Probabilidade e Estatística

Projeto de Curso

Engenharia de Computação e Informação UFRJ

Zuilho Rodrigues Castro Segundo 122064877

Link do Repositório: https://github.com/ZuilhoSe/Probest

1 Introdução

O presente relatório tem como propósito realizar uma análise aprofundada de um conjunto de dados fornecidos por um provedor de Internet de médio porte, com foco nas taxas de dados enviados (taxa de upload) e recebidos (taxa de download) por dois dispositivos específicos: Smart-TV e Chromecast. Este projeto visa aplicar os conceitos e teorias discutidos em sala de aula, proporcionando uma compreensão mais ampla e prática dos princípios aprendidos.

2 Tratamento dos dados

Para a leitura dos dados foi utilizada a estrutuda de dados de um pandas dataframe, que funciona muito parecido com uma tabela. Seguindo a recomendação de trabalhar com os dados em escala logaritmica, uma nova coluna foi adicionada, reescalonando os dados para logarítmo na base 10, uma para upload e outra para download. Ao aplicar a operção de logarítmo, surge o problema com os valores 0, que se tornam menos infinito. Na nova coluna, todos esses dados são reajustados para 0.

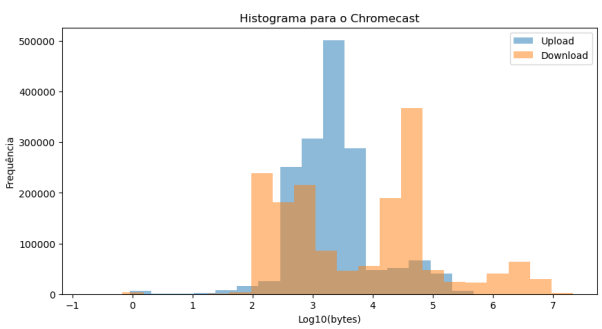
3 Estatísticas Descritivas

Nessa etapa, realizamos os cálculos das Estatísticas Gerais e focamos na construção de gráficos que retratem isso. Para os histogramas, foi utilizado o método de Sturges para definir o número de bins.

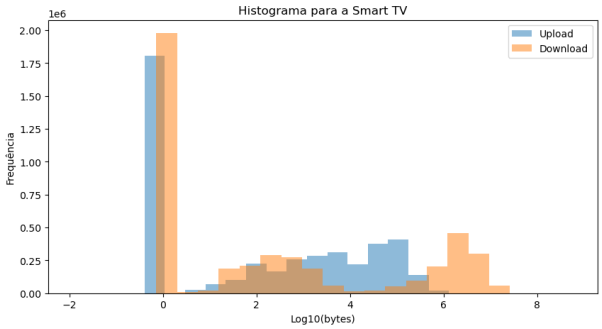
Estatística	Chromecast		SmartTV	
	Up	Down	Up	Down
Média	3.3497	3.7993	2.1566	2.3502
Mediana	3.3323	4.0250	2.1616	1.6498
D.Padrão	0.6794	1.2907	2.0281	2.5931
Variância	0.4616	1.6660	4.1131	6.7239

Tabela 1: Tabela de Estatísticas

Em resumo, os dados sugerem que o Chromecast tem, em média, taxas de upload e download mais altas em comparação com a Smart TV. Além disso, a variabilidade nos dados da Smart TV é mais pronunciada, indicando possíveis comportamentos mais diversos em relação ao uso da internet pelos usuários desse dispositivo. Essas informações podem ser úteis para o provedor de Internet na otimização e planejamento de sua infraestrutura, considerando as demandas específicas desses dispositivos.



(a) ChromeCast



(b) SmartTV

Figura 1: Histogramas

Como podemos ver nos histogramas, os dados baixados pela SmartTV em sua maioria são muito baixos, enquanto o Chromecast ou faz download de coisas leves, ou pesas, sem meio termos, enquanto de upload, parece seguir uma distribuição normal, com média em 3.5.

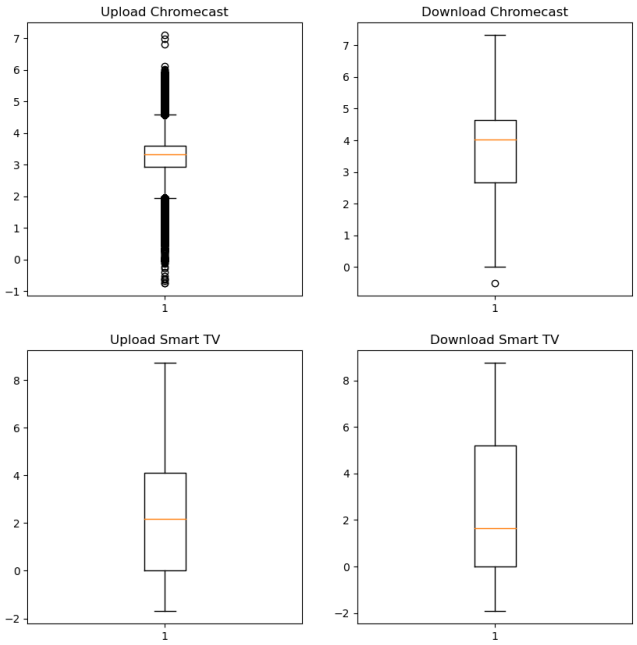


Figura 2: Boxplot Estatísticas Gerais

Analisando os boxplots, vemos o que foi falado, onde temos uma grande variância dos dados na smarttv, e uma grande quantidade de outliers para o upload do chromecast, pois na média, as coisas ficam pelo meio, mas temos uma grande parte de dados sendo gerados nas extremidades.

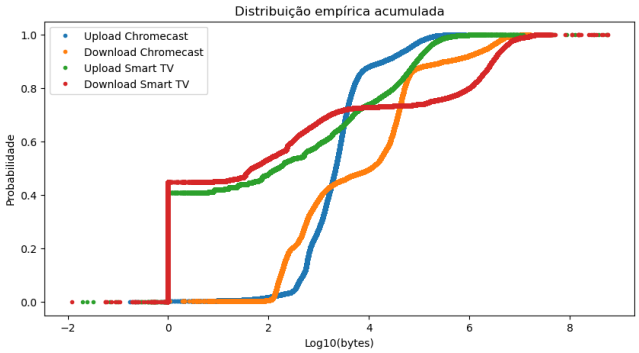
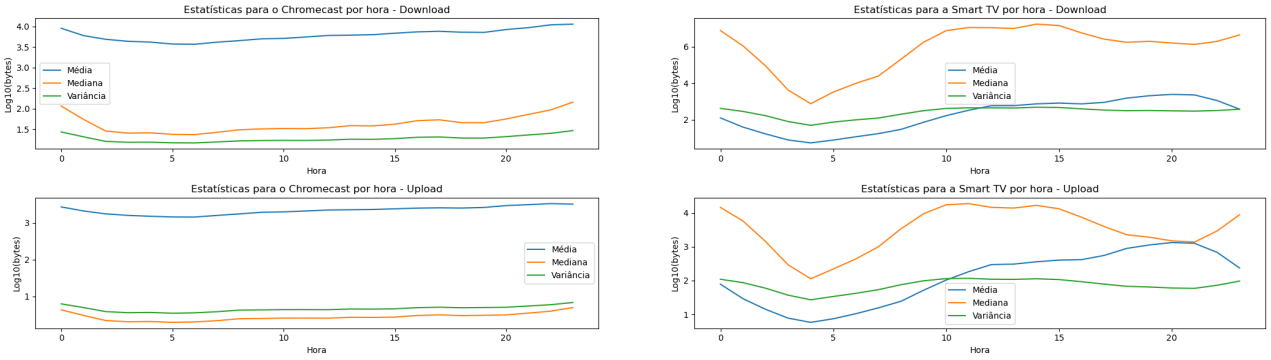


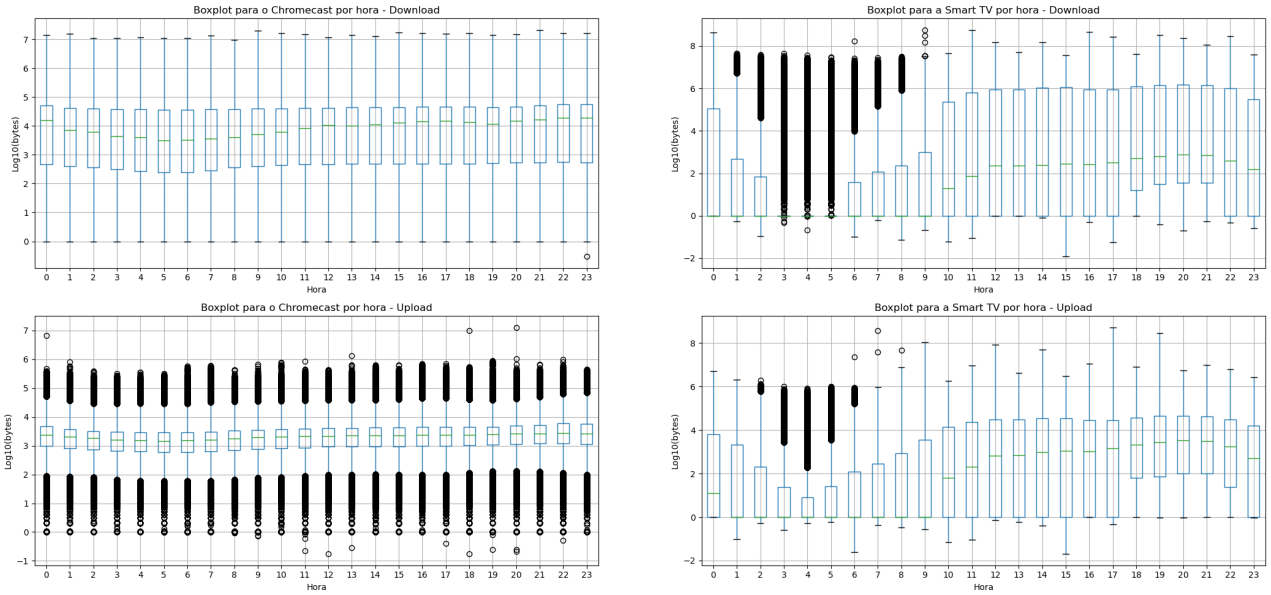
Figura 3: Distribuição Empírica

4 Estatísticas por Hora



(a) ChromeCast (b) SmartTV

Figura 4: Estatíticas por Horário



(a) ChromeCast (b) SmartTV

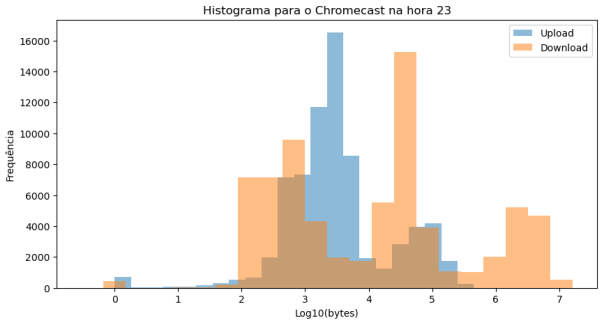
Figura 5: Boxplots por Horário

Como podemos ver, para o Chromecast os valores tem baixa variância e se mantém constante para quase todos os horários. Agora para a SmartTV, constatamos a grande variação, já citada, e percebemos os horários da madrugada como sendo aqueles com as menores médias. Além disso, para a SmartTV temos alguns outliers no período da madrugada, ressaltando como esse horário não é muito utilizado, mas que pode em algum momento ser requisitado. Uma coisa que fica clara é a constância do consumo de dados do Chromecast durante o dia, mas ainda e possível visualizar os outliers que acontecem em momentos atípicos do dia.

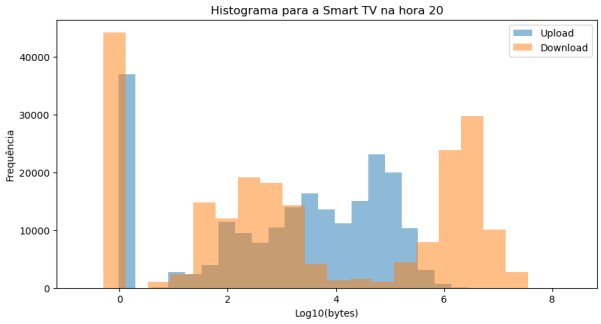
5 Tráfego

O primeiro passo foi definir os horários com maior tráfego, o que foi definido pela maior média, que para o Chromecast foi as 23h e para a SmartTV foi as 20h. Para o chromecast, o tráfego de download é diferente do tráfego de upload, por isso, foi utilizado como maior hora de tráfego a hora de download.

Como podemos perceber, a SmartTV apresenta duas corcundas, indicando uma variação durante o dia do que está acontecendo, tendo os picos em meados da manha e da noite. Para o chromecast, temos um upload que aumenta e decresce no decorrer do dia, e um donwload que tem picos du uso, provavelmente em momentos de utilização e acesso a novos conteúdos. Os uploads dos dois dispositivos seguem distribuições parecidas,



(a) ChromeCast



(b) SmartTV

Figura 6: Histogramas de horário de maior tráfego

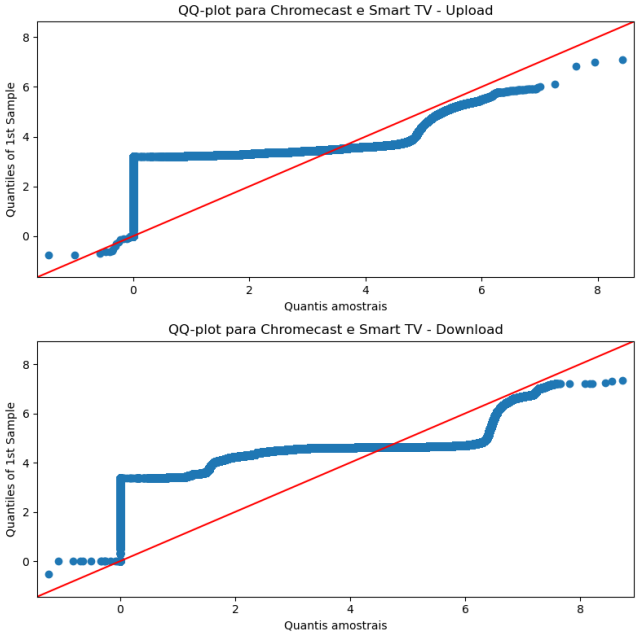
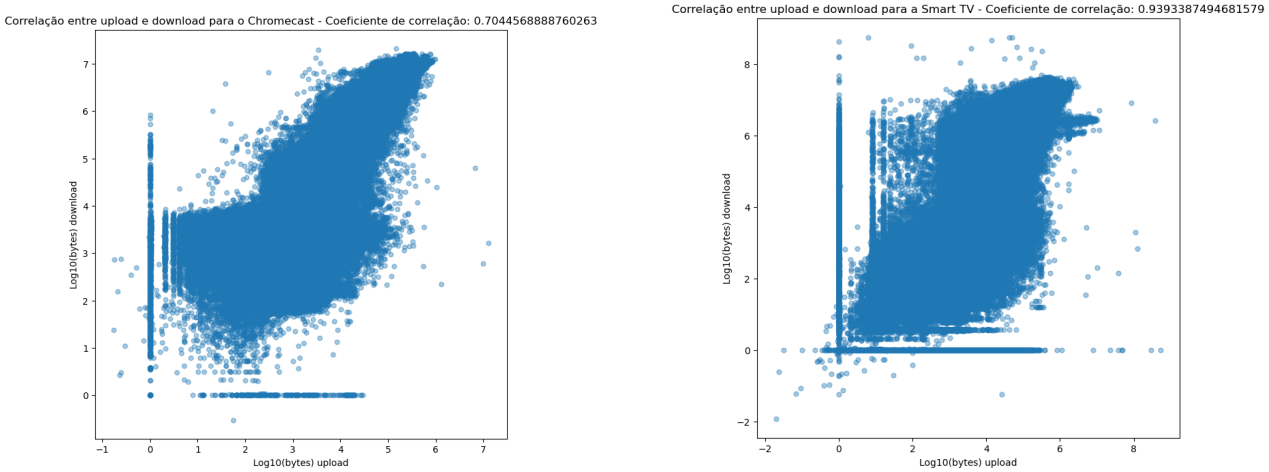


Figura 7: QQ-Plot

enquanto o download é bem diferente. No caso dos uploads, provavelmente são apenas os que precisam para enviar relatórios e logs para o servidor, que deve ser algo parecido para os dois.

Para o QQ-Plot, podemos perceber que durante vários momentos do dia, temos a presença de outliers, que são aqueles pontos que se afastam da linha de 45^{o}. Já nos momentos mais finais do dia, podemos perceber que a distirbuição se comporta mais da meneira esperada, pois segue, de alguma forma a inclinação da linha de referência.

6 Correlação



(a) ChromeCast

(b) SmartTV

Figura 8: Correlação

Nesses gráficos podemos induzir que existe uma correlação forte, de 0.70 para o Chromecast e para 0.93 para a SmartTV. O gráfico indica uma correlação quase que linear entre as variáveis de download e upload, indicando que os momentos de alto download implicam e se relacionam com os momentos de alto upload e vice-versa.

7 Conclusão

Os dados nos dizem muito e ajudam a perceber padrões de consumo, como momentos de download e upload, como esses de relacionam, momentos do dia de maior consumo e qual dispositvo mais consome, o que foi deixado claro durante o relatório. Observando isso, seria possível ajustar a banda de consumo baseada, por exemplo no horário de uso do cliente e ainda analisar atividades suspeitas, como utilização de dispositivos para ataques DDOS por exemplo, algo que poderia ser feito pelo próproi servidor.