Estatística e Modelos Probabilísticos

Trabalho Final da Disciplina
2023.2

Luiz Guilherme de A. Pires
121070338

Relatório referente ao
trabalho final da matéria de Estatística e Modelos Probabilísticos

Contents

1 Introdução **3**

2 Estatísticas Gerais **3**
2.1 Pré-Tratamento . . . 3
2.2 Histograma . . . 3
2.2.1 Chromecast . . . 4
2.2.2 Smart TV . . . 5
2.3 Função Distribuição Empírica . . . 6
2.3.1 Chromecast . . . 7
2.3.2 Smart TV . . . 9
2.4 Blox Plot . . . 10
2.4.1 Chromecast . . . 11
2.4.2 Smart TV . . . 12
2.5 Média, Variância e Desvio Padrão . . . 12
2.5.1 Chromecast . . . 12
2.5.2 Smart TV . . . 13
2.6 Análise dos Resultados . . . 13

3 Estatísticas por horário **13**
3.1 Blox Plot . . . 14
3.1.1 Chromecast . . . 14
3.1.2 Smart TV . . . 26
3.2 Média, Variância e Desvio Padrão . . . 38
3.2.1 Chromecast . . . 39
3.2.2 Smart TV . . . 40
3.3 Análise dos Resultados . . . 41

4 Caracterizando os horários com maior valor de tráfego **41**
4.1 Horários . . . 42
4.1.1 Chromecast . . . 42
4.1.2 Smart TV . . . 42
4.2 Histograma . . . 42
4.2.1 Chromecast . . . 42
4.2.2 Smart TV . . . 43
4.3 Q-Q Plot . . . 44
4.3.1 Chromecast . . . 45
4.3.2 Smart TV . . . 46
4.4 Análise dos Resultados . . . 47

5 Análise da correlação entre as taxas de upload e download para os horários com o maior valor de tráfego **47**
5.1 Coeficientes de correlação . . . 47
5.2 Scatter Plot . . . 48
5.2.1 Chromecast . . . 48

5.2.2 Smart TV 49
5.3 Análise dos Resultados 49

1 Introdução

A partir de um conjunto de dados reais, foi aplicado teorias aprendidas durante a disciplina COE241 - Estatística e Modelos Probabilísticos para gerar um conjunto de análises, assim como uma análise crítica dos resultados obtidos. Segue abaixo o link para o código feito em Python, utilizando o ambiente Jupyter, que produziu os dados mostrados durante o trabalho:

- https://github.com/ziuLGAP/Probest.2023.2

2 Estatísticas Gerais

2.1 Pré-Tratamento

Como foi dito no texto do projeto, a fim de obter uma escala de grandeza mais fácil de ser visualizada, era necessário reescalonar os dados para Log_{10}, tendo em vista que o número de bytes em upload e download podem chegar na ordem de 7 ou mais casas decimais, e para isso foi utilizado a função log10 da biblioteca numpy a fim de criar novas colunas nos dois datasets. Vale ressaltar que, como não é possível realizar o log para valores iguais a 0, foi somado 1 à estes valores, ficando assim $log(1)$, que é igual a zero, desta forma evitando a perda destes valores.

Outro tratamento que foi realizado, foi a criação de uma coluna contendo apenas a hora em que um dado foi obtido.

2.2 Histograma

Antes de criar os histogramas, foi utilizado uma forma de definir a quantidade ideal de colunas utlizando o método de Sturges:

$$k = 1 + log_2(N)$$

que no código ficou como a função *get_bin*, além disso foi utilizado o método *hist* da biblioteca plt do Python.

Tendo vista isso, segue abaixo os histogramas gerados para ambos os datasets.

2.2.1 Chromecast

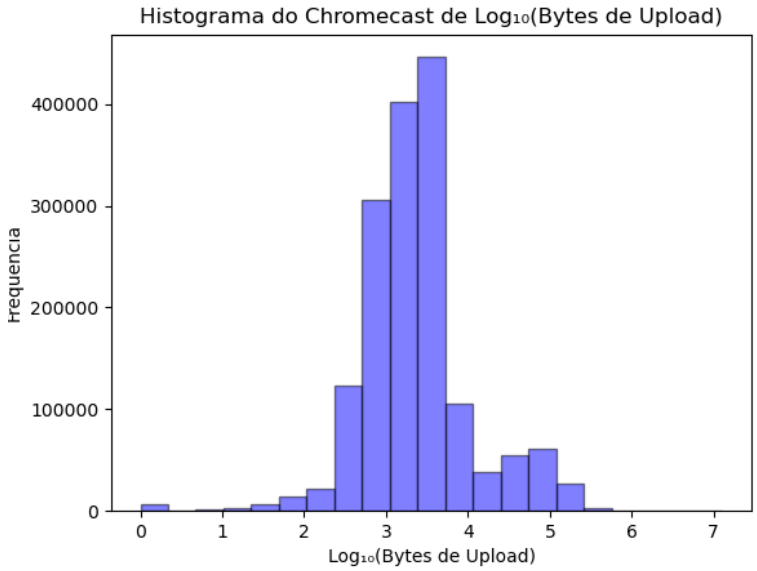

Figure 1: Histograma de Log_{10}(taxa de upload) para o Chromecast

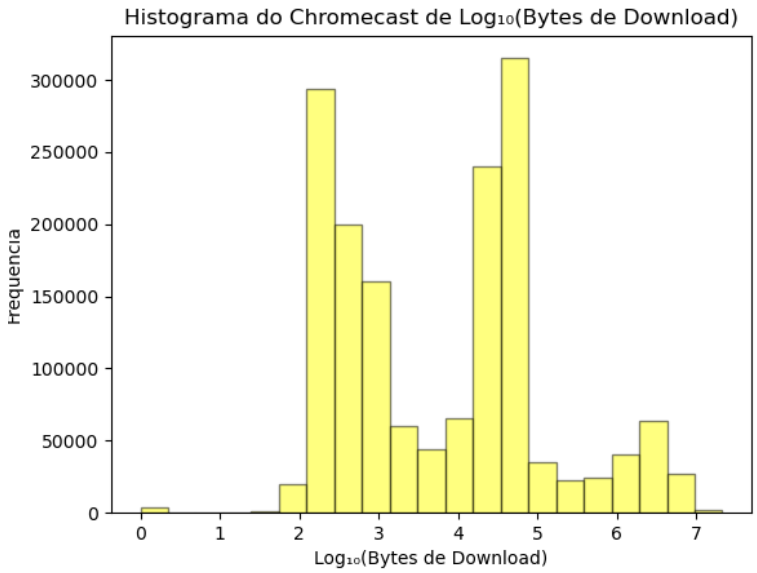

Figure 2: Histograma de Log_{10}(taxa de download) para o Chromecast

2.2.2 Smart TV

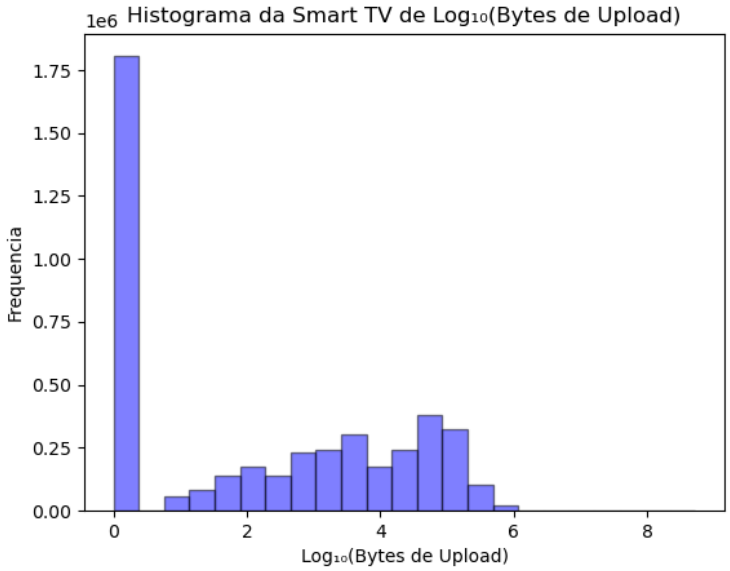

Figure 3: Histograma de Log_{10}(taxa de upload) para a Smar-TV

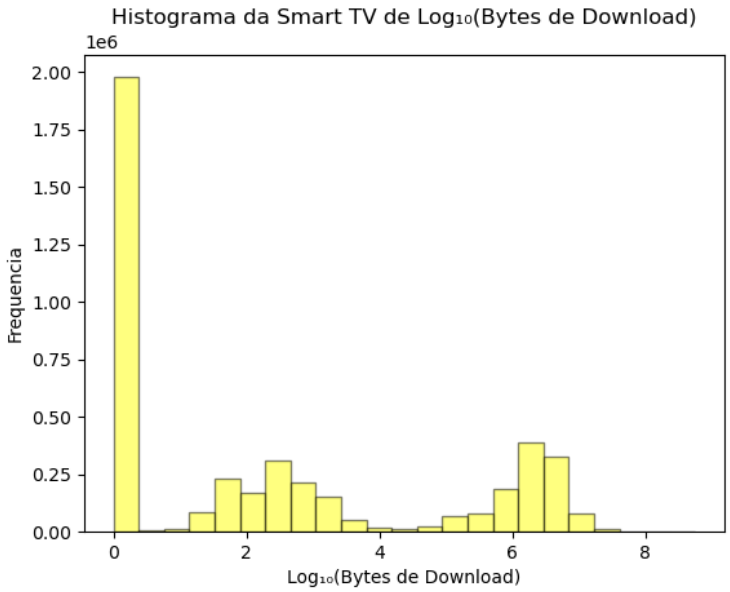

Figure 4: Histograma de Log_{10}(taxa de download) para a Smart-TV

2.3 Função Distribuição Empírica

Assim como no histograma, foi utilizada a biblioteca plt, mais especificamente a função plot para plotar as função distribuição empírica, onde para o eixo X foram utilizados os valores da coluna de interesse, já para o eixo Y, foi gerado um array de valores entre 0 e 1 com o número de dados sendo o número de dados do dataframe, por meio do método *linspace*, ficando assim com os seguintes gráficos.

2.3.1 Chromecast

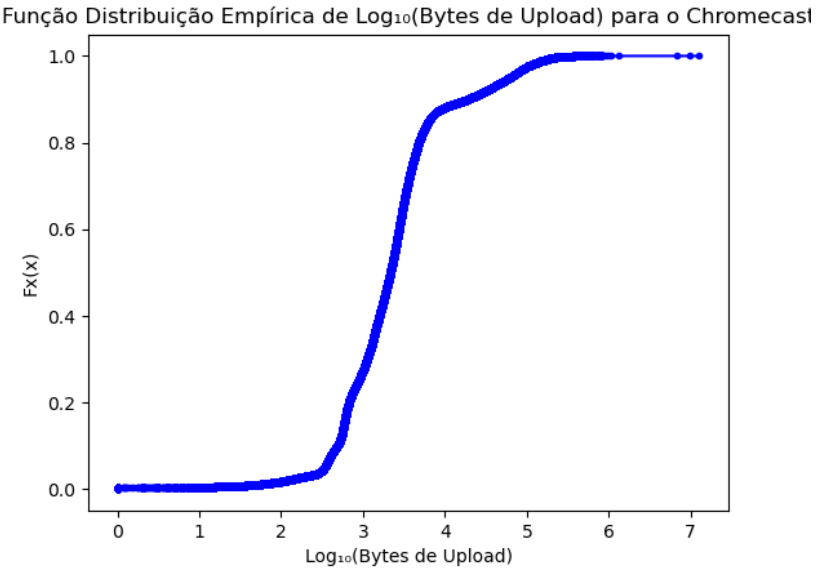

Figure 5: Função Distribuição Empírica de Log_{10}(taxa de upload) para o Chromecast

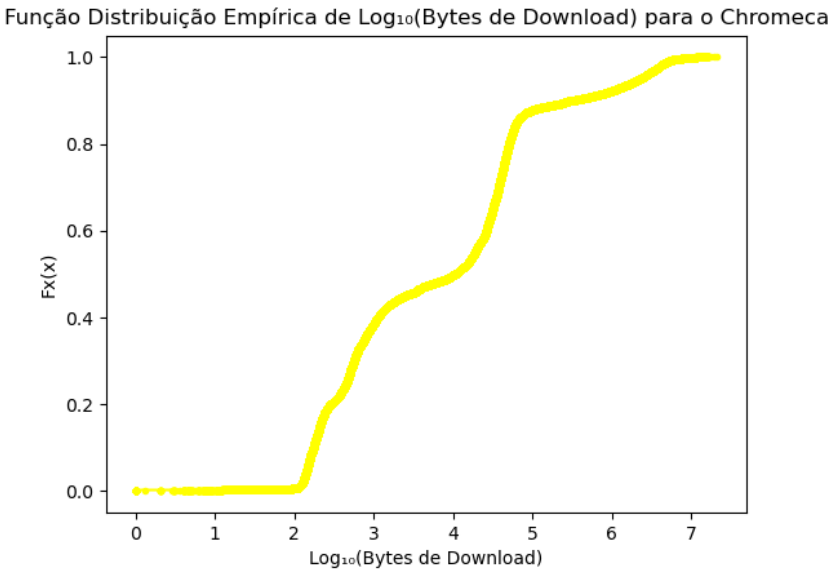

Figure 6: Função Distribuição Empírica de Log_{10}(taxa de download) para o Chromecast

2.3.2 Smart TV

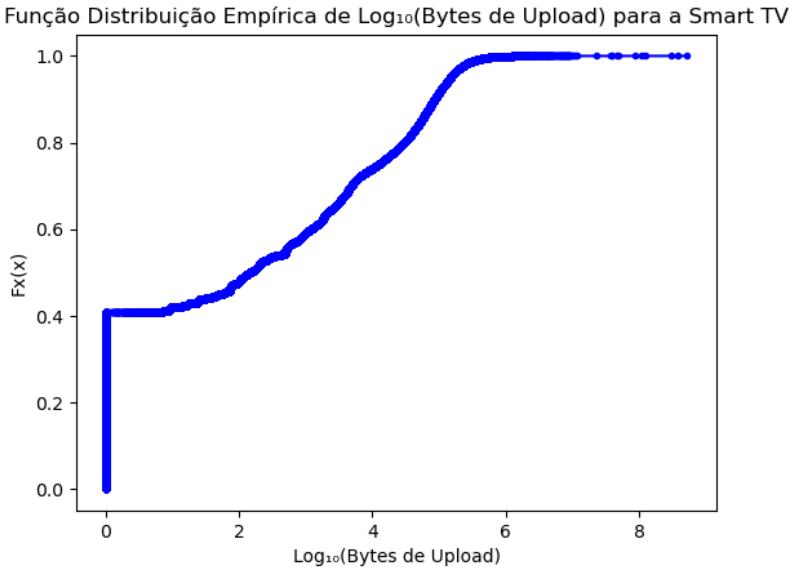

Figure 7: Função Distribuição Empírica de Log10(taxa de upload) para a Smart TV

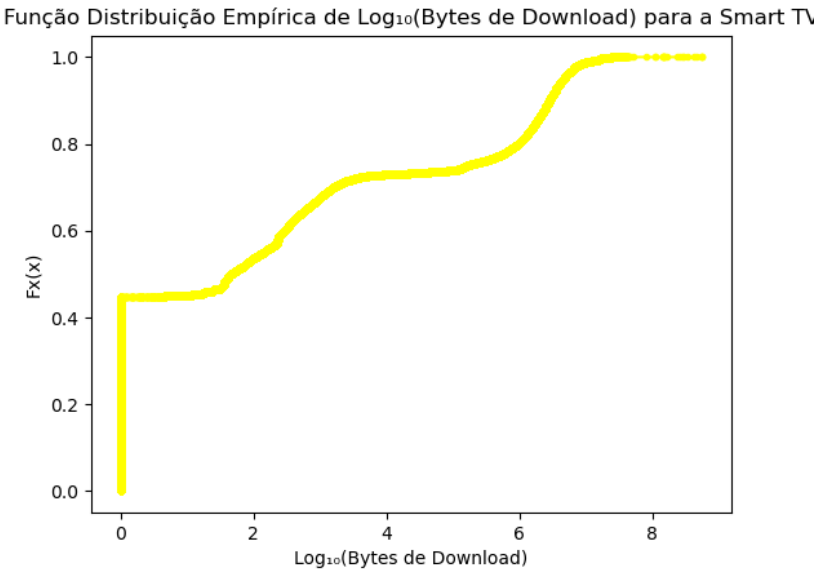

Figure 8: Função Distribuição Empírica de Log10(taxa de download) para a Smart TV

2.4 Blox Plot

Para a geração destes boxplots, foi utilizado o método *boxplot* da biblioteca plt.

2.4.1 Chromecast

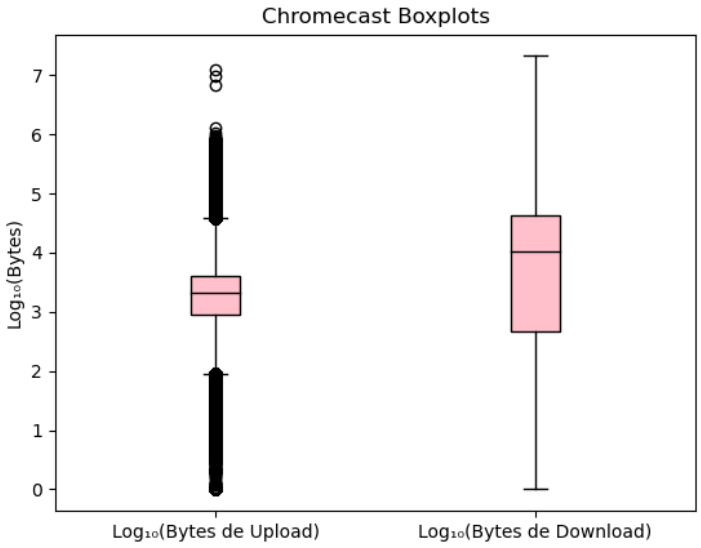

Figure 9: Boxplot de Log10(taxa de upload e download) para o Chromecast

2.4.2 Smart TV

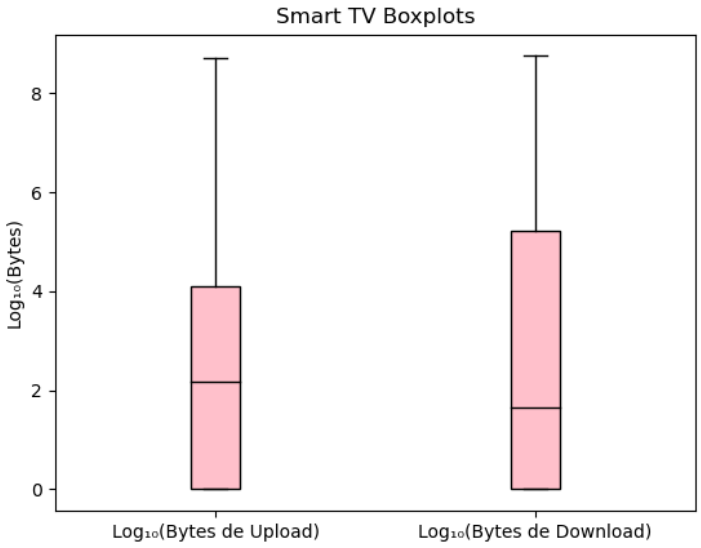

Figure 10: Boxplot de Log10(taxa de upload e download) para a Smart TV

2.5 Média, Variância e Desvio Padrão

Foram utilizados métodos *mean*, *var* e *std* da biblioteca pandas para realizar esta análise estatística.

2.5.1 Chromecast

Estatísticas	Log_{10}(Bytes de Upload)	Log_{10}(Bytes de Download)
Média	3.3503	3.80004
Variância	0.459969	1.6639
Desvio Padrão	0.67821	1.28992

Table 1: Estatísticas do Chromecast

2.5.2 Smart TV

Estatísticas	Log_{10}(Bytes de Upload)	Log_{10}(Bytes de Download)
Média	2.15829	2.35168
Variância	4.11014	6.72132
Desvio Padrão	2.02735	2.5925

Table 2: Estatísticas da Smart TV

2.6 Análise dos Resultados

Como podemos ver, ao comparar os histogramas de download e upload entre dispositivos, observamos muitos dados com frequência zero de upload e download em smart TVs, enquanto esses momentos são menos comuns no Chromecast. Isso significa que você pode esperar que o dispositivo se comporte de maneira diferente ao fazer download ou upload de dados.

Se observarmos o histograma de velocidades de upload e download de cada dispositivo, poderá ver que o Chromecast se comporta de maneira diferente nas faixas de download e upload, pois os picos máximos de upload estão na faixa de 1000 bytes por segundo. Esse fato pode ser devido à capacidade de internet do aparelho. As faixas de download, por outro lado, têm um pico um pouco mais alto. Comportamento semelhante também é observado na smart TV.

Adicionalmente, também pode ser observada uma clara complementaridade entre as distribuições de upload e download no que diz respeito ao histograma do dispositivo, o que pode fazer com que o dispositivo se comporte de tal forma que as duas ações não atinjam o pico, o que levanta uma suspeita. Isso é fácil de observar tanto no Chromecast quanto em smart TVs entre os valores $Log_{10}(2)$ e $Log_{10}(6)$, ignorando o valor da taxa de download de 0.

Agora comparando os box plots de upload e download entre os dispositivos, pode-se perceber que a dispersão entre eles em termos de velocidade de download é muito diferente, a concentração de dados também nos logs é diferente devido à altura do terceiro quartil e do primeiro quartil entre, os dois quadrados, além do quadrado da caixa do Chromecast, têm a presença de outliers muito claros. Para o box plot de download, ambos têm a área do terceiro quartil próxima, mas a área do primeiro quartil e do segundo quartil estão bem espaçadas.

Quanto ao comportamento da função de distribuição empírica entre dispositivos, é muito variável, apresentando um aumento gradual no Chromecast, ao contrário das smart TVs que têm um aumento muito claro em devido ao grande número de vezes em que o dispositivo não realizou download nem upload.

3 Estatísticas por horário

Para essa parte do trabalho, foi realizado um loop, onde a cada iteração era gerado dois box plots referentes à download e upload ocorridos naquele determinado horário, ficando com os gráficos mostrados a seguir.

3.1 Blox Plot

3.1.1 Chromecast

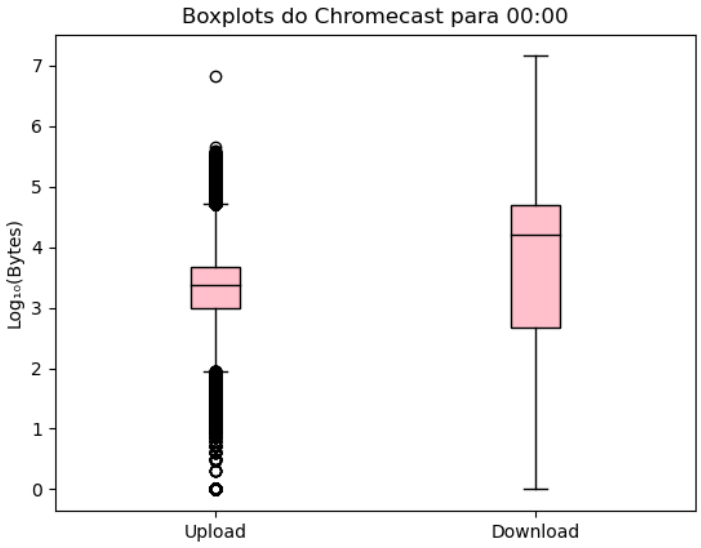

Figure 11: Chromecast na Hora 00:00

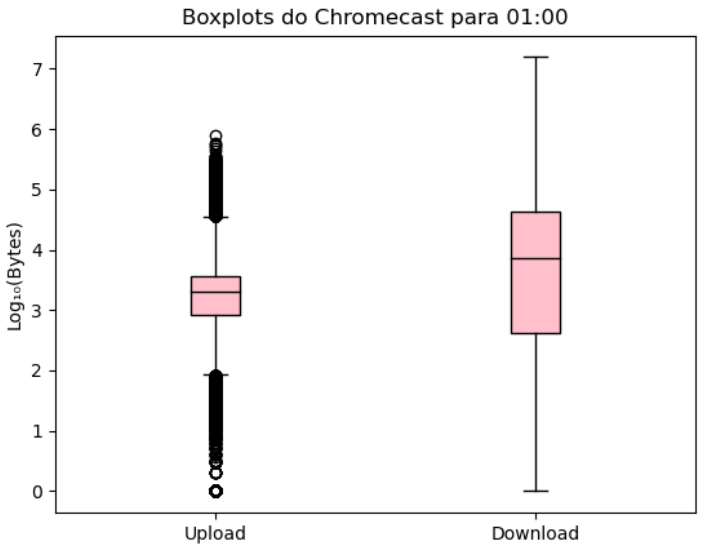

Figure 12: Chromecast na Hora 01:00

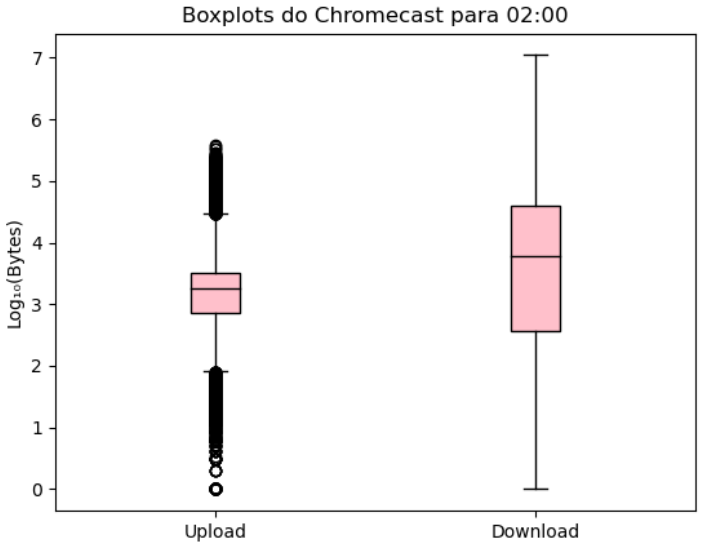

Figure 13: Chromecast na Hora 02:00

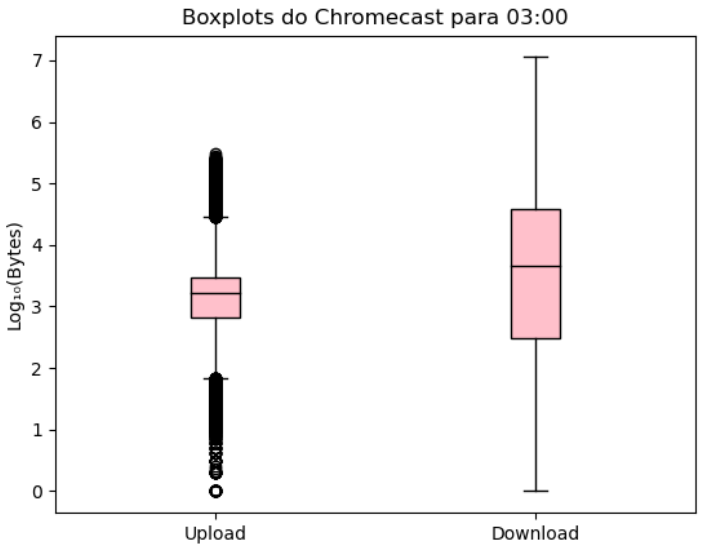

Figure 14: Chromecast na Hora 03:00

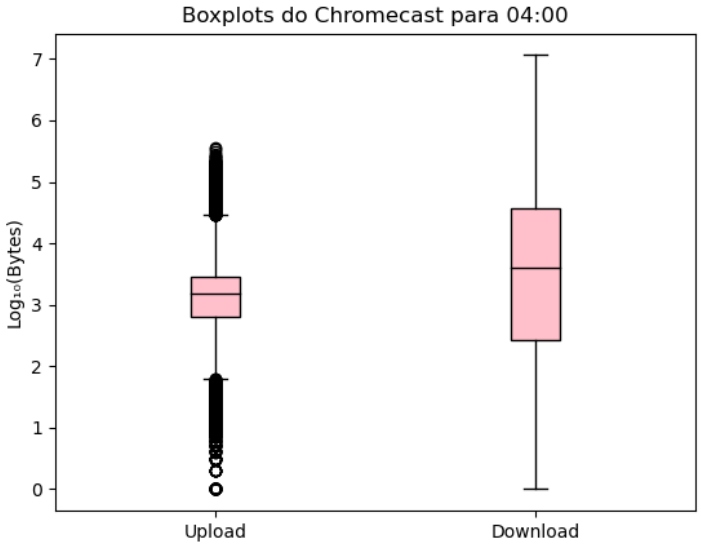

Figure 15: Chromecast na Hora 04:00

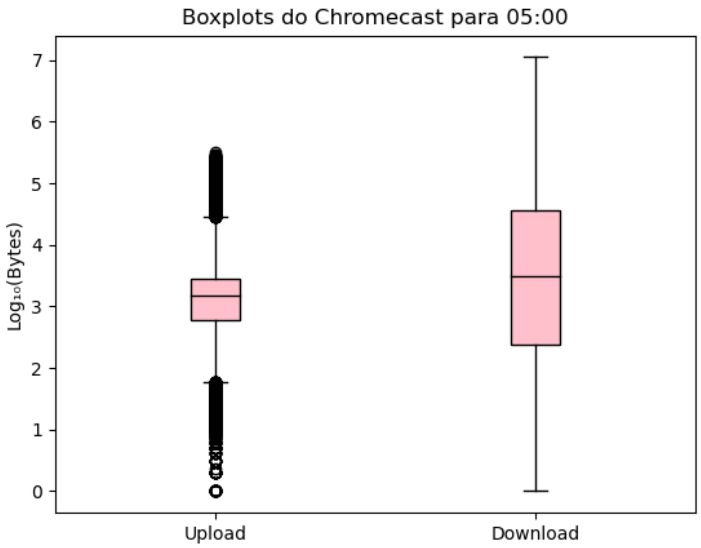

Figure 16: Chromecast na Hora 05:00

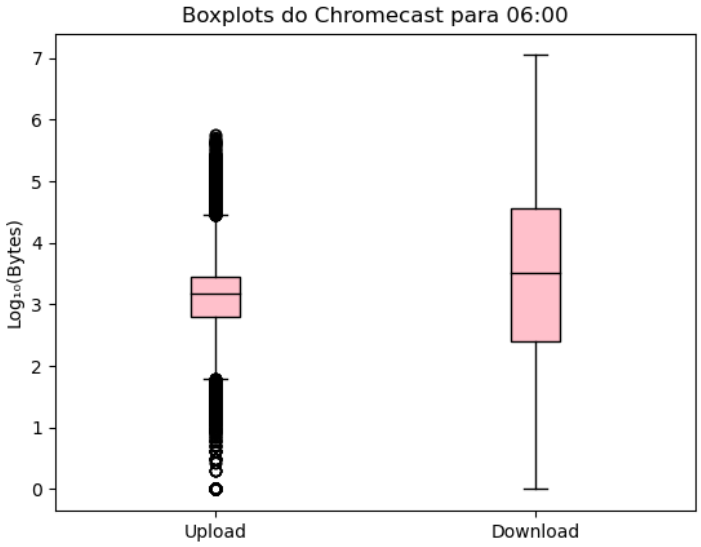

Figure 17: Chromecast na Hora 06:00

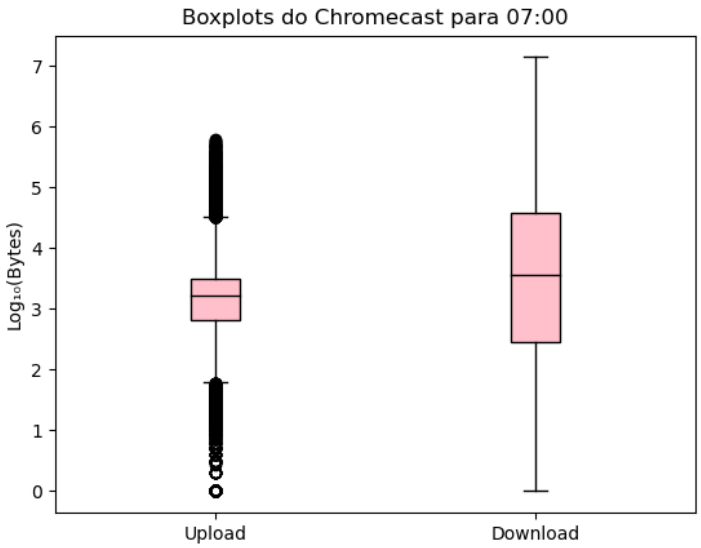

Figure 18: Chromecast na Hora 07:00

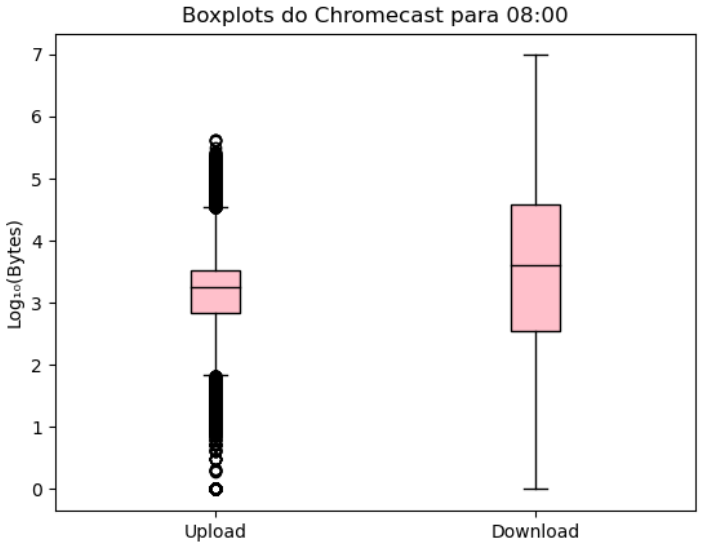

Figure 19: Chromecast na Hora 08:00

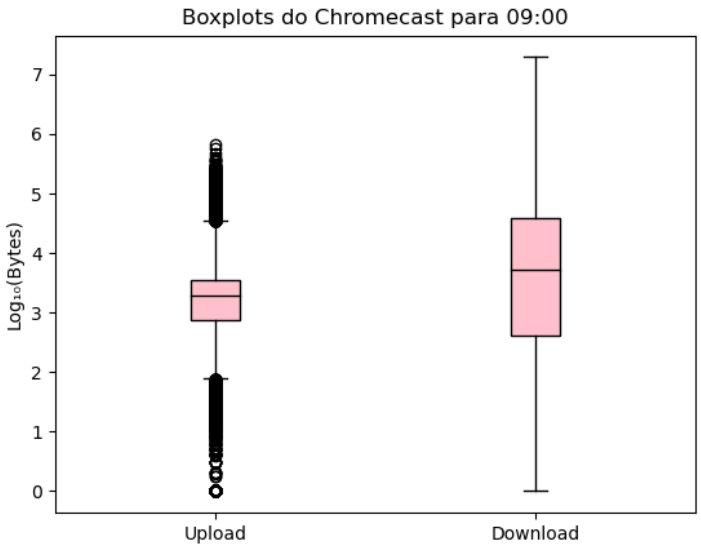

Figure 20: Chromecast na Hora 09:00

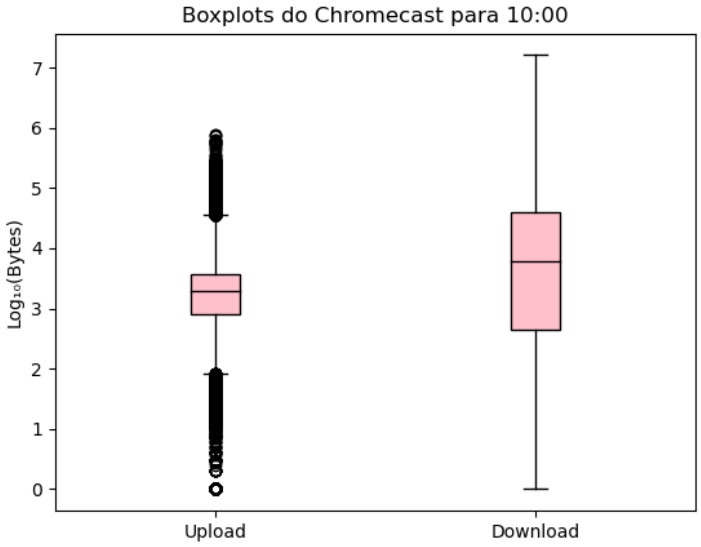

Figure 21: Chromecast na Hora 10:00

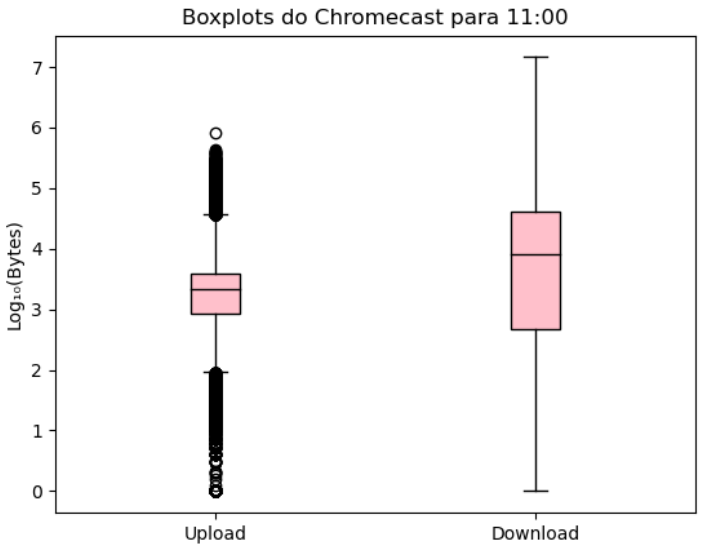

Figure 22: Chromecast na Hora 11:00

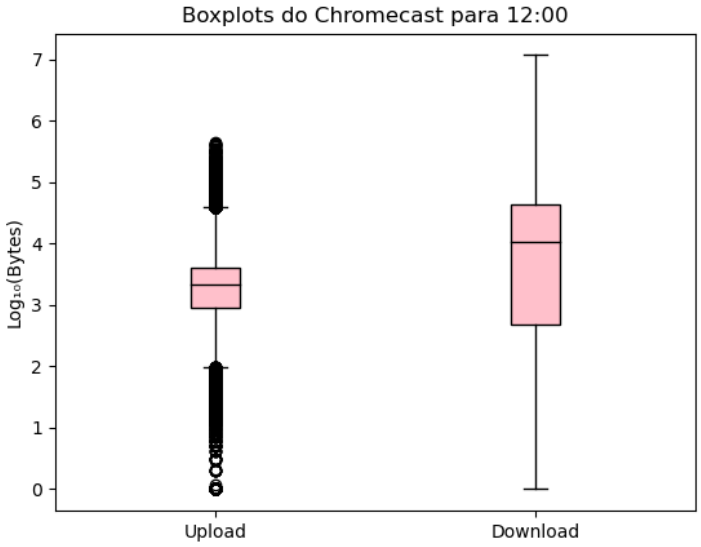

Figure 23: Chromecast na Hora 12:00

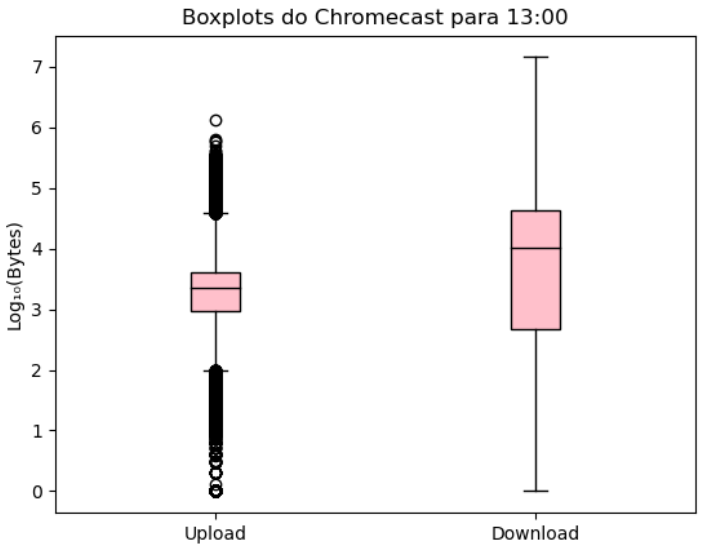

Figure 24: Chromecast na Hora 13:00

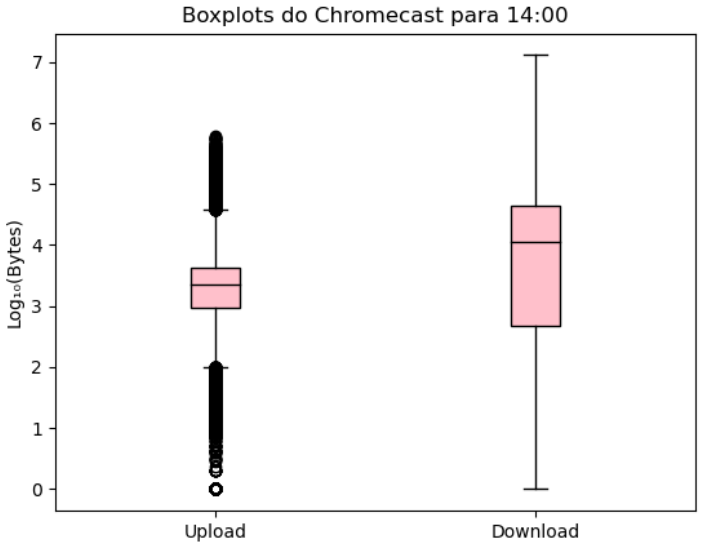

Figure 25: Chromecast na Hora 14:00

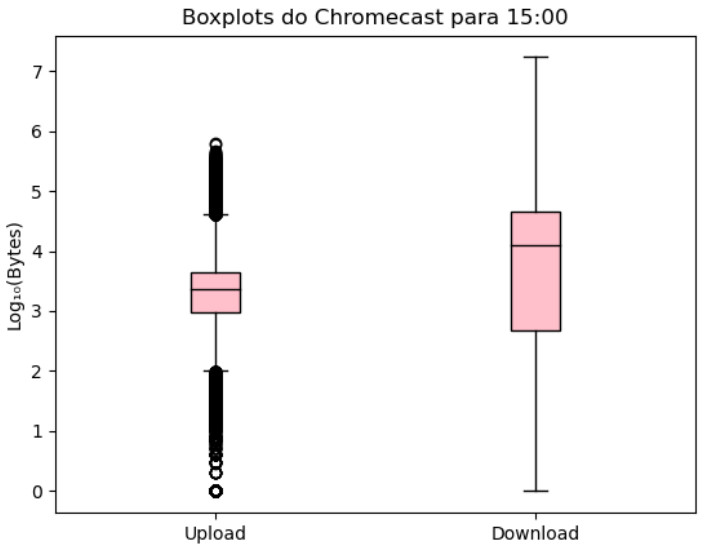

Figure 26: Chromecast na Hora 15:00

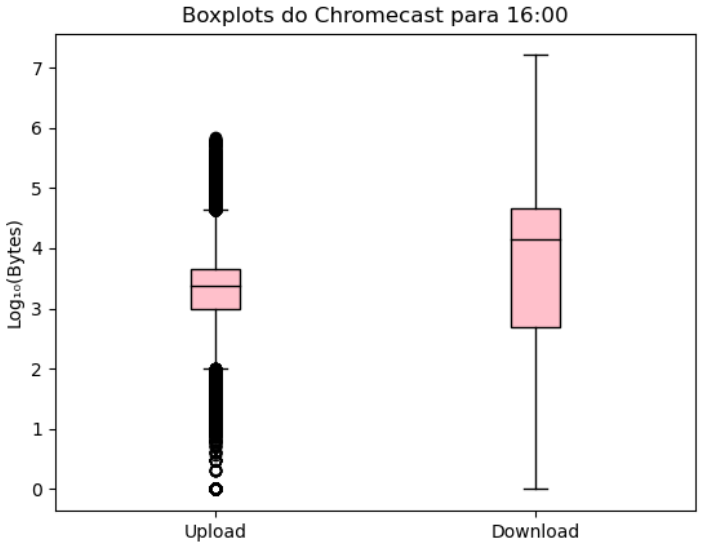

Figure 27: Chromecast na Hora 16:00

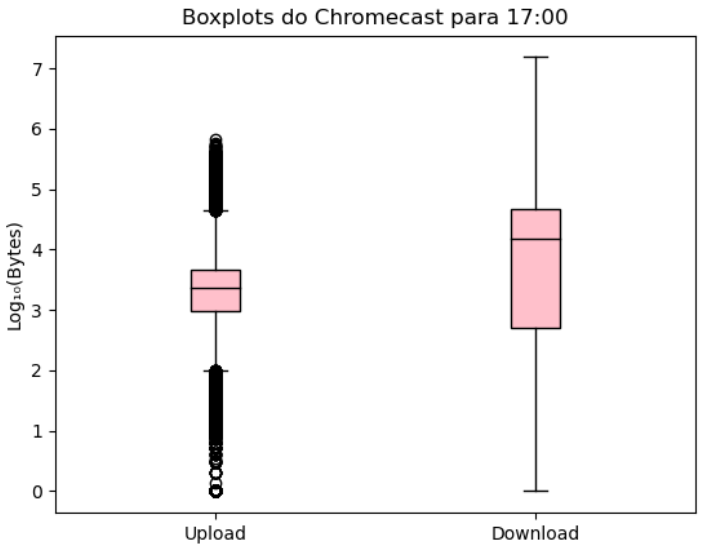

Figure 28: Chromecast na Hora 17:00

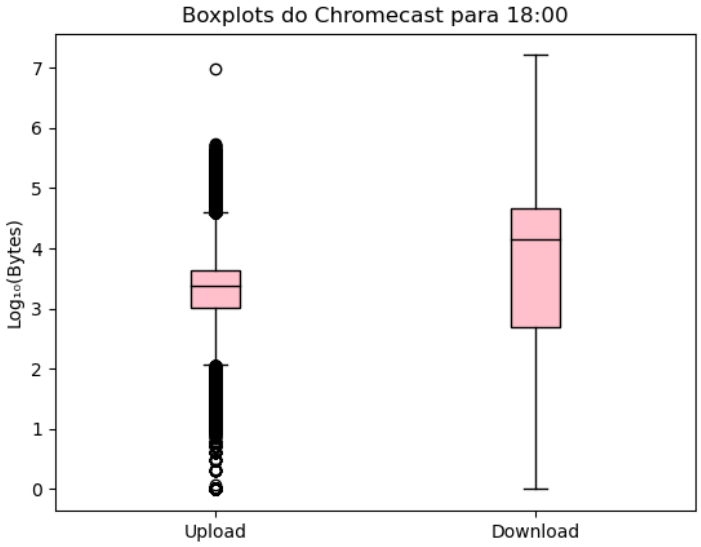

Figure 29: Chromecast na Hora 18:00

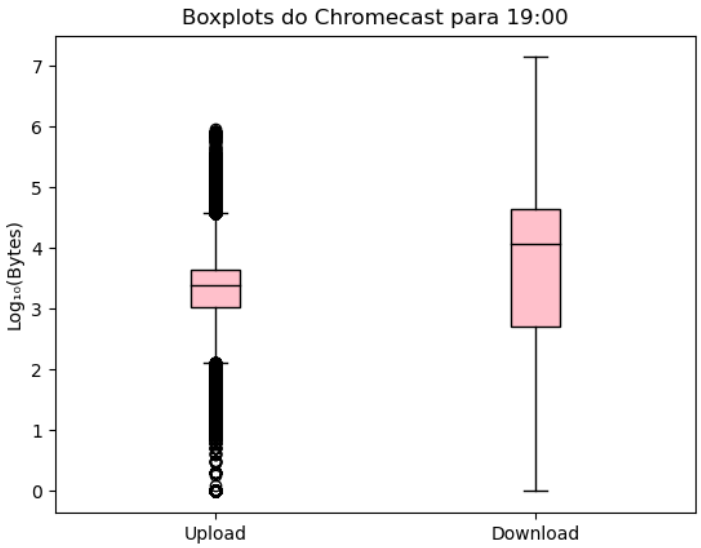

Figure 30: Chromecast na Hora 19:00

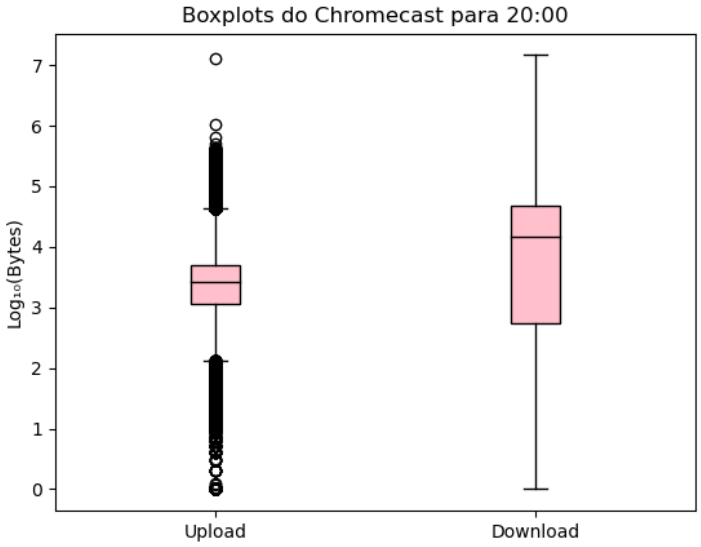

Figure 31: Chromecast na Hora 20:00

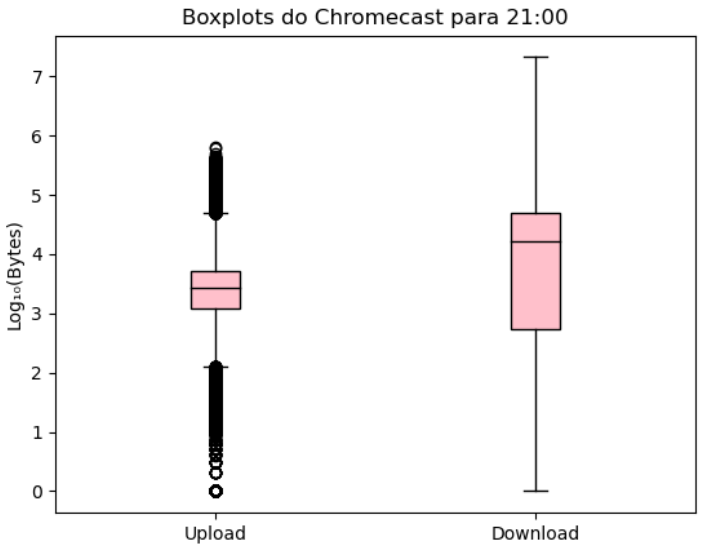

Figure 32: Chromecast na Hora 21:00

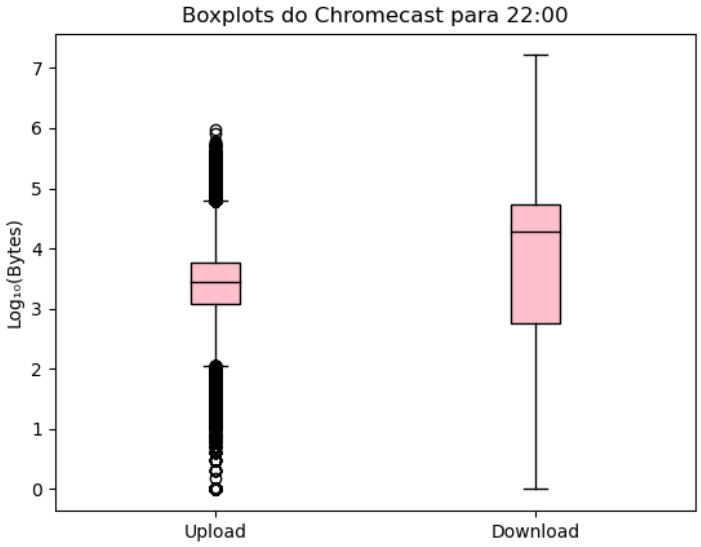

Figure 33: Chromecast na Hora 22:00

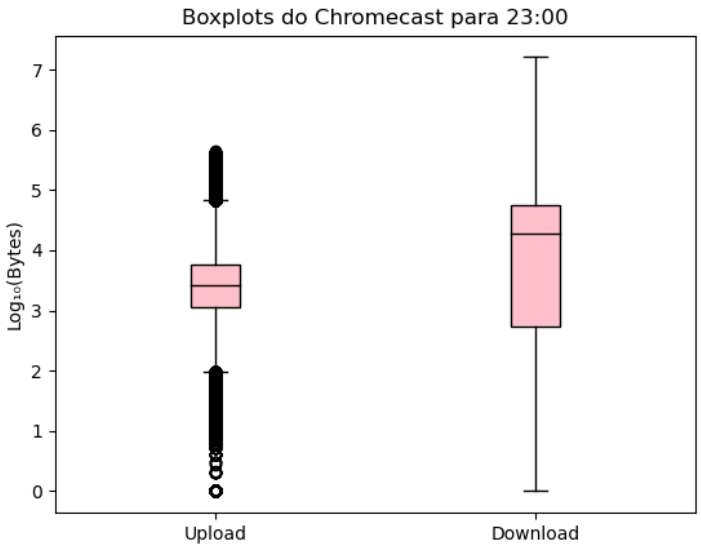

Figure 34: Chromecast na Hora 23:00

3.1.2 Smart TV

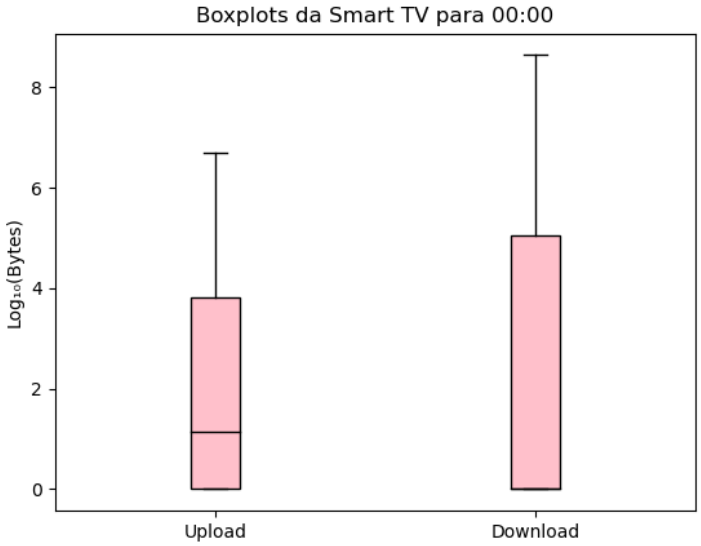

Figure 35: Smart TV na Hora 00:00

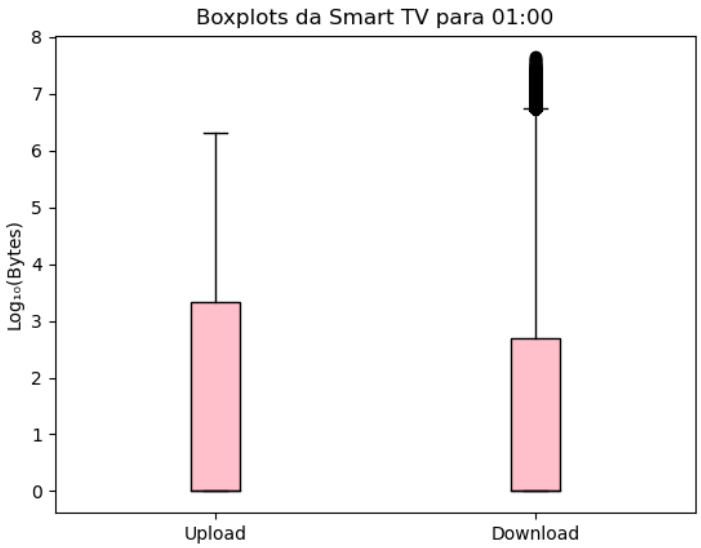

Figure 36: Smart TV na Hora 01:00

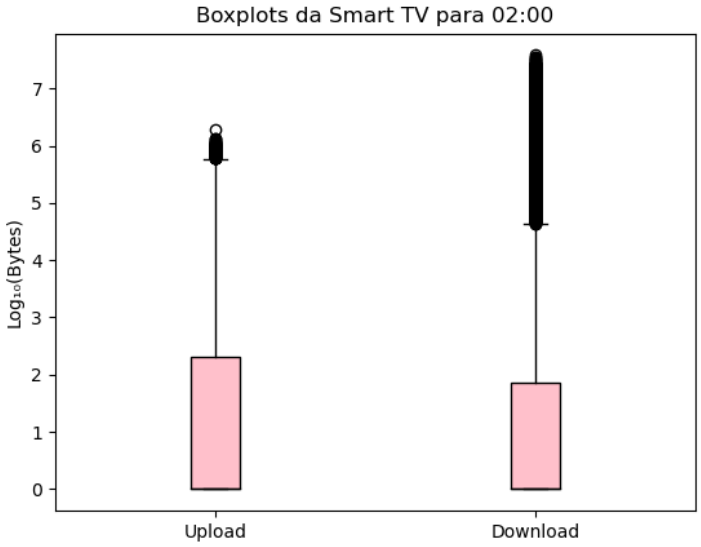

Figure 37: Smart TV na Hora 02:00

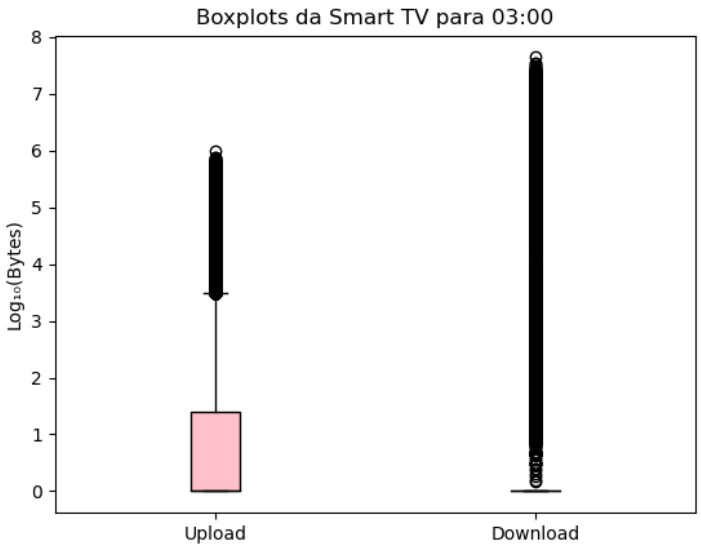

Figure 38: Smart TV na Hora 03:00

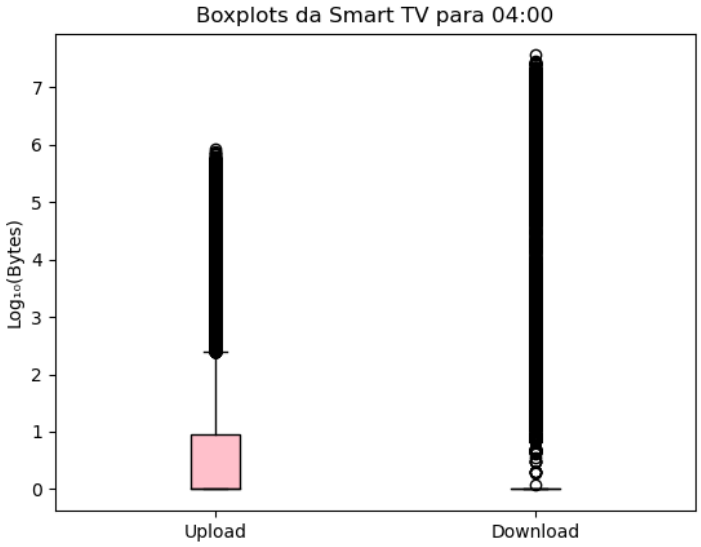

Figure 39: Smart TV na Hora 04:00

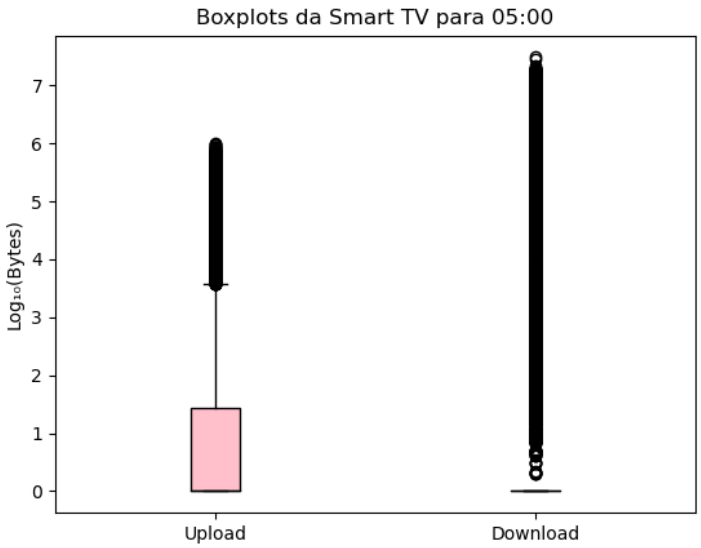

Figure 40: Smart TV na Hora 05:00

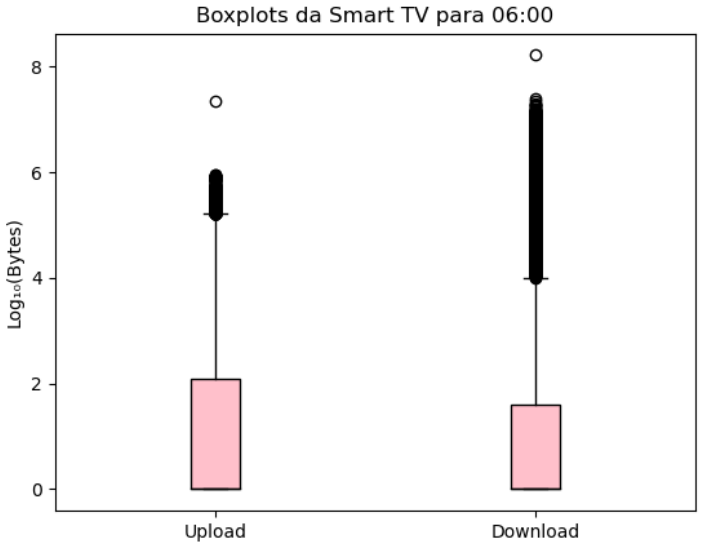

Figure 41: Smart TV na Hora 06:00

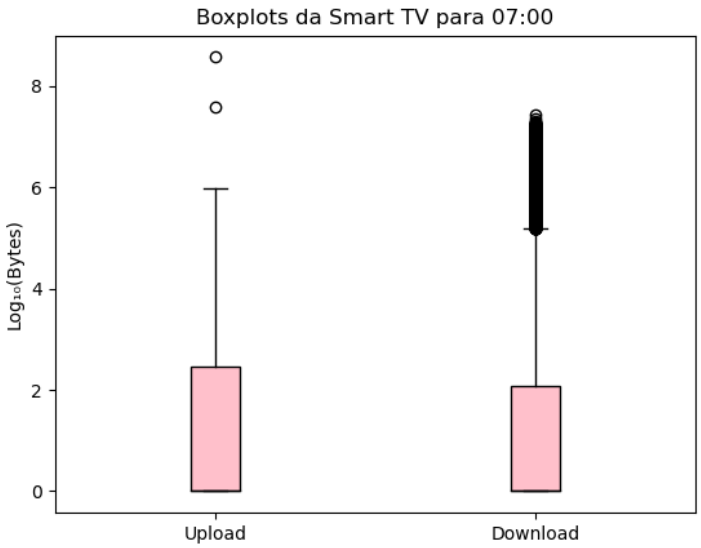

Figure 42: Smart TV na Hora 07:00

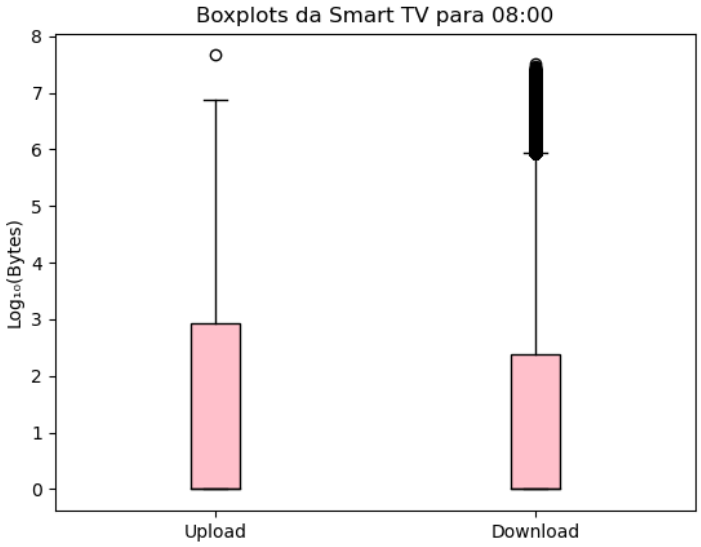

Figure 43: Smart TV na Hora 08:00

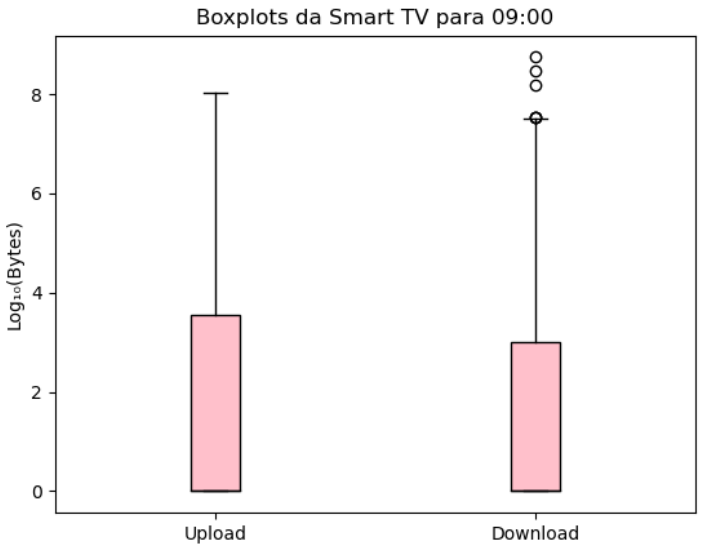

Figure 44: Smart TV na Hora 09:00

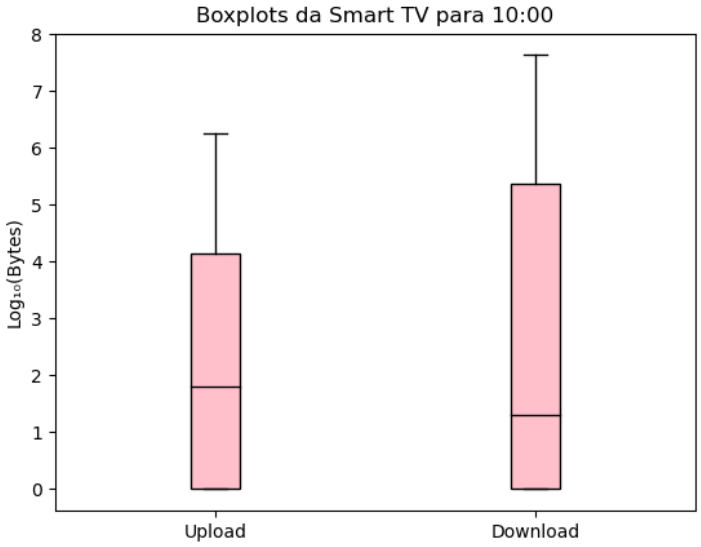

Figure 45: Smart TV na Hora 10:00

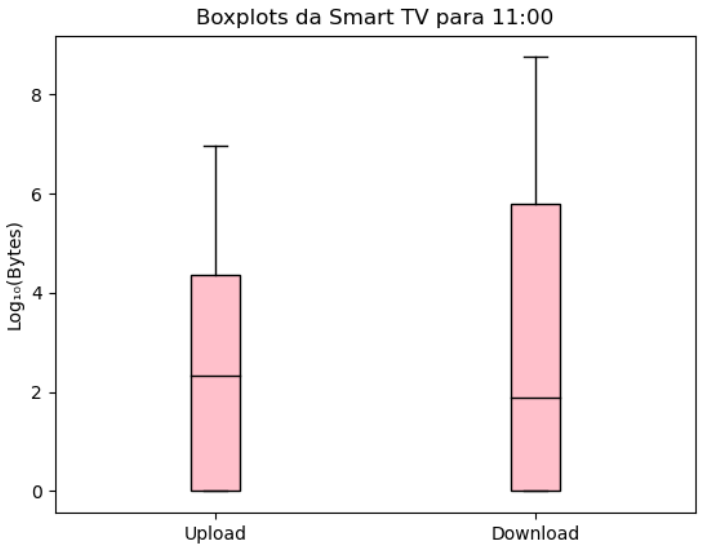

Figure 46: Smart TV na Hora 11:00

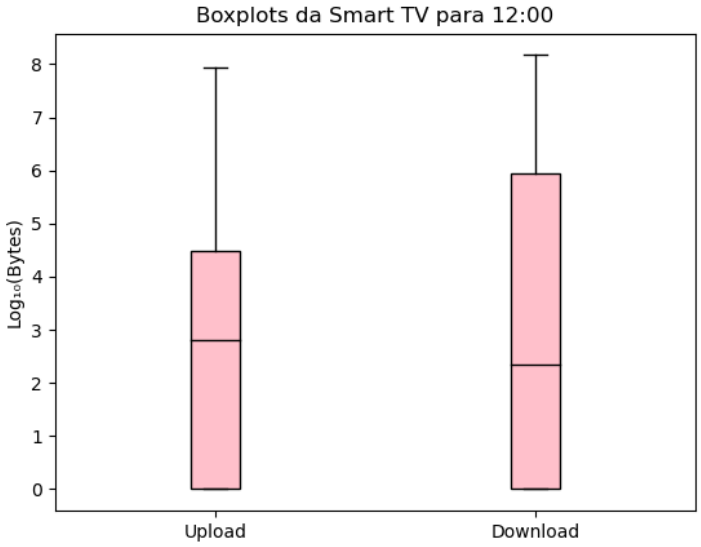

Figure 47: Smart TV na Hora 12:00

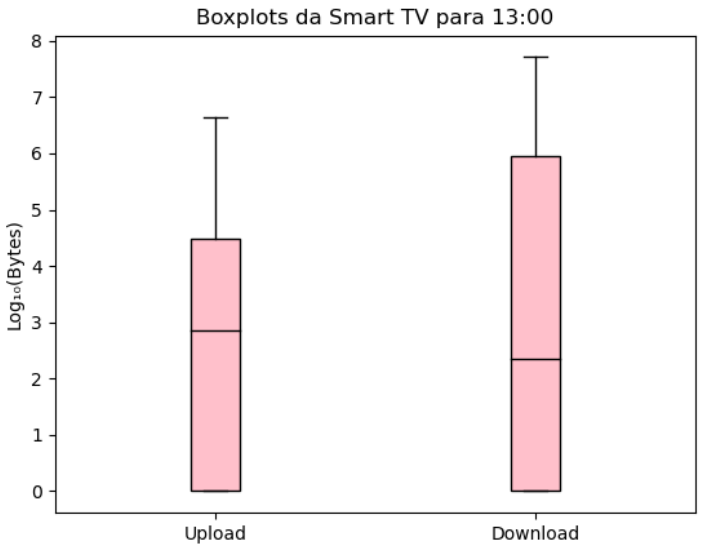

Figure 48: Smart TV na Hora 13:00

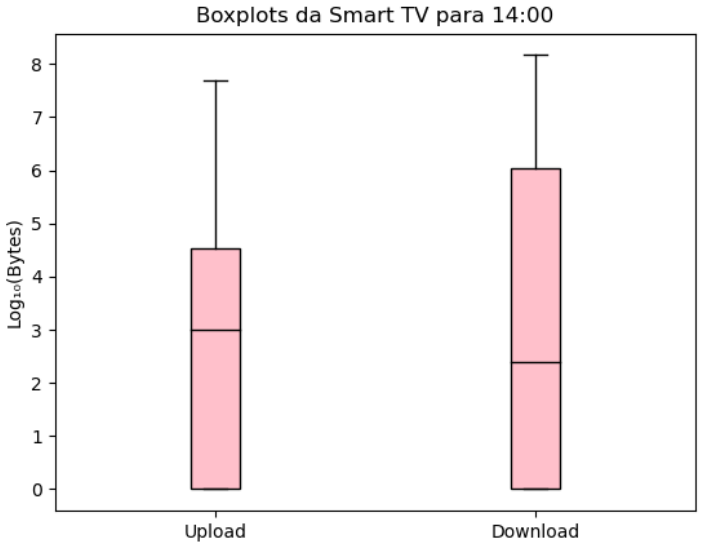

Figure 49: Smart TV na Hora 14:00

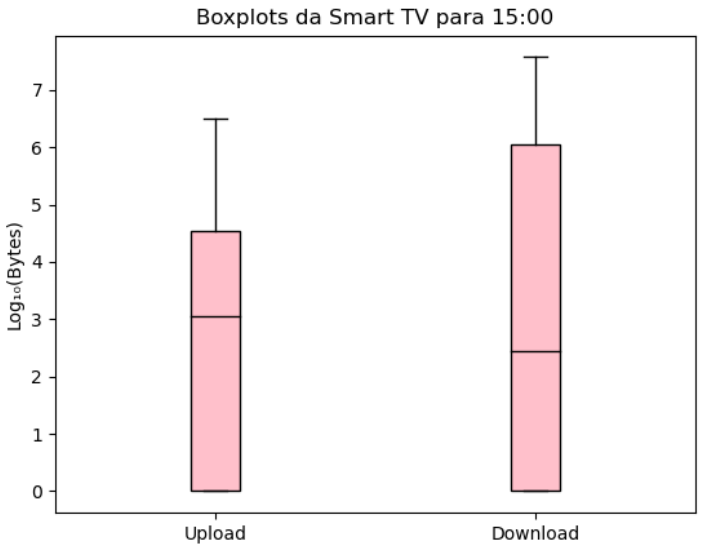

Figure 50: Smart TV na Hora 15:00

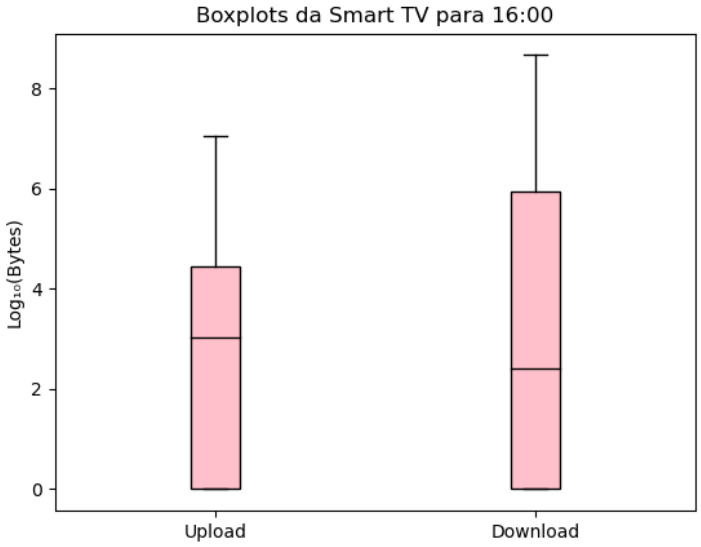

Figure 51: Smart TV na Hora 16:00

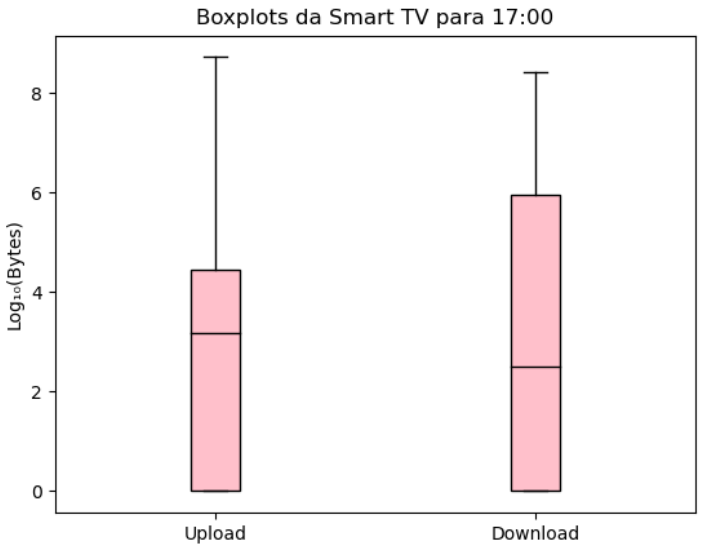

Figure 52: Smart TV na Hora 17:00

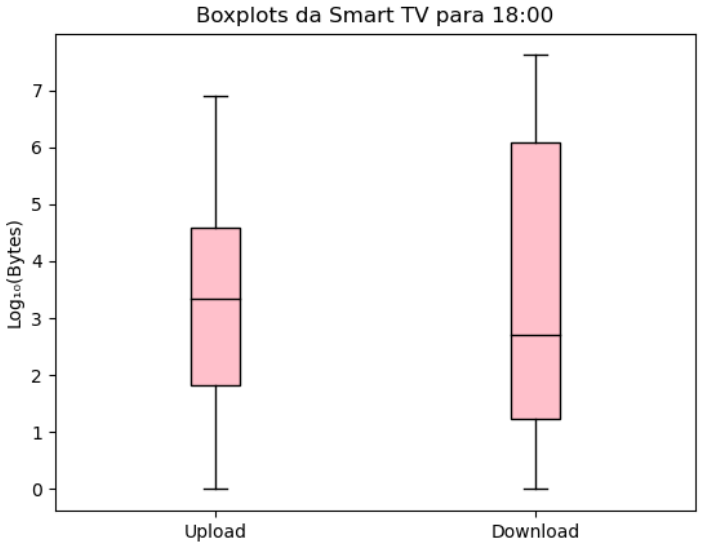

Figure 53: Smart TV na Hora 18:00

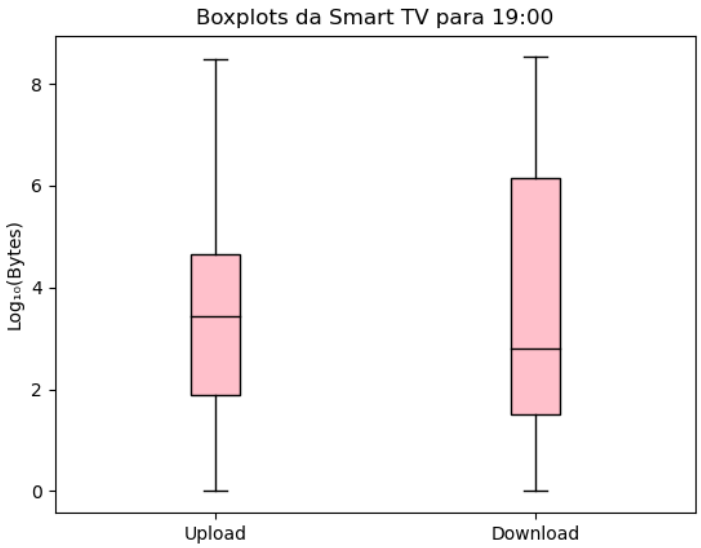

Figure 54: Smart TV na Hora 19:00

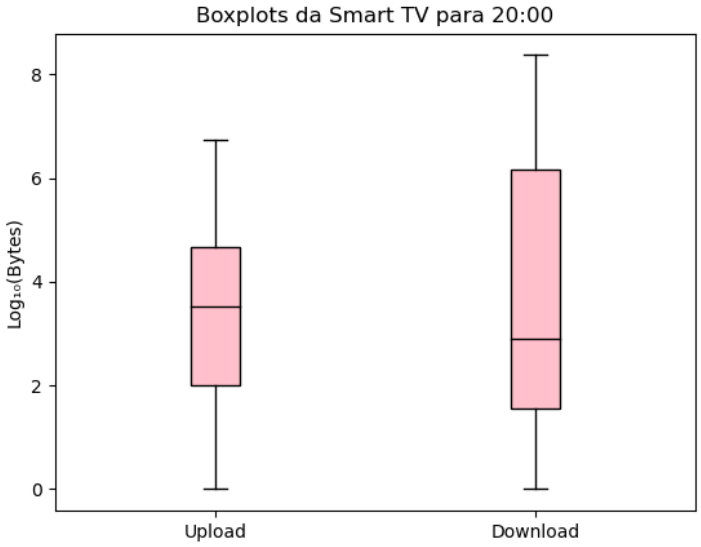

Figure 55: Smart TV na Hora 20:00

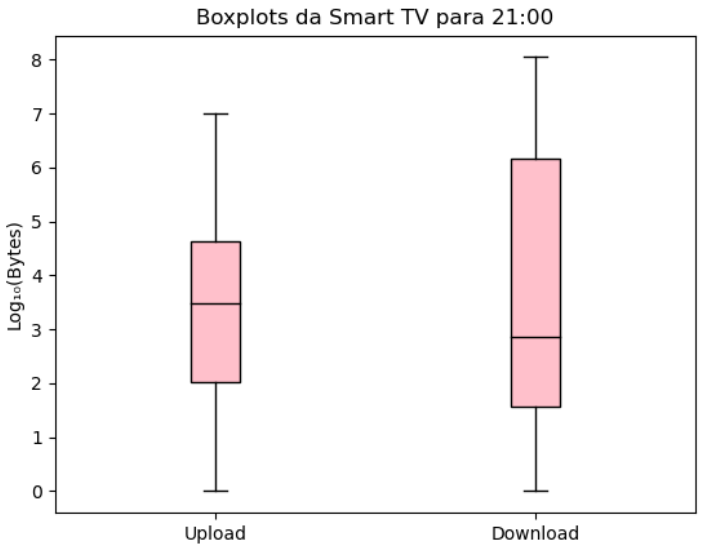

Figure 56: Smart TV na Hora 21:00

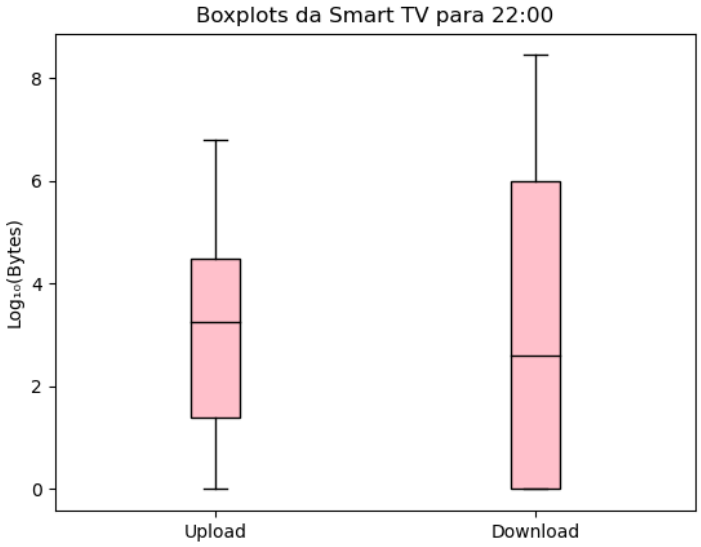

Figure 57: Smart TV na Hora 22:00

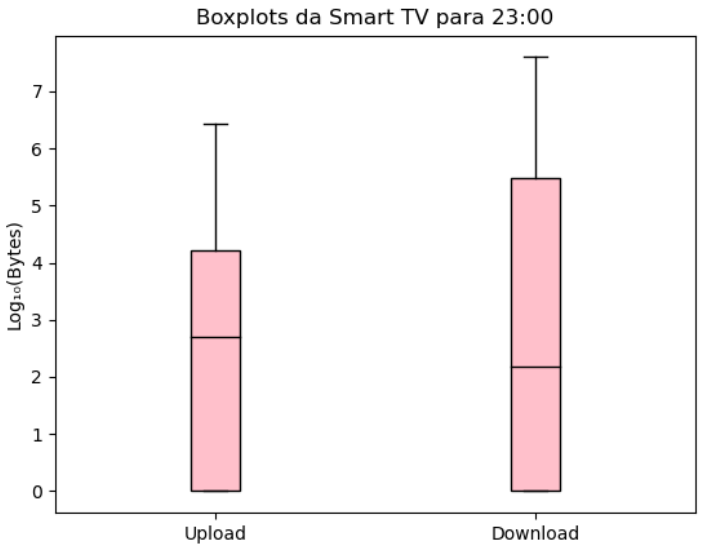

Figure 58: Smart TV na Hora 23:00

3.2 Média, Variância e Desvio Padrão

Para o plot das estatísticas variando por hora, foi utilizado o método groupby a fim de agrupar os dados das estatísticas por hora. Desta forma gerando um gráfico onde o eixo x é a hora e o eixo y é a estatística observada para a coluna de interesse.

3.2.1 Chromecast

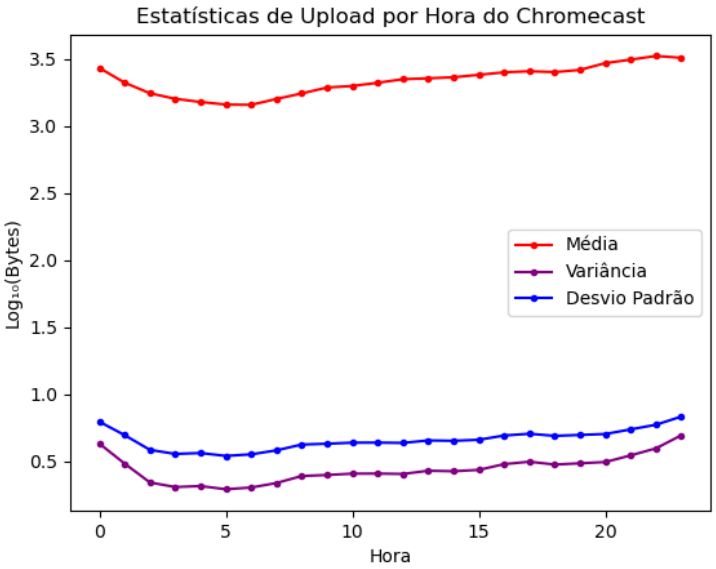

Figure 59: Estatísticas de Upload por Hora Chromecast

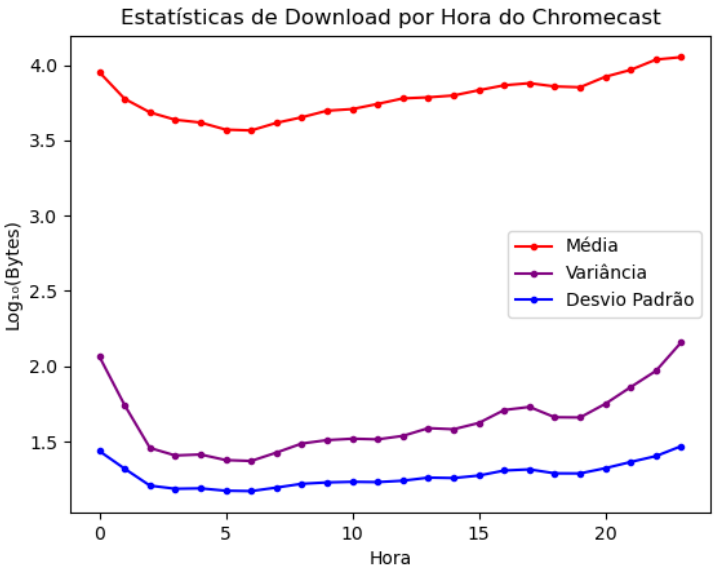

Figure 60: Estatísticas de Download por Hora Chromecast

3.2.2 Smart TV

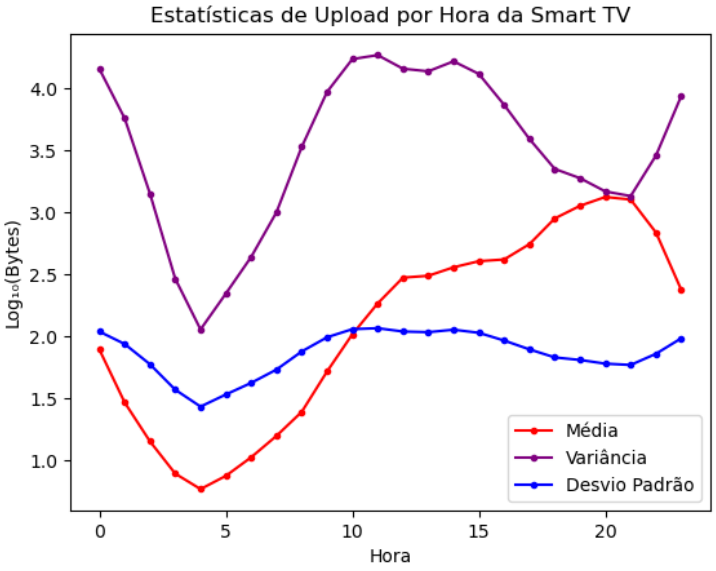

Figure 61: Estatísticas de Upload por Hora Smart TV

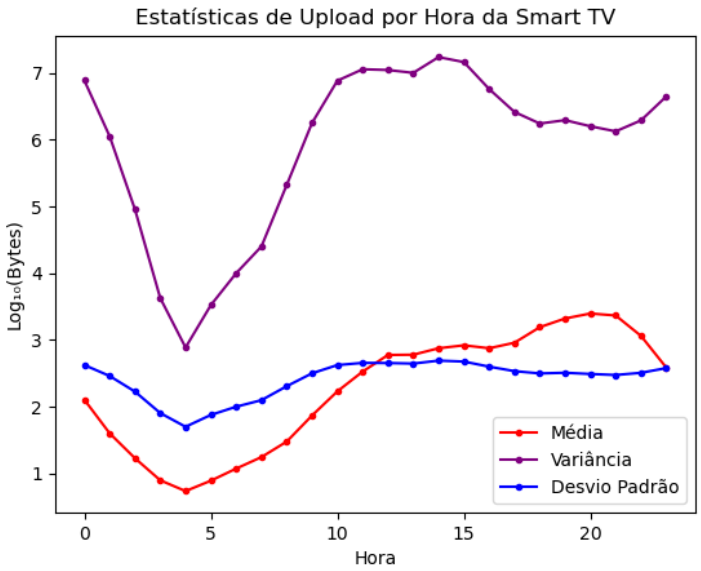

Figure 62: Estatísticas de Download por Hora Smart TV

3.3 Análise dos Resultados

Analisando os resultados,pode-se ver que todos os boxplots de taxa de upload do Chromecast têm outliers, enquanto os boxplots de upload das 22h e 23h têm menos outliers na faixa superior da porção. Para downloads envolvendo o Chromecast , quase não existe outliers, tirando um, localizado no plot das 23h.

Além disso, a análise Chromecast revelou que seu funcionamento é contínuo, o que significa que, diferentemente da SmartTV, as taxas de download e upload são sempre altas. Este segundo dispositivo tende a ser baseado no uso do pelo usuário, pois a média de download tende a aumentar proporcionalmente à taxa de upload durante o horário de acesso compartilhado (aproximadamente das 10h às 20h).

4 Caracterizando os horários com maior valor de tráfego

4.1 Horários

4.1.1 Chromecast

	Hora da Média Máxima
Upload	22
Download	23

Table 3: Horas de mediana e média máxima para o Chromecast

4.1.2 Smart TV

	Hora da Média Máxima
Upload	20
Download	20

Table 4: Horas de mediana e média máxima para a Smart TV

4.2 Histograma

4.2.1 Chromecast

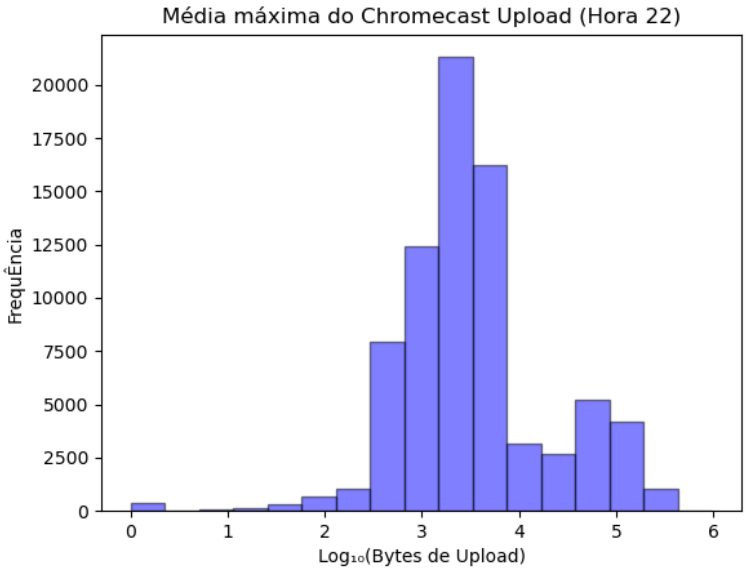

Figure 63: Histograma de upload na hora de maior média para o Chromecast

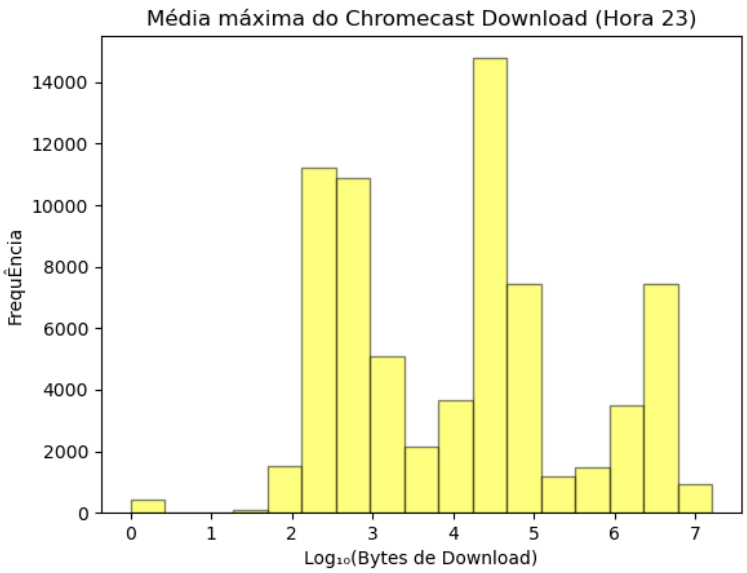

Figure 64: Histograma de upload na hora de maior mediana para o Chromecast

4.2.2 Smart TV

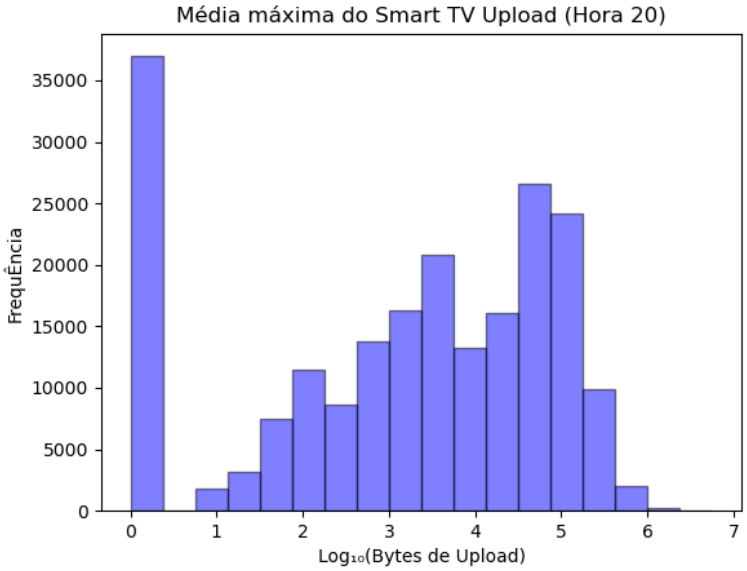

Figure 65: Histograma de upload na hora de maior média para a Smart TV

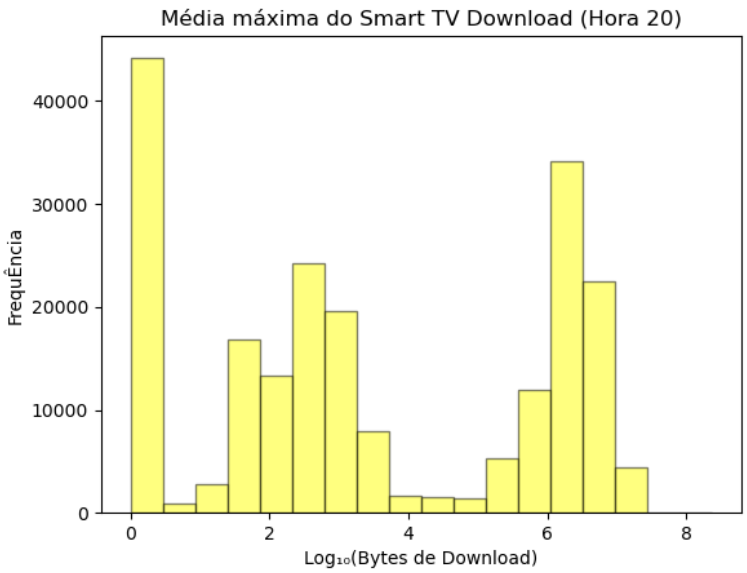

Figure 66: Histograma de download na hora de maior média para a Smart TV

4.3 Q-Q Plot

Para esta parte foi usado a função *probplot* da biblioteca *scipy*, juntamente com a biblioteca pylab para plotar a função, usando a distribuição Gaussina como paramentro, gerou os resultados a seguir.

4.3.1 Chromecast

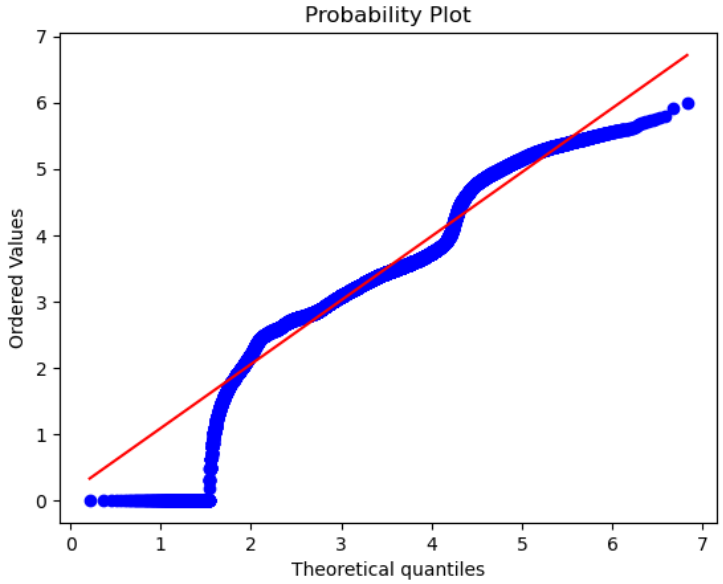

Figure 67: Q-Q Plot de upload da hora de maior média do Chromecast

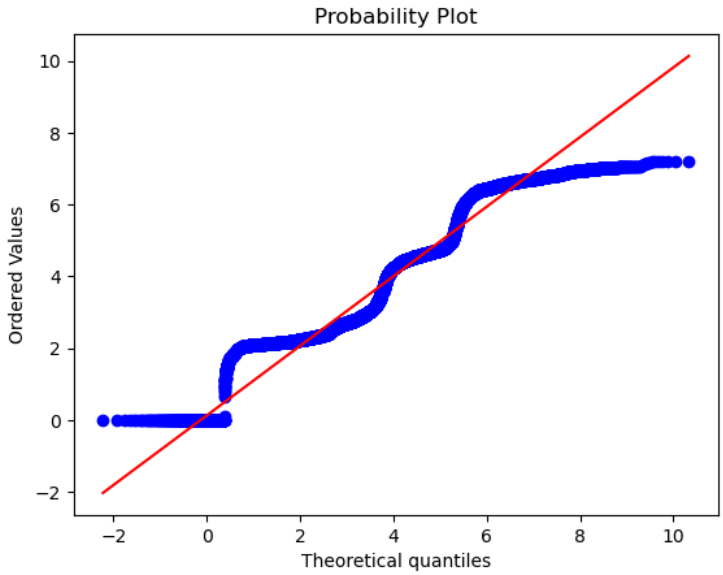

Figure 68: Q-Q Plot de download da hora de maior média do Chromecast

4.3.2 Smart TV

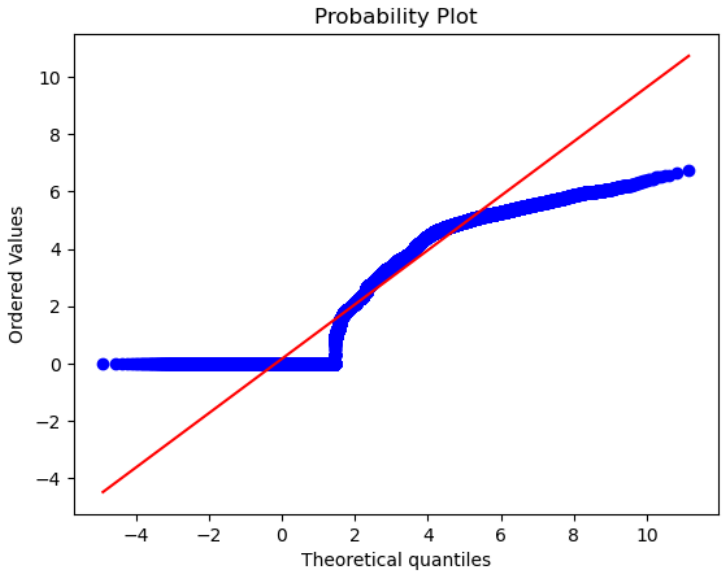

Figure 69: Q-Q Plot de upload da hora de maior média da Smart TV

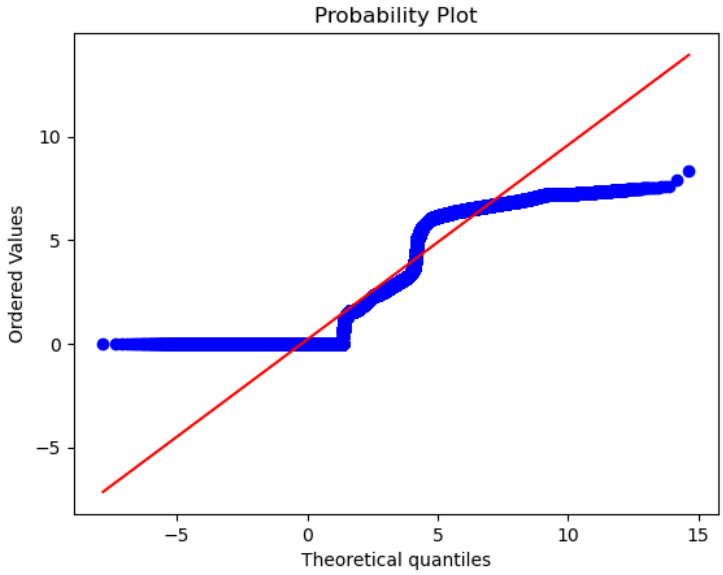

Figure 70: Q-Q Plot de download da hora de maior média da Smart TV

4.4 Análise dos Resultados

Como vimos na Seção 4.1, os horários de maior média de upload do Chromecast foi as 22:00, e para o download foi as 23:00. Já para a Smart TV a hora foi igual para tanto para upload quanto para download, sendo ela às 20:00 horas.

Podemos verificar nos Histogramas do chromecast, que o aparelho possui uma frequência maior em quantidade maior de bytes baixados que com relação a upload, evidenciando um possível comportamento do dispositivo quanto ao funcionamento do upload de dados.

Já com relação a Smart-TV, o zero ainda se mantém evidente, normalmente ressaltando que mesmo no horário com a maior média de download ainda há muitos dispositivos que não fazem a transmissão e ou download de dados, porém, os horários de maiores pico da taxa de download ou upload, se comportam de forma diferente.

Os datasets 3 e 4, possuem uma distribuição bem similar, variando apenas nas regiões entre o 2° quantil e 6° quantil. Já para os datasets 1 e 2, eles possuem uma distribuição parecida nos quantis antes do 0 e após o 8°, já entre eles é possível verificar uma certa discrepância.

Com os Q-Q Plots do chromecast é possível dizer que o dataset que upload bytes para a maior média do pode ser mapeado para uma gaussiana, enquanto para o de download não é possível afirmar.

Já para a Smart TV, não é possível caracterizar ambos os datasets de upload e download com uma distribuição Gaussiana, mostrando as diferenças de comportamento dos dados entre os datasets mencionados.

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

5.1 Coeficientes de correlação

Para o cálculo de correlação foi utilizado o coeficiente de pearson, atraves do método *personr* da biblioteca Scipy, obtendo a seguinte tabela:

	Coeficiente de pearson
Chromecast	0.003436
Smart TV	0.915609

Table 5: Coeficientes de correlação entre os aparelhos

5.2 Scatter Plot

Para gerar o scatter plot foi utilizado o método *scatter* da biblioteca matplotlib.

5.2.1 Chromecast

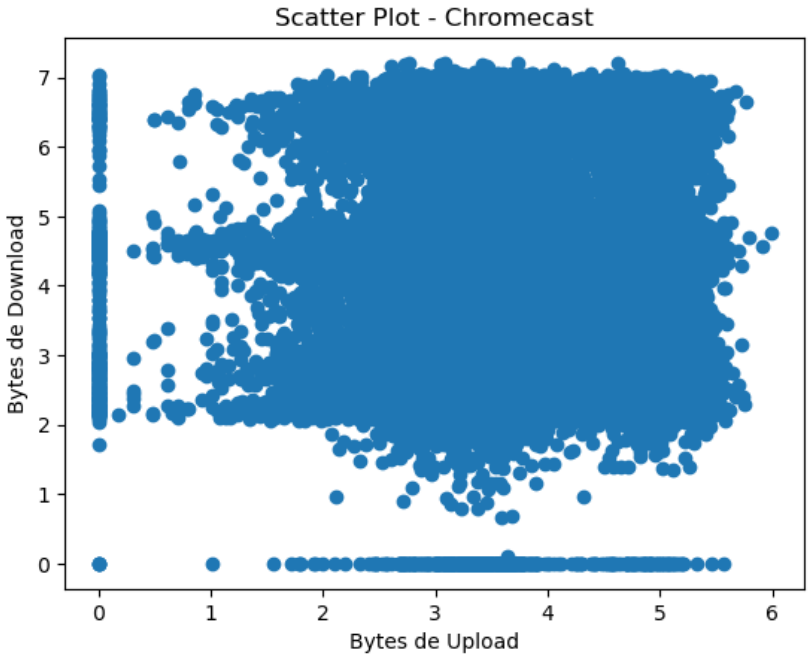

Figure 71: Scatter Plot - Chromecast

5.2.2 Smart TV

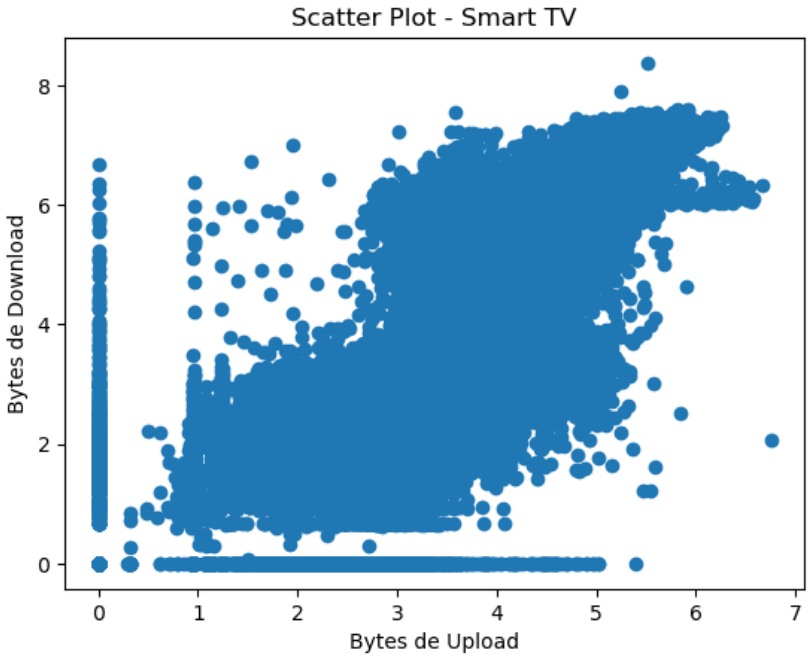

Figure 72: Scatter Plot - Smart TV

5.3 Análise dos Resultados

Podemos observar, através dos dados mostrados na tabela 5, que os dados relacionados à Smart TV possuem uma correlação, algo que não ocorre pros dados do chromecast, tendo um coeficente de correlação bem próximo do zero.