07 de janeiro de 2022

Universidade Federal do Rio de Janeiro

Estatística e Modelos Probabilísticos

Trabalho final da disciplina

Bruno Dantas de Paiva - 118048097

Sumário

1 Estatísticas gerais

1.1 Pré-Tratamento

Antes de iniciar efetivamente a análise dos dados é necessário conhecer um pouco mais a respeito dos dados e verificar se é necessário o pré-tratamento destes. Como sugestão do próprio texto apresentando o projeto, era necessário reescalonar os dados para log_10 a fim de obter uma escala de grandeza mais fácil de ser visualizada, tendo em vista que o número de bytes em upload e download podem chegar na ordem de 7 ou mais casas decimais.

Para isto, foi aplicada uma função log10 da biblioteca numpy, a fim de criar novas colunas em ambos os datasets, log_bytes_up e log_bytes_down. Além disso, considerando que este método retorna um tipo $-np.inf$ para os logs de valores igual a 0, bastou somar 1 à estes valores, a fim de ainda sabermos quando aconteceu uma taxa de 0 bytes de upload ou download na escala log. Tal caso vale ressaltar que o zero poderia ser considerado um outlier tendo em vista que zero de upload ou de download não impactaria na análise, porém como existe uma frequência alta para um dos dispositivos de momentos de taxa de bytes zero, foi optado por manter esse valor a fim de ser importante para a caracterização do dispositivo.

Além desse tratamento, uma outra coluna foi criada *hour* que consiste em obter a hora em que um dado foi obtido, como o dado possui um formato $date\_hour$, foi necessário somente extrair a hora deste formato usando o método *split* do python e armazenando a hora num formato inteiro.

Tendo em vista este pré-processamento, agora é possível seguir com a análise para os dados de ambos dispositivos.

1.2 Histograma

Para a obtenção dos histogramas foi utilizado o método hist da biblioteca *plt* do python, a fim de gerar os histogramas para cada uma das colunas de interesse para os datasets e interesse.

Além disso, um dos parâmetros para este método é o bin, para isso foi desenvolvida uma função calculate_bin que implementa o método de sturges para obter o bin.

Tendo em vista isso, abaixo obteremos todos os histogramas gerados tanto para Chromecast quanto para Smart-TV.

1.2.1 Chromecast

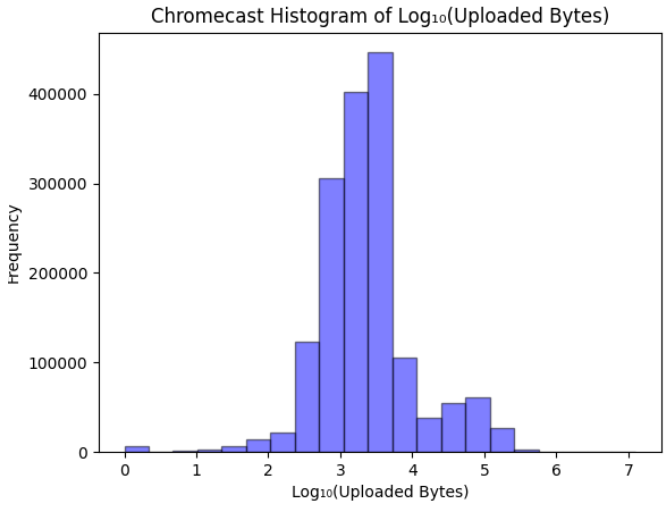

Figura 1: Histograma de $Log10$(taxa de upload) para o Chromecast

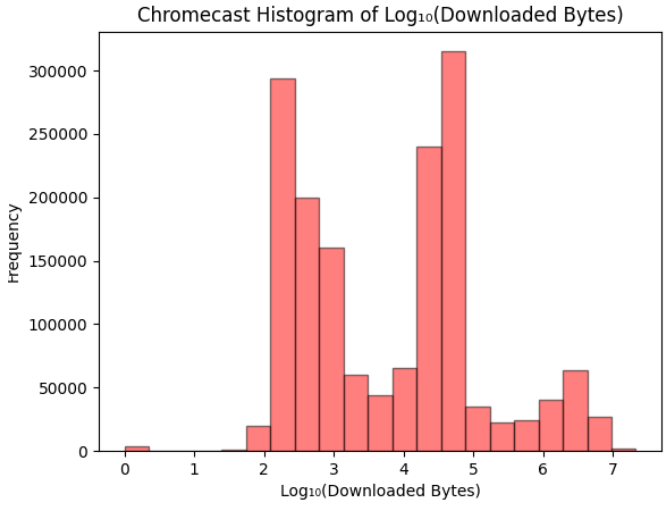

Figura 2: Histograma de $Log10$(taxa de download) para o Chromecast

1.2.2 Smart-TV

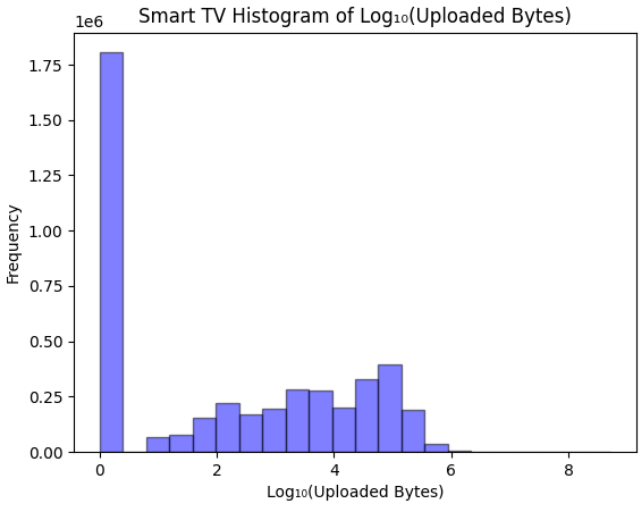

Figura 3: Histograma de $Log10$(taxa de upload) para a Smart-TV

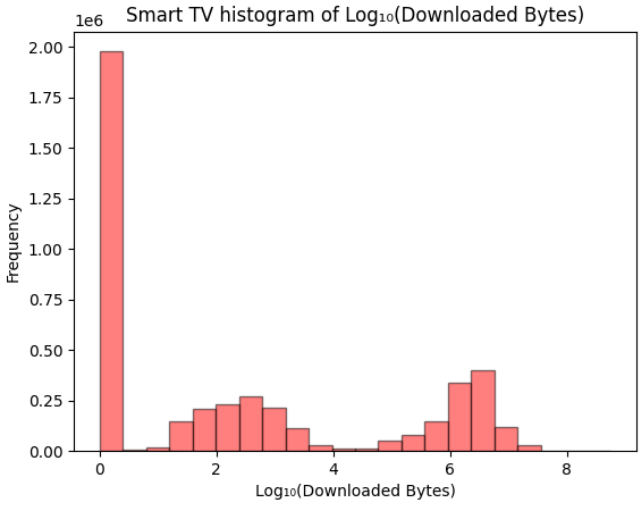

Figura 4: Histograma de $Log10$(taxa de download) para a Smart-TV

1.3 Função Distribuição Empírica

A função distribuição empírica foi plotada por meio da função plot da biblioteca plt. Da o eixo X foram utilizados os valores da coluna de interesse ordenados de forma crescente, enquanto o eixo y foi gerado por meio do método *linspace*, gerando um array de valores entre 0 e 1 com o número de dados sendo o número de dados do dataframe.

Tendo em vista isso, podemos obter os gráficos abaixo.

1.3.1 Chromecast

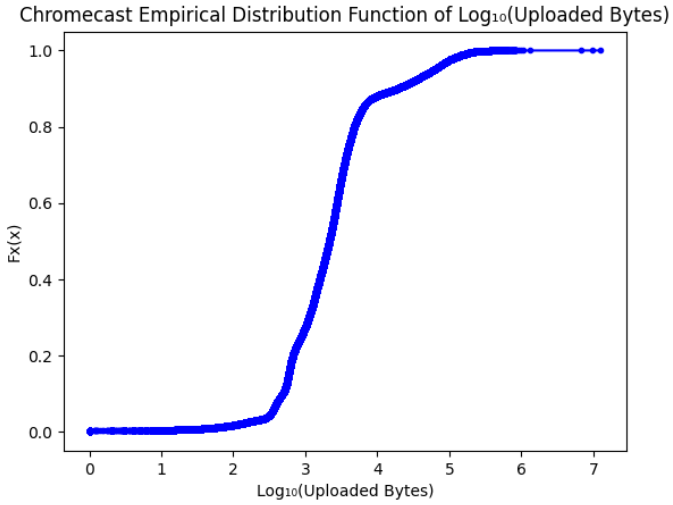

Figura 5: Função Distribuição Empírica de $Log10$(taxa de upload) para o Chromecast

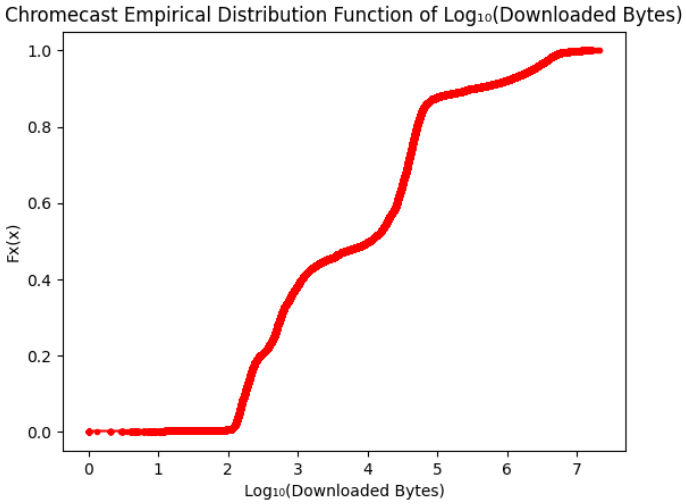

Figura 6: Função Distribuição Empírica de $Log10$(taxa de download) para o Chromecast

1.3.2 Smart-TV

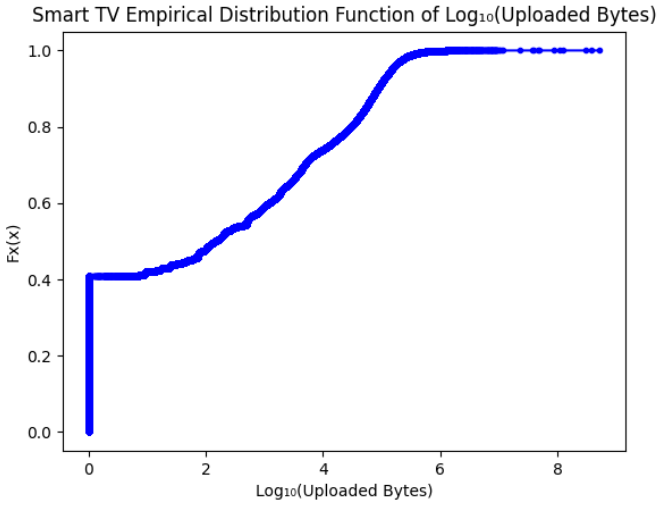

Figura 7: Função Distribuição Empírica de $Log10$(taxa de upload) para a Smart-TV

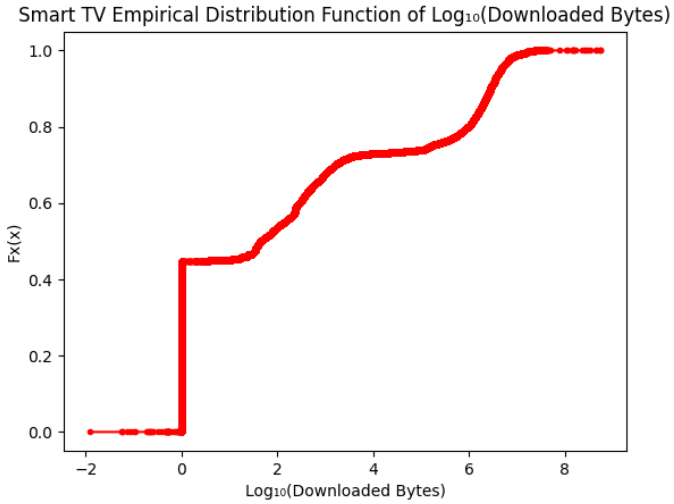

Figura 8: Função Distribuição Empírica de $Log10$(taxa de download) para a Smart-TV

1.4 Box Plot

Os boxplots foram gerados em pares, a fim de que seja possível ter lado a lado, para um mesmo dispositivo, os boxplots e download e upload.

Para a geração destes boxplots, foi utilizado o método *boxplot* da biblioteca plt com os parâmetros sendo os dados da coluna de interesse para cada um dos dispositivos.

1.4.1 Chromecast

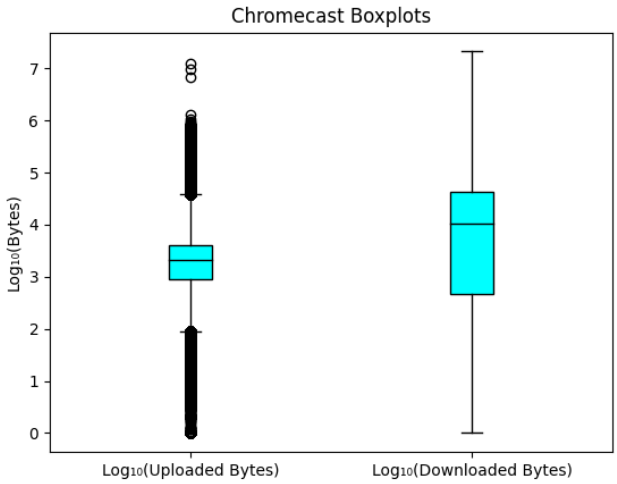

Figura 9: Boxplot de $Log10$(taxa de download e upload) para o Chromecast

1.4.2 Smart-TV

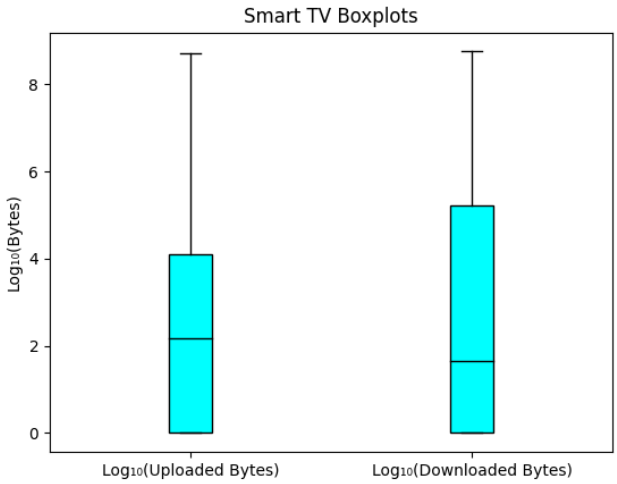

Figura 10: Boxplot de $Log10$(taxa de download e upload) para a Smart-TV

1.5 Média, Variância e Desvio Padrão

Para esta análise estatística, foram utilizados os métodos do dataframe pandas de média, variância e desvio padrão para gerar esses dados, a forma de output gerada foi somente uma tabela em markdown.

1.5.1 Chromecast

Statistics	Log_{10}**(Uploaded Bytes)**	Log_{10}**(Downloaded Bytes)**
Mean	3.3503	3.80005
Variance	0.459969	1.6639
Standard Deviation	0.67821	1.28992

Tabela 1: Média, Variância e Desvio Padrão de Log_{10}(Bytes) para o Chromecast

1.5.2 Smart-TV

Statistics	Log_{10}**(Uploaded Bytes)**	Log_{10}**(Downloaded Bytes)**
Mean	2.15829	2.35168
Variance	4.11014	6.72132
Standard Deviation	2.02735	2.59255

Tabela 2: Média, Variância e Desvio Padrão de Log_{10}(Bytes) para a Smart-TV

1.6 Análise de resultados

Como é possível observar, ao comparar os histogramas de download e upload entre os dispositivos, podemos observar um alto número de dados em que a frequência de upload e download é zero na Smart-TV, enquanto no chromecast estes momentos tem uma menor ocorrência. Desta forma, é possível assumir que o comportamento dos dispositivos se dá de forma diferente ao baixar os dados ou realizar o upload dos dados.

Observando já o histograma da taxa de upload e download em cada dispositivo, podemos observar que o comportamento do chromecast é diferente na faixa de download e upload, pois o pico máximo de upload se dá numa faixa de 1000 bytes por segundo, tal fato pode ser por conta da capacidade de internet dos dispositivos. Enquanto a faixa de download chega seu pico

um pouco mais acima. No qual um comportamento similar também pode ser observado na Smart-TV.

Além disso, também quanto ao histograma dos dispositivos, é possível observar uma aparente complementariedade entre as distribuições de upload e download, deixando a suspeitar de um possível comportamento do dispositivo de não manter as duas ações em pico, o que pode ser observado mais facilmente entre os valores de $Log_{10}(2)$ e $Log_{10}(6)$, tanto para o chromecast quanto para a smart-tv, isso desconsiderando o valor de 0 de taxa de download.

Já comparar os box plots de download e upload entre os dispositivos, é possível perceber que a dispersão entre eles para a taxa de upload acaba sendo bem diferente entre si, a concentração dos dados também então em Logs diferentes devido a altura do terceiro quartil e primeiro quartil entre os dois boxplots além do boxplot do chromecast ter a presença de outliers bem evidentes. Já para o boxplot de download, ambos possuem uma região de terceiro quartil próxima, porém uma região de primeiro quartil e segundo quartil bem distante entre-si, além disso, similar ao gráfico de upload, o chromecast também possui outliers.

Com relação ao comportamento da função distribuição empírica entre os dispositivos, é bem diferente, apresentando um aumento bem gradual no chromecast, diferente da smart-tv que tem um aumento bem evidente no 0, devido ao grande número de momentos em que este dispositivo não realiza download ou upload.

2 Estatísticas por horário

2.1 Box Plot

Para a geração do boxplot de hora em hora, também foi utilizado o método *boxplot* da biblioteca plt, porém, agora os dados foram filtrados com base no horário de interesse. Para isso foi gerado um loop variando de hora em hora e os dados foram filtrado para essas horas de parâmetro.

Com isto, foram gerados os boxplots abaixo.

2.1.1 Chromecast

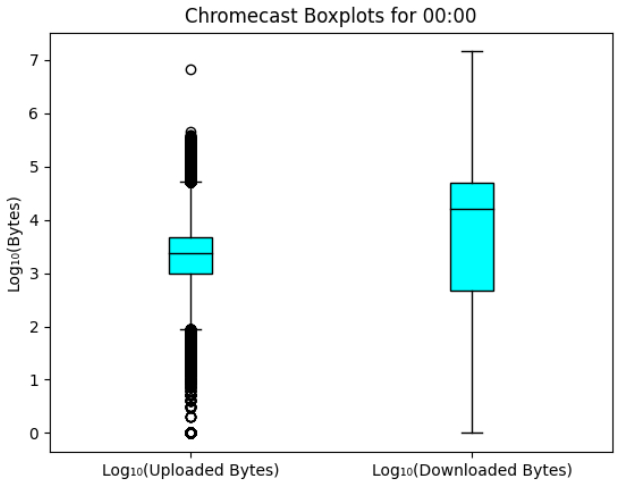

Figura 11: Boxplot de $Log10$(taxa de download e upload) para o Chromecast na Hora 00:00

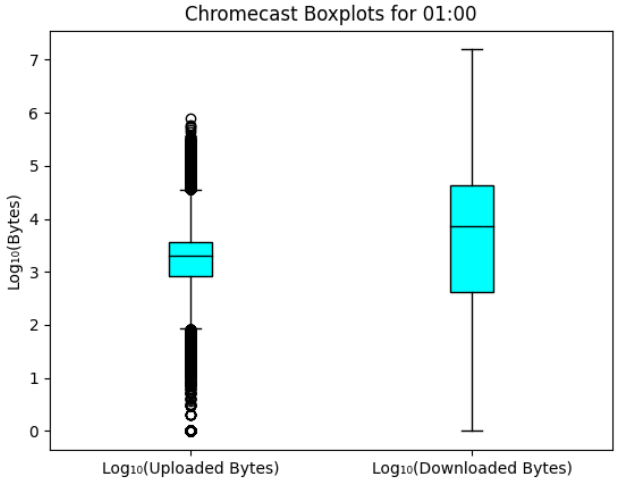

Figura 12: Boxplot de $Log10$(taxa de download e upload) para o Chromecast na Hora 01:00

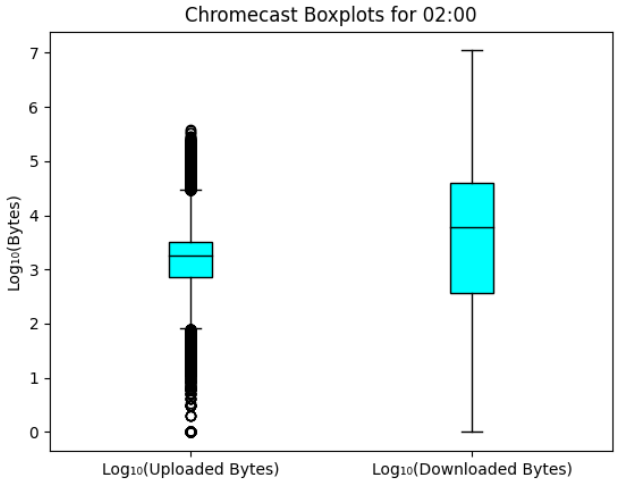

Figura 13: Boxplot de $Log10$(taxa de download e upload) para o Chromecast na Hora 02:00

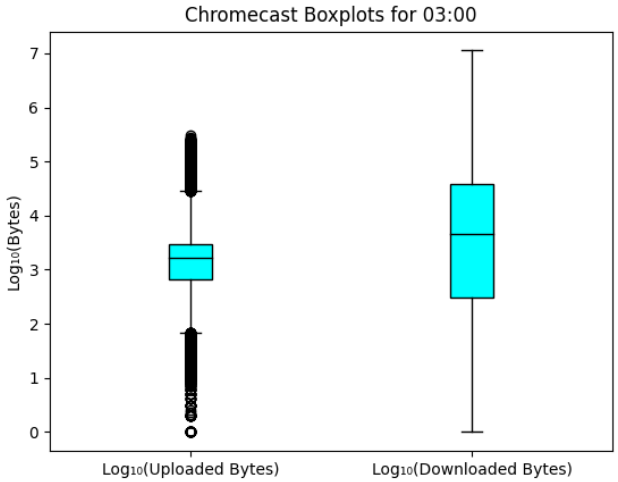

Figura 14: Boxplot de $Log10$(taxa de download e upload) para o Chromecast na Hora 03:00

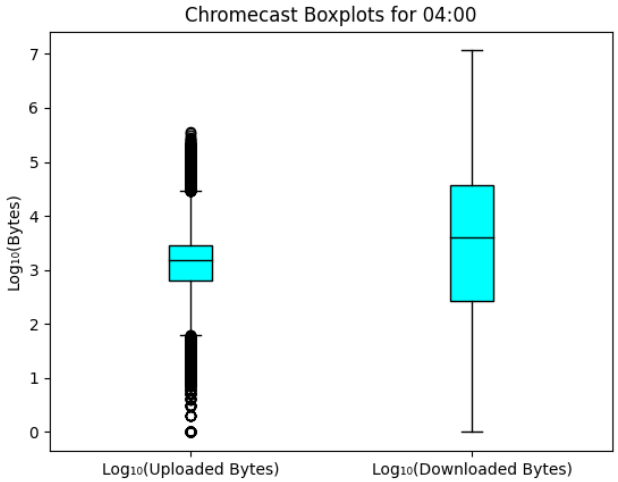

Figura 15: Boxplot de $Log10$(taxa de download e upload) para o Chromecast na Hora 04:00

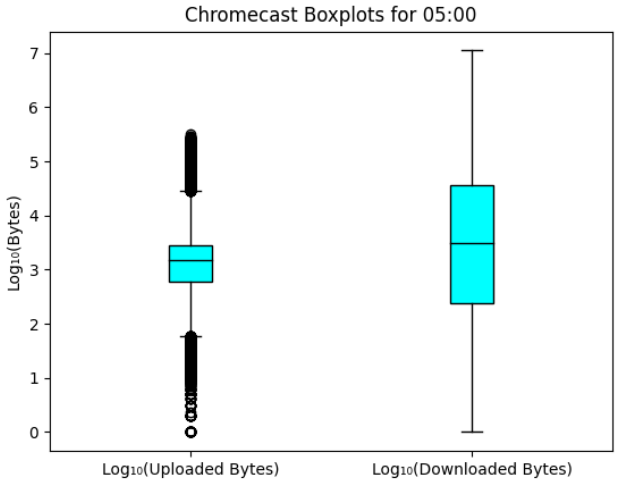

Figura 16: Boxplot de $Log10$(taxa de download e upload) para o Chromecast na Hora 05:00

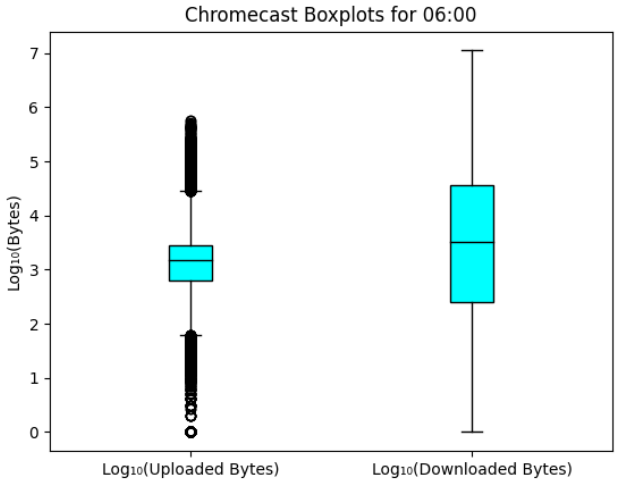

Figura 17: Boxplot de $Log10$(taxa de download e upload) para o Chromecast na Hora 06:00

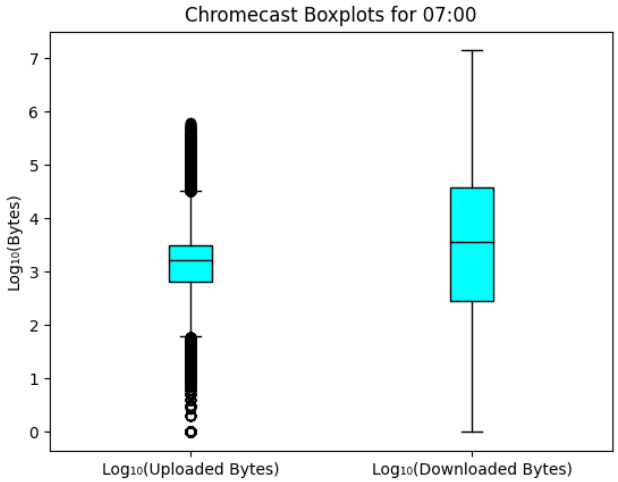

Figura 18: Boxplot de $Log10$(taxa de download e upload) para o Chromecast na Hora 07:00

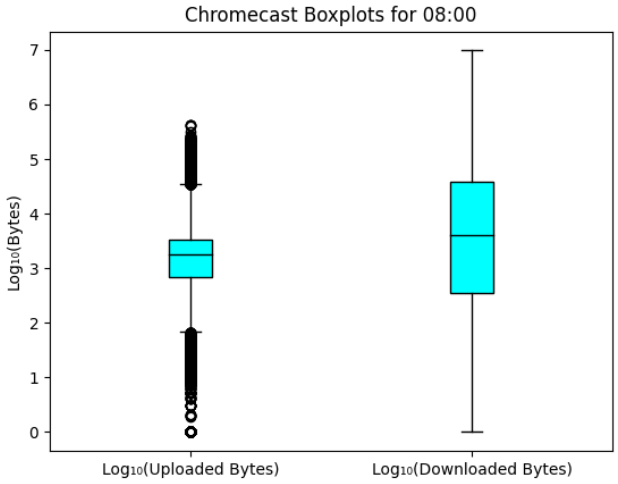

Figura 19: Boxplot de $Log10$(taxa de download e upload) para o Chromecast na Hora 08:00

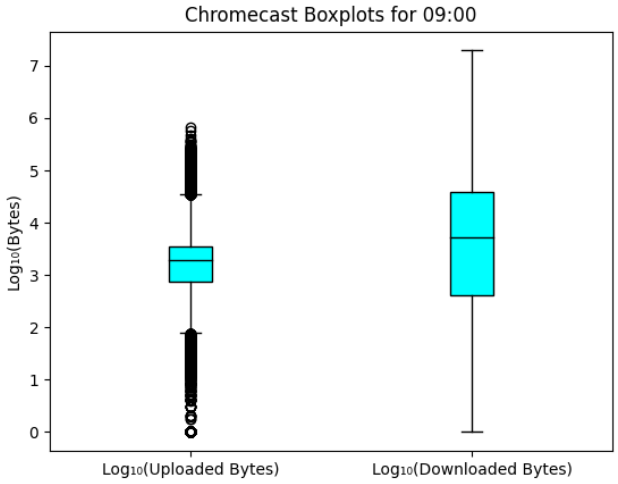

Figura 20: Boxplot de $Log10$(taxa de download e upload) para o Chromecast na Hora 09:00

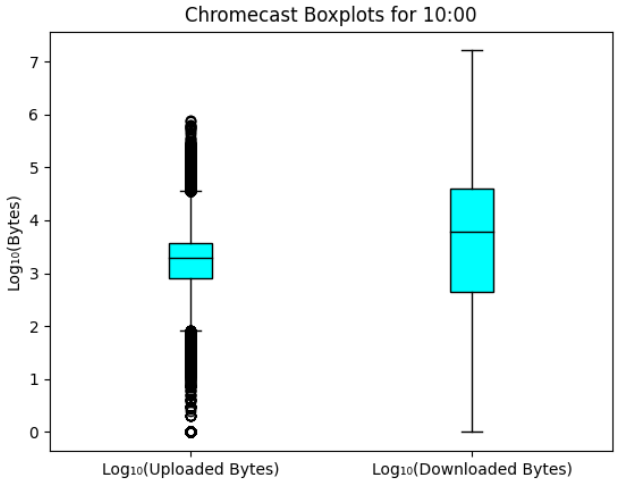

Figura 21: Boxplot de $Log10$(taxa de download e upload) para o Chromecast na Hora 10:00

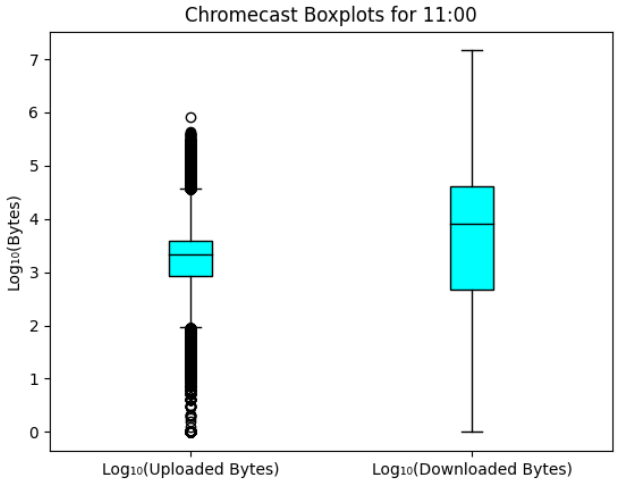

Figura 22: Boxplot de $Log10$(taxa de download e upload) para o Chromecast na Hora 11:00

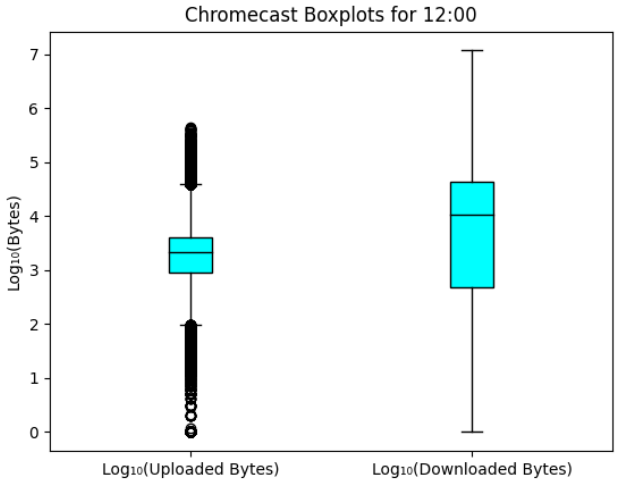

Figura 23: Boxplot de $Log10$(taxa de download e upload) para o Chromecast na Hora 12:00

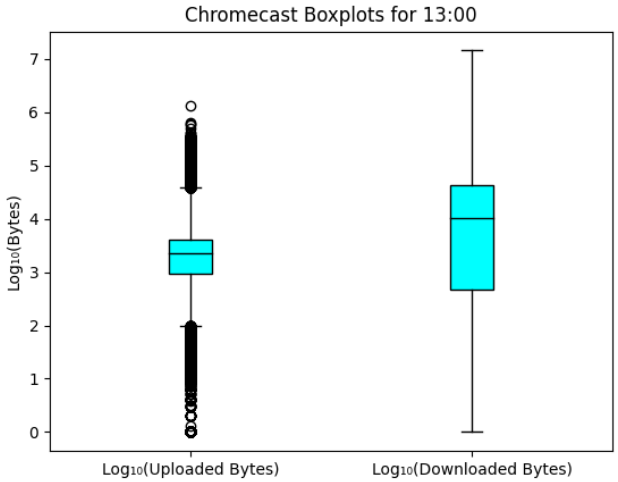

Figura 24: Boxplot de $Log10$(taxa de download e upload) para o Chromecast na Hora 13:00

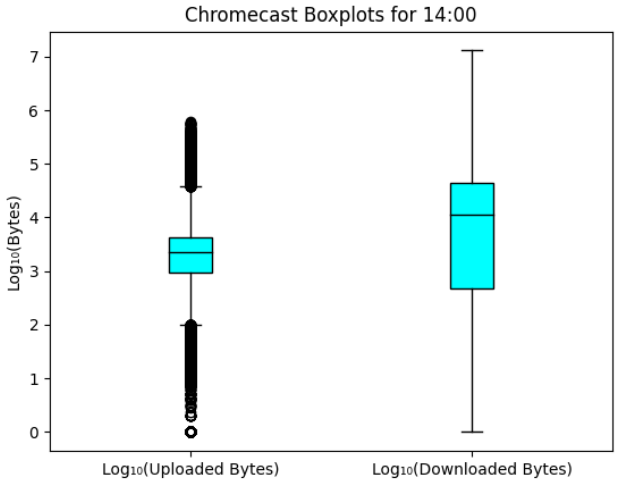

Figura 25: Boxplot de $Log10$(taxa de download e upload) para o Chromecast na Hora 14:00

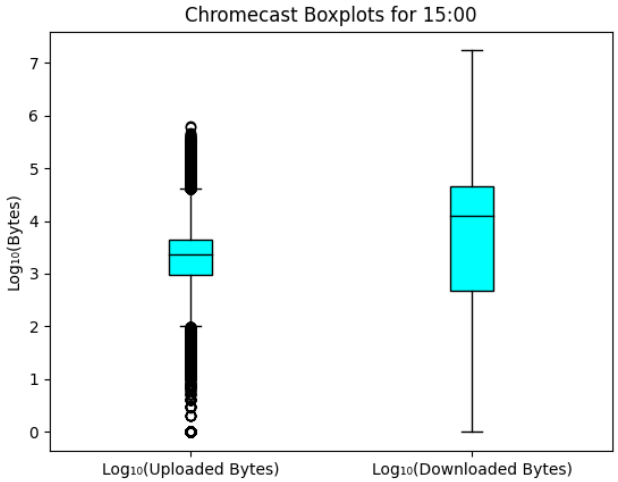

Figura 26: Boxplot de $Log10$(taxa de download e upload) para o Chromecast na Hora 15:00

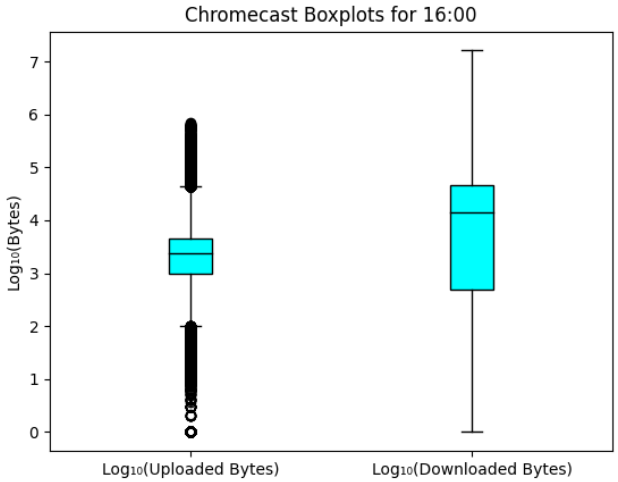

Figura 27: Boxplot de $Log10$(taxa de download e upload) para o Chromecast na Hora 16:00

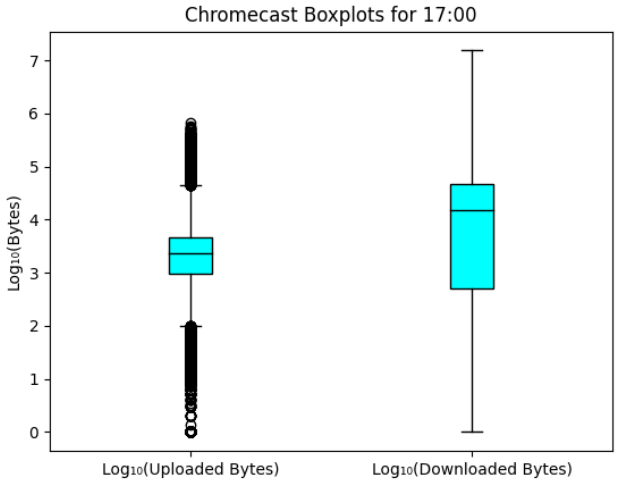

Figura 28: Boxplot de $Log10$(taxa de download e upload) para o Chromecast na Hora 17:00

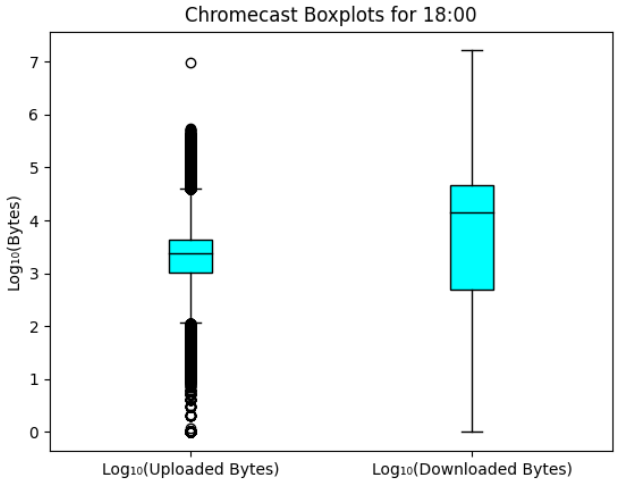

Figura 29: Boxplot de $Log10$(taxa de download e upload) para o Chromecast na Hora 18:00

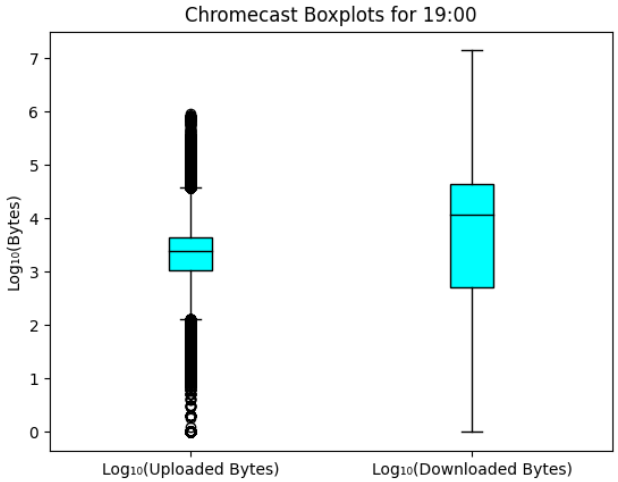

Figura 30: Boxplot de $Log10$(taxa de download e upload) para o Chromecast na Hora 19:00

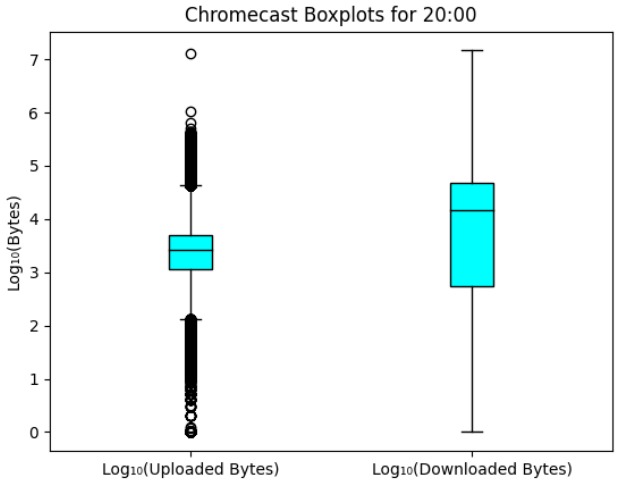

Figura 31: Boxplot de $Log10$(taxa de download e upload) para o Chromecast na Hora 20:00

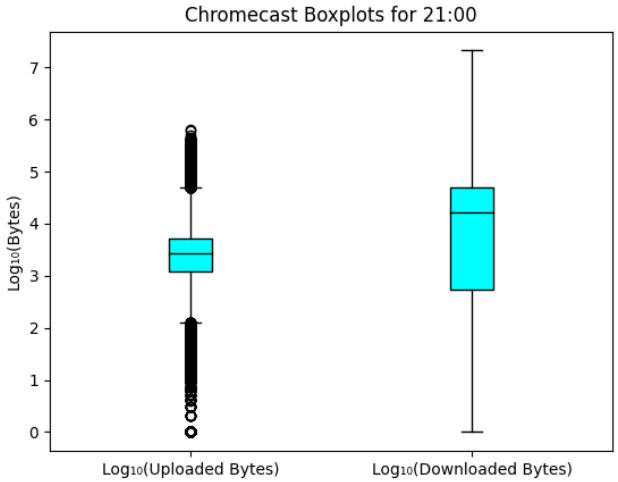

Figura 32: Boxplot de $Log10$(taxa de download e upload) para o Chromecast na Hora 21:00

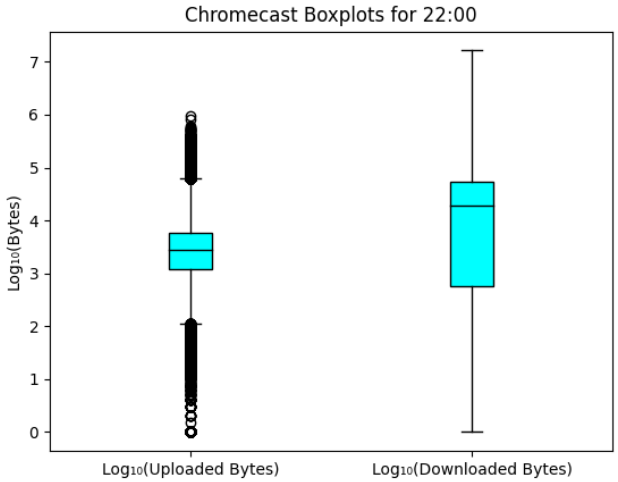

Figura 33: Boxplot de $Log10$(taxa de download e upload) para o Chromecast na Hora 22:00

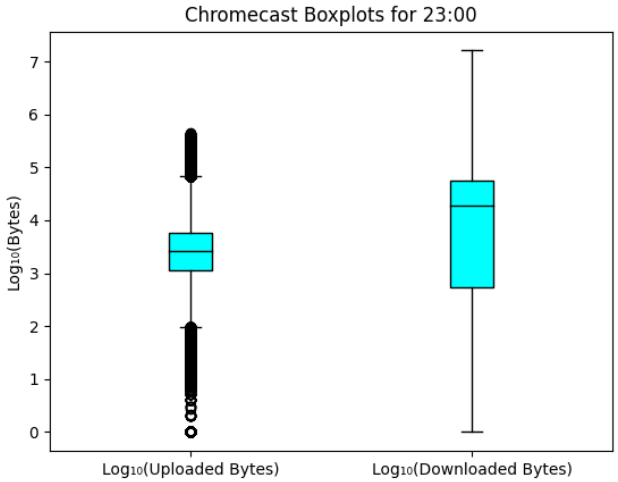

Figura 34: Boxplot de $Log10$(taxa de download e upload) para o Chromecast na Hora 23:00

2.1.2 Smart-TV

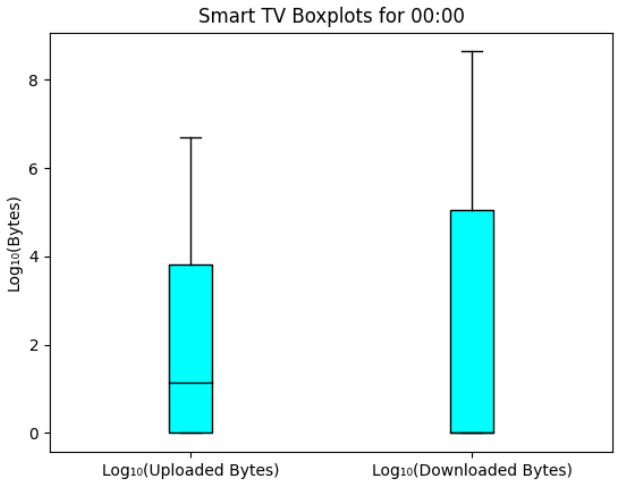

Figura 35: Boxplot de $Log10$(taxa de download e upload) para a Smart-TV na Hora 00:00

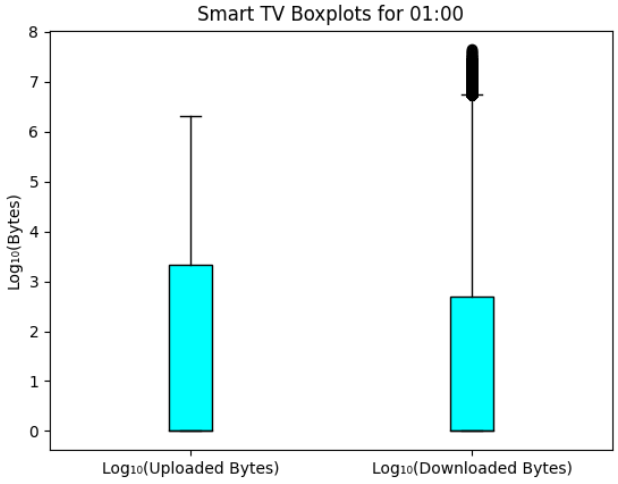

Figura 36: Boxplot de $Log10$(taxa de download e upload) para a Smart-TV na Hora 01:00

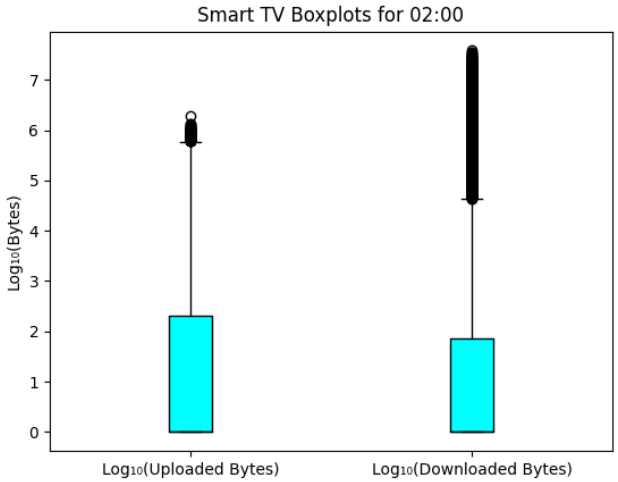

Figura 37: Boxplot de $Log10$(taxa de download e upload) para a Smart-TV na Hora 02:00

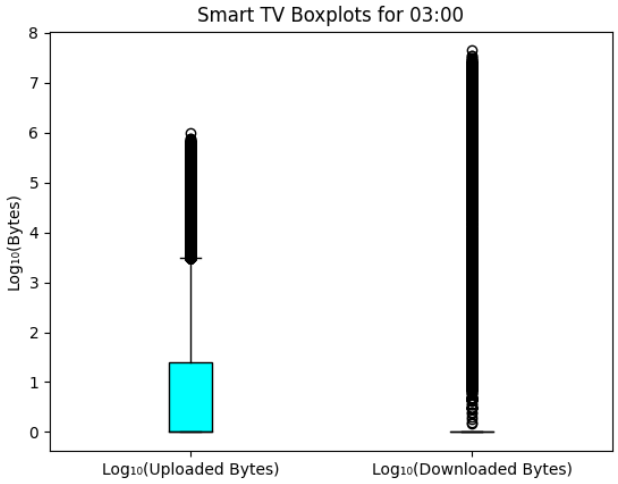

Figura 38: Boxplot de $Log10$(taxa de download e upload) para a Smart-TV na Hora 03:00

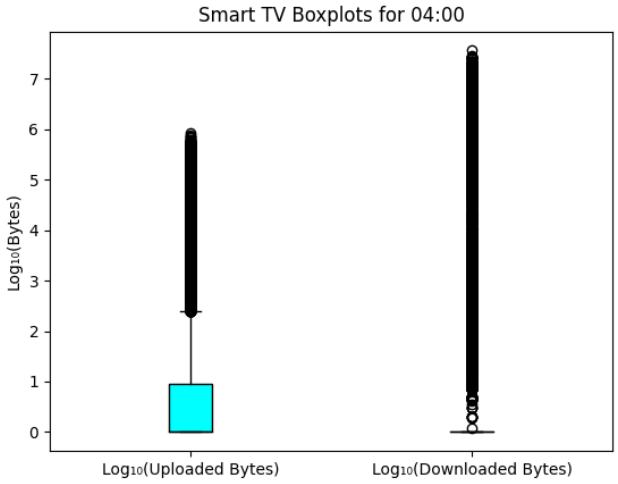

Figura 39: Boxplot de $Log10$(taxa de download e upload) para a Smart-TV na Hora 04:00

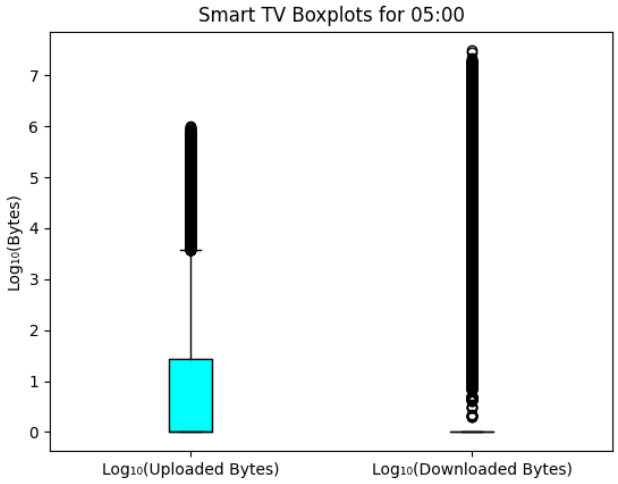

Figura 40: Boxplot de $Log10$(taxa de download e upload) para a Smart-TV na Hora 05:00

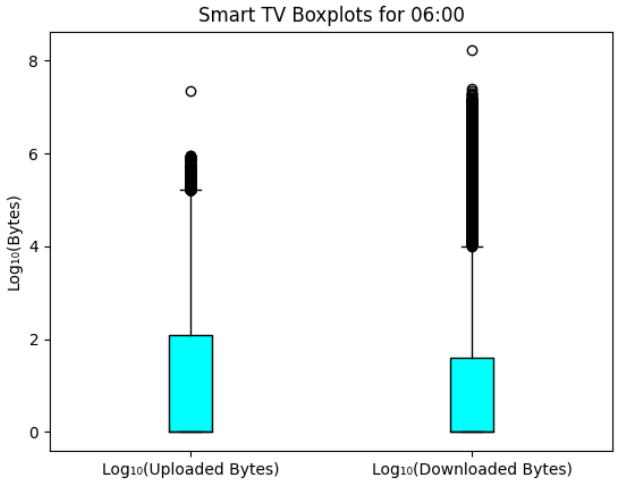

Figura 41: Boxplot de $Log10$(taxa de download e upload) para a Smart-TV na Hora 06:00

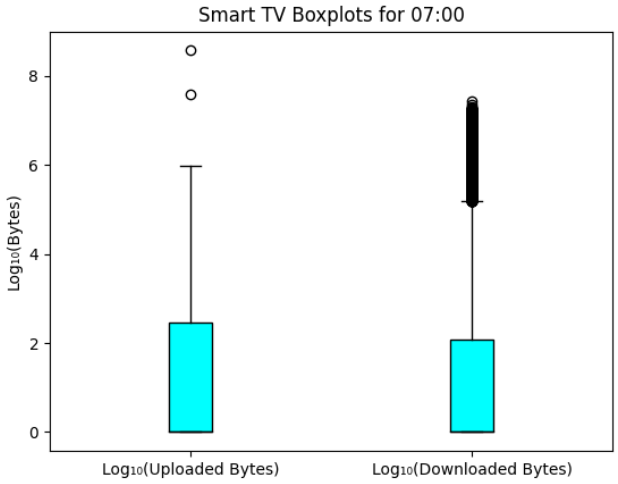

Figura 42: Boxplot de $Log10$(taxa de download e upload) para a Smart-TV na Hora 07:00

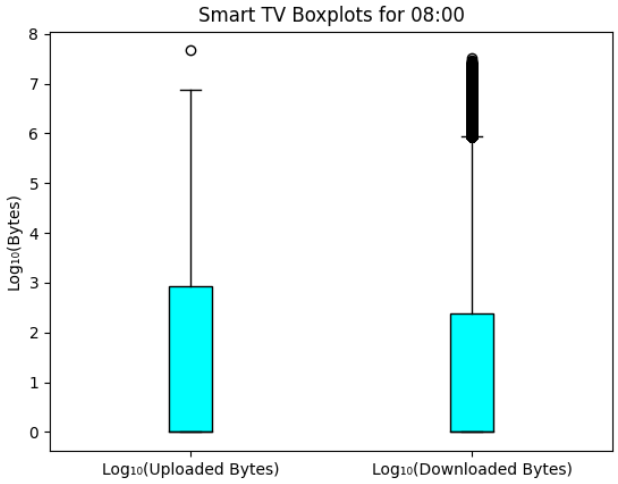

Figura 43: Boxplot de $Log10$(taxa de download e upload) para a Smart-TV na Hora 08:00

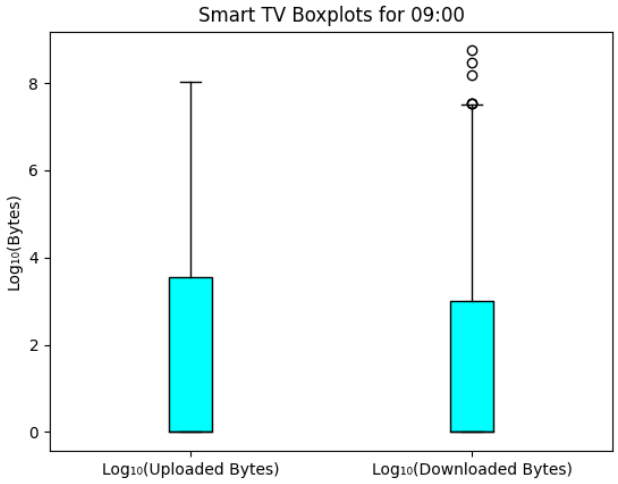

Figura 44: Boxplot de $Log10$(taxa de download e upload) para a Smart-TV na Hora 09:00

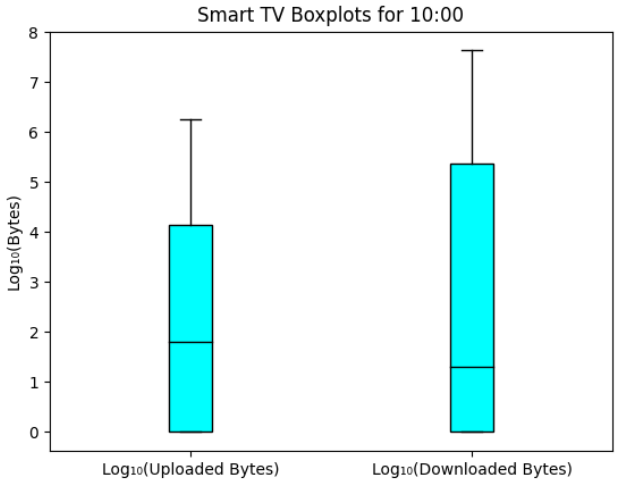

Figura 45: Boxplot de $Log10$(taxa de download e upload) para a Smart-TV na Hora 10:00

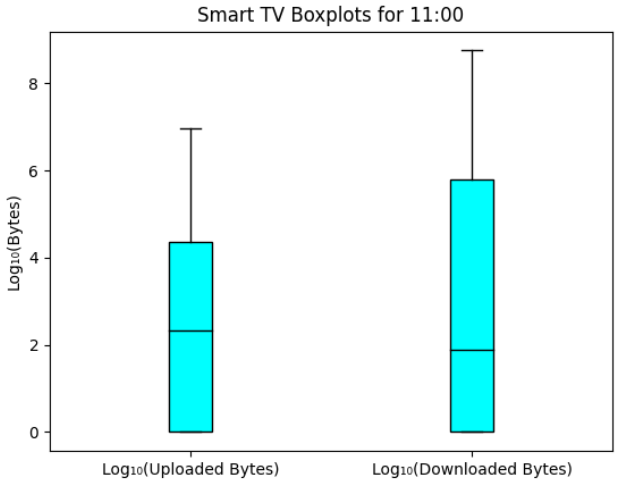

Figura 46: Boxplot de $Log10$(taxa de download e upload) para a Smart-TV na Hora 11:00

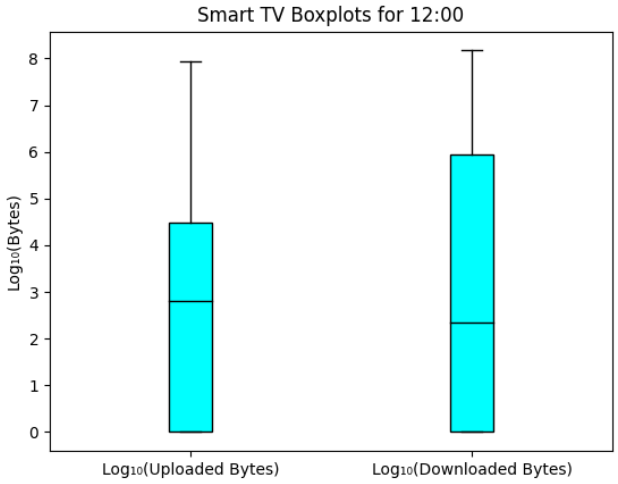

Figura 47: Boxplot de $Log10$(taxa de download e upload) para a Smart-TV na Hora 12:00

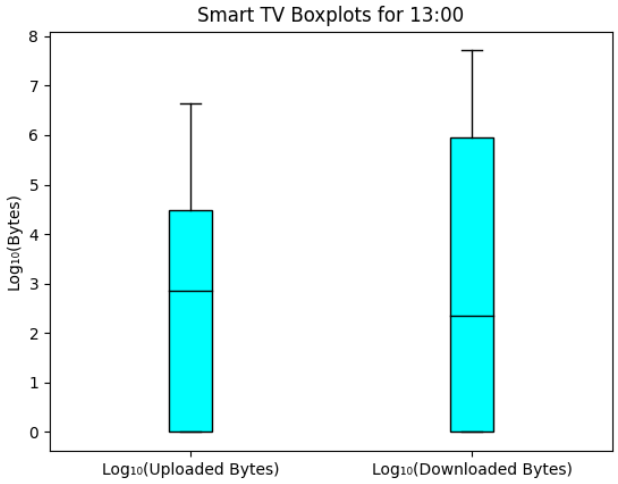

Figura 48: Boxplot de $Log10$(taxa de download e upload) para a Smart-TV na Hora 13:00

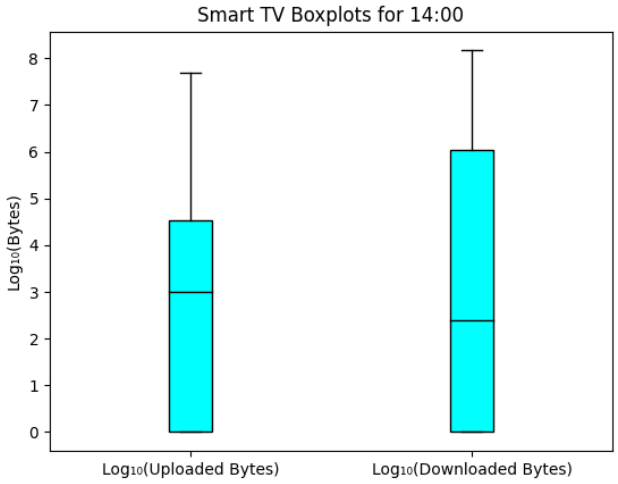

Figura 49: Boxplot de $Log10$(taxa de download e upload) para a Smart-TV na Hora 14:00

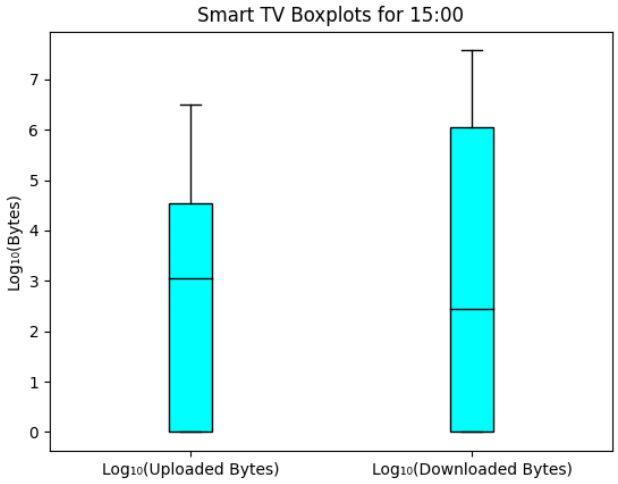

Figura 50: Boxplot de $Log10$(taxa de download e upload) para a Smart-TV na Hora 15:00

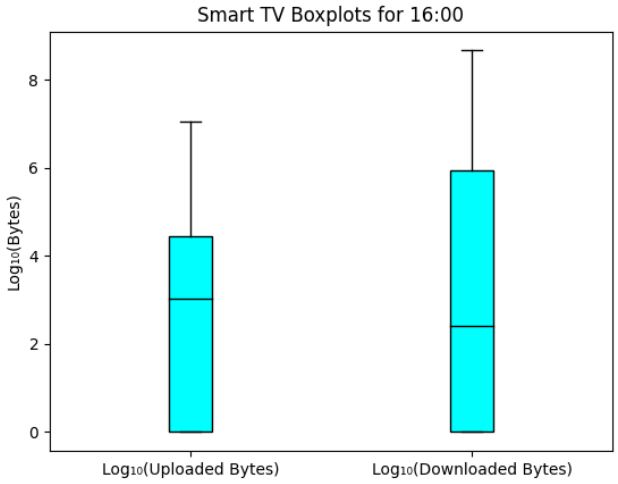

Figura 51: Boxplot de $Log10$(taxa de download e upload) para a Smart-TV na Hora 16:00

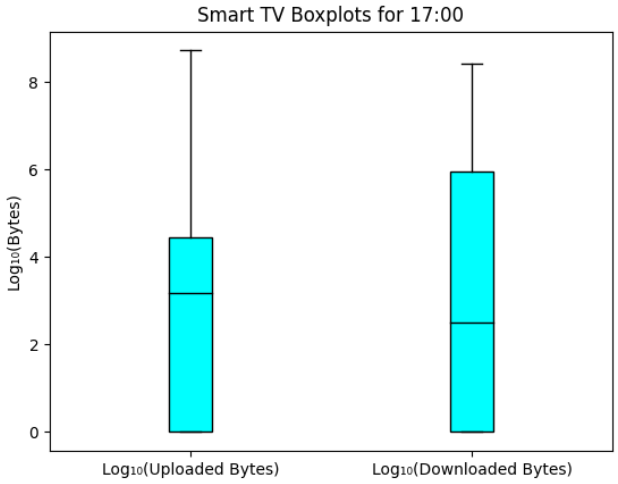

Figura 52: Boxplot de $Log10$(taxa de download e upload) para a Smart-TV na Hora 17:00

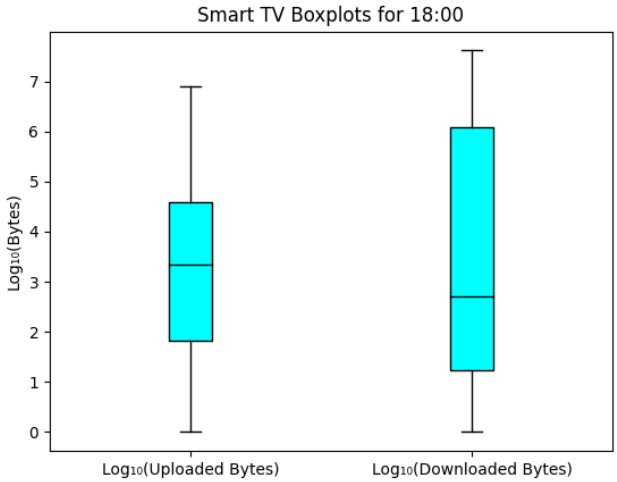

Figura 53: Boxplot de $Log10$(taxa de download e upload) para a Smart-TV na Hora 18:00

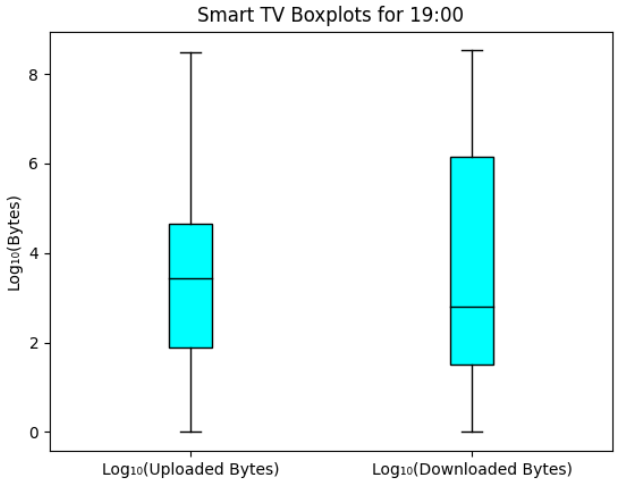

Figura 54: Boxplot de $Log10$(taxa de download e upload) para a Smart-TV na Hora 19:00

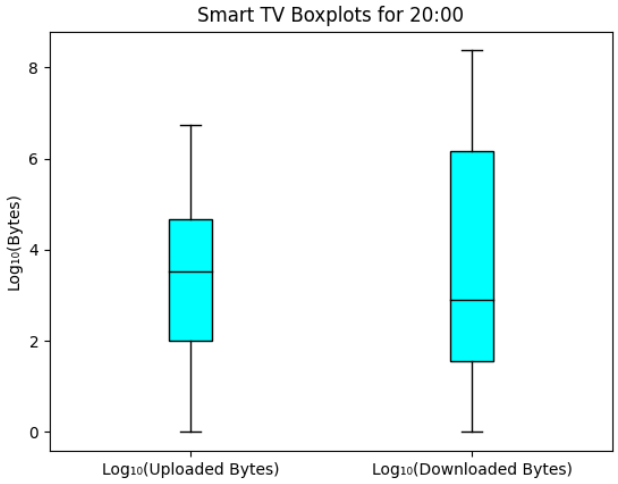

Figura 55: Boxplot de $Log10$(taxa de download e upload) para a Smart-TV na Hora 20:00

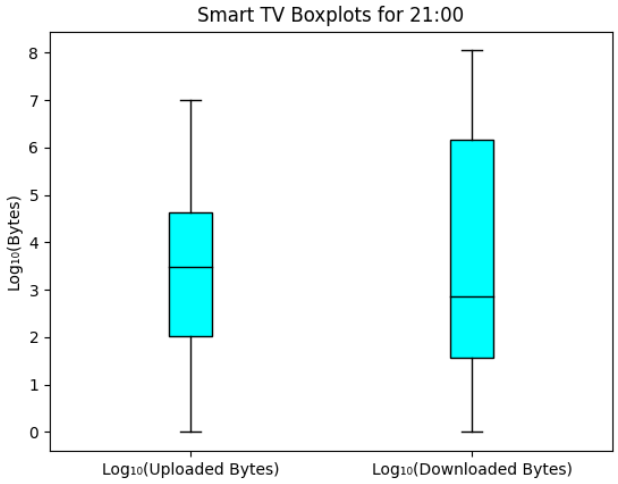

Figura 56: Boxplot de $Log10$(taxa de download e upload) para a Smart-TV na Hora 21:00

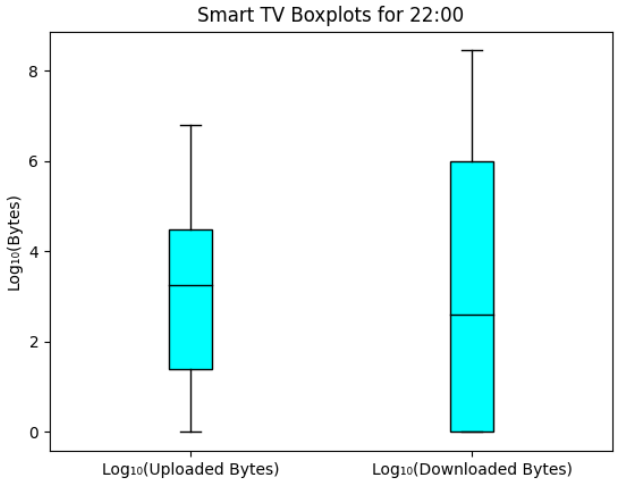

Figura 57: Boxplot de $Log10$(taxa de download e upload) para a Smart-TV na Hora 22:00

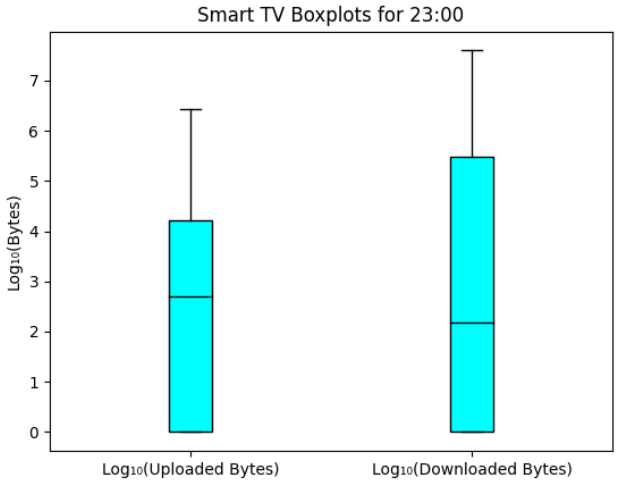

Figura 58: Boxplot de $Log10$(taxa de download e upload) para a Smart-TV na Hora 23:00

2.2 Análise Estatística

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

Desta forma, foram gerados os gráficos abaixo.

2.2.1 Chromecast

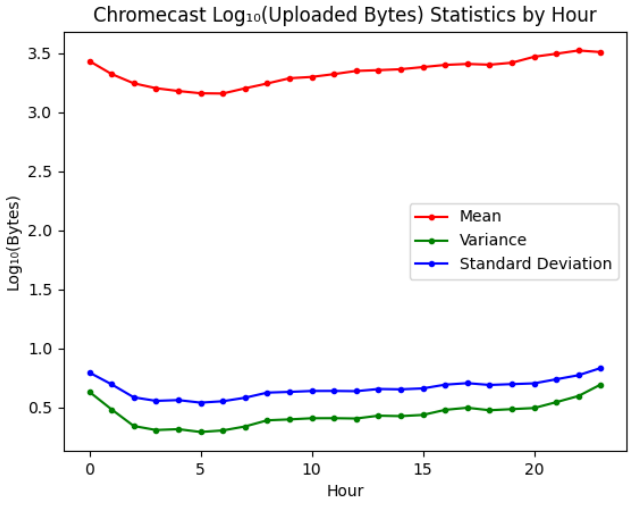

Figura 59: Gráfico de Média, Variância e Desvio padrão para o log_{10}(taxa de upload) Chromecast x Hora

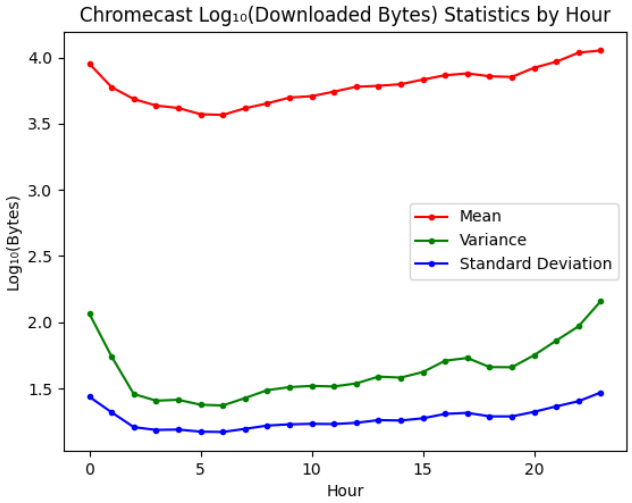

Figura 60: Gráfico de Média, Variância e Desvio padrão para o log_{10}(taxa de download) Chromecast x Hora

2.2.2 Smart-TV

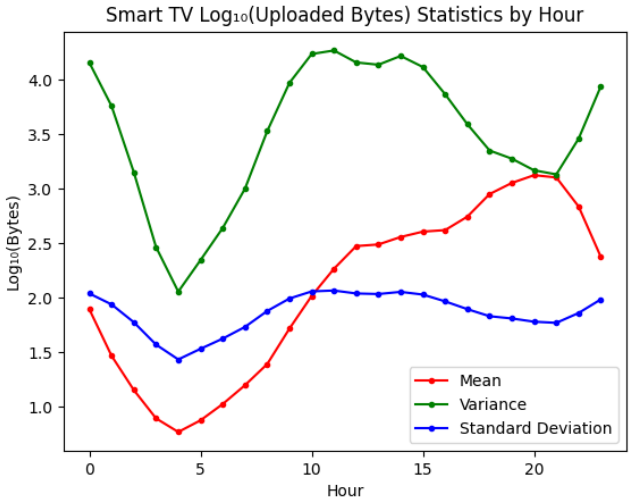

Figura 61: Gráfico de Média, Variância e Desvio padrão para o log_{10}(taxa de upload) Smart-TV x Hora

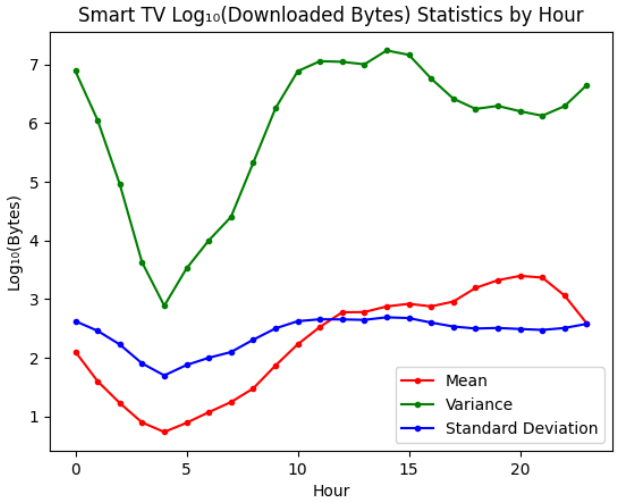

Figura 62: Gráfico de Média, Variância e Desvio padrão para o log_{10}(taxa de download) Smart-TV x Hora

2.3 Análise de resultados

Durante a análise dos resultados foi possível perceber, para os box plots da taxa de upload do chromecast que todos possuem outliers, porém, para o boxplot de upload para 22h e 23h possuem um menor número de outliers na parte superior. Já para Downloads, também para o chromecast, foi possível perceber que o outlier da parte inferior está localizado exatamente no horário de 23h.

Para a análise estatística do chromecast, foi possível perceber que certos horários possuem uma média com um maior desvio padrão em horários de 20:00 às 03:00, tal fato pode ser porque a utilização do dispositivo nestes intervalos de tempo possa ser mais baixa. Além disso, o chromecast deixou evidente que seu comportamento é contínuo, deixando a entender que ele está sempre com uma alta taxa de download e upload diferentemente da smart-tv. Este segundo dispositivo, tende a agir mais conforme a utilização do usuário pois em horários de acesso comum (por volta de 10 às 20h) a média de download tende a aumentar proporcionalmente com a taxa de upload.

3 Caracterizando os horários com maior valor de tráfego

3.1 Horas escolhidas

3.1.1 Chromecast

Metrics	Hour of Max Median	Hour of Max Mean
Log_{10} **(Uploaded Bytes)**	22	22
Log_{10} **(Downloaded Bytes)**	23	23

Tabela 3: Hora de mediana e média máxima para o Chromecast

3.1.2 Smart-TV

Metrics	Hour of Max Median	Hour of Max Mean
Log_{10} **(Uploaded Bytes)**	20	20
Log_{10} **(Downloaded Bytes)**	20	20

Tabela 4: Hora de mediana e média máxima para a Smart-TV

3.2 Histograma

Para a geração os histogramas foram mantidos quase todos os parâmetros dos histogramas gerados anteriorment. A única diferença são os dataframes que passamos como parâmetros, no qual estes dataframes são os gerados na etapa anterior, sendo filtrados pelas horas de maior média e mediana para download e upload.

Assim, foram gerados os gráficos abaixo.

3.2.1 Chromecast

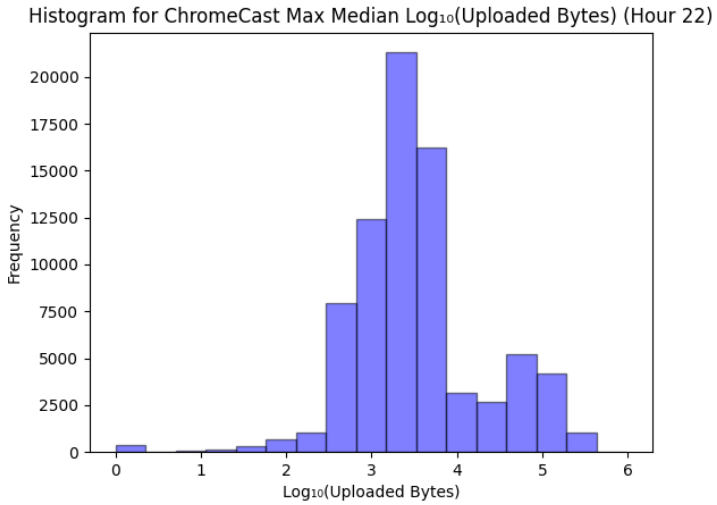

Figura 63: Histograma de $Log10$(taxa de upload) na hora de maior mediana para o Chromecast

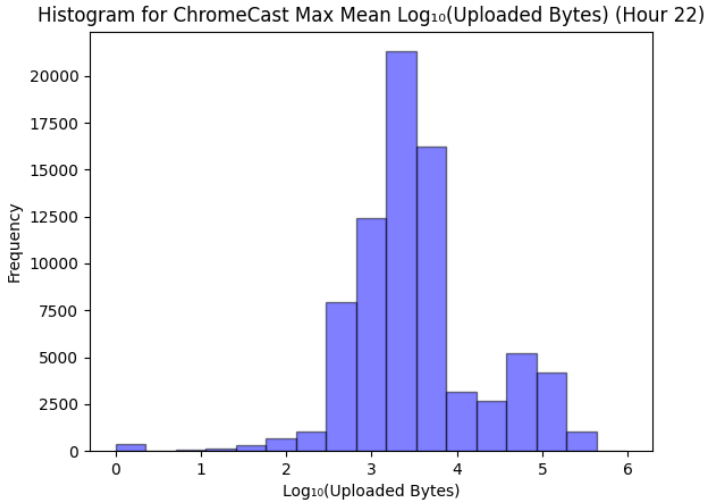

Figura 64: Histograma de $Log10$(taxa de upload) na hora de maior média para o Chromecast

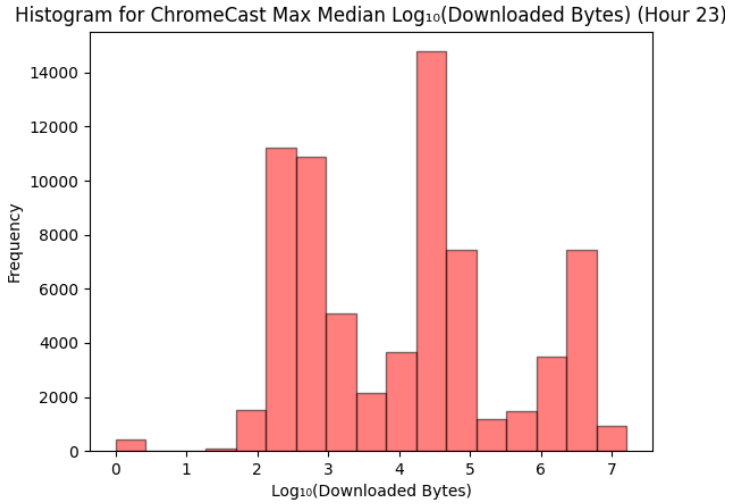

Figura 65: Histograma de $Log10$(taxa de download) na hora de maior mediana para o Chromecast

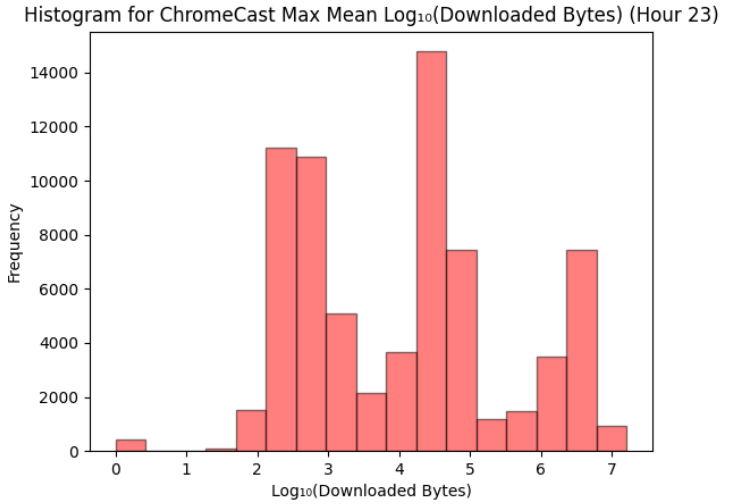

Figura 66: Histograma de $Log10$(taxa de download) na hora de maior média para o Chromecast

3.2.2 Smart-TV

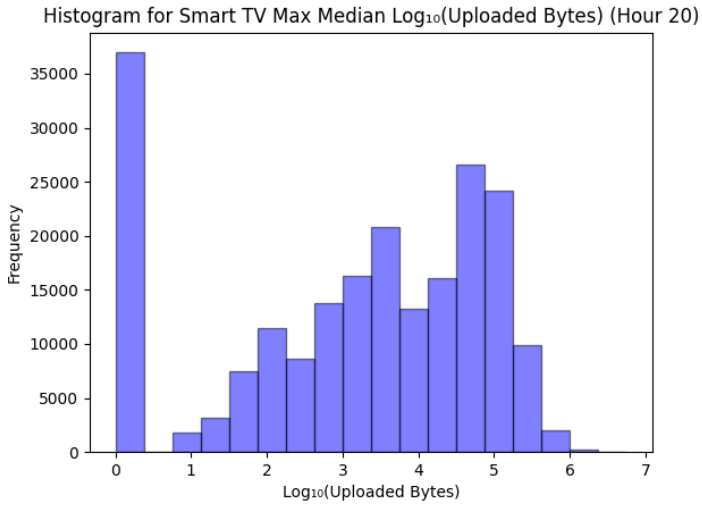

Figura 67: Histograma de $Log10$(taxa de upload) na hora de maior mediana para a Smart-TV

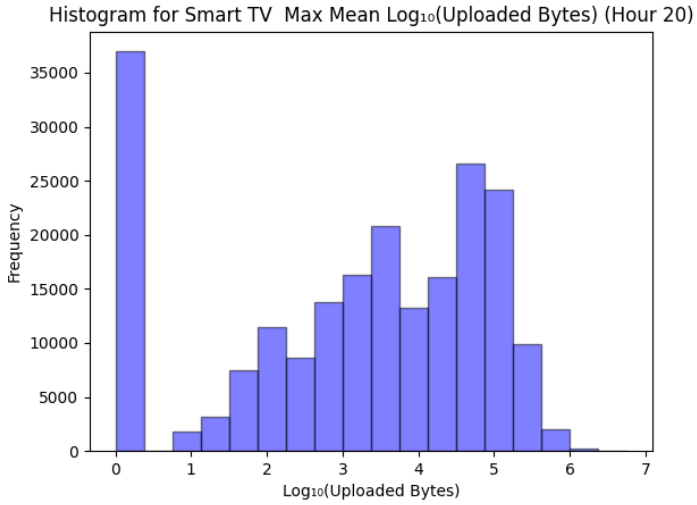

Figura 68: Histograma de $Log10$(taxa de upload) na hora de maior média para a Smart-TV

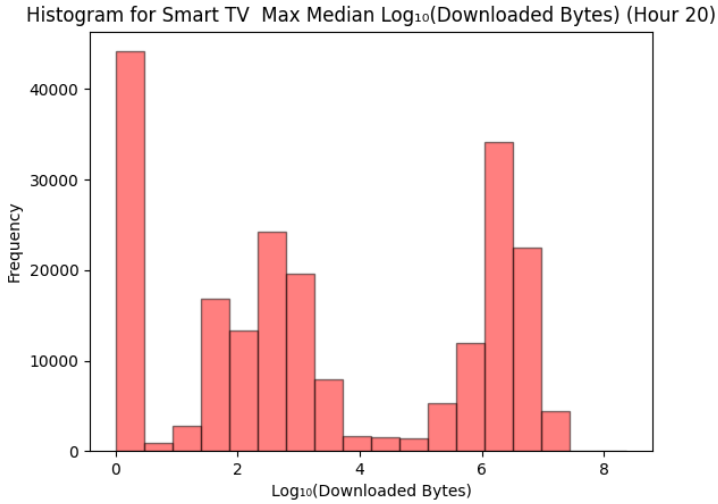

Figura 69: Histograma de $Log10$(taxa de download) na hora de maior mediana para a Smart-TV

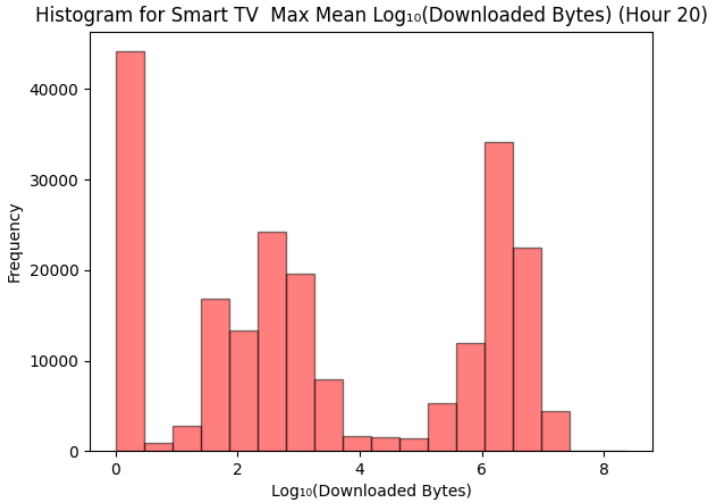

Figura 70: Histograma de $Log10$(taxa de download) na hora de maior média para a Smart-TV

3.3 MLE

Para calcular o MLE das distribuições de interesse, é possível que obtenhamos isto matematicamente. Para isso basta obter a função likelihood da distribuição e para obter o maximum likelihood, basta derivar e igualar a 0, a fim de obter as variáveis que maximizem esta nova função.

Também é possível aplicar log a fim de facilitar estes cálculos e, como as duas funções abaixo são funções de duas variáveis, é necessário que façamos uma derivada parcial a fim de obter o valor que maximize a função para os dois parâmetros.

Porém, a linguagem de programação utilizada já possuem bibliotecas prontas que facilitam o processo de cálculo e estas foram utilizadas para facilitar o processo de análise destes dados.

3.3.1 Gamma

Para a distribuição gamma, foi utilizado o método gamma da biblioteca *scipy.stats*, desta forma é possível obter os parâmetros que maximizem a função gamma para os parâmetros das colunas de interesse. A função gamma possui como parâmetros o shape e o scale, porém a função utilizada nos retorna 3 parâmetros, o parâmetro a mais é o loc, que define a localização da distribuição, deslocando para a direita ou esquerda conforme o valor passado como parâmetro.

Desta forma, foram obtidos os seguintes valores:

3.3.1.1 Chromecast

Metric	**shape**	**loc**	**scale**	**Hour**
Max Median Log10(Uploaded Bytes)	3148.88	-39.809	0.0137607	22
Max Median Log10(Downloaded Bytes)	27.1301	-3.63137	0.28323	23
Max Mean Log10(Uploaded Bytes)	3148.88	-39.809	0.0137607	22
Max Mean Log10(Downloaded Bytes)	27.1301	-3.63137	0.28323	23

Tabela 5: Tabela com o MLE para a distribuição Gamma para o Chromecast

3.3.1.2 Smart-TV

Metric	shape	loc	scale	hour
Max Median Log10(Uploaded Bytes)	217.147	-23.8596	0.124245	20
Max Median Log10(Downloaded Bytes)	896.547	-71.0622	0.00830499	20
Max Mean Log10(Uploaded Bytes)	217.147	-23.8596	0.124245	20
Max Mean Log10(Downloaded Bytes)	896.547	-71.0622	0.00830499	20

Tabela 6: Tabela com o MLE para a distribuição Gamma para a Smart-TV

3.3.2 Gaussiana

Para a distribuição gaussiana, foi utilizado o próprio cálculo da média e da gaussiana para a coluna passada como parâmetro, desta forma, é possível realizar a descoberta do MLE sem a utilização de alguma biblioteca auxiliar.

Desta forma, foram obtidos os seguintes valores:

3.3.2.1 Chromecast

Metric	mean	median	hour
Max Median Log10(Uploaded Bytes)	3.52155	3.4438	22
Max Median Log10(Downloaded Bytes)	4.0527	4.28566	23
Max Mean Log10(Uploaded Bytes)	3.52155	3.4438	22
Max Mean Log10(Downloaded Bytes)	4.0527	4.28566	23

Tabela 7: Tabela com o MLE para a distribuição Gaussiana para o Chromecast

3.3.2.2 Smart-TV

Metric	mean	median	hour
Max Median Log10(Uploaded Bytes)	3.12426	3.53052	20
Max Median Log10(Downloaded Bytes)	3.39609	2.88961	20
Max Mean Log10(Uploaded Bytes)	3.12426	3.53052	20
Max Mean Log10(Downloaded Bytes)	3.39609	2.88961	20

Tabela 8: Tabela com o MLE para a distribuição Gaussiana para o Chromecast

3.4 Gráfico com todas as curvas + Histograma

Para plotar o histograma junto dos gráficos, tendo em vista a diferença no eixo y, foi utilizado o parâmetro *density* do método *hist* da biblioteca matplotlib a fim de normalizar a altura do histograma com o restante das distribuições.

Com isso, foi possível gerar as seguintes visualizações.

3.4.1 Chromecast

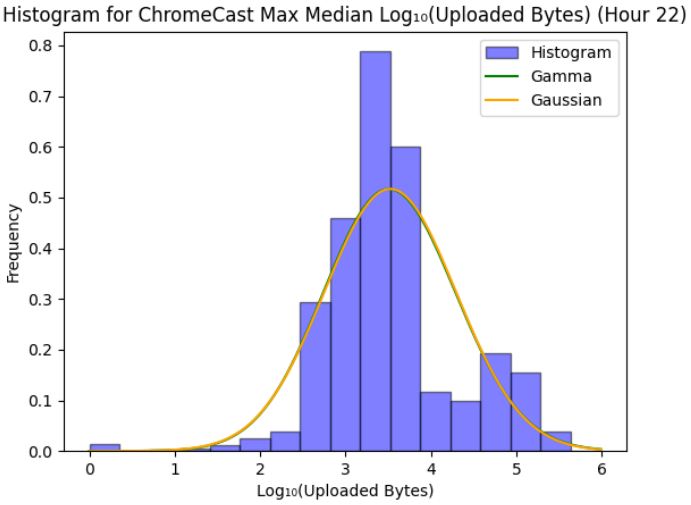

Figura 71: Histograma + distribuições de $Log10$(taxa de upload) na hora de maior mediana para a Chromecast

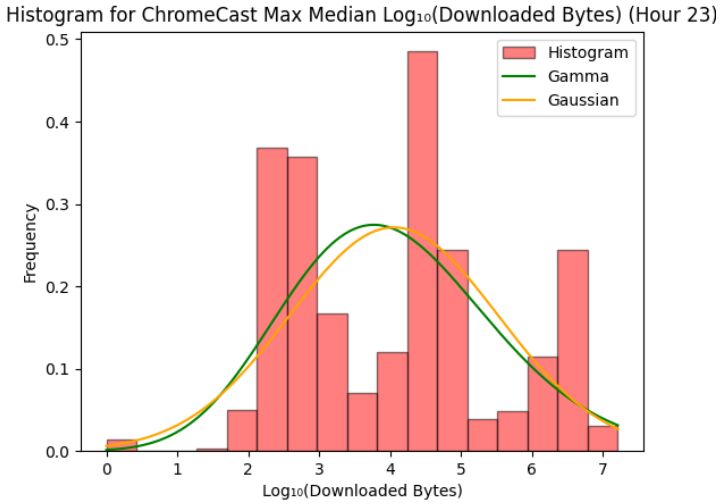

Figura 72: Histograma + distribuições de $Log10$(taxa de download) na hora de maior mediana para a Chromecast

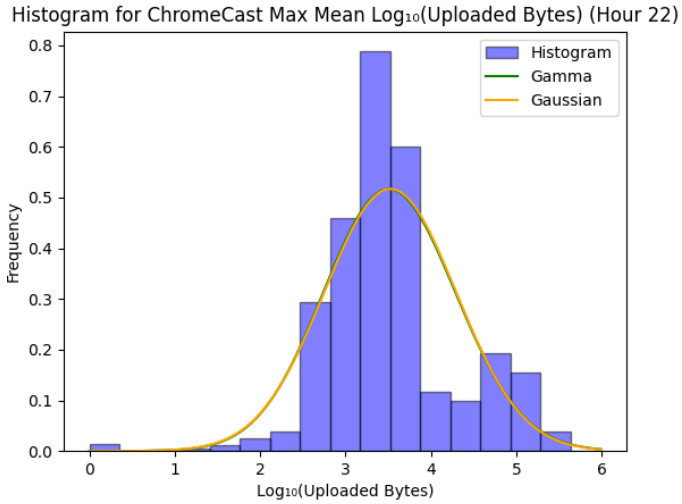

Figura 73: Histograma + distribuições de $Log10$(taxa de upload) na hora de maior média para a Chromecast

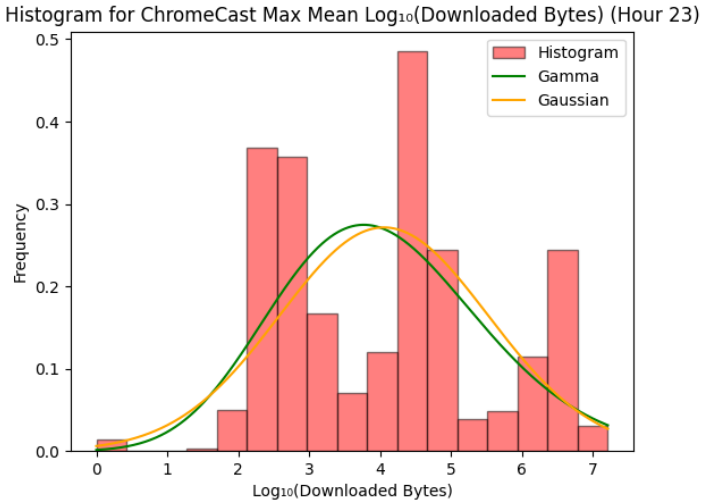

Figura 74: Histograma + distribuições de $Log10$(taxa de download) na hora de maior média para a Chromecast

3.4.2 Smart-TV

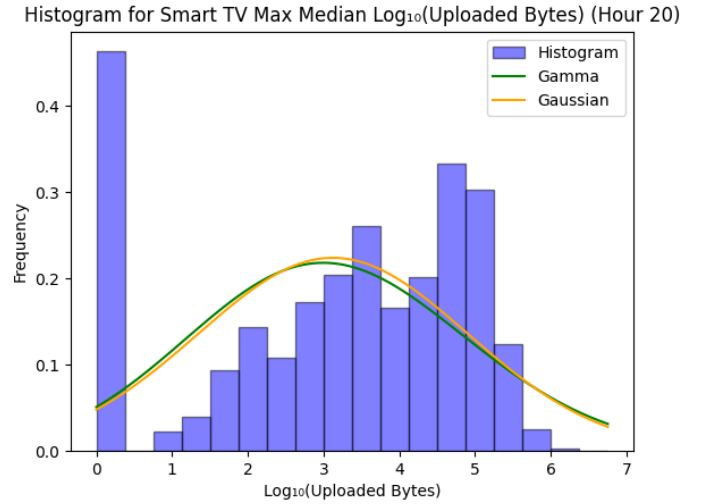

Figura 75: Histograma + distribuições de $Log10$(taxa de upload) na hora de maior mediana para a Smart-TV

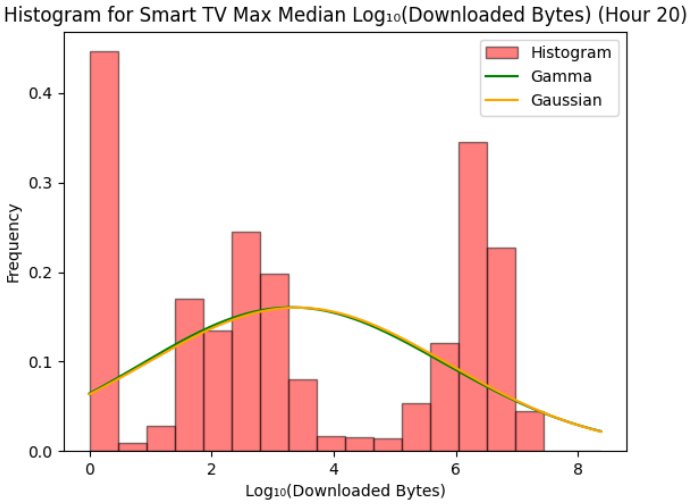

Figura 76: Histograma + distribuições de $Log10$(taxa de download) na hora de maior mediana para a Smart-TV

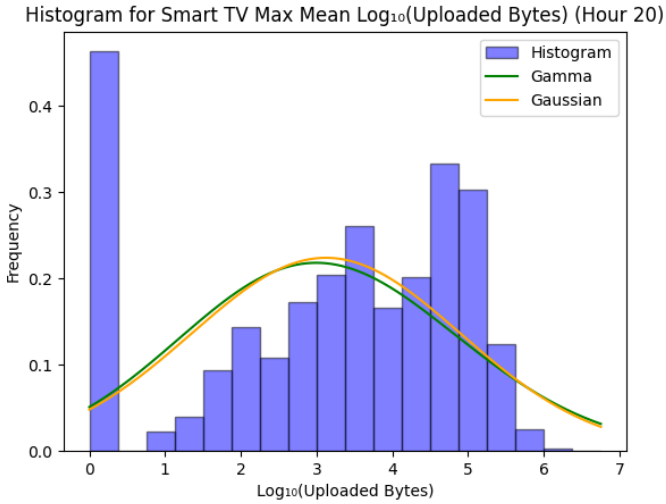

Figura 77: Histograma + distribuições de $Log10$(taxa de upload) na hora de maior média para a Smart-TV

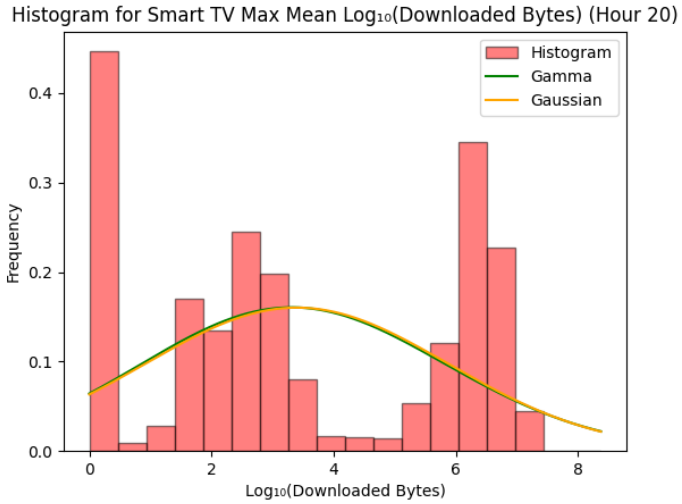

Figura 78: Histograma + distribuições de $Log10$(taxa de download) na hora de maior média para a Smart-TV

3.5 Probability Plot

Para gerar o probability plot foi usado o método probplot de cada uma das distribuições cujo os parâmetros são os MLEs das distribuições gamma e gaussiana.

3.5.1 Chromecast

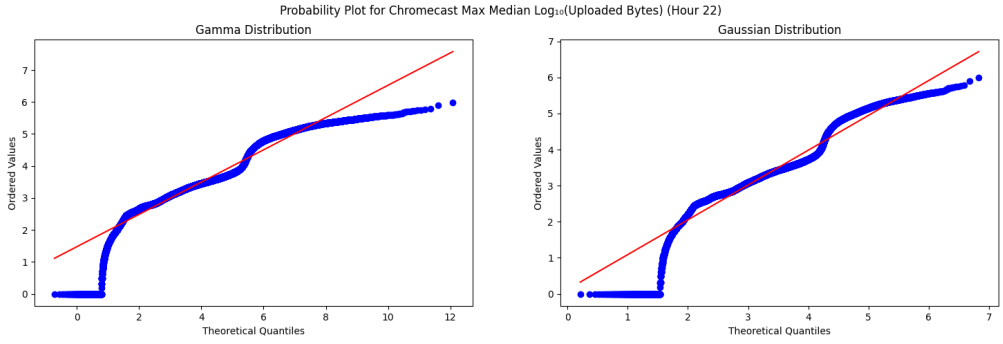

Figura 79: Probability Plot de $Log10$(taxa de upload) na hora de maior mediana para o Chromecast

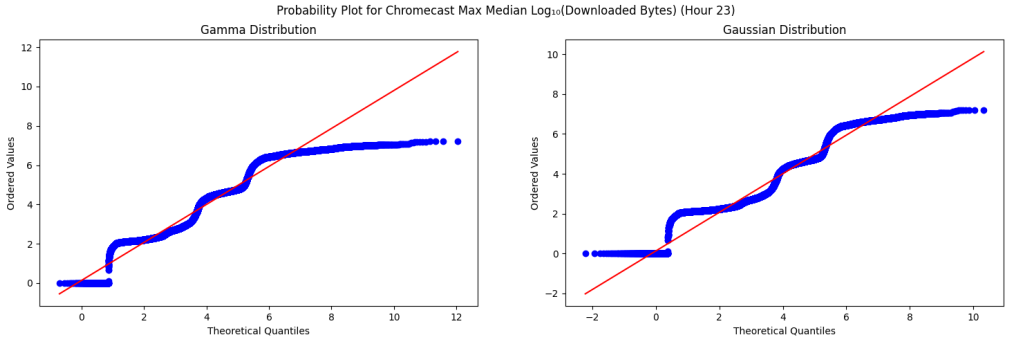

Figura 80: Probability Plot de $Log10$(taxa de download) na hora de maior mediana para o Chromecast

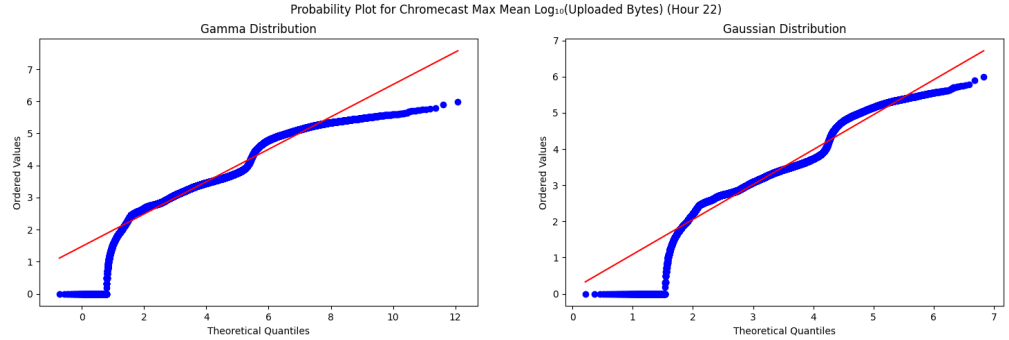

Figura 81: Probability Plot de $Log10$(taxa de upload) na hora de maior média para o Chromecast

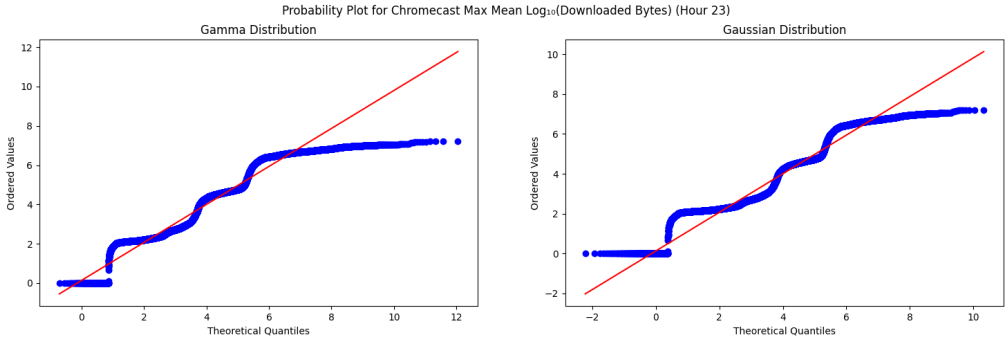

Figura 82: Probability Plot de $Log10$(taxa de download) na hora de maior média para o Chromecast

3.5.2 Smart-TV

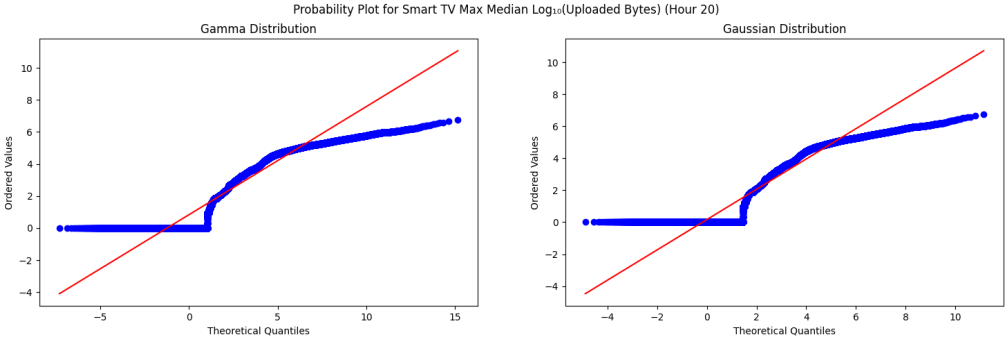

Figura 83: Probability Plot de $Log10$(taxa de upload) na hora de maior mediana para a Smart-TV

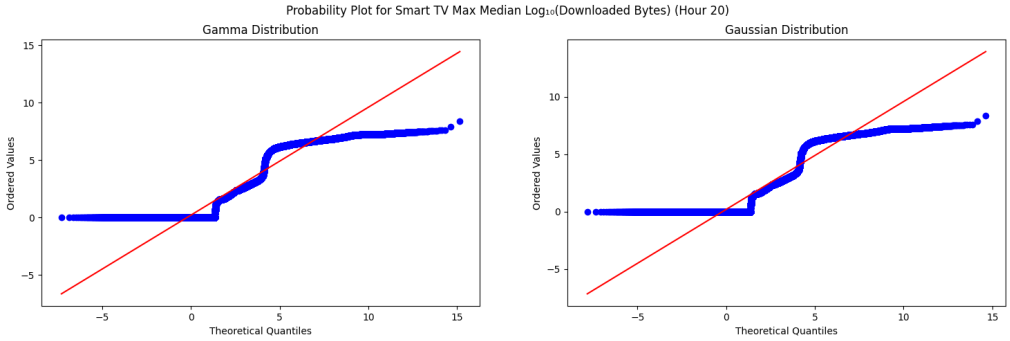

Figura 84: Probability Plot de $Log10$(taxa de download) na hora de maior mediana para a Smart-TV

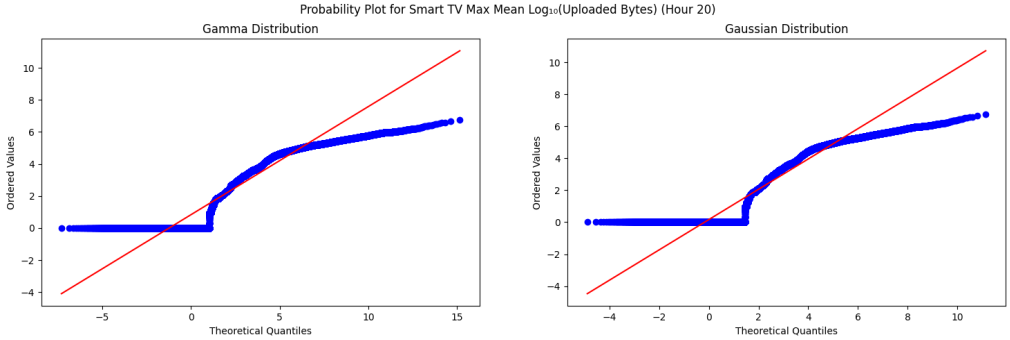

Figura 85: Probability Plot de $Log10$(taxa de upload) na hora de maior média para a Smart-TV

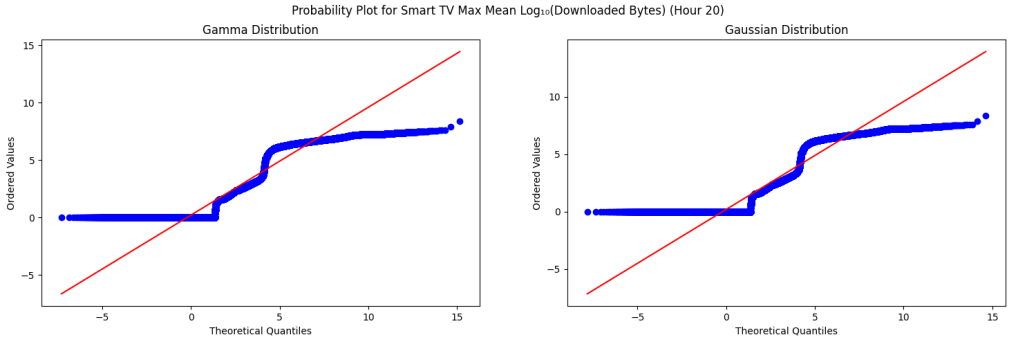

Figura 86: Probability Plot de $Log10$(taxa de download) na hora de maior média para a Smart-TV

3.6 Análise de Resultados

Para a escolha dos horários, o chromecast teve seus horários escolhidos de 22:00 para a taxa de upload, enquanto para download foi o horário de 23:00, tanto para os horários de mediana máxima quanto média máxima. Já para a smart-TV, para todas as taxas em horário de média máxima ou mediana, os horários escolhidos foram de 20:00.

Tanto para a maior média, quanto para a maior mediana, nos histogramas de chromecast foram iguais, devido ao fato de ser o mesmo horário de upload. O mesmo é possível dizer para o histograma de chromecast para maior média e mediana para download. Comparando ambos entre si, é possível perceber que o chromecast possuem 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 mediana 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.

O download tem uma zona que varia entre 1 e 3,5 log_{10} de bytes baixados e tem outra zona que varia de 6 até 7 log_{10} de bytes baixados, talvez as smarts tvs possuam um sistema que, caso a internet do usuário esteja baixa eles reduzem a capacidade a fim de manter a entrega do vídeo para o usuário. Já para a taxa de upload, as frequência variam de forma crescendo, representando uma transmissão de dados de forma constante até um certo

limite, que em nosso caso é 6 log_{10} de bytes baixados.

Os datasets 1 e 2, 3 e 4, 5 e 6 e 7 e 8 tem comportamentos iguais entre si, tendo em vista que, para os horários escolhidos, a média e mediana possuem o mesmo gráfico para ambos dispositivos, desta forma não há diferença entre eles por filtrarmos somente pelo horário.

É possível dizer que o dataset que upload bytes para a maior média do chromecast pode ser mapeado tanto para uma gaussiana quanto por uma distribuição gamma. Já para os datasets de download isso não é possível afirmar.

Já para a smart-tv, ambas as distribuições utilizadas não caracterizam o dataset. O motivo se dá similarmente pelo comportamento dos histogramas de download e upload com relação a forma do gráfico das distribuições, é possível observar que a estrutura do histograma não acompanha a curva, desta forma trazendo o indício da não caracterização.

Quanto aos gráficos de Probability Plot, é possível perceber justamente o que fora comentado acima. A taxa de download do chromecast segue um pouco o comportamento da curva em azul, mostrando que é possível mapear este dataset em variáveis aleatórias da literatura.

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

4.1 Cálculo dos coeficientes de correlação

Para o cálculo de correlação foi utilizado o coeficiente de pearson para medir a direção do relacionamento entre duas variáveis, no caso, duas colunas de dois datasets distintos, no qual esta correlação pode ser medida num intervalo entre -1 e 1, no qual 1 indica uma forte correlação, -1 uma forte correlação negativa e 0 não indica relação entre as variáveis.

Vale ressaltar que para a correlação foi levado em consideração os horários de download cuja mediana era máxima e os horários de download cuja média era máxima para o dataset do chromecast, isto pois os horários de download são diferentes de upload.

Com isso, foi gerada a seguinte tabela com estas correlações:

name	pearson coeficient	p value
Smart-TV Max Median Log_{10}(Bytes)	0.915477	0
Smart-TV Max Mean Log_{10}(Bytes)	0.915477	0
Chromecast Max Median Log_{10}(Bytes)	0.791959	0
Chromecast Max Median Log_{10}(Bytes)	0.791959	0

Tabela 9: Dados de correlação entre os dispositivos

Além disso, é importante frisar que as tabelas acima ambas possuem o o $p - valor = 0$, o que significa que os resultados acima possuem uma grande significância.

4.2 Scatter Plot

Para gerar o scatter plot foi utilizado o método *scatter* também da biblioteca matplotlib, cujos parâmetros foram exatamente os dataframes exigidos pelo trabalho com relação a suas taxas de download e upload.

Tais gráficos podem ser observados abaixo.

4.2.1 Smart-TV Horário de Maior Mediana

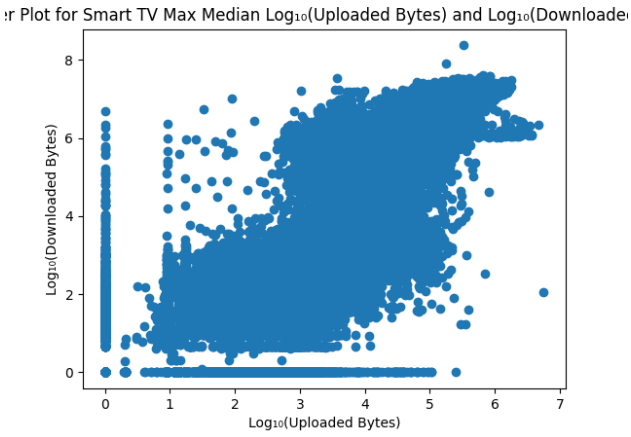

Figura 87: Scatter Plot de Log_{10}(taxa de download e upload) para a Smart-TV maior mediana

4.2.2 Smart-TV Horário de Maior Média

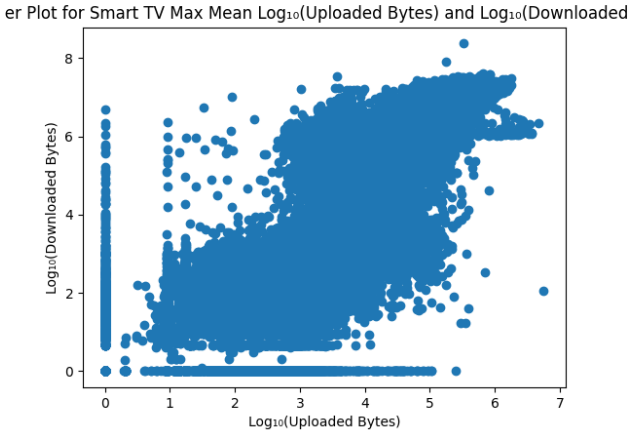

Figura 88: Scatter Plot de Log_{10}(taxa de download e upload) para a Smart-TV maior média

4.2.3 Chromecast Horário de Maior Mediana

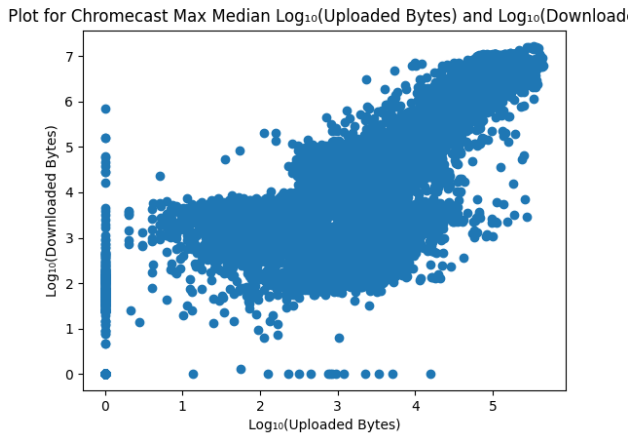

Figura 89: Scatter Plot de Log_{10}(taxa de download e upload) para o Chromecast maior mediana

4.2.4 Chromecast Horário de Maior Média

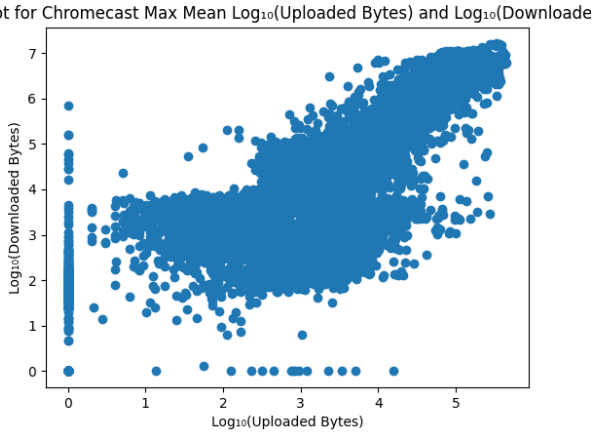

Figura 90: Scatter Plot de Log_{10}(taxa de download e upload) para o Chromecast maior média

4.3 Análise dos Resultados

Como é possível observar com os dados da tabela, podemos observar que os dados possuem sim uma correlação quanto ao dispositivo de Smart-TV, enquanto os dados de chromecast também possuem correlação entre si, ainda que a sua taxa do coeficiente de pearson seja menor que a taxa entre os dispositivos de Smart-TV.

Além disso, é importante ressaltar que houve a mudança da escolha de horários de download e upload para o Chromecast, tal fato se dá pois a correlação entre a taxa de upload e download entre dois horários destintos era inexistente, ou seja, o coeficiente de pearson dava próximo de 0.

5 Comparação dos dados gerados pelos dispositivos Smart-TV e Chromecast

Para gerar os dados do teste G, foi necessário inicialmente segmentar os dados em bins iguais a fim de conseguir comparar os dados ordenados e categorizados por cada um dos bins gerados.

Após isso, em segunda instância, foi utilizado o método *power_divergence* que serve para calcular o teste-g e o valor p para os dataframes de input.

Com isso, fori gerada a tabela abaixo.

name	g_test	p_value
smart_tv_max_median_up with chromecast_max_median_up	1.74065	0.999996
smart_tv_max_mean_up with chromecast_max_mean_up	1.74065	0.999996
smart_tv_max_median_down with chromecast_max_median_down	2.35922	0.999967
smart_tv_max_mean_down with chromecast_max_mean_down	2.35922	0.999967

Tabela 10: Comparação entre a Smart-TV e Chromecast para download e upload

Para comparar e analisar esses valores, optei por inicialmente observar o p-valor e, como podemos perceber está num valor próximo de 1. Quando tal caso acontece, podemos afirmar que o teste não possui uma grande significância, assim nada é possível afirmar por meio destes valores obtidos no g-teste.