de uma amostra correspondem aos quantis esperados de outra distribuição. Se os pontos se alinharem aproximadamente em uma linha reta de 45 graus, indica que as distribuições são semelhantes. Eles são extremamente úteis para comparar distribuições entre diferentes conjuntos de dados.

O primeiro QQ-Plot compara as Taxas de Upload entre os dispositivos. Ele revela que não há uma semelhança notável entre as distribuições das taxas em dispositivos diferentes, especialmente na região próxima a x = 0, onde existem uma grande concentração na Smart TV. A uniformidade inicial sugere uma concordância próxima ao longo de diferentes valores quantílicos, mas não sugere que os dados de diferentes dispositivos venham de uma mesma distribuição.

Já no segundo QQ-Plot, que compara as Taxas de Download, é revelado um alinhamento ainda mais caótico entre os Datasets. As múltiplas interseções com a linha de referência e as oscilações que assemelham-se a uma função trigonométrica indicam que os valores estão constantemente se afastando e se aproximando em diferentes momentos. A distância maior da linha de referência em pontos específicos sugere uma congruência ainda menos provável entre as distribuições.

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

O coeficiente de correlação de Pearson é uma métrica estatística que avalia a relação linear entre duas variáveis. Variando de -1 a 1, o coeficiente indica a força e direção da associação: 1 representa uma correlação positiva perfeita, -1 uma correlação negativa perfeita e 0 uma ausência de correlação linear. Essa métrica é sensível apenas a relações lineares, não capturando associações não lineares.

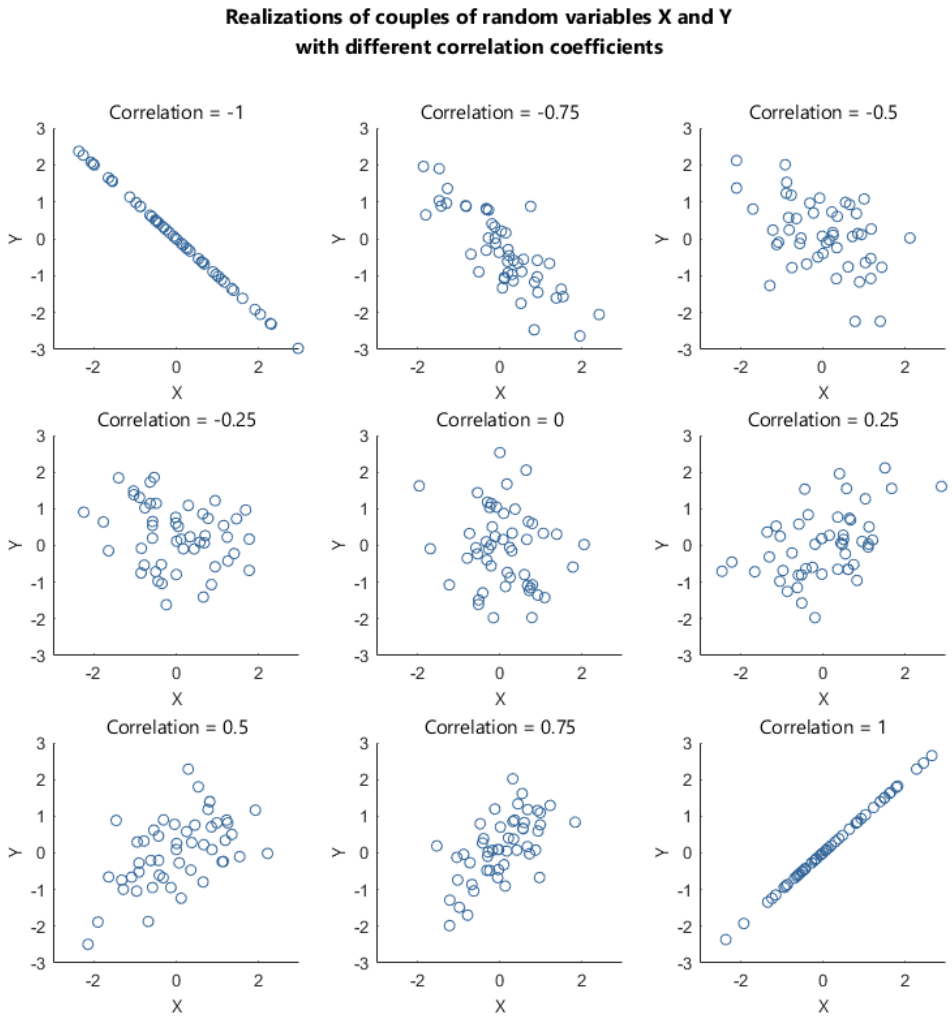

Figura 22: Exemplos de gráficos de variáveis com coeficientes de Pearson diferentes.

Ao examinar Scatter Plots, é possível identificar visualmente a força e natureza da relação entre as variáveis. Contudo, é importante observar que um gráfico pode sugerir uma forte associação sem ser linear. No contexto dos dispositivos Chromecast e Smart TV, foi feita a inclusão de uma linha de regressão linear para identificar correlações lineares dominantes. No entanto, é importante destacar que se a relação entre as variáveis for melhor descrita por uma função não linear, como uma parábola, por exemplo, nuances de padrões não lineares podem ser subestimadas ou até mesmo ignoradas.

Por isso, também foi feita a adição de uma linha de referência de segunda ordem. Embora o coeficiente de correlação linear de Pearson capture a força da relação linear, a linha de ordem 2 mostra que uma relação quadrática também pode ser um ajuste adequado. Essa observação é relevante para interpretar os gráficos dos dispositivos mencionados, indicando que, apesar de uma correlação linear ser evidente, uma dependência mais complexa não pode ser descartada. Essa compreensão mais abrangente ressalta a importância de uma análise visual cuidadosa e da consideração de diferentes modelos de ajuste ao interpretar Scatter Plots.

5.1 Chromecast

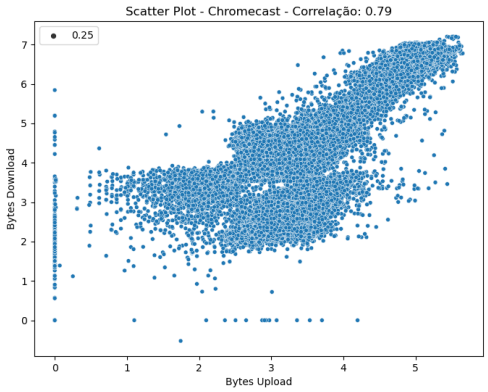

Figura 23: Gráfico Scatter para o Chromecast.

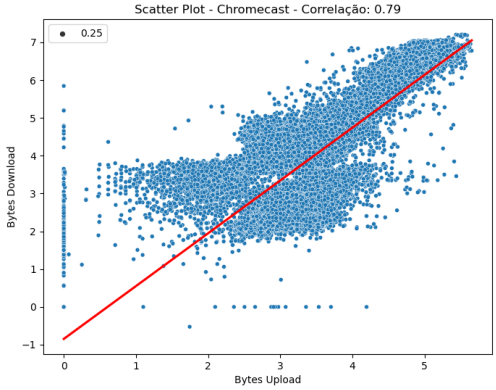

Figura 24: Gráfico Scatter para o Chromecast com linha de regressão linear.

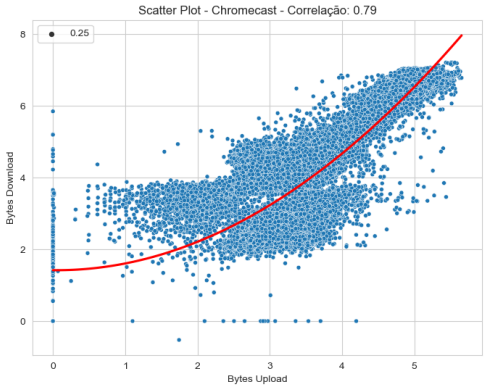

Figura 25: Gráfico Scatter para o Chromecast com linha de regressão de segunda ordem.

Para o Chromecast, o Scatter Plot revela uma relação positiva entre as taxas de download e upload. O coeficiente de correlação de 0.79 sugere fortemente que há uma correlação entre as variáveis. A inclusão de uma linha de regressão linear na segunda imagem destaca essa relação, principalmente na segunda metade, à medida que as taxas ficam mais altas. No entanto, a análise visual também sugere uma possível relação de segunda ordem, principalmente no início.

5.2 Smart TV

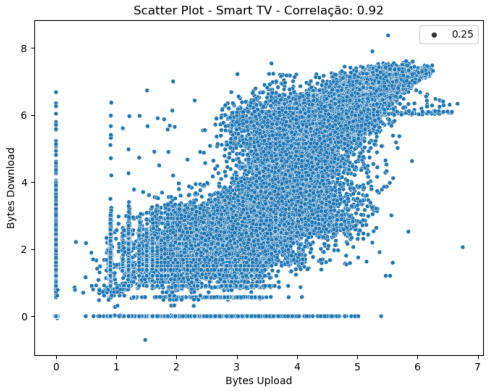

Figura 26: Gráfico Scatter para Smart TV.

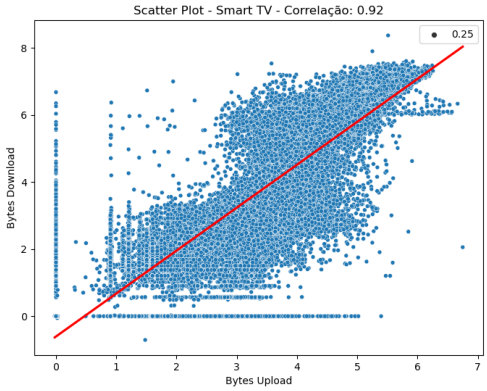

Figura 27: Gráfico Scatter para Smart TV com linha de regressão linear.

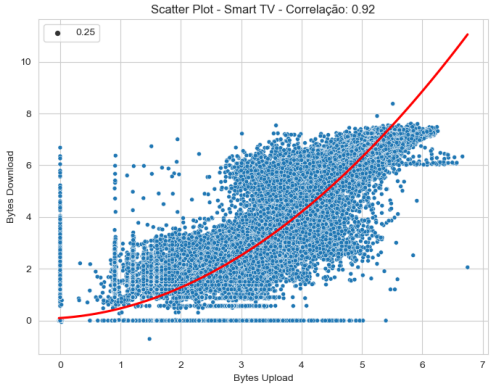

Figura 28: Gráfico Scatter para a Smart TV com linha de regressão de segunda ordem.

No caso da Smart TV, o Scatter Plot mostra uma relação ainda mais forte entre as Taxas de Download e Upload, evidenciada pelo altíssimo coeficiente de correlação de 0.92. A segunda imagem, com a linha de regressão linear, reforça a natureza linear dessa relação. A distribuição dos pontos em relação à linha sugere uma dependência linear robusta entre as variáveis, indicando que, em horários de maior tráfego (e muito provavelmente em todos os outros), as variações em uma taxa estão fortemente associadas às variações na outra.

6 Conclusão

Este projeto proporcionou uma análise detalhada das Taxas de Upload e Download nos dispositivos Chromecast e Smart TV, especialmente em seus horários de maior tráfego. Os diversos métodos estatísticos empregados, como histogramas, boxplots, funções de distribuição empírica, e análise de correlação, permitiram uma análise abrangente e profunda desses conjuntos de dados.

Os resultados obtidos revelam padrões distintos de comportamento nos dispositivos. A análise dos boxplots ofereceram insights sobre a presença de outliers e a variabilidade nos dados ao longo do dia. A aplicação de funções de distribuição empírica proporcionou uma compreensão mais aprofundada da probabilidade associada a diferentes taxas, e foi um dos dados que melhor relacionou ambos os dispositivos entre si, com suas distribuições semelhante.

Os histogramas, ao oferecerem uma representação visual das médias das taxas, desempenham um papel fundamental na capacidade dos provedores de prever e atender às necessidades do público. Juntamente com os Boxplots, essas representações gráficas permitem uma compreensão clara das tendências de consumo de largura de banda, revelando os horários de pico e indicando quando o público demanda maiores velocidades de Download ou Upload. A análise das médias apresentadas possibilita aos provedores antecipar padrões de tráfego, otimizando a alocação de recursos para garantir um desempenho consistente e satisfatório durante os

períodos de maior demanda. Essa abordagem proativa, baseada nas informações fornecidas pelos histogramas, não apenas melhora a eficiência operacional, mas também contribui para uma experiência de usuário mais fluida e alinhada às expectativas, fortalecendo a satisfação do cliente.

No contexto específico dos horários de pico, a análise de correlação evidenciou associações lineares entre as Taxas de Upload e Download para ambos os dispositivos, mas em especial para a Smart TV. Essa observação também é crucial para provedores de serviços de internet pois, sabendo que a uniformidade nas estatísticas do Chromecast sugere que as variações nas taxas podem ser independentes dos padrões diários, enquanto que a variabilidade na Smart TV sugere uma possível correlação mais forte com o comportamento dos usuários, temos que, para a Smart TV, as variações nas Taxas de Upload podem ser preditivas das variações nas Taxas de Download, e vice-versa.

Em última análise, os resultados deste estudo não apenas aprofundam a compreensão dos padrões de uso desses dispositivos, mas também oferecem *insights* valiosos para provedores de serviço de internet. Ao incorporar essas descobertas em suas estratégias de gerenciamento de rede, os provedores podem aprimorar a eficiência operacional e garantir uma experiência de streaming cada vez mais ótima para seu público.