R Notebook

Contents

Preliminary CloudFront Logs Analysis

Logs were taken from tf-front-logs-production/master for the periods from 2020-01-01 to 2020-05-14

Analysis Overview

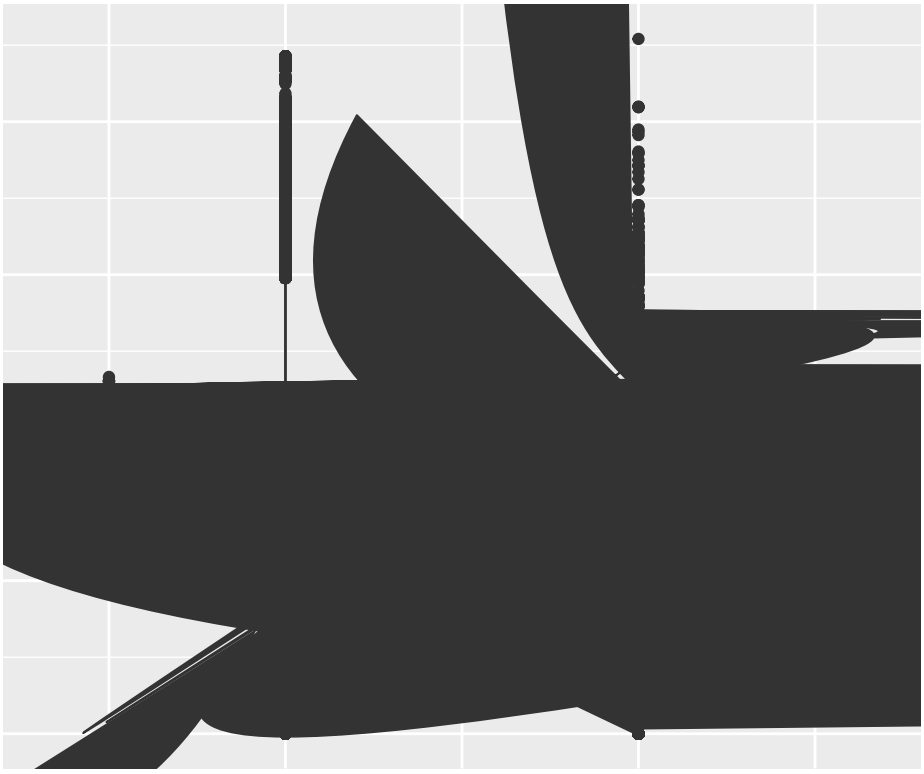
1. Majority of requests are tiny *~1KB*
2. Cache efficiency is high enough *~40%*
3. Top3 error types: *Error* (doesnt fit any of the other categories), *ClientCommError* (due to a communication problem between CloudFront and the viewer), *OriginError* (origin returned an incorrect response)
4. Majority of latencies are small ~1s due to efficient caching. But there are huge outliers ~10^6 seconds.
5. Latency outliers related to *China* and *United States*.
6. There is strong relationship between latency outliers and bytes consumed (incoming traffic).

Fields

Field	Description
date	The date on which the event occurred in the format YYYY-MM-DD
time	The time when the CloudFront server finished responding to the request (in UTC), for example, 01:42:39
x.edge.location	The edge location that served the request. Each edge location is identified by a three-letter code and an arbitrarily assigned number, for example, DFW3. The three-letter code typically corresponds with the International Air Transport Association airport code for an airport near the edge location.
sc.bytes	The total number of bytes that CloudFront served to the viewer in response to the request, including headers, for example, 1045619.
c.ip	The IP address of the viewer that made the request
cs.method	The HTTP request method: DELETE, GET, HEAD, OPTIONS, PATCH, POST, or PUT
Host	The domain name of the CloudFront distribution
cs.uri.stem	The portion of the URI that identifies the path and object, for example, /images/cat.jpg
sc.status	One of the following values: An HTTP status code; 000, which indicates that the viewer closed the connection (for example, closed the browser tab) before CloudFront could respond to a request
Referer	The name of the domain that originated the request

Field	Description
User.Agent	The value of the User-Agent header in the request
cs.uri.query	The query string portion of the URI, if any
Cookie	The cookie header in the request, including name-value pairs and the associated attributes
x.edge.result.type	How CloudFront classifies the response after the last byte left the edge location: Hit, RefreshHit, Miss, LimitExceeded, CapacityExceeded, Error, Redirect
x.edge.request.id	An encrypted string that uniquely identifies a request
x.host.header	The value that the viewer included in the Host header for this request
cs.protocol	The protocol that the viewer specified in the request: http, https, ws, or wss
cs.bytes	The number of bytes of data that the viewer included in the request, including headers
time.taken	The number of seconds (to the thousandth of a second, for example, 0.002) between the time that a CloudFront edge server receives a viewer's request and the time that CloudFront writes the last byte of the response to the edge server's output queue as measured on the server
x.forwarded.for	If the viewer used an HTTP proxy or a load balancer to send the request, the value of c-ip in field 5 is the IP address of the proxy or load balancer. In that case, this field is the IP address of the viewer that originated the request. This field contains IPv4 addresses (such as 192.0.2.44) and IPv6 addresses, as applicable.
ssl.protocol	Possible values include the following: SSLv3, TLSv1, TLSv1.1, TLSv1.2
ssl.cipher	Possible values include the following: ECDHE-RSA-AES128-GCM-SHA256, ECDHE-RSA-AES128-SHA256, ECDHE-RSA-AES128-SHA, ECDHE-RSA-AES256-GCM-SHA384, ECDHE-RSA-AES256-SHA384, ECDHE-RSA-AES256-SHA, AES128-GCM-SHA256, ...
x.edge.response.result.type	How CloudFront classified the response just before returning the response to the viewer. See also x-edge-result-type in field 14.
cs.protocol.version	Possible values include: HTTP/0.9, HTTP/1.0, HTTP/1.1, HTTP/2.0
fle.status	When field-level encryption is configured for a distribution, this field contains a code that indicates whether the request body was successfully processed. If field-level encryption is not configured for the distribution, the value of this field is a hyphen (-).
fle.encrypted.fields	The number of fields that CloudFront encrypted and forwarded to the origin
c.port	The port number of the request from the viewer
time.to.first.byte	The number of seconds between receiving the request and writing the first byte of the response, as measured on the server
x.edge.detailed.result.type	When x-edge-result-type (field 14) is not Error, this field contains the same value as x-edge-result-type. When x-edge-result-type is Error, this field contains the specific type of error: AbortedOrigin, ClientCommError, ClientGeoBlocked, ClientHungUpRequest, Error, InvalidRequest, InvalidRequestBlocked, ...
sc.content.type	The value of the HTTP Content-Type header of the response
sc.content.len	The value of the HTTP Content-Length header of the response
sc.range.start	When the response contains the HTTP Content-Range header, this field contains the range start value
sc.range.end	When the response contains the HTTP Content-Range header, this field contains the range end value

KB per reuest

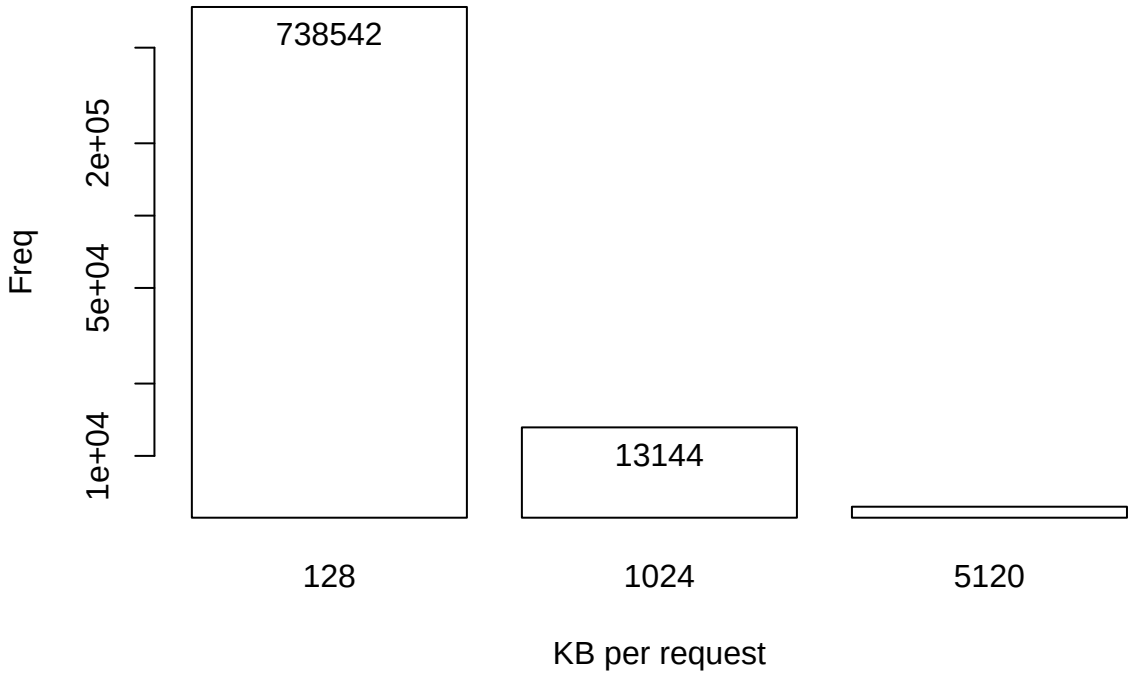


As expected majority of outliers inside GET and POST requests.

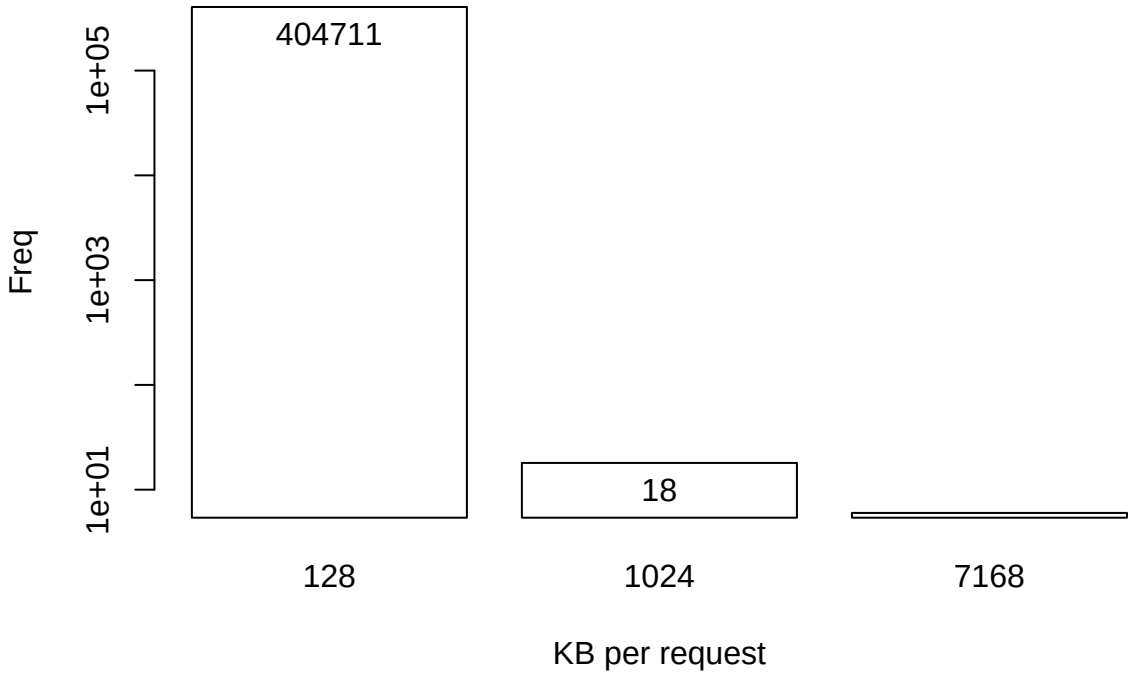
And what about incoming traffic: _

There are lots of GET requests with perceptible incoming traffic.

GET KB per reuest



POST KB per reuest



CloudFront Cache Efficiency

Error	Hit	Miss	Redirect	RefreshHit
6381	320365	430190	830	65

Level of cache *Hit* events is relatively low compared to level of *Miss* events:

```
[1] "42.7%"
```

CloudFront Errors

Error rate:

```
[1] "0.7%"
```

ClientCommError	ClientHungUpRequest	Error	Hit
2190	46	5760	320365
InvalidRequest	Miss	OriginConnectError	OriginError
1	1010068	8	1121
Redirect	RefreshHit		
928	71		

Top 3:

1. *Error* - An error occurred for which the error type doesnâ€™t fit any of the other categories. This error type can occur when CloudFront serves an error response from the CloudFront cache.

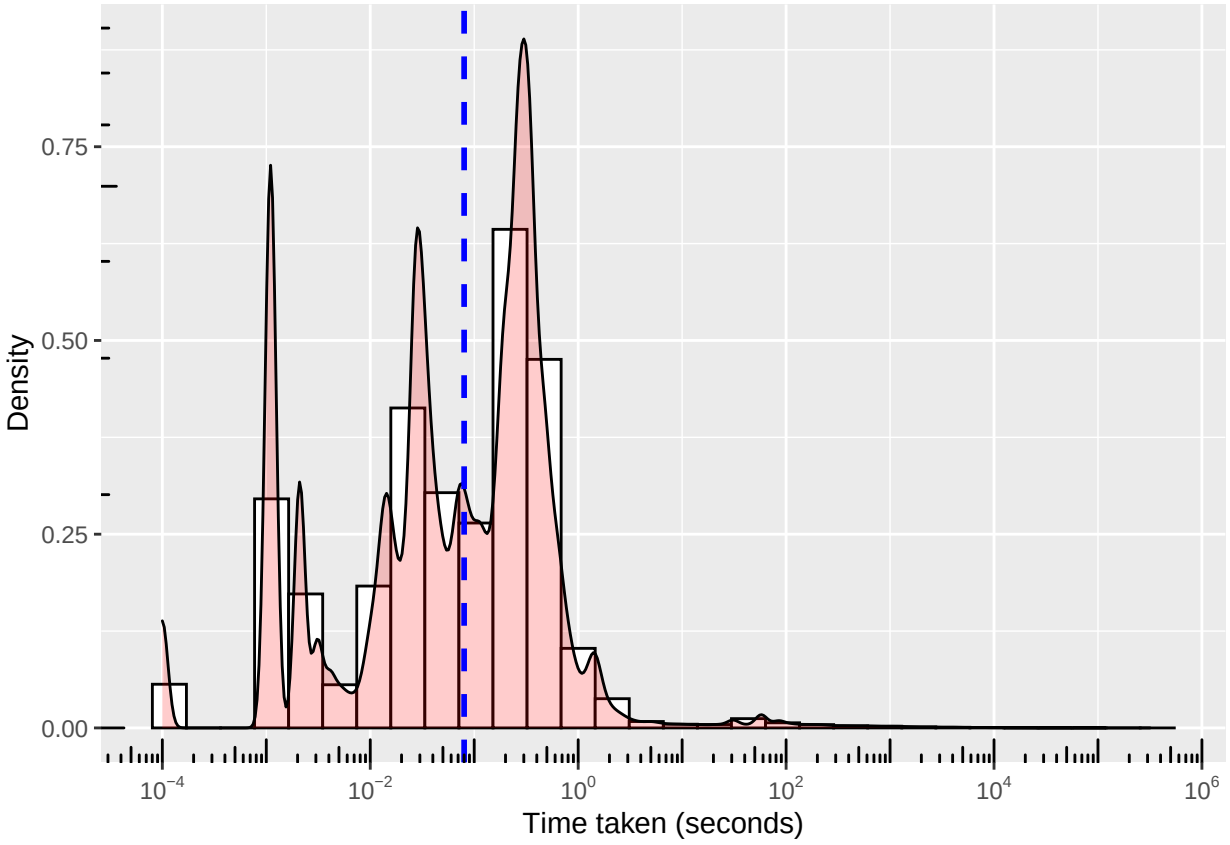
2. *ClientCommError* - The response to the viewer was interrupted due to a communication problem between CloudFront and the viewer.
3. *OriginError* - The origin returned an incorrect response.

No cases of *LimitExceeded* or *CapacityExceeded*.

Latency

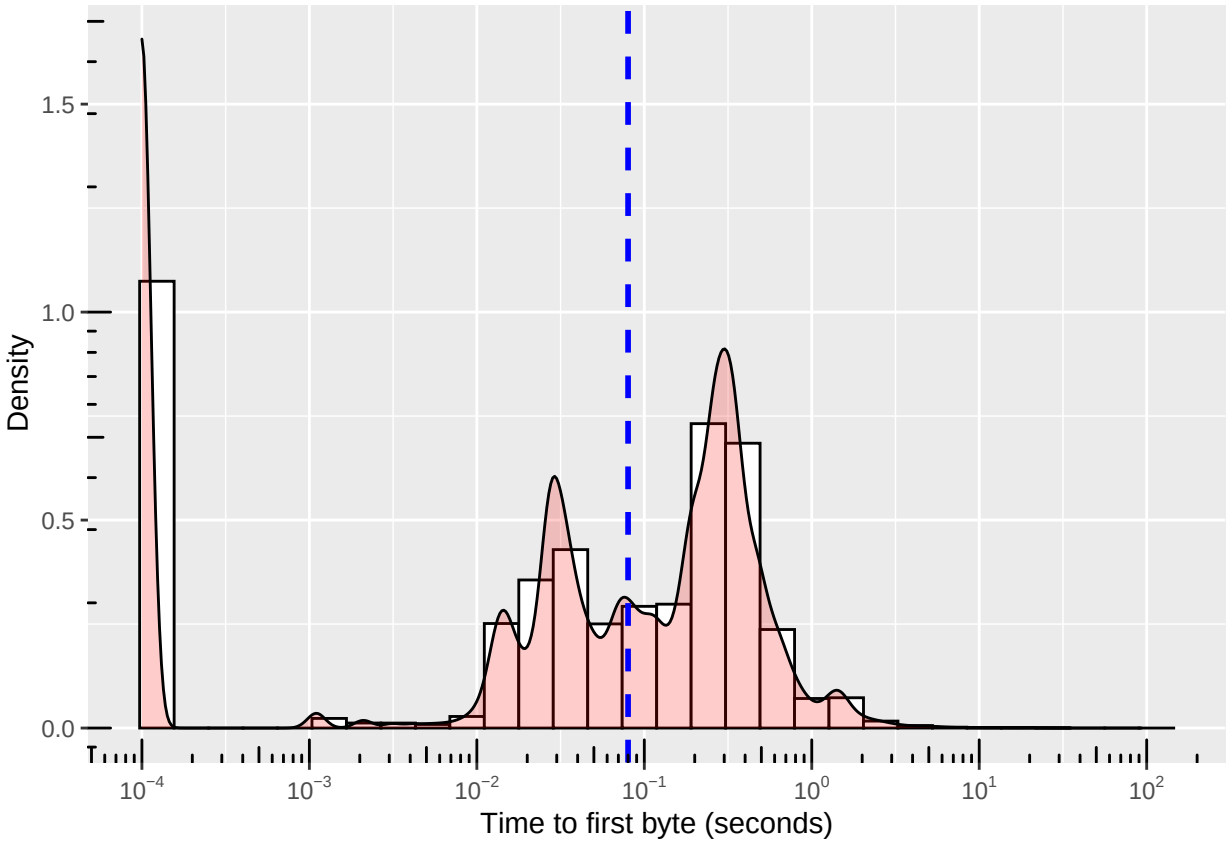
Time taken:

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

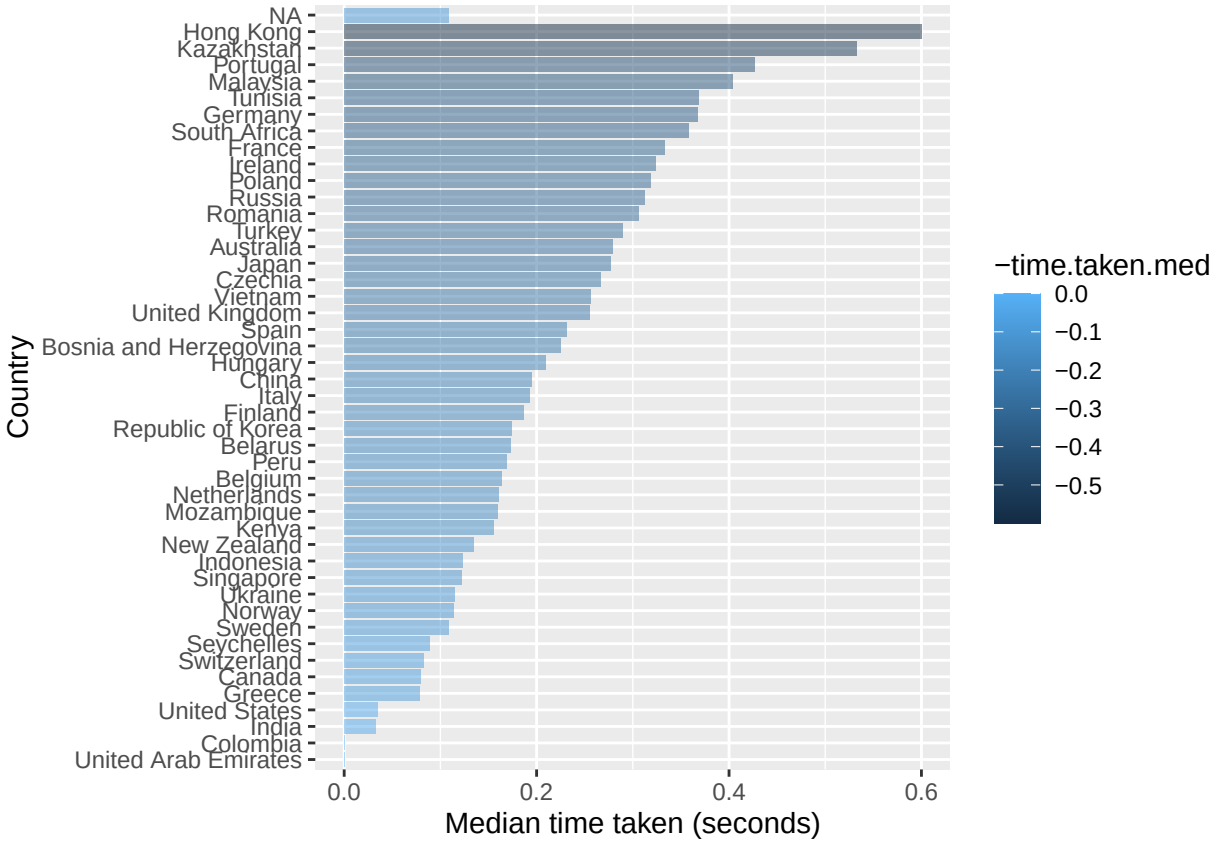


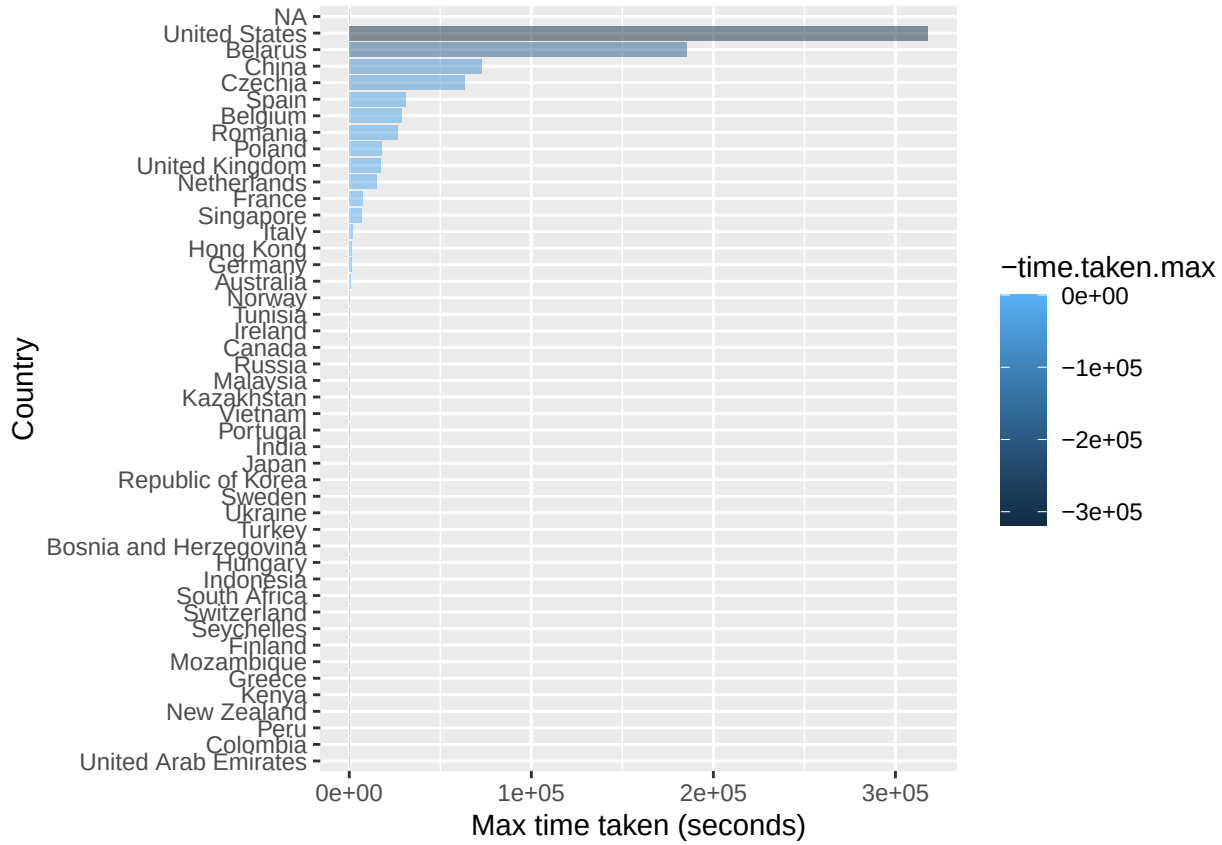
Time to first byte:

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


Density
1.5
1.0
0.5
0.0
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
Time to first byte (seconds)

Latency by country





There are lots of outliers in “United State”, “Belarus”, “China”. They should be analized separately.

Latency by result type

Latency outliers

Lets focus on values greater than 99% percentile 30.18843 (seconds).

Time taken >99%:

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
    30.2     56.2     86.7    698.4    294.7 317895.5
```

Countries >99%:

```
        China  United States         Poland         France        Belarus
         5565           4201           1656            727            451
  Netherlands        Czechia          Spain United Kingdom        Belgium
          381            204             71             61             33
        Italy      Singapore        Germany        Romania      Hong Kong
           17             11             10              9              5
    Australia         Norway        Tunisia
            2              1              1
```

Result type >99%:

```
 Miss Error   Hit
13331    71     4
```

So most of the >99% latencies are due to absence of entries in cache ("Miss"). It makes sense to eject "Error" and "Hit" categories to find out relationships between latency (time.taken) and other variables.

Latency and bytes consumed

```
Call:
lm(formula = time.taken ~ cs.bytes, data = logs.data.out.miss)

Residuals:
   Min     1Q Median     3Q    Max
-45836   -224   -107    155   9208

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.682e+03  8.569e+00  -312.9   <2e-16 ***
cs.bytes     2.743e+00  4.293e-03   639.0   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 778 on 13329 degrees of freedom
Multiple R-squared:  0.9684,    Adjusted R-squared:  0.9684
F-statistic: 4.083e+05 on 1 and 13329 DF,  p-value: < 2.2e-16
```

R-squared is close to 1.0 and p-value is small enough <0.001 to talk about statistically significant linear relationship between latency and bytes consumed (incoming traffic):

```
`geom_smooth()` using formula 'y ~ x'
```