Análise e Sumarização de dados do Kernel versão 3.12.2 e dados de logs da Copa de 98

Lucas Gonçalves Abreu[1], **Mateus Gabriel Dias**[1]

[1] Graduando em Ciência da Computação - Departamento de Computação
Universidade Federal de Ouro Preto (UFOP)
Ouro Preto - Minas Gerais - Brasil

{lucasgabreu,matsgdias}@gmail.com

Resumo. *Este trabalho apresenta a caracterização da distribuição de arquivos no código fonte do Kernel Linux (versão 3.12.2) e das requisições feitas à página do Mundial de 1998 nos dias 4, 5 e 6 de Junho de 1998. Assim como a comparação entre essas distribuições.*

1. Resenha Artigo: A Workload Characterization Study of the 1998 World Cup Web Site

A copa do mundo de 1998, realizada na França, foi o evento com maior cobertura da mídia da história. O público que assistiu aos jogos pela TV foi estimado em 40 bilhões, mais que o dobro da audiência dos Jogos Olímpicos de Verão que aconteceu em Atlanta no ano de 1996. O site da copa do mundo de 1998 também obteve um número enorme de visitantes, superando 1 bilhão de acessos em cerca de 3 meses de uso.

O artigo, em sua essência, apresenta a caracterização do detalhamento da carga de trabalho. Caracterização essa que tem um papel muito importante, visto que, feita de uma forma correta passa a servir de referências e estudos para o futuro dos algoritmos que tratam dessas situações.

O artigo compara os resultados que os pesquisadores obtiveram estudando a caracterização da carga de trabalho do servidor web da Copa de 98, com outros resultados obtidos anteriormente para determinar como as cargas de trabalho dos servidores veem mudando ao longo do tempo. Além disso, a grande carga que o site da Copa de 98 teve nos permite prever o que pode acontecer aos servidores web do futuro, para que assim possamos planejar e nos preparar para grandes demandas. Caracterização da carga de trabalho de um servidor é apenas uma das medidas necessárias para a compreensão das mudanças que ocorrem com o tráfego na web.

A Copa do Mundo é realizada uma vez a cada quatro anos e seu objetivo é determinar a melhor seleção do mundo. Devido ao grande número de países que desejam participar do torneio, uma eliminatória é realizada para selecionar as equipes que vão disputar o título de melhor seleçãoo do mundo. Dos 172 países que participam das eliminatórias apenas 30 são selecionados para competir a Copa do Mundo, juntamente com o país anfitrião, França, e o atual campão, Brasil.

O site da Copa de 98 contou com uma ampla rede de informações. Além de acessar os pontos e as outras estatísticas de cada seleção em tempo real, o usuário também podia contar com estatísticas e biografias dos jogadores, histórias das equipes, informações

sobre os estádios, fotos da partida, entrevistas com os jogadores entre outras coisas. Durante o torneiro, 30 servidores foram utilizados. Toda a criação e atualização da página ocorreu na França.

O conjunto de dados utilizado nesse estudo é composto de registro de acessos (contador), recolhidos a partir de cada servidor utilizado no site da Copa do Mundo. Os contadores de cada servidor foram arquivados diariamente em uma base de dados. Para este estudo todos registros de acesso, desde 01 de maio até 23 de julho de 1998, foram analisados.

Após os dados terem sido dados coletados, a primeira preocupação foi com o tamanho dos arquivos dos contadores: 125 Gbytes no total, 14 Gbytes quando compactados, a fim de fazer análises mais eficientes o arquivo foi convertido em um binário mais compacto. Utilizando algumas estratégias envolvendo o arquivo de acesso binário, o tamanho do arquivo foi reduzido para 25 Gbytes, 9 quando comprimido. Além disso, cada pedido está agora numa estrutura de tamanho fixo, o que também ajuda a melhorar a eficiência de nossas análises. Apesar da grande quantidade de dados que foram recolhidos por cada dos servidores, uma série de informações interessantes ainda não está disponível.

A primeira análise examinou a versão de protocolo de transferência de hipertexto (HTTP) suportado pelo cliente quando o mesmo solicita uma requisição. Como esperado, verificou-se que é ainda HTTP/1.0 a versão utilizada pela maioria, 78,7 porcento. No entanto, mais de 20 porcento do tráfego vieram de clientes que suportam HTTP/1.1. Os dados mostram que a maior parte dos pedidos resultou em sucesso na transferência do objeto. O sucesso nas transferências bem sucedidas foi responsável por quase todo o conteúdo solicitado, 97, 86 porcento, transferidos a partir do site para o cliente.

Alguns dados mostram colapso nas respostas do tipo do arquivo que foi requisitado pelo cliente. Para a maioria das respostas a extensão do arquivo foi utilizada para determinar o tipo do arquivo, por exemplo, arquivos que terminam com .gif ou .jpg foram colocados na categoria de imagens.

Os dados também mostraram que quase todos os requisitos dos usuários, 98,01 porcento, foram para o HTML, 9,85 porcento, ou arquivos de imagem, 88,16 porcento. Uma característica similar foi observada nas primeiras cargas de trabalho. Os arquivos HTML tiveram maior impacto do que os arquivos de imagens na quantidade de arquivos transferidos ao site (38,60 porcento para os arquivos HTML e 35,02 porcento para as imagens).

Desde o começo de maio até o inicio da Copa em junho, o tráfego do site estava tranquilo, embora tenha começado a se intensificar antes do inicio do evento. No dia em que a Copa começou, 10 de junho, o tráfego cresceu enormemente e esta marca foi mantida por um certo tempo. O site se tornou rapidamente muito popular e manteve-se assim por um período curto de tempo, após isso caiu em uma obscuridade profunda. Embora o volume de tráfego diário ser muito inconstante ele sempre ficou maior do que o volume de tráfego anterior ao começo do evento. No dia 30 de junho o site registrou a sua maior ocupação, sendo essa maior do que 73 milhões de usuários.

Muitas variáveis afetavam a hora que o site tinha mais acessos. Uma delas era que o site registrava maior número de usuários quando as partidas estavam acontecendo, e o volume de tráfego diminuía quando as partidas acabavam. Esses picos representavam

sobrecarga de servidores em pequena escala. O volume de tráfego também era afetado pelos times que estavam jogando no momento. As tradicionais potências do futebol, como Brasil e Alemanha por exemplo, geravam um grande número de visitas ao site no momento da partida, não só de brasileiros e alemães, como de outros fãs do bom futebol. A diferença entre os fusos horários dos países envolvidos na Copa ajudaram o site a se manter em pé. Se todos os usuários fossem acessar de um mesmo lugar, o site teria picos enormes durante o dia, e teoricamente, pela noite e madrugada afora ficaria com pouco tráfego.

O tráfego do site era muito baixo nos fins de semana, embora muitas partidas terem ocorrido aos sábados e domingos. Os dados da carga de trabalho do site possuem duas características importantes nos arquivos de referência: localidade temporal e concentração de referências. Localidade temporal quer dizer que o arquivo de referência referenciado anteriormente venha a ser novamente referenciado em um curto período de tempo. A heurística era baseada em algoritmos de pilha. Para os dados de carga de trabalho coletados a pilha média era muito menor do que os acessos a arquivos únicos. Isso indicava que o decrescimento da localidade temporal nas cargas de trabalho é muito forte. O decrescimento da localidade temporal era mais forte quando os usuários estavam focados em um só tema especifico, por exemplo, o placar do jogo corrente. Os estudos da segunda característica, baseado no foco de concentração de referências, constatou que existe um padrão n˜ao uniforme dos arquivos da web. Isso que significa que existe um pequeno numero de arquivos em sites da web que são muito populares, recebendo assim, a maioria das requisições provenientes daquele site, enquanto muitos outros arquivos em sites da web quase não são acessados.

As análises de pico da carga de trabalho notaram facilmente que os picos mais altos ocorriam quando as partidas de futebol estavam acontecendo. Foi analizado um perío odo de superlotação que durou 15 minutos, durante este período aconteciam as cobranças de penaltis que classificariam Argentina ou Inglaterra para a próxima fase do torneio. Muitos outros dados foram coletados e contribuíram para a pesquisa. Este trabalho foi de grande importância para todas as pessoas envolvidas com a área da computação. O trabalho mostrou padrões de carga de trabalho nunca estudados antes, padrões estes muito importantes para prévias de padrões futuros. Lendo este artigo ressalta-se a importância da computação nas nuvens (cloud computing), vendo que em 1998 tiveram que montar 30 servidores espalhados em vários países para suportar um site que teria uma imensidão de acessos durante apenas três meses. Com isso se gasta muito dinheiro investindo em hardware que com pouco tempo de uso não terá mais utilidade. Muitos dos temas abordados no artigo requerem estudos mais aprofundados para melhor serem entendidos e trabalhados pelos pesquisadores.

2. Sumarização dos Dados

Na Tabela 2 são apresentados alguns dados sobre a distribuição dos arquivos nos dois conjuntos de dados avaliados, como: nº de arquivos, nº de tamanhos de arquivos únicos, média do tamanho dos arquivos, variância no tamanho dos arquivos, desvio padrão, coenficiente de variabilidade e mediana. Nas Tabelas 2 e 2 são exibidos os quartis da distribuição e alguns percentis. Através do coeficiente de variabilidade podemos notar que a distribuição do tamanho dos arquivos do Kernel do Linux é mais comportada.

3. Visualização dos Dados

Nesta sessão os dados são apresentados em formas de gráficos, assim facilitando a sua análise. Os tamanhos dos dados estão sendo representandos em potências de 10, para tornar a análise melhor, dado que em escala linear os dados se concentravem em pequenas regiões e tornava a análise incompreensivel. Podemos notar que a maioria dos dados possuem tamanho entre 2ª e a 5ª potência de 10 em bytes.

3.1. Histogramas

Podemos observar que as distribuições dos tamanhos de arquivos são bem diferentes. No Kernel do Linux a distribuição dos tamanhos dos arquivos em escala logaritmica se comporta de forma suave e possue maior concentração entre a 3ª e a 5ª potência de 10, como podemos observar na Figura 1. Enquanto na distribuição do tamanho das repostas das requisições à página da Copa do Mundo, a distribuição seria bastante similar, se não houve picos entre a 2ª e a 3ª potência de 10, como podemos ver na Figura 2.

3.2. Função Densidade de Probabilidade - PDF

Dado que a PDF é o histograma contínuo mapeado como uma função de probabilidade, o comportamento é o mesmo que do histograma como podemos observar nas Figuras 3 e 4. Na Figura 5 podemos analisar a diferença entre as curvas das duas PDFs. As Figuras 6 e 7 são as CDFs das distribuições, que é a integral das PDFs, e representa a probabilidade do tamanho estar abaixo daquele valor, e seu comportamento é resultado delas. Já as Figuras 8 e 9 são as CCDFs, que é a CDF vista de trás pra frente com relação ao eixo x.

3.3. Função Distribuição Acumulada - CDF

3.4. Função Distribuição Acumulada Complementar - CCDF

4. Amostragem

Uma amostra que contém 5dos dois conjuntos de dados. E abaixo são apresentados os dados da estimativa do total, utilizando intervalo de confiança, os dados são apresentados abaixo. Na Tabela 4 podemos analisar a média das amostras, e os intervalos de confiança com 90médias do todo, e os intervalos no caso dos arquivos do Kernel do Linux não se afastam muito da média. Já os intervalos dos arquivos da Copa do Mundo se afastam muito da média, isso é resultado a própria distribuição, com variância muito alta, além de ser bimodal.

5. Anexos

	Linux Kernel	World Cup
Nº de arquivos	47460	18036
Nº de tamanhos únicos de arquivos	19232	1682
Média	10956.19	1239245887
Variância	3.786993e+18	583101122
Desvio Padrão	24147.49	1946019848
Coeficiente de Variabilidade	2.204004	1.570326
Mediana	4096	332

Table 1. Dados Gerais

	1º Quartil	2º Quartil	3º Quartil	4º Quartil
Linux Kernel	1540.00	4096.00	10773.25	988691.00
World Cup	207	332	429496729	429496729

Table 2. Quartis

	1º Percentil	10º Percentil	90º Percentil	99º Percentil
Linux Kernel	42.59	490.00	26015.40	101633.17
World Cup	207	207	4294967295	4294967295

Table 3. Percentis

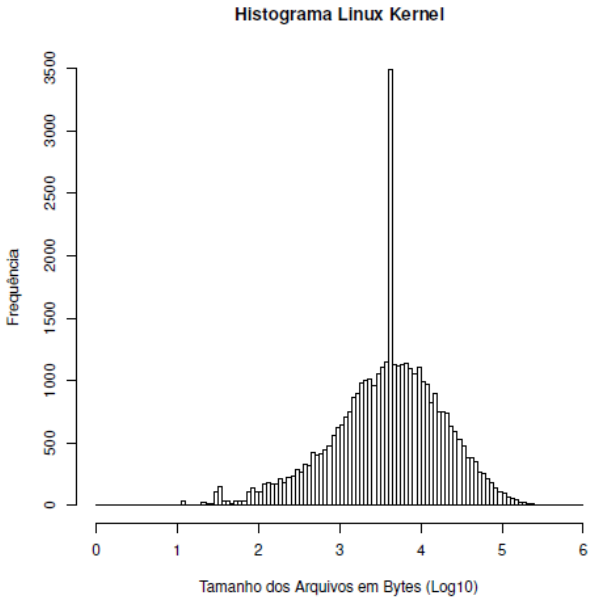

Figure 1. Histograma Linux Kernel

	Linux Kernel	World Cup
Média da Amostra	11445.08	1315663898
Intervalo de Confiança de 90%	10451.52 - 12438.64	1207000689 - 1424327107
Intervalo de Confiança de 95%	10261.04 - 12629.12	1186142793 - 1445185003
Intervalo de Confiança de 99%	9888.524 - 13001.631	1145311989 - 1486015807

Table 4. Percentis

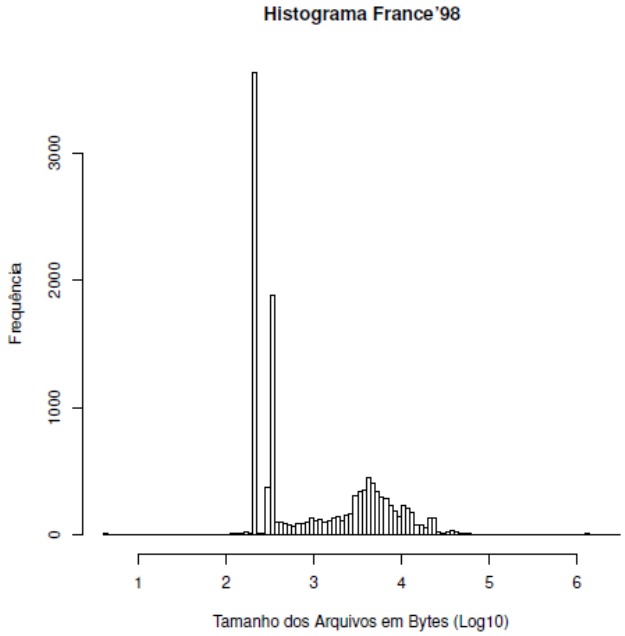

Figure 2. Histograma France 98

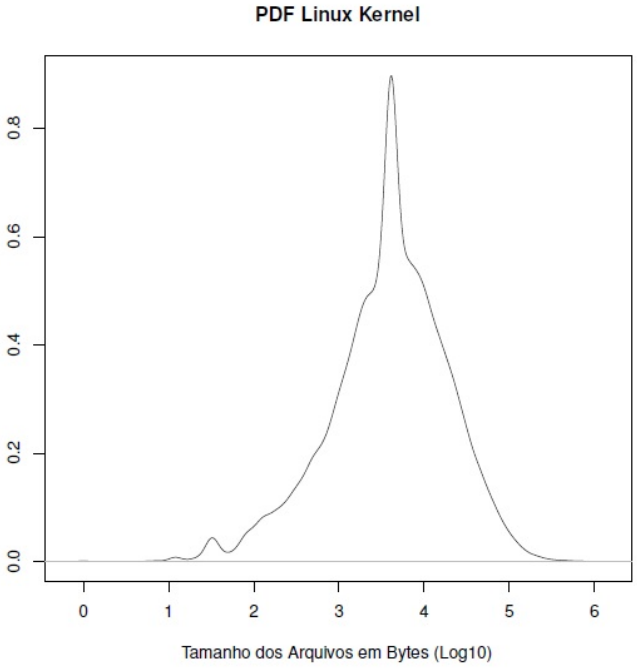

Figure 3. PDF France 98

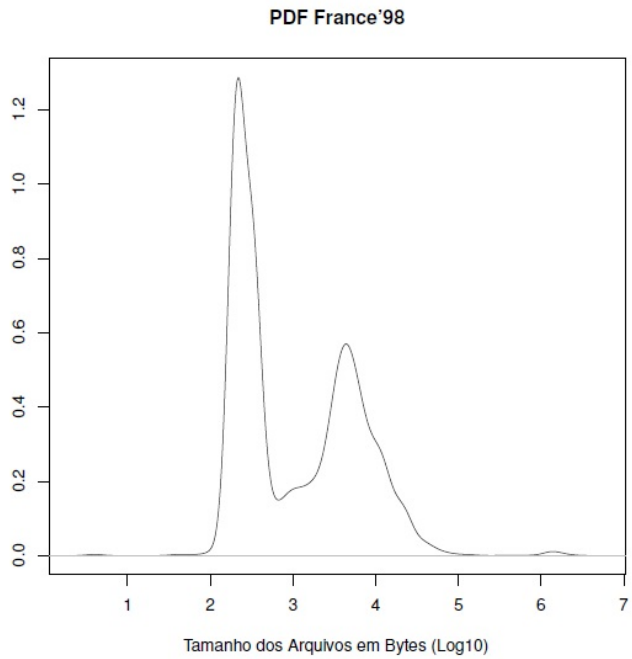

Figure 4. PDF World Cup

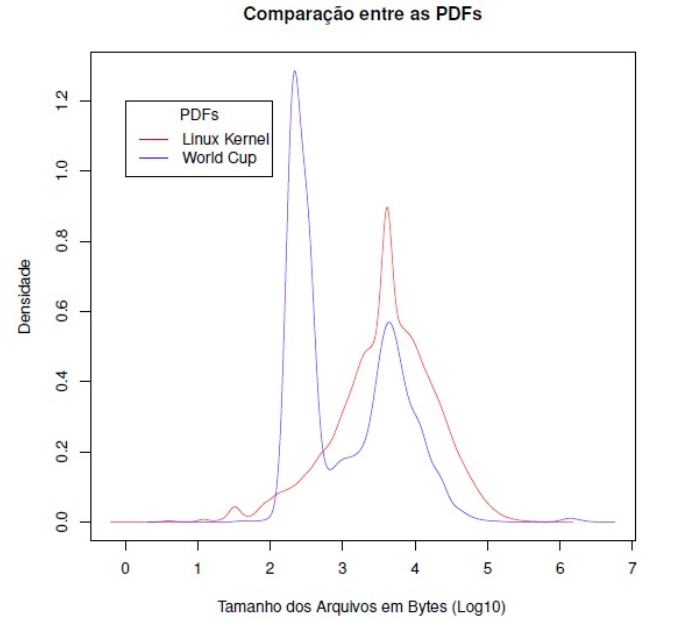

Figure 5. Comparação entre as PDFs

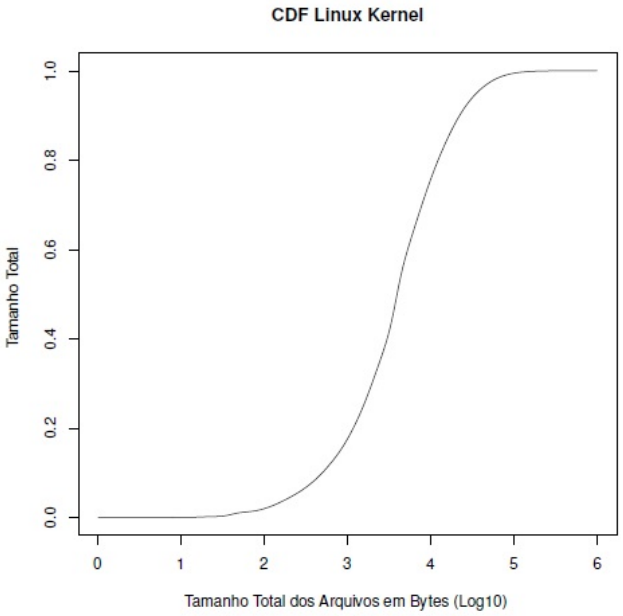

Figure 6. CDF Linux Kernel

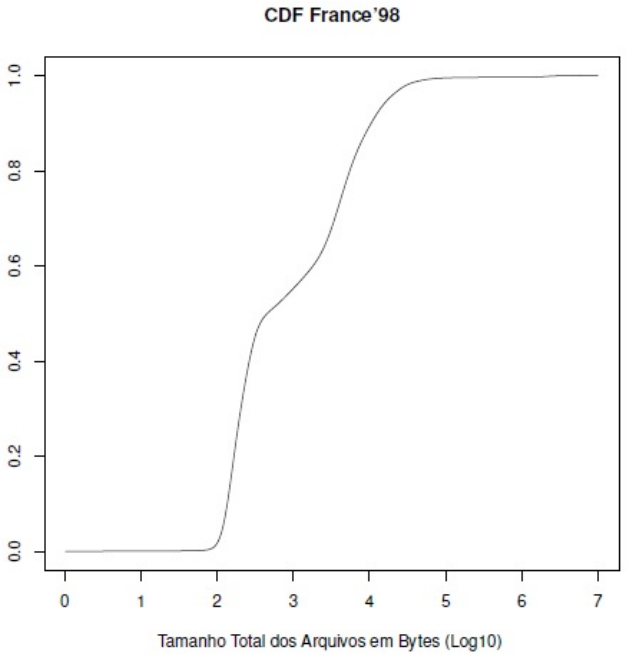

Figure 7. CDF World Cup

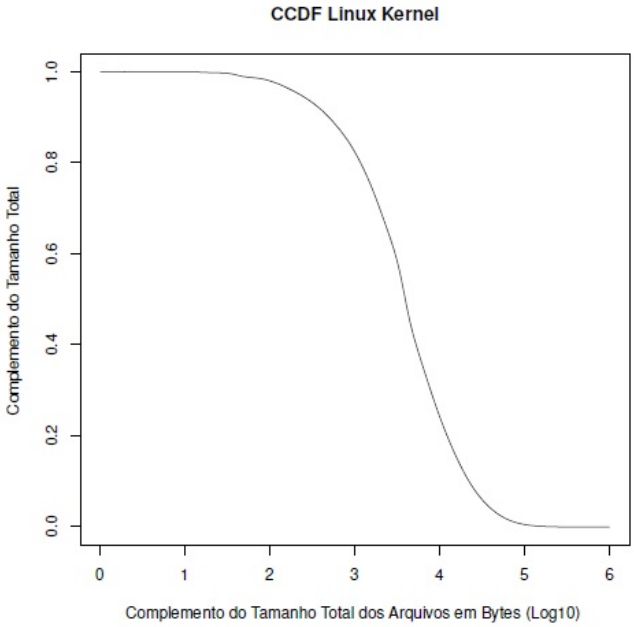

Figure 8. CCDF Linux Kernel

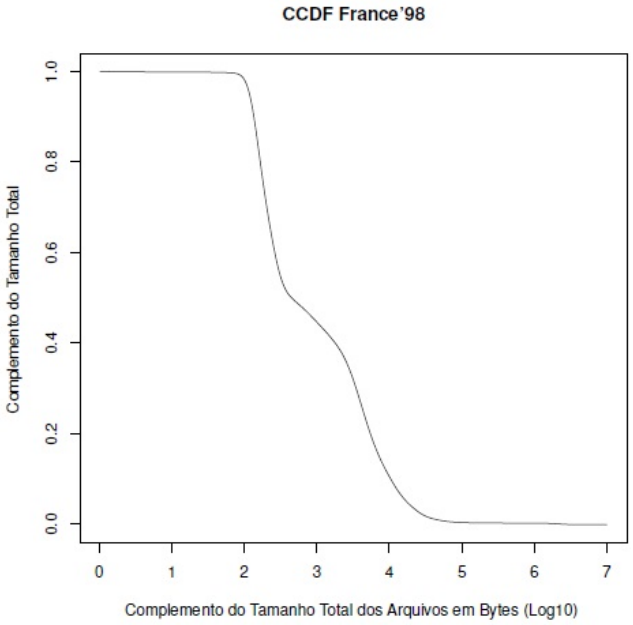

Figure 9. CCDF World Cup

```
############################ Linux Kernel ############################

file = "saida_kernel"
print(file)
lk_vector = vector()
csv_file = read.csv(file=file, head=FALSE)
size = length(csv_file$V1)

for(i in 1:size)
    lk_vector[i]=csv_file$V1[i]

######################### World Cup ############################

file = "saida"

wc_vector = vector()
print(file)
csv_file = read.csv(file=file, head=FALSE, sep=" ")
size = length(csv_file$V1)
    for(i in 1:size){
        if (is.na(wc_vector[csv_file$V2[i]+1]) || csv_file$V1[i]>wc_vector[csv_
            wc_vector[csv_file$V2[i]+1]=csv_file$V1[i]
}
wc_vector = wc_vector[!is.na(wc_vector)]
wc_vector = wc_vector[wc_vector>0]

########################### Analysis ################################
##
### Number of files
lk_numfiles = length(lk_vector)
wc_numfiles = length(wc_vector)
```

```r
### Number of unique sizes
lk_uniq_size_files = length(unique(lk_vector))
wc_uniq_size_files = length(unique(wc_vector))

### Mean
lk_mean = mean(lk_vector)
wc_mean = mean(wc_vector)

### Variance
lk_var = var(lk_vector)
wc_var = var(wc_vector)

### Standart deviation
lk_sd = sd(lk_vector)
wc_sd = sd(wc_vector)

### Coefficient of variability
lk_cv = lk_sd/lk_mean
wc_cv = wc_sd/wc_mean

### Median
lk_median = median(lk_vector)
wc_median = median(wc_vector)

### Quartiles
lk_quartiles = quantile(lk_vector)
wc_quartiles = quantile(wc_vector)

### Percentiles
lk_percentiles = quantile(lk_vector, c(.01, .10, .90, .99))
wc_percentiles = quantile(wc_vector, c(.01, .10, .90, .99))
```

```
##################### Visualization ############################

### Histogram
hist(log10(lk_vector), breaks=100, main="Histograma Linux Kernel", xlab="Tamanho dos Arquivos em Bytes (Log10)", ylab="Frequência")
dev.copy2eps(file="HistogramaLK.eps")
hist(log10(wc_vector), breaks=100, main="Histograma France'98", xlab="Tamanho dos Arquivos em Bytes (Log10)", ylab="Frequência")
dev.copy2eps(file="HistogramaWC.eps")

### PDF
lk_pdf = density(log10(lk_vector))
plot(lk_pdf, main="PDF Linux Kernel", xlab="Tamanho dos Arquivos em Bytes (Log10)", ylab="Densidade")
dev.copy2eps(file="PDF_LK.eps")
wc_pdf = density(log10(wc_vector))
plot(wc_pdf, main="PDF France'98", xlab="Tamanho dos Arquivos em Bytes (Log10)", ylab="Densidade")
dev.copy2eps(file="PDF_WC.eps")

### CDF
lk_cdf = cumsum(c(0,diff(lk_pdf$x))*lk_pdf$y)
plot(lk_cdf, main="CDF Linux Kernel", xlab="Tamanho Total dos Arquivos em Bytes (Log10)", ylab="Tamanho Total ", type="lines", xaxt="n")
axis(1, at=seq(0,512,512/6), labels=(0:6))
dev.copy2eps(file="CDF_LK.eps")
wc_cdf = cumsum(c(0,diff(wc_pdf$x))*wc_pdf$y)
plot(wc_cdf, main="CDF France'98", xlab="Tamanho Total dos Arquivos em Bytes (Log10)", ylab="Tamanho Total", type="lines", xaxt="n")
axis(1, at=seq(0,512,512/7), labels=(0:7))
dev.copy2eps(file="CDF_WC.eps")


### CCDF
lk_ccdf = 1-lk_cdf
plot(lk_ccdf, main="CCDF Linux Kernel", xlab="Complemento do Tamanho Total dos Arquivos em Bytes (Log10)", ylab="Complemento do Tamanho Total",
type="lines", xaxt="n")
axis(1, at=seq(0,512,512/6), labels=(0:6))
```

```
type="lines", xaxt="n")
axis(1, at=seq(0,512,512/6), labels=(0:6))
dev.copy2eps(file="CCDF_LK.eps")

wc_ccdf = 1-wc_cdf
plot(wc_ccdf, main="CCDF France'98", xlab="Complemento do Tamanho Total dos Arquivos em Bytes (Log10)", ylab="Complemento do Tamanho Total", type="lines",
xaxt="n")
axis(1, at=seq(0,512,512/7), labels=(0:7))
dev.copy2eps(file="CCDF_WC.eps")


### Compare the PDFs
plot(range(lk_pdf$x, wc_pdf$x), range(lk_pdf$y, wc_pdf$y), type = "n", main="Comparação entre as PDFs", xlab = "Tamanho dos Arquivos em Bytes (Log10)"
, ylab = "Densidade")
lines(lk_pdf, col = "red")
lines(wc_pdf, col = "blue")
legend(0, 1.2, c("Linux Kernel", "World Cup"), cex=1, col=c("red","blue"), lty=1, title="PDFs")
dev.copy2eps(file="ComparePDFs.eps")

######################## Samples ###################################

lk_sample = sample(lk_vector, .05*length(lk_vector))
wc_sample = sample(wc_vector, .05*length(wc_vector))

t.test(lk_sample, conf.level = 0.90)
t.test(lk_sample, conf.level = 0.95)
t.test(lk_sample, conf.level = 0.99)

t.test(wc_sample, conf.level = 0.90)
t.test(wc_sample, conf.level = 0.95)
t.test(wc_sample, conf.level = 0.99)
```