Projeto do Curso
COE241 - Semestre 2023/2

Gabriel Gazola Milan

28 de dezembro de 2023

Introdução

O objetivo desse trabalho é analisar um conjunto de dados fornecidos por um provedor de Internet de médio porte. Os dados representam a taxa de dados enviados (taxa de upload) e taxa de dados recebidos (taxa de download) em bps de/por um dispositivo. Existem dois tipos de dispositivos: Smart TVs e Chromecasts.

A organização desse relatório segue a estrutura de seções fornecida no enunciado do projeto do curso.

O código utilizado para geração de imagens e valores aqui demonstrados, escrito em Python, pode ser encontrado em `https://github.com/gabriel-milan/coe241`.

Seção 1 - Dataset (Análise exploratória e tratamento inicial dos dados)

Foram fornecidos dois arquivos de dados para o desenvolvimento desse trabalho: um correspondente aos dispositivos do tipo Smart TV e outro aos dispositivos do tipo Chromecast. Após abrir os arquivos e realizar uma breve exploração (observar uma amostra, tipagem das colunas, quartis, valores mínimos, máximos e médios) notou-se a necessidade de converter a coluna `date_hour` para o `datetime` do Python, que iria facilitar operações nessa coluna no futuro.

Também, como foi sugerido reescalonar os dados para log_{10}, houve uma preocupação com os valores de taxa de upload/download zero. Caso fosse considerado $log_{10}(0) = 0$, um valor $x \in (0, 1)$ teria $log_{10}(x) < 0$ e, dessa

forma, a análise deixaria de fazer sentido nessa escala. Por isso, e ainda evitando ter valores infinitamente negativos após reescalonar, os valores zero foram substituídos por valores arbitrários. Os valores arbitrários selecionados foram 10% do menor valor de cada taxa, sendo todos maiores que zero e menores que 0.03.

Por fim, foi aplicado log_{10} em todas as taxas, conforme sugerido no enunciado.

Seção 2 - Estatísticas gerais

De agora em diante, houve a tentativa de manter um padrão de cores para as imagens envolvidas nesse trabalho. O código de cores, então, é o seguinte:

- Azul: Taxa de upload do Chromecast
- Laranja: Taxa de download do Chromecast
- Verde: Taxa de upload da Smart TV
- Vermelho: Taxa de download da Smart TV

Primeiramente, os histogramas podem ser observados na Figura 1. A primeira coisa notada é que, nas Smart TVs, existem várias ocorrências de tráfego ocioso (barras mais à esquerda), ao passo que nos Chromecasts a ociosidade praticamente não ocorre.

Seguindo para a Função Distribuição Empírica (FDE) demonstrada na Figura 2, é possível notar que, apesar dos períodos ociosos, as taxas de download parecem ter dois picos de ocorrência, independentemente do tipo de dispositivo. Enquanto isso, a taxa de upload tende a um valor central de maior ocorrência, principalmente nos Chromecasts.

Não existe informação nos dados que demonstre se essas métricas são coletadas exclusivamente enquanto os dispositivos estão ligados ou não. Para o caso de serem coletadas somente durante o funcionamento dos dispositivos (doravante referida por "Premissa 1"), essa ociosidade nas Smart TVs poderia ser devida ao tempo de tela da TV digital já que, para assistí-la, não é necessária uma conexão com a Internet. Por outro lado, os Chromecasts dependem exclusivamente de uma conexão com a Internet para operar, explicando a falta de ociosidade em seus histogramas.

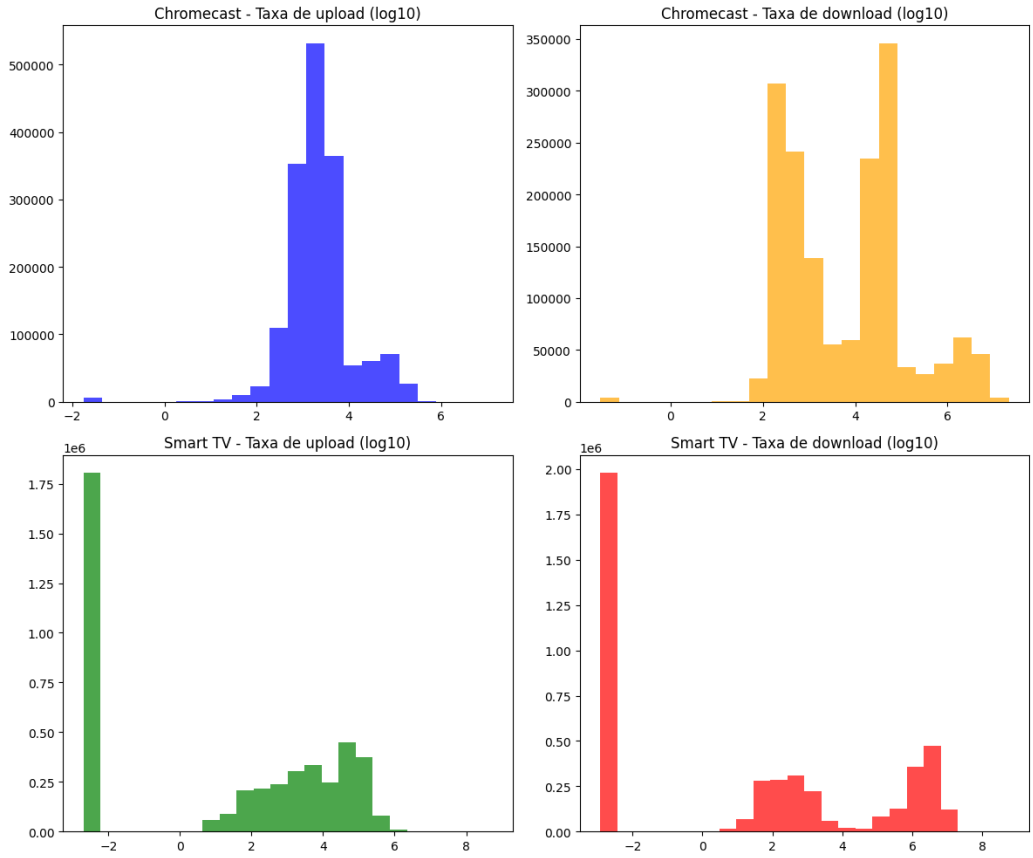

Figura 1: Histogramas

Observando os Box Plots, na Figura 3, notam-se valores medianos maiores para o Chromecast, além de claramente uma distância inter-quartis menor, principalmente na taxa de upload devido à alta concentração de amostras, classificando diversas amostras como outliers. Como nessa etapa não houve nenhum tipo de filtragem (horário, dia da semana, etc.) possivelmente não seria correto afirmar que esses pontos são de fato outliers.

Por fim, os valores de média, variância e desvio padrão para cada taxa são os seguintes:

- Chromecast Upload
  - Média: 3.34308631091325
  - Variância: 0.517277337886133

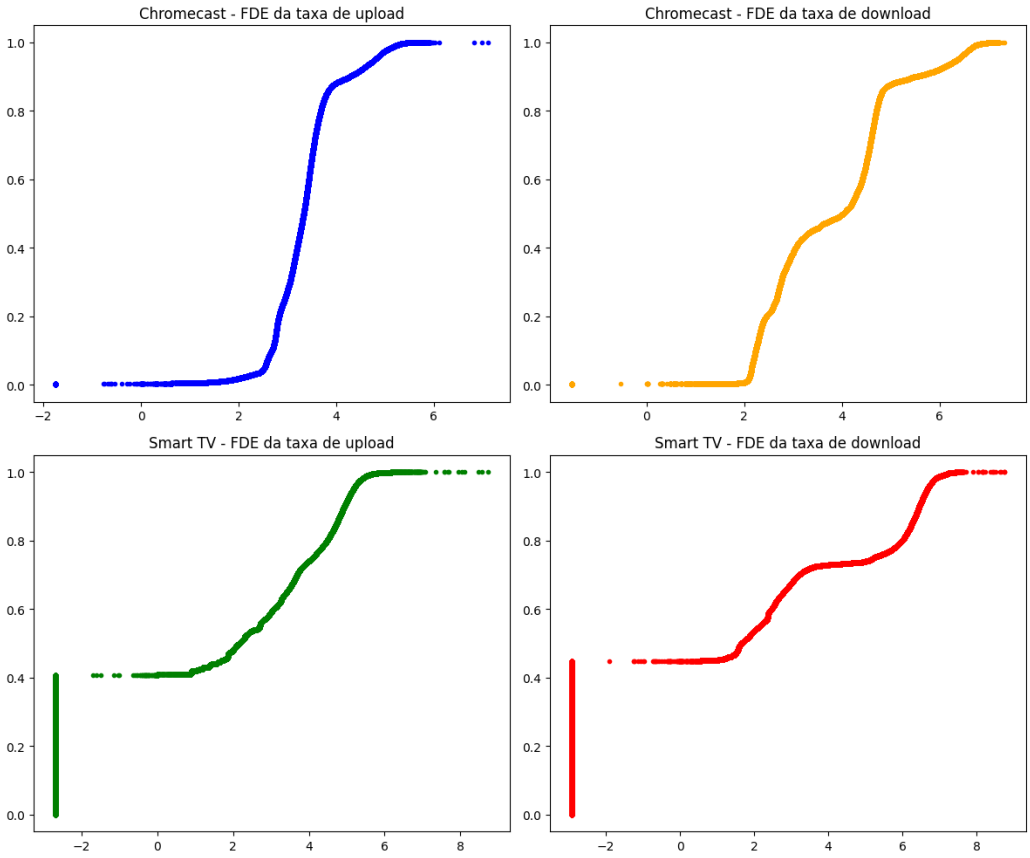

Figura 2: FDEs para todas as taxas

  - Desvio padrão: 0.719219950978929

- Chromecast Download
  - Média: 3.7954787766229656
  - Variância: 1.7011502790548356
  - Desvio padrão: 1.3042815183290897
- Smart TV Upload
  - Média: 1.0545533293386347
  - Variância: 10.62633460224527

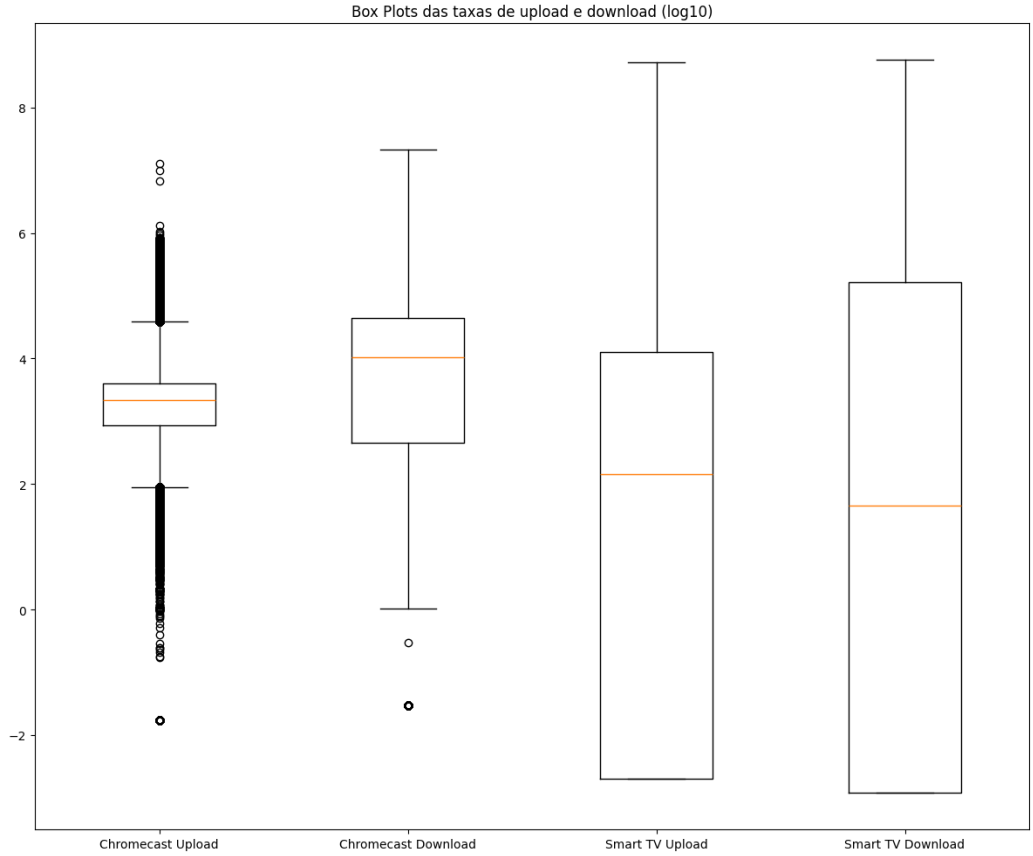

Figura 3: Box plots para todas as taxas

  - Desvio padrão: 3.2598059148123024

- Smart TV Download
  - Média: 1.0415747226804573
  - Variância: 14.986444953944806
  - Desvio padrão: 3.871233001763754

Como é possível notar, a variabilidade dos dados de Smart TVs é muito maior que nos Chromecasts. Isso, no entanto, é altamente afetado pelos períodos de ociosidade mencionados anteriormente. Pelo mesmo motivo, as médias, que são mais baixas, são afetadas.

Seção 3 - Estatísticas por horário

A seguir, os dados foram agrupados por hora do dia. Nas Figuras 4 e 5 é possível observar os valores de média e desvio padrão para as taxas a cada hora do dia. A opção de exibir somente essas duas estatísticas pretende auxiliar na visualização, colocando então o desvio padrão como valor de erro no eixo Y.

Também, nas Figuras 6 e 7 é possível observar os Box Plots para os mesmos dados.

Observando conjuntamente todas as figuras, é possível notar que os dispositivos Chromecast possuem baixa variação em seus valores médios e medianos a depender do horário do dia. No entanto, a taxa de upload desses dispositivos possui alta concentração dos valores próximos à mediana, assim diminuindo muito a distância inter-quartis, como é possível visualizar na Figura 6.

Também é possível observar, nas linhas de média para as taxas, um comportamento muito similar em todas as taxas. Claro que existe uma atenuação dos efeitos de variação das taxas nos dispositivos Chromecast, como sugerido pelos próprios Box Plots da Figura 6, mas o aumento quase que constante a partir da manhã até o final do dia é bem característico. Nota-se também que esse comportamento está levemente "empurrado"à direita nos Chromecasts, sugerindo um início (e fim) mais tardio do uso desses dispositivos.

Algo que chama atenção na Figura 7 é a mediana tender ao valor arbitrário definido para o "zero"nos períodos de madrugada. Isso, tomando como verdadeira a Premissa 1, pode caracterizar o uso de Smart TVs durante a madrugada somente para TV digital.

Seção 4 - Caracterizando os horários com maior valor de tráfego

Com base nos gráficos expostos na seção anterior, é possível concluir os seguintes horários como pico:

- Chromecast - Horário de pico para upload: 22h
- Chromecast - Horário de pico para download: 23h
- Smart TV - Horário de pico para upload: 20h
- Smart TV - Horário de pico para download: 20h

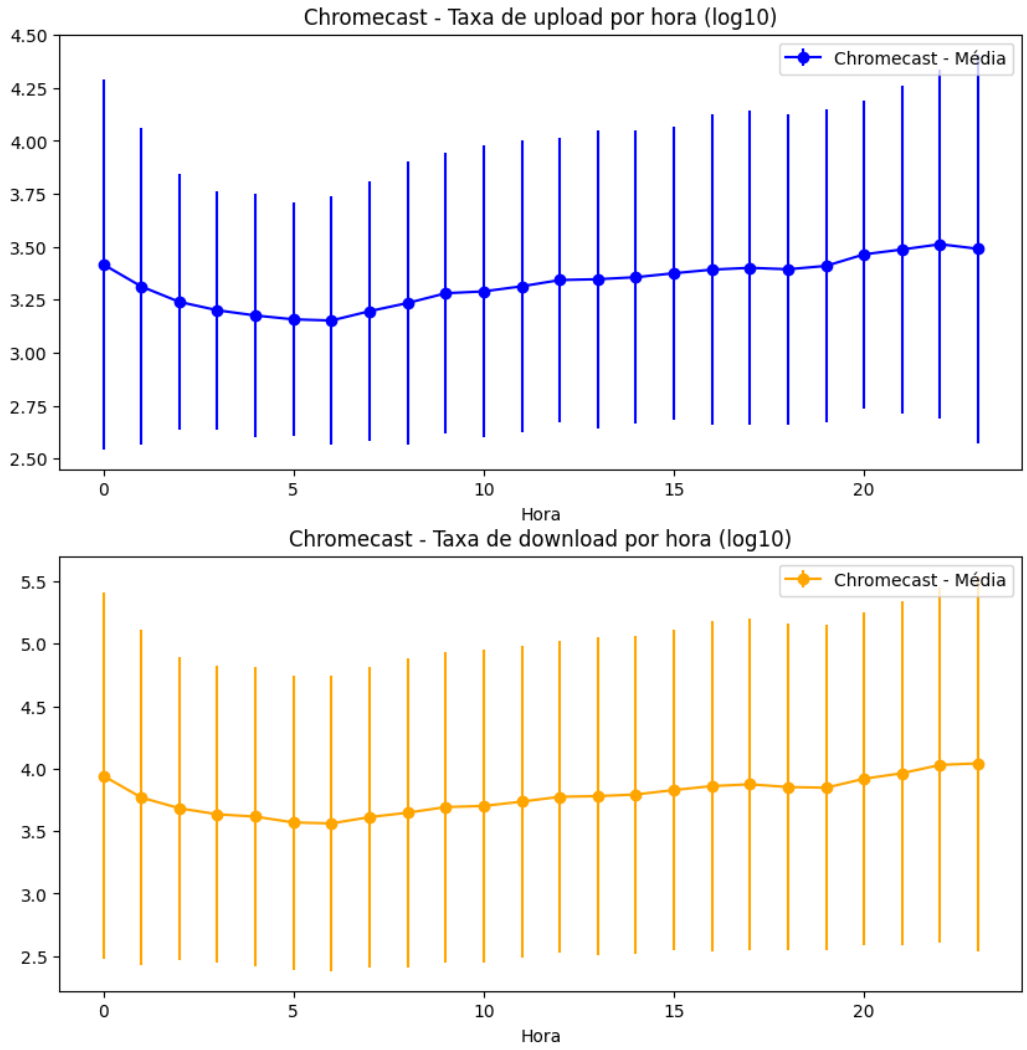

Figura 4: Média e desvio padrão para as taxas de Chromecasts a cada hora do dia

Na Figura 8 é possível observar os histogramas de cada taxa em seu respectivo horário de pico. Nota-se que, ainda no horário de pico, as Smart TVs possuem uma quantidade considerável de ociosidade, possivelmente sugerindo que exista um público de constante uso de TV digital. Uma análise levando em consideração os identificadores de cada dispositivo poderia trazer informações relevantes a respeito disso.

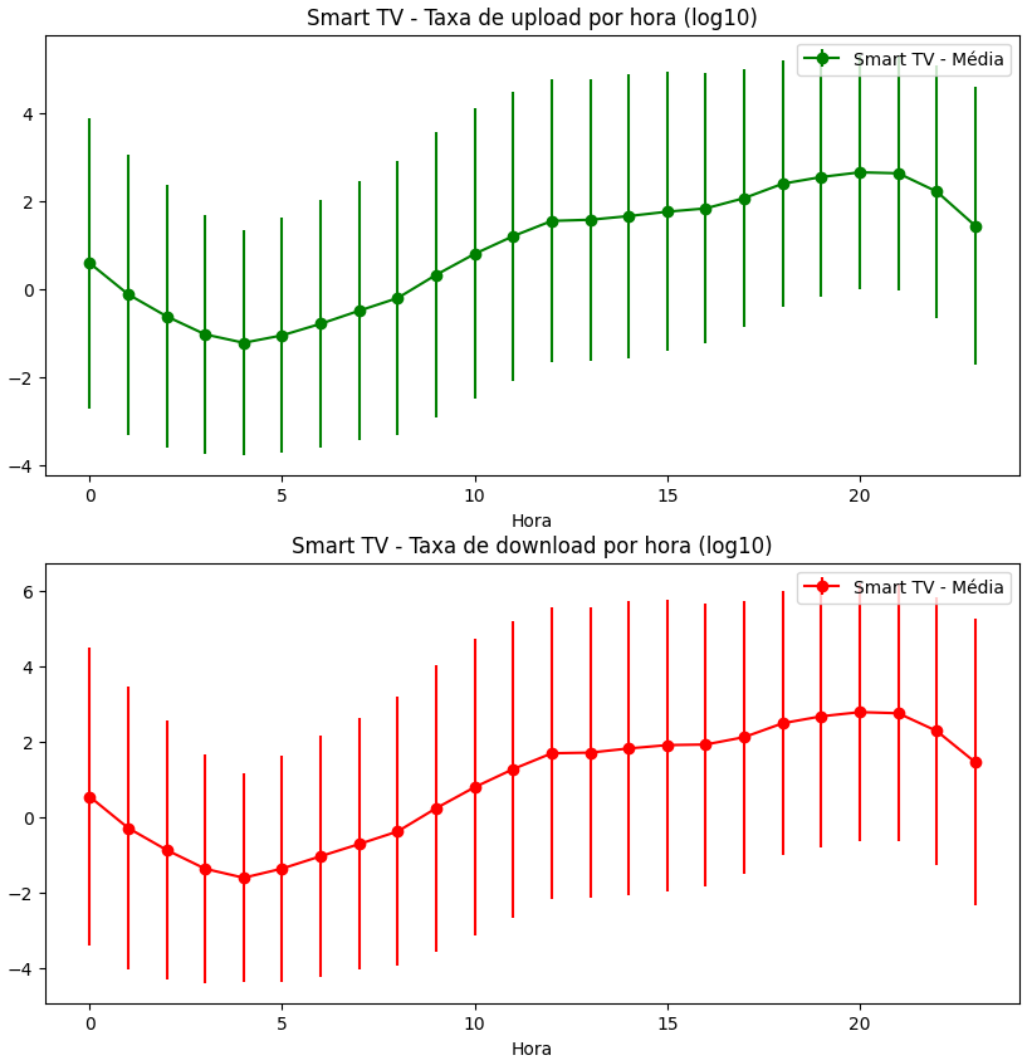

Figura 5: Média e desvio padrão para as taxas de Smart TVs a cada hora do dia

Apesar dessa mudança na ociosidade das Smart TVs, o perfil dos histogramas se assimila aos demonstrados anteriormente na Figura 1, sugerindo que talvez seja uma representação razoável do horário de maior tráfego na rede.

Por fim, é realizada uma comparação das distribuições das taxas de upload e download entre os dispositivos através de QQ Plots. Esses podem ser

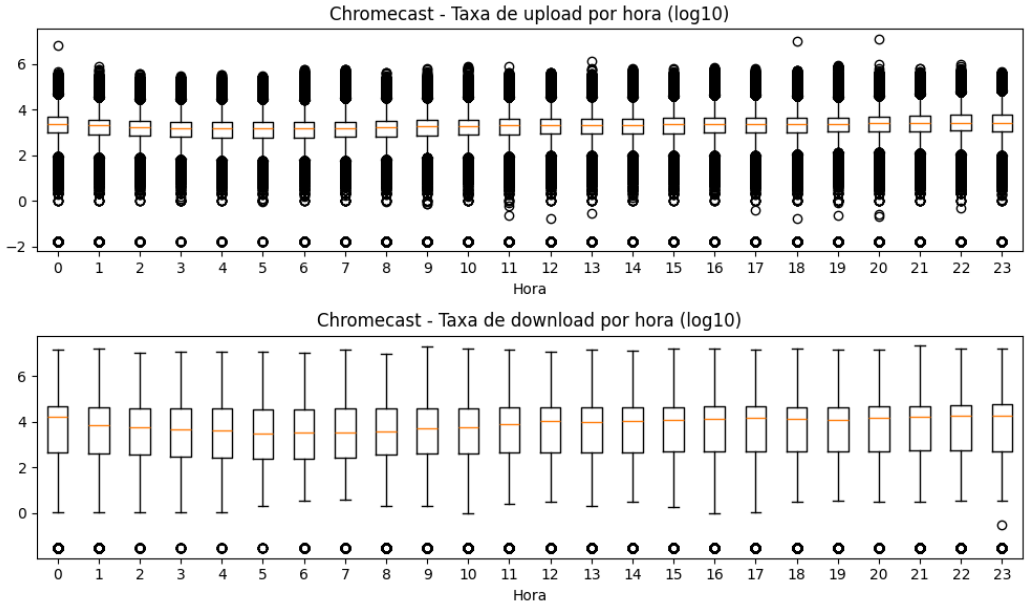

Figura 6: Box Plots para as taxas de Chromecasts a cada hora do dia

observados nas Figuras 9 e 10.

Os QQ Plots denotam uma similaridade entre as distribuições das taxas em seus respectivos horários de pico. A proximidade com a linha vermelha é a característica que mostra isso. Vale salientar também que os períodos de ociosidade presentes nas Smart TVs contribuem para o deslocamento à direita da curva. Uma eventual caracterização do público ocioso e filtragem dele para análise poderia talvez aproximar as distribuições, por, quanto ao uso da Internet, se tratarem de dispositivos com propósitos muito similares.

Pensando pela perspectiva do provedor de Internet, possivelmente caracterizar os tipos de clientes seria mais adequado para entender as demandas. O fato de haver uma mistura de Smart TVs extremamente ociosas e não-ociosas pode não colaborar tanto.

Seção 5 - Análise da correlação entre as taxas para os horários de pico

As Figuras 11 e 12 exibem o scatter plot da taxa de upload vs. a taxa de download para cada um dos tipos de dispositivo.

Uma breve observação das figuras sugere uma correlação positiva às

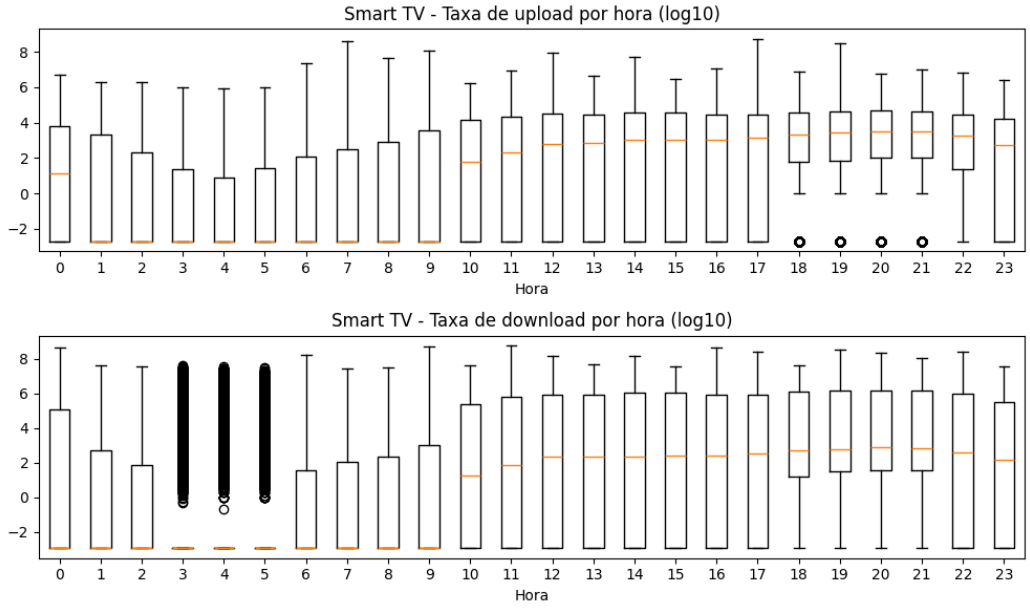

Figura 7: Box Plots para as taxas de Smart TVs a cada hora do dia

variáveis, dada a orientação diagonal e positiva. Isso implicaria que o aumento de uma taxa poderia estar correlacionado ao aumento da outra. Ao calcular os valores dos coeficientes de correlação entre as taxas de upload e download, são obtidos os seguintes resultados:

- Chromecast: 0.7791542846112808
- Smart TV: 0.9050720519252451

Esses valores indicam uma forte correlação positiva, que é natural no uso da Internet.

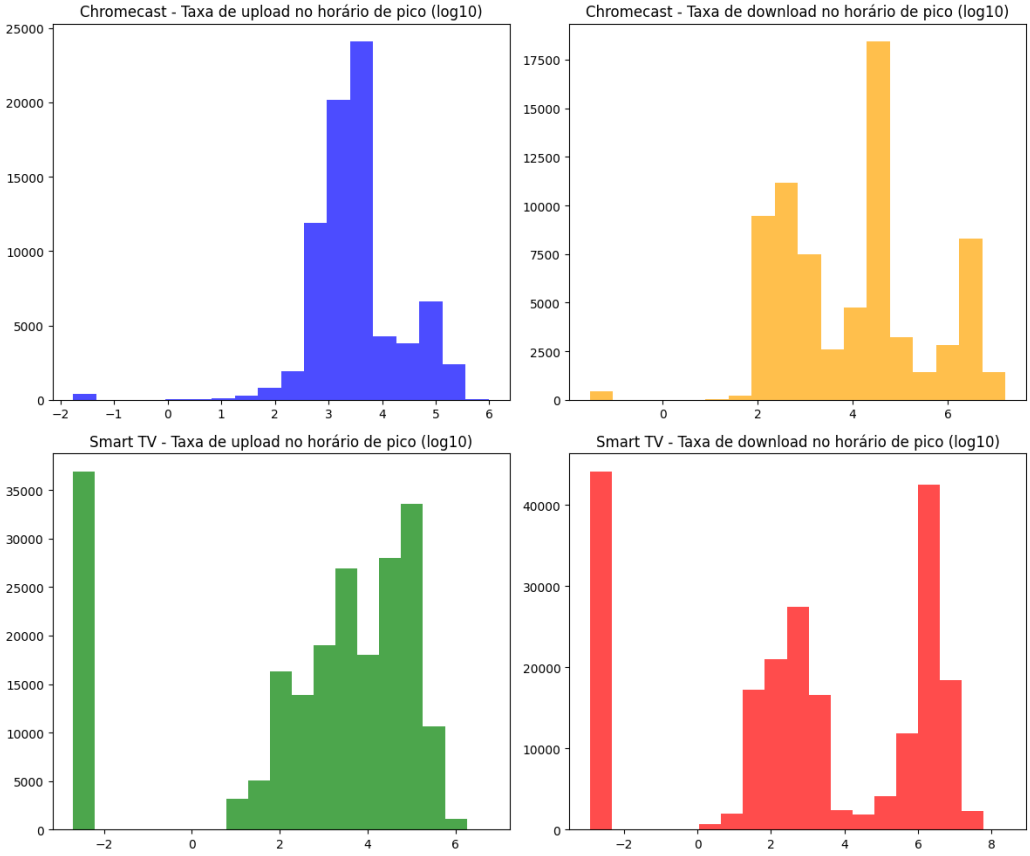

Figura 8: Histogramas para as taxas em seus respectivos horários de pico

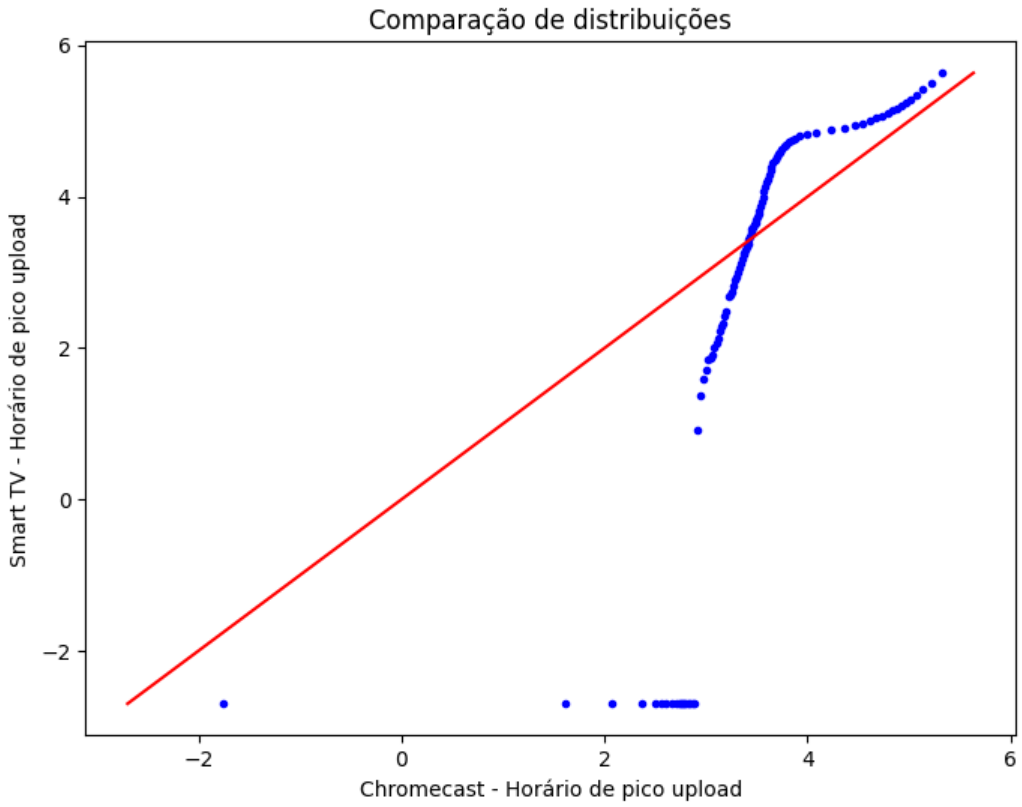

Figura 9: Comparação das distribuições das taxas de upload entre tipos de dispositivos

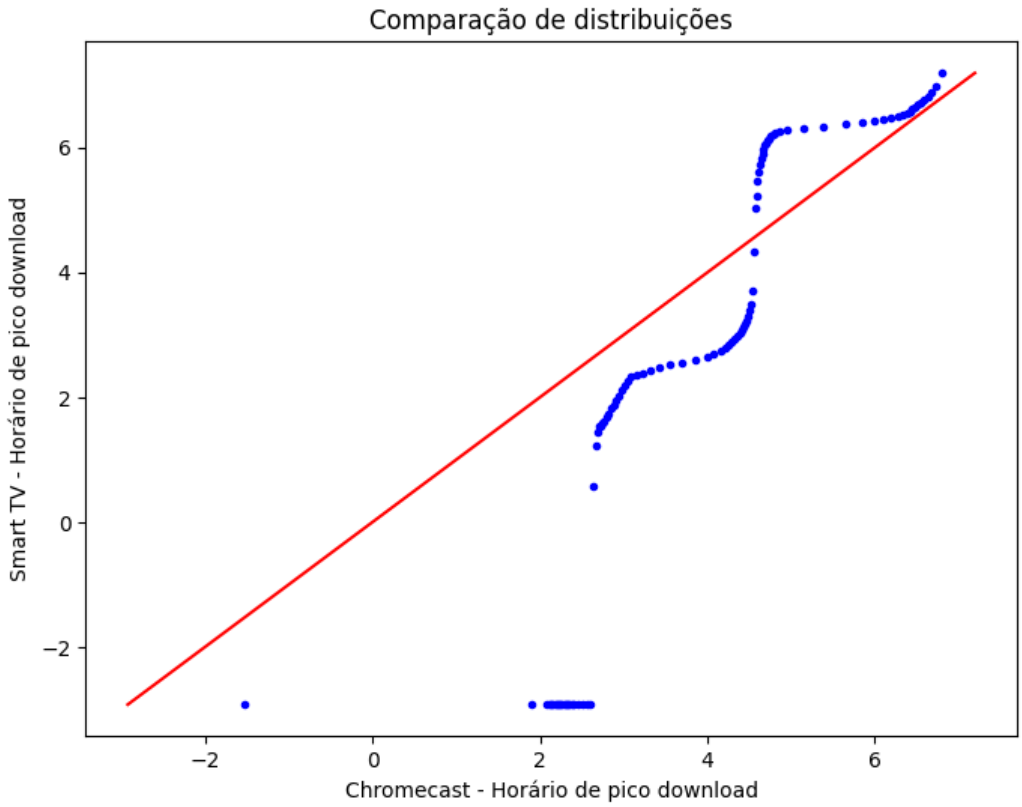

Figura 10: Comparação das distribuições das taxas de download entre tipos de dispositivos

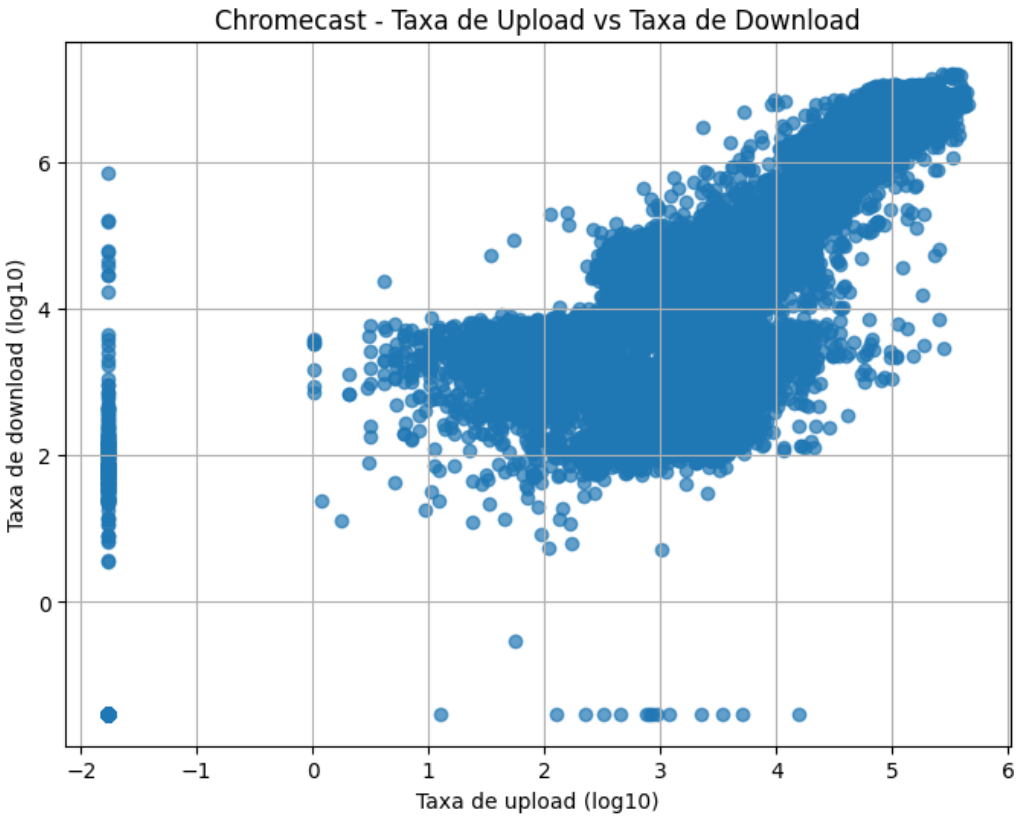

Figura 11: Chromecast - Taxa de upload vs. Taxa de download

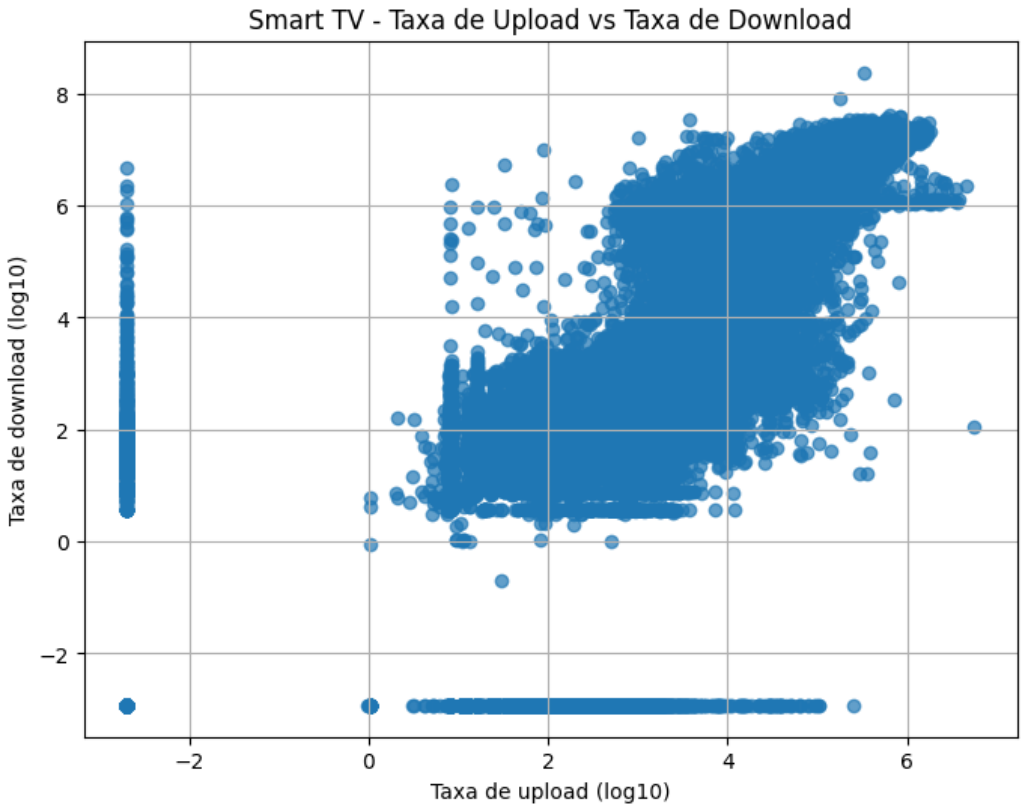

Figura 12: Smart TV - Taxa de upload vs. Taxa de download

Conclusão e observações finais

O trabalho permitiu a observação e caracterização de dois tipos de dispositivos perante o uso da Internet. No entanto, a validação da Premissa 1 pareceu importante para a continuidade do estudo. Também, caso essa seja confirmada, a separação de Smart TVs em mais classes de dispositivos (a sugestão inicial seria separar as que possuem maior ociosidade das outras) poderia auxiliar o provedor a compreender melhor suas demandas. Uma breve análise levando em consideração os IDs dos dispositivos, individualmente, poderia possivelmente auxiliar a separação dessas classes.

De qualquer forma, foi interessante observar que, desconsiderando os períodos de ociosidade, os dispositivos possuem perfis de uso de rede similares e, principalmente, que os usuários de Chromecast aqui explicitados tendem a usá-lo em períodos mais tardes que os usuários de Smart TVs.