Apache logs analysis

Load file:

```
logs <- read.csv("apache_logs.txt",stringsAsFactors = FALSE, header = F, sep= " ", quote = "\"")
head(logs)
```

```
##             V1 V2 V3                     V4     V5
## 1 83.149.9.216  -  - [17/May/2015:10:05:03 +0000]
## 2 83.149.9.216  -  - [17/May/2015:10:05:43 +0000]
## 3 83.149.9.216  -  - [17/May/2015:10:05:47 +0000]
## 4 83.149.9.216  -  - [17/May/2015:10:05:12 +0000]
## 5 83.149.9.216  -  - [17/May/2015:10:05:07 +0000]
## 6 83.149.9.216  -  - [17/May/2015:10:05:34 +0000]
##                                                                                V6
## 1      GET /presentations/logstash-monitorama-2013/images/kibana-search.png HTTP/1.1
## 2  GET /presentations/logstash-monitorama-2013/images/kibana-dashboard3.png HTTP/1.1
## 3 GET /presentations/logstash-monitorama-2013/plugin/highlight/highlight.js HTTP/1.1
## 4        GET /presentations/logstash-monitorama-2013/plugin/zoom-js/zoom.js HTTP/1.1
## 5         GET /presentations/logstash-monitorama-2013/plugin/notes/notes.js HTTP/1.1
## 6          GET /presentations/logstash-monitorama-2013/images/sad-medic.png HTTP/1.1
##    V7     V8
## 1 200 203023
## 2 200 171717
## 3 200  26185
## 4 200   7697
## 5 200   2892
## 6 200 430406
##                                                    V9
## 1 http://semicomplete.com/presentations/logstash-monitorama-2013/
## 2 http://semicomplete.com/presentations/logstash-monitorama-2013/
## 3 http://semicomplete.com/presentations/logstash-monitorama-2013/
## 4 http://semicomplete.com/presentations/logstash-monitorama-2013/
## 5 http://semicomplete.com/presentations/logstash-monitorama-2013/
## 6 http://semicomplete.com/presentations/logstash-monitorama-2013/
##
## 1 Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/32.0.
## 2 Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/32.0.
## 3 Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/32.0.
## 4 Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/32.0.
## 5 Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/32.0.
## 6 Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/32.0.
```

```
str(logs)
```

```
## 'data.frame':    10000 obs. of  10 variables:
##  $ V1 : chr  "83.149.9.216" "83.149.9.216" "83.149.9.216" "83.149.9.216" ...
##  $ V2 : chr  "-" "-" "-" "-" ...
##  $ V3 : chr  "-" "-" "-" "-" ...
##  $ V4 : chr  "[17/May/2015:10:05:03" "[17/May/2015:10:05:43" "[17/May/2015:10:05:47" "[17/May/2015:1
##  $ V5 : chr  "+0000]" "+0000]" "+0000]" "+0000]" ...
##  $ V6 : chr  "GET /presentations/logstash-monitorama-2013/images/kibana-search.png HTTP/1.1" "GET /p
##  $ V7 : int  200 200 200 200 200 200 200 200 200 200 ...
##  $ V8 : chr  "203023" "171717" "26185" "7697" ...
```

```
##  $ V9 : chr  "http://semicomplete.com/presentations/logstash-monitorama-2013/" "http://semicomplete.
##  $ V10: chr  "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_1) AppleWebKit/537.36 (KHTML, like Gecko) 
```

```
summary(logs)
```

```
##       V1                 V2                 V3           
##  Length:10000       Length:10000       Length:10000      
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##                                                           
##                                                           
##                                                           
##       V4                 V5                 V6                 V7        
##  Length:10000       Length:10000       Length:10000       Min.   :200.0  
##  Class :character   Class :character   Class :character   1st Qu.:200.0  
##  Mode  :character   Mode  :character   Mode  :character   Median :200.0  
##                                                           Mean   :210.8  
##                                                           3rd Qu.:200.0  
##                                                           Max.   :500.0  
##       V8                 V9                 V10           
##  Length:10000       Length:10000       Length:10000      
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##                                                           
##                                                           
##                                                           
```

In the file are two fields unused, and a data partial import, so we can just remove it from the DataSet:

```
logs = logs[,-c(2,3,5)]
logs$V4<-substring(logs$V4,2)
colnames(logs) = c("IP", "Date", "URL", "Status", "Size", "Referer", "User_Agent")
#Factors
logs$IP <- as.factor(logs$IP)
logs$URL <- as.factor(logs$URL)
logs$Status <- as.factor(logs$Status)
logs$Referer <- as.factor(logs$Referer)
logs$User_Agent <- as.factor(logs$User_Agent)
summary(logs)
```

```
##               IP           Date          
##  66.249.73.135 : 482   Length:10000      
##  46.105.14.53  : 364   Class :character  
##  130.237.218.86: 357   Mode  :character  
##  75.97.9.59    : 273                     
##  50.16.19.13   : 113                     
##  209.85.238.199: 102                     
##  (Other)       :8309                     
##                                          URL           Status      
##  GET /favicon.ico HTTP/1.1                : 747   200    :9126  
##  GET /style2.css HTTP/1.1                 : 531   304    : 445  
##  GET /reset.css HTTP/1.1                  : 524   404    : 213  
##  GET /images/jordan-80.png HTTP/1.1       : 521   301    : 164  
##  GET /images/web/2009/banner.png HTTP/1.1 : 504   206    :  45  
##  GET /blog/tags/puppet?flav=rss20 HTTP/1.1: 488   500    :   3  
##  (Other)                                  :6685   (Other):   4  
```

```
##      Size
##  Length:10000
##  Class :character
##  Mode  :character
##
##
##
##
##                                                                 Referer
##  -                                                                 :4073
##  http://semicomplete.com/presentations/logstash-puppetconf-2012/: 689
##  http://www.semicomplete.com/projects/xdotool/                   : 656
##  http://semicomplete.com/presentations/logstash-scale11x/        : 406
##  http://www.semicomplete.com/articles/dynamic-dns-with-dhcp/     : 335
##  http://www.semicomplete.com/                                    : 228
##  (Other)                                                         :3613
##
##  Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/32.0.1700.107 Safa
##  Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/33.0.1
##  UniversalFeedParser/4.2-pre-314-svn +http://feedparser.org/
##  Mozilla/5.0 (Windows NT 6.1; WOW64; rv:27.0) Gecko/20100101 Firefox/27.0
##  Mozilla/5.0 (iPhone; CPU iPhone OS 6_0 like Mac OS X) AppleWebKit/536.26 (KHTML, like Gecko) Versio
##  Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/32.0.1700.107 Safari/5
##  (Other)
```

```
freq_logs <- table(logs$URL)
print("Top 10 requested pages followed by number or requests ")
```

```
## [1] "Top 10 requested pages followed by number or requests "
```

```
print(sort(freq_logs, decreasing=TRUE)[1:10])
```

```
##
##                 GET /favicon.ico HTTP/1.1
##                                        747
##                  GET /style2.css HTTP/1.1
##                                        531
##                   GET /reset.css HTTP/1.1
##                                        524
##        GET /images/jordan-80.png HTTP/1.1
##                                        521
##  GET /images/web/2009/banner.png HTTP/1.1
##                                        504
## GET /blog/tags/puppet?flav=rss20 HTTP/1.1
##                                        488
##                 GET /?flav=rss20 HTTP/1.1
##                                        217
##           GET /projects/xdotool/ HTTP/1.1
##                                        208
##                  GET /robots.txt HTTP/1.1
##                                        157
##                             GET / HTTP/1.1
##                                        151
```

```
print("Top ten responses")
```

```
## [1] "Top ten responses"
```

```
status <- table(logs$Status)
print(sort(status, decreasing=TRUE))
```

```
##
##  200  304  404  301  206  500  403  416
## 9126  445  213  164   45    3    2    2
```

```
cat("\n\n")
```

```
print("Percentatge of top ten responses")
```

```
## [1] "Percentatge of top ten responses"
```

```
print(sort(round(100*prop.table(status), digits = 2), decreasing=TRUE))
```

```
##
##   200   304   404   301   206   500   403   416
## 91.26  4.45  2.13  1.64  0.45  0.03  0.02  0.02
```

```
success <- c('200','206','301','304')
failure <- c('403','404','416','500')
logs_success<-logs[logs$Status %in% success,]
logs_failure<-logs[logs$Status %in% failure,]

cat("\nTotal successful requests: ", nrow(logs_success)/nrow(logs)*100,"% \n")
```

```
##
## Total successful requests:  97.8 %
```

```
cat("Total unsuccessful requests: ", nrow(logs_failure)/nrow(logs)*100,"% \n")
```

```
## Total unsuccessful requests:  2.2 %
```

```
freq_logs_un <- table(logs_failure$URL)
print("Top 10 unsuccessful requested pages followed by number or requests ")
```

```
## [1] "Top 10 unsuccessful requested pages followed by number or requests "
```

```
print(sort(freq_logs_un, decreasing=TRUE)[1:10])
```

```
##
##                                          GET /files/logstash/logstash-1.3.2-monolithic.jar HTTP/1.1
##                                                                                           61
## GET /presentations/logstash-puppetconf-2012/images/office-space-printer-beat-down-gif.gif HTTP/1.1
##                                                                                           31
##                                                                      GET /blog/wp-admin/ HTTP/1.1
##                                                                                            6
##                                                                           GET /wp-admin/ HTTP/1.1
##                                                                                            6
##                                                        GET /wp-login.php?action=register HTTP/1.0
##                                                                                            6
##                                                                        GET /wp/wp-admin/ HTTP/1.1
##                                                                                            6
##                                                                  GET /wordpress/wp-admin/ HTTP/1.1
##                                                                                            5
##                                                                           GET /admin.php HTTP/1.1
##                                                                                            4
```

```
##                                                                                 GET /administrator/ HTTP/1.1
##                                                                                                            4
##                                                                              GET /image/logstash.png HTTP/1.1
##                                                                                                            4
```

```
cat("\n\n")
```

```
freq_ip <- table(logs$IP)
print("Top 10 IPs requesting followed by number or requests ")
```

```
## [1] "Top 10 IPs requesting followed by number or requests "
```

```
print(sort(freq_ip, decreasing=TRUE)[1:10])
```

```
##
##  66.249.73.135   46.105.14.53 130.237.218.86     75.97.9.59    50.16.19.13
##            482            364            357            273            113
## 209.85.238.199 68.180.224.225  100.43.83.137 208.115.111.72 198.46.149.143
##            102             99             84             83             82
```