Comparing the Dutch and Egyptian Client RTT's of the UK ISP's that Served Egypt during January of 2011

Thursday, April 16, 2015

Finding Egypt's ISP's for the January, 2011 time period

I acquired the January 2011 Egypt data set from Google BigQuery as described in the "Kinga Farkas Initial Contribution" write up.
I read in the .csv file into R and kept only the relevant columns using the "dataConverter" function (see https://github.com/lilbludot/OutreachyProject).

```
setwd("~/M_LabProject")
source("dataConverterFunction.R")
dfJan <- read.csv("Egypt2011JanRTTComplete.csv", colClasses="character")
mainJanDF <- dataConverter(dfJan)
```

```
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Given the whittled down data frame "mainJanDF", I created and printed a list of the unique IP addressess of the servers grouped by country, using the function "uniqueServerFinder".

```
uniqueServerFinder <- function(mainDF){
        library(dplyr)
        mainDF$date <- as.character(mainDF$date)
        serverCountries <- unique(mainDF$serverCountry)
        serverCountries
        serverListPerCountry <- list()
        for (i in serverCountries){
            dummy <- filter(mainDF, serverCountry == i)
            x <- unique(dummy$serverIP)
            serverListPerCountry[[i]]<-x
            print(i)
            print(x)

        }
       return(serverListPerCountry)
}

JanServerList <- uniqueServerFinder(mainJanDF)
```

```
## [1] "Greece"
## [1] "83.212.4.37" "83.212.4.10" "83.212.4.24"
## [1] "United States"
##  [1] "64.9.225.141"  "38.98.51.45"   "38.106.70.173" "74.63.50.32"
##  [5] "38.107.216.45" "74.63.50.47"   "64.9.225.167"  "74.63.50.19"
##  [9] "38.98.51.33"   "38.102.0.97"   "64.9.225.154"  "38.106.70.147"
## [13] "4.71.254.147"  "4.71.210.237"  "38.98.51.20"   "4.71.251.173"
## [17] "38.107.216.17" "38.107.216.32" "38.106.70.160" "4.71.210.211"
## [21] "4.71.254.160"  "4.71.251.160"  "4.71.210.224"  "38.102.0.109"
## [1] "Europe"
## [1] "80.239.168.216" "80.239.168.234" "80.239.168.203" "80.239.142.233"
## [5] "80.239.142.203" "80.239.142.216"
## [1] "United Kingdom"
## [1] "217.163.1.101"   "213.244.128.164" "217.163.1.75"    "213.244.128.152"
## [5] "213.244.128.139" "217.163.1.88"
## [1] "Australia"
## [1] "203.5.76.151" "203.5.76.140" "203.5.76.166"
## [1] "Netherlands"
## [1] "72.26.217.88"  "72.26.217.103"
```

Getting the Egyptian and Dutch Clients' Data for the UK servers during January, 2011

At some earlier point, while I plotted the locations of the Egyptian ISP's, I noticed that quite a bit of Egypt's traffic went through the ISP located in the UK. Since I did not have the time to look at each one of Egypt's ISP's, I ended up selecting the UK one, especially after I realized that the very same ISP also had the Netherlands for a client. So, I took the six IP addressess of the UK servers and I queried Google BigQuery's m-lab data set:

SELECT *, connection_spec.client_geolocation.continent_code, connection_spec.client_geolocation.country_name, connection_spec.client_geolocation.city, connection_spec.server_geolocation.continent_code, connection_spec.server_geolocation.country_name, connection_spec.server_geolocation.city

FROM [measurement-lab:m_lab.2011_02] WHERE (connection_spec.server_ip == '217.163.1.101' OR connection_spec.server_ip == '213.244.128.164' OR connection_spec.server_ip == '217.163.1.75' OR connection_spec.server_ip == '213.244.128.152' OR connection_spec.server_ip == '213.244.128.139' OR connection_spec.server_ip == '217.163.1.88') AND (connection_spec.client_geolocation.country_name == "Egypt" OR connection_spec.client_geolocation.country_name == "Netherlands") AND IS_EXPLICITLY_DEFINED(web100_log_entry.connection_spec.remote_ip) AND IS_EXPLICITLY_DEFINED(web100_log_entry.connection_spec.local_ip) AND IS_EXPLICITLY_DEFINED(web100_log_entry.snap.HCThruOctetsAcked) AND IS_EXPLICITLY_DEFINED(web100_log_entry.snap.SndLimTimeRwin) AND IS_EXPLICITLY_DEFINED(web100_log_entry.snap.SndLimTimeCwnd) AND IS_EXPLICITLY_DEFINED(web100_log_entry.snap.SndLimTimeSnd) AND IS_EXPLICITLY_DEFINED(project) AND project = 0 AND IS_EXPLICITLY_DEFINED(connection_spec.data_direction) AND connection_spec.data_direction = 1 AND IS_EXPLICITLY_DEFINED(web100_log_entry.is_last_entry) AND web100_log_entry.is_last_entry = True AND web100_log_entry.snap.HCThruOctetsAcked >= 8192 AND (web100_log_entry.snap.SndLimTimeRwin + web100_log_entry.snap.SndLimTimeCwnd + web100_log_entry.snap.SndLimTimeSnd) >= 9000000 AND (web100_log_entry.snap.SndLimTimeRwin + web100_log_entry.snap.SndLimTimeCwnd +

web100_log_entry.snap.SndLimTimeSnd) < 3600000000 AND
IS_EXPLICITLY_DEFINED(web100_log_entry.snap.MinRTT) AND
IS_EXPLICITLY_DEFINED(web100_log_entry.snap.CountRTT) AND web100_log_entry.snap.CountRTT > 0
AND (web100_log_entry.snap.State == 1 OR (web100_log_entry.snap.State >= 5 AND
web100_log_entry.snap.State <= 11));

This way I got all the Dutch and Egyptian client data for the six UK servers for the January, 2011 time period.
I then downloaded and cleaned up the data a bit:

```
EgyptNethJan2011DF<-read.csv("UKServers.csv")
Jan2011DF <- dataConverter(EgyptNethJan2011DF)
Jan2011DF$date <- as.character(Jan2011DF$date)
head(Jan2011DF)
```

```
##          logTime RTT clientContinent clientCountry clientCity
## 75455 1293840255  17              EU   Netherlands
## 75460 1293840309  13              EU   Netherlands Roosendaal
## 75454 1293840370  41              EU   Netherlands
## 75452 1293840447  31              EU   Netherlands
## 75451 1293840454   9              EU   Netherlands
## 75459 1293840496   6              EU   Netherlands
##       serverContinent  serverCountry serverCity        serverIP
## 75455              EU United Kingdom         NA 213.244.128.139
## 75460              EU United Kingdom         NA 213.244.128.139
## 75454              EU United Kingdom         NA 213.244.128.139
## 75452              EU United Kingdom         NA 213.244.128.164
## 75451              EU United Kingdom         NA 213.244.128.152
## 75459              EU United Kingdom         NA 213.244.128.139
##                      date   moDayYear
## 75455 2011-01-01 00:04:15 Jan 01 2011
## 75460 2011-01-01 00:05:09 Jan 01 2011
## 75454 2011-01-01 00:06:10 Jan 01 2011
## 75452 2011-01-01 00:07:27 Jan 01 2011
## 75451 2011-01-01 00:07:34 Jan 01 2011
## 75459 2011-01-01 00:08:16 Jan 01 2011
```

I separated the data frame into two parts: one containing all the Egyptian client data, the other the Dutch data.

```
library(dplyr)
EgyptJan2011DF <- filter(Jan2011DF, clientCountry=="Egypt")
NethJan2011DF <-filter(Jan2011DF, clientCountry =="Netherlands")
```

Comparing the Dutch and Egyptian Data Sets

First of all, how did the size of the data sets compare?

Well,

```
nrow(EgyptJan2011DF)
```

```
## [1] 605
```

```
nrow(NethJan2011DF)
```

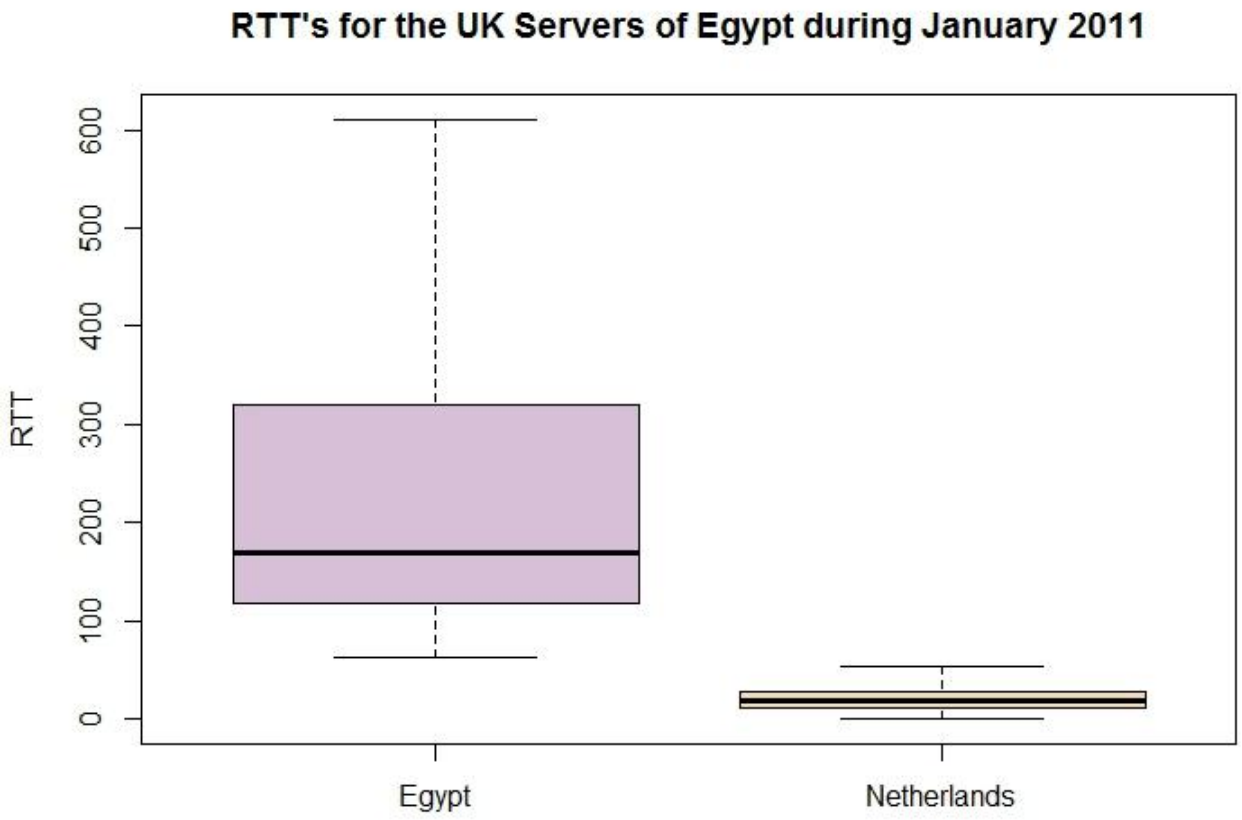
```
## [1] 83449
```

The number of tests run in Egypt (`605`) is but a fraction of the tests run in the Netherlands (`83449.`)
I had to be careful when comparing values other than averages, median and IQR's, specifically it was not fair to plot the outliers of the Dutch data set side by side the Egyptian data set.

Boxplots

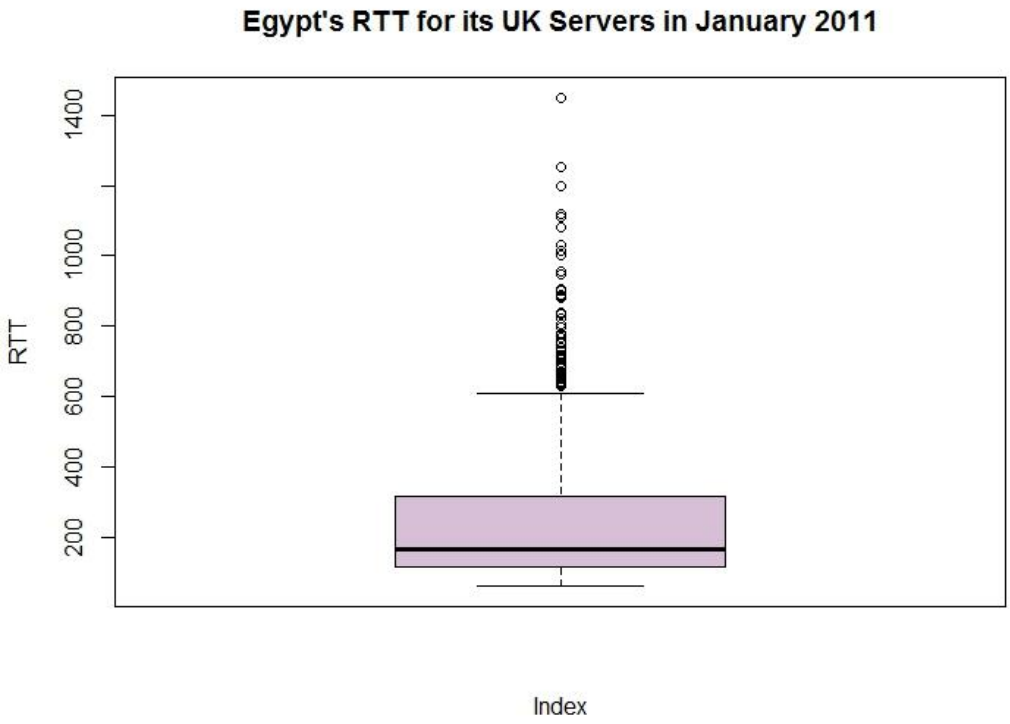
I created side by side boxplots without the outliers:

```
boxplot(EgyptJan2011DF$RTT , NethJan2011DF$RTT,  main="RTT's for the UK Servers of Egypt
during January 2011", xlab="", ylab="RTT",
names=c("Egypt","Netherlands"),col=c("thistle","wheat"), outline=FALSE)
```

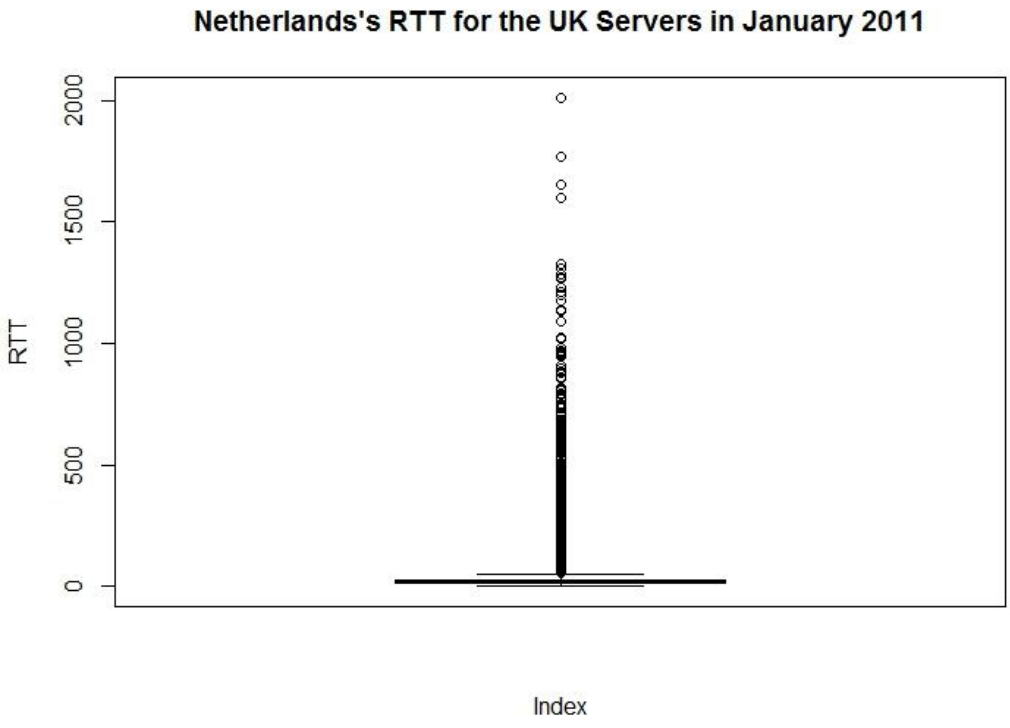


Then I plotted the individual boxplots, now with the outliers:

```
boxplot(EgyptJan2011DF$RTT , main="Egypt's RTT for its UK Servers in January 2011",
        xlab="Index", ylab="RTT", col="thistle")
```




```
boxplot(NethJan2011DF$RTT , main="Netherlands's RTT for the UK Servers in January 2011",
        xlab="Index", ylab="RTT", col="wheat")
```



Finally, in order to get a bit deeper understanding of the data I calculated the averages and medians of the RTT's for both Egypt and the Netherlands:

```r
median(EgyptJan2011DF$RTT)
```

```
## [1] 169
```

```r
mean(EgyptJan2011DF$RTT)
```

```
## [1] 265.0496
```

```r
median(NethJan2011DF$RTT)
```

```
## [1] 19
```

```r
mean(NethJan2011DF$RTT)
```

```
## [1] 24.25665
```

The Egyptian distribution is more right skewed than the Dutch one. And, more obviously, the average Egyptian RTT for the UK servers is more than ten times larger than the average Dutch RTT.

Conclusions

The fact that the mean Egyptian RTT is $\frac{265}{24} \approx 11$ times larger than the average Dutch RTT - while clients of the same British ISP - lead me to suspect that the slow internet speeds in Egypt (as evidenced by the large Egyptian RTT's) are not due to the British ISP's poor performance.
However, further and more precises analyses must be done in order to be certain.