CSDS 325 Project 5 Report

Abhinav Khanna
axk1312

Web Server performance Analysis for different times of the day

I. Introduction:

The theme of my project is to explore the theme of web server performance by delving into a comprehensive analysis of key network metrics - Round Trip Times(RTT), Time to First Byte(TTFB), Packet loss. I collect data for a selection of 35 websites and examine these metrics at different times of the day(Morning, Evening, and Night). The report aims to analyze the data to find correlations within the data with other factors such as geographical location, etc. These metrics are crucial for network administrators and developers to optimize their systems to ensure a smoother and reliable online experience for the users and because of my background in web development, I wanted to explore this topic of web server performance.

II. Procedure:

A. Gathering Data:

I used three main tools for data collection: Ping, curl and wget. Ping was employed to measure the Average Round Trip Time (RTT) across different networks, providing insights into the latency and speed of data transmission. Ping also collects data about packet loss which was also used in this analysis. Curl, on the other hand, was instrumental in determining the Time to First Byte (TTFB), a critical metric for assessing the responsiveness of a network. Wget was used to measure the page load time for each website. These tools were chosen for their reliability, widespread use in the industry, and the detailed metrics they offer, making them ideal for a comprehensive analysis of network performance.

Ping response(ping_script.py):

For the ping response, I sent 5 ping requests to each website URL and returned the response, and saved it into a text file which was parsed later to extract relevant information. In each execution of the script, one iteration on the list of 35 websites sent 5 pings to each and recorded the response. The script had 5 iterations of the main loop and the script was executed 5 times during the morning, evening, and night each. Hence the total number of responses recorded is 5 pings x 35 websites x 5 times in execution * 5 times during each time of day * 3 times of the day = 13125 data points. The ping response data is stored in a txt file called ping_results.txt.

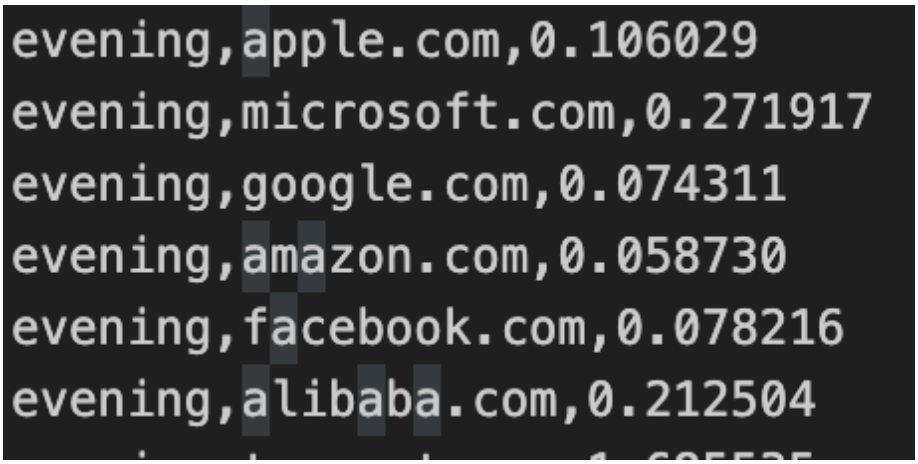
Each ping response is given in the following format: Time of the day, name of website, ping response

```
evening,apple.com,PING apple.com (17.253.144.10): 56 data bytes
64 bytes from 17.253.144.10: icmp_seq=0 ttl=55 time=42.407 ms
64 bytes from 17.253.144.10: icmp_seq=1 ttl=55 time=26.211 ms
64 bytes from 17.253.144.10: icmp_seq=2 ttl=55 time=22.185 ms
64 bytes from 17.253.144.10: icmp_seq=3 ttl=55 time=18.219 ms
64 bytes from 17.253.144.10: icmp_seq=4 ttl=55 time=22.831 ms

--- apple.com ping statistics ---
5 packets transmitted, 5 packets received, 0.0% packet loss
round-trip min/avg/max/stddev = 18.219/26.371/42.407/8.411 ms
```

Curl response(ttfb_script.py):

Similarly, to collect information on the Time to First Bytes, I used the curl command to extract the time to start the transfer using the response from curl. The curl response was saved to a text file called ttfb_results.txt. in the following format to specifically store the time to first byte for ease of analysis: Time of the day, name of the website, time to first byte



wget response(page_download_script.py):

I used Wget to collect the information regarding page download time. I used the wget command to extract the time taken to download the webpage files. I do not save the web files but only record the time taken to download the web pages. The wget response was saved in the following format: Time of the day, name of the website, and time taken to download the webpage content. I have a 20 second timeout for each wget call and in the case when the request timeouts, I simply save “timeouts” instead of the time. The results are stored in the text file called wget_results.txt.

```
night, apple.com, 0.352127s
night, microsoft.com, timeouts
night, google.com, 0.233267s
night, amazon.com, 0.324917s
night, facebook.com, 0.317709s
night, alibaba.com, 0.945363s
```

B. Analyzing data:

ping_visualize.py file:
For the round trip time, I used regex patterns to scrape and break down the data into two lists: results and overall_averages. The results list stores the average round trip time for each website during the different times of the day(morning, evening, and night). The overall_averages list stores the average round trip time for each website across the entire day. This gives me multiple visualizations and data to analyze and show which websites perform the best during different times of the day and which ones perform the worst during different times of the day.

Packet_loss_visualize.py file:
I used the same raw data from ping results on the 35 different websites to extract the website name, time of the day, and packet loss in each iteration. This allowed me to use the packet loss data to visualize it for the heat map to represent the average packet loss across websites and times of the day. For this, I only used the data points where there was some packet loss and not 0% packet loss. This was to show the average packet loss in the cases when there was some packet loss.

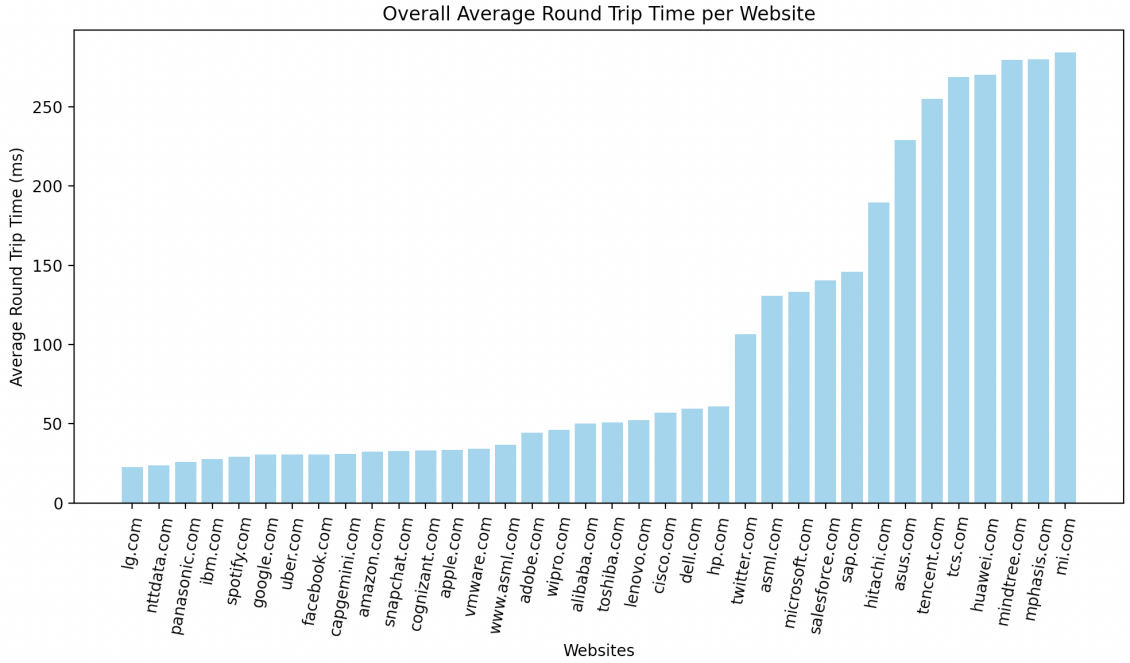
ttfb_visualize.py and wget_visualize.py:
For the Time To First Byte(TTFV) and page download time, I used the curl and wget tools respectively and used the in-built formatting available in the tools to extract the information into a text file. In the Python files, I used pandas to extract the information and store it in a data frame and then performed analysis and visualization using libraries including matplotlib, numpy, etc.

Ttfb_correlations.py and wget_correlations.py: These scripts contain the code to find the correlation coefficients between Time to First Byte(TTFB), page download time and the time of the day. The results are discussed later but I did this to have quantitative measure of how the time of day affects the factors such as page download time and Time to First Byte.

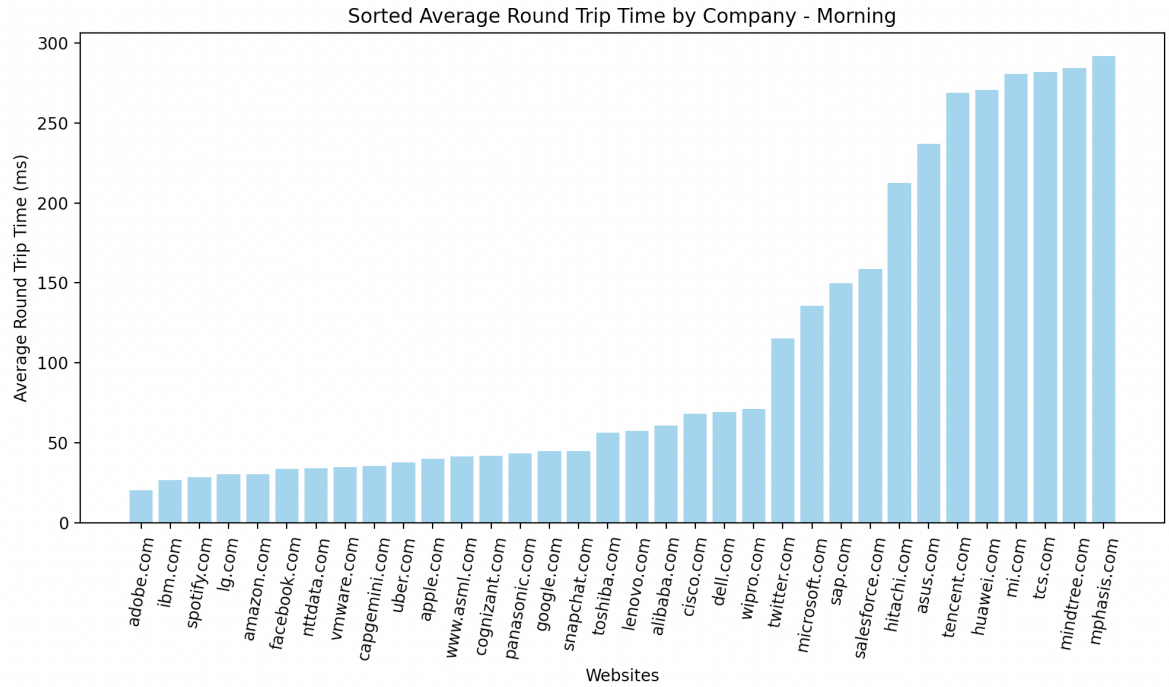
III. Results:

I. Average Round Trip Time(RTT) by Company

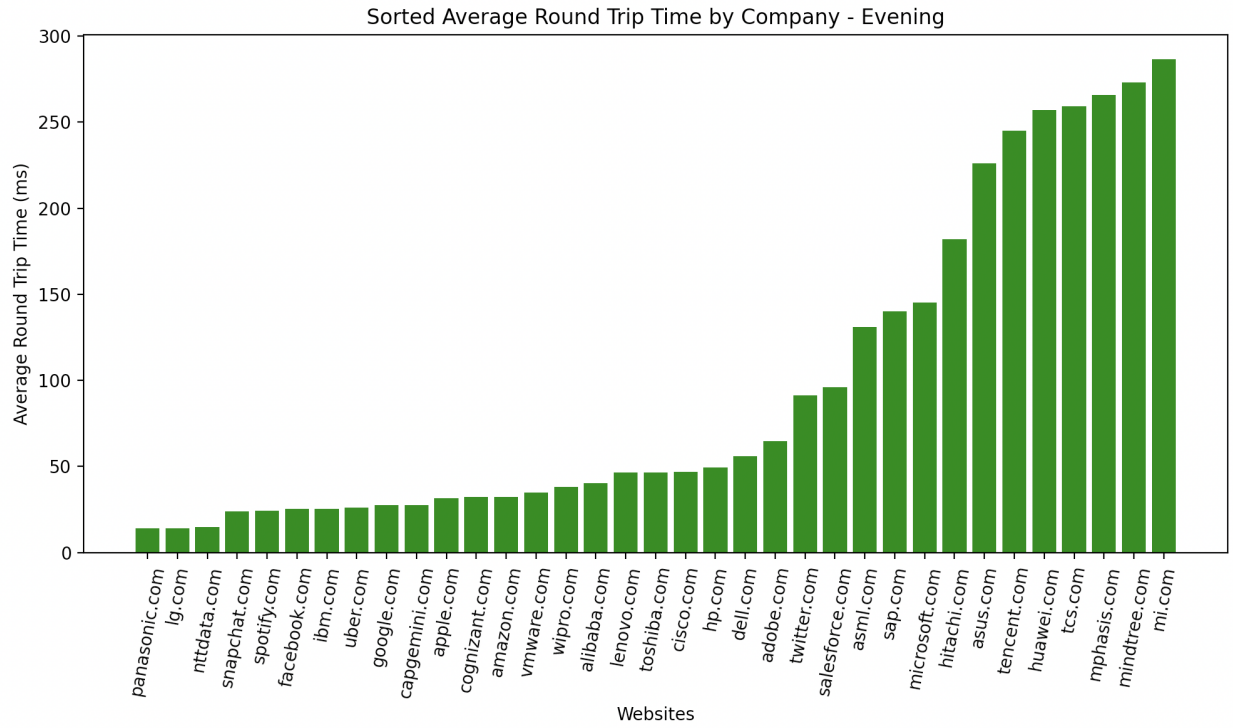
The entire day(Includes morning, evening and night):



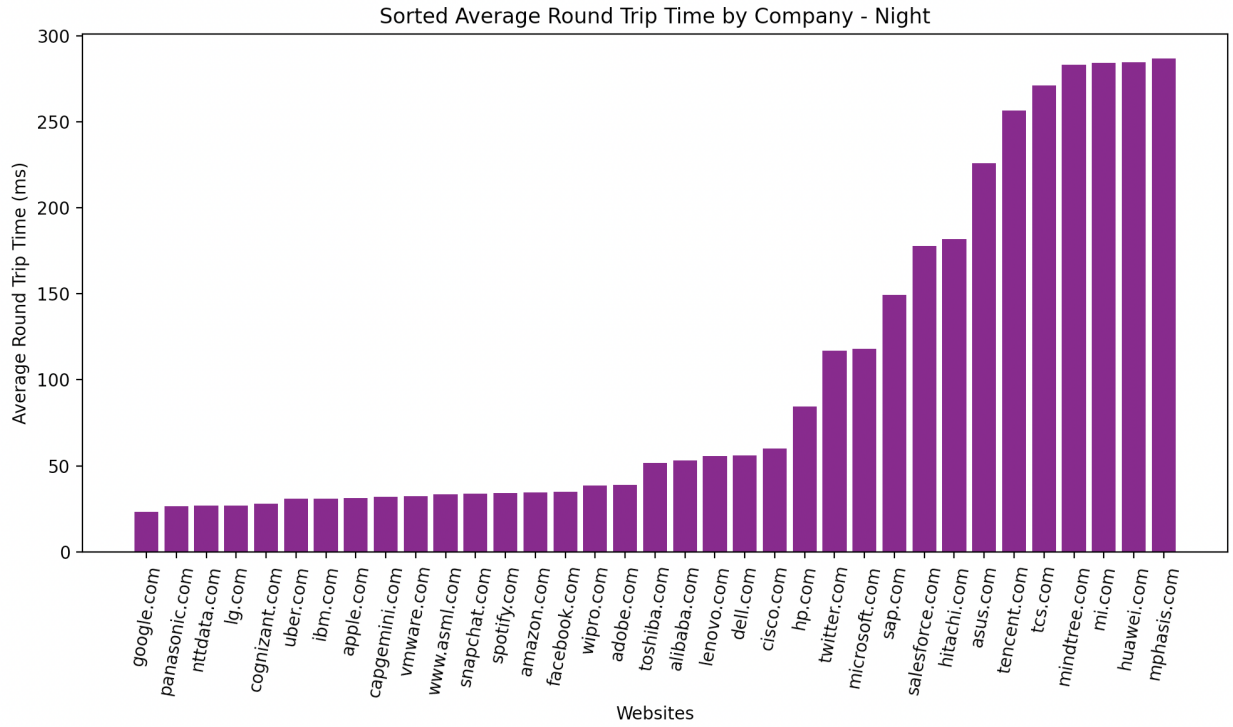
Morning:



Evening:



Night:

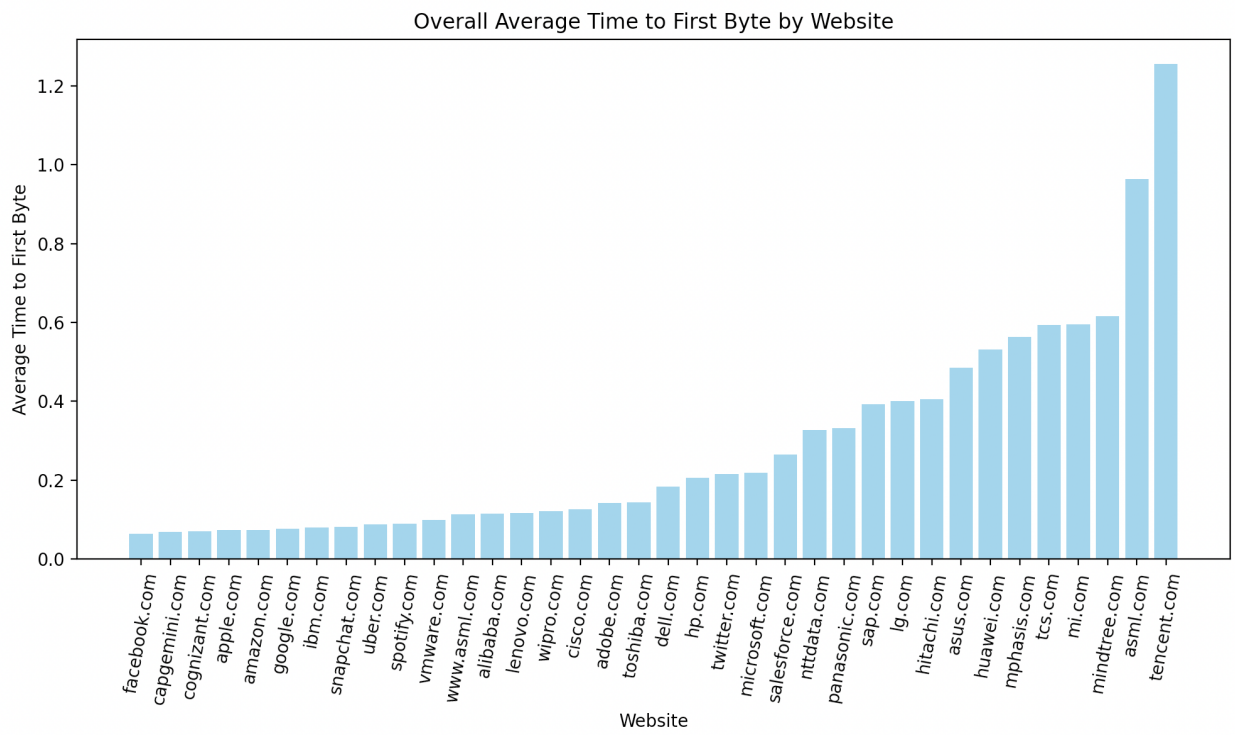


From the graphs above, we can see that most of the websites mentioned in the list that are US-based companies have a lower Round Trip Rime than most of the non-US-based companies.

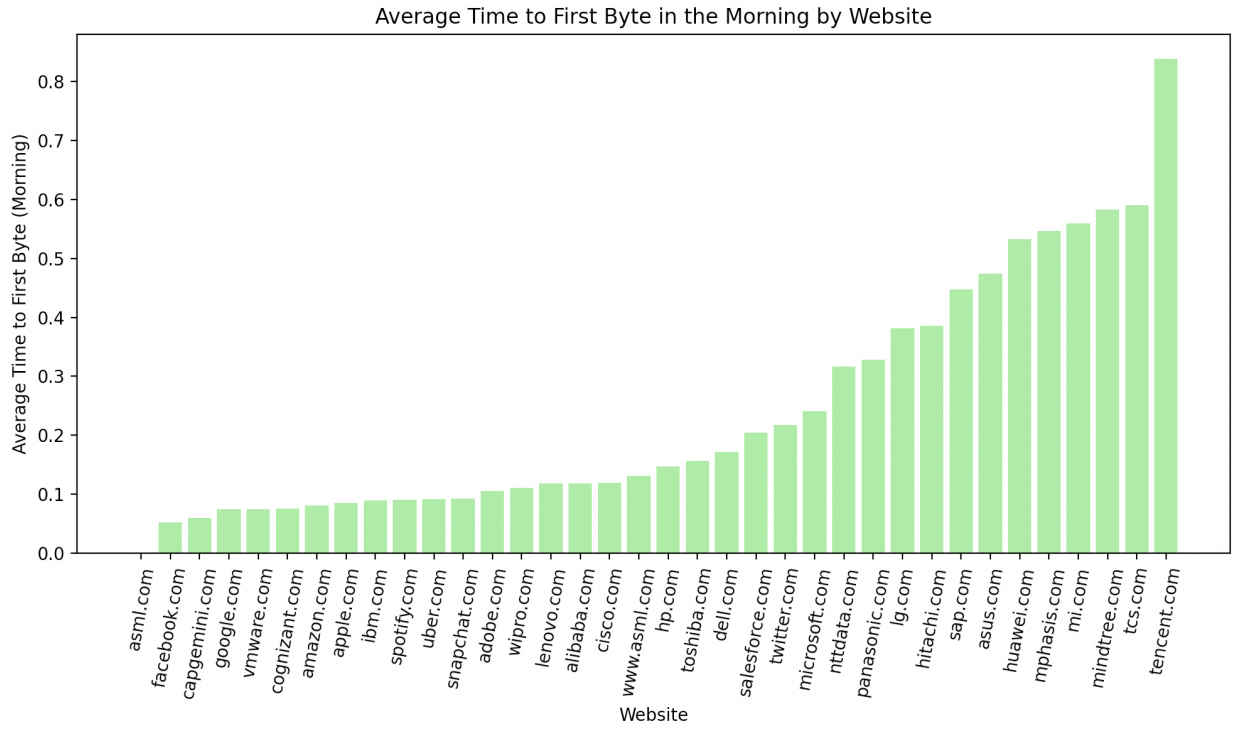
Companies like mi.com(China), mphasis.com(India), mindtree.com(India), and Huwaei(China) continuously have the largest time taken for Round Trip Time. This could also be attributed to the geographical location and because mi.com's servers are located farther away from other US-based companies, then the Round Trip Time also increases. However, we have the exception of NTTData which is a Japanese company but has a better relationship and presence in the US. NTTData operates 7 data centers in the US which might be one of the reasons why their Round Trip Time is faster as compared to other international companies who might not have their servers in the US.

II. Average Time to First Byte(TTFB) by Company

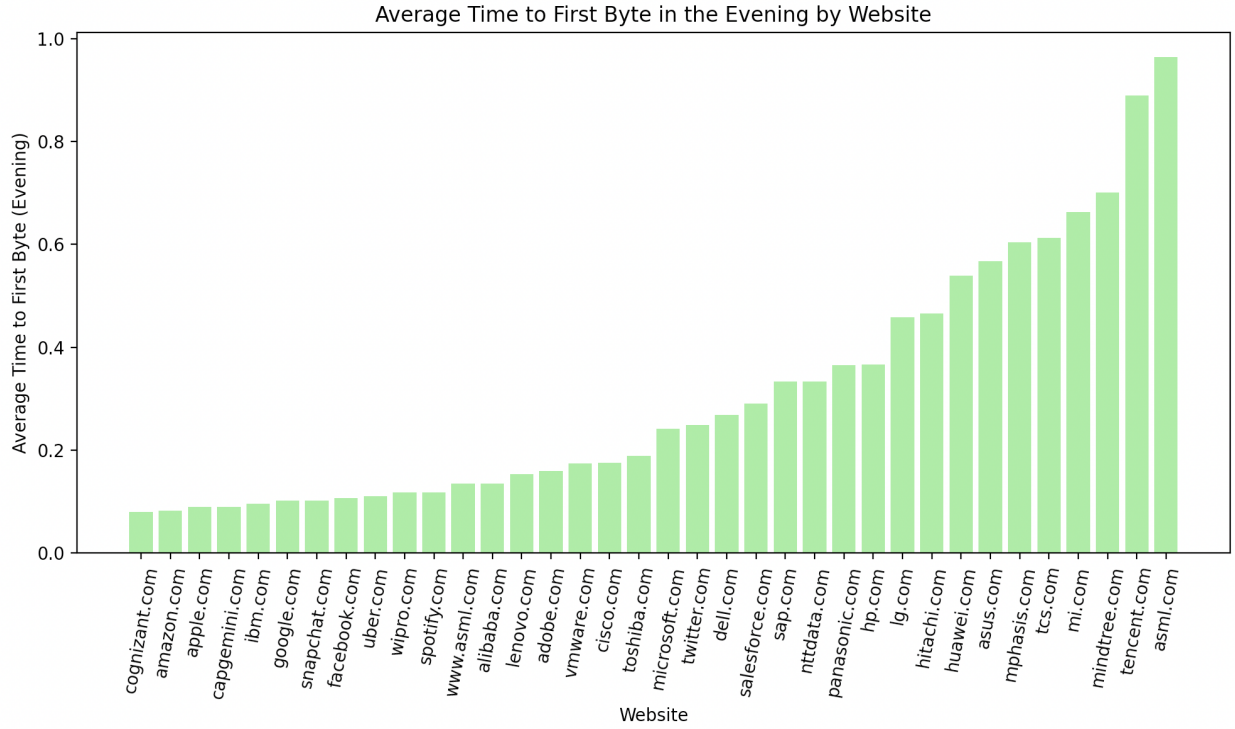
The entire day:



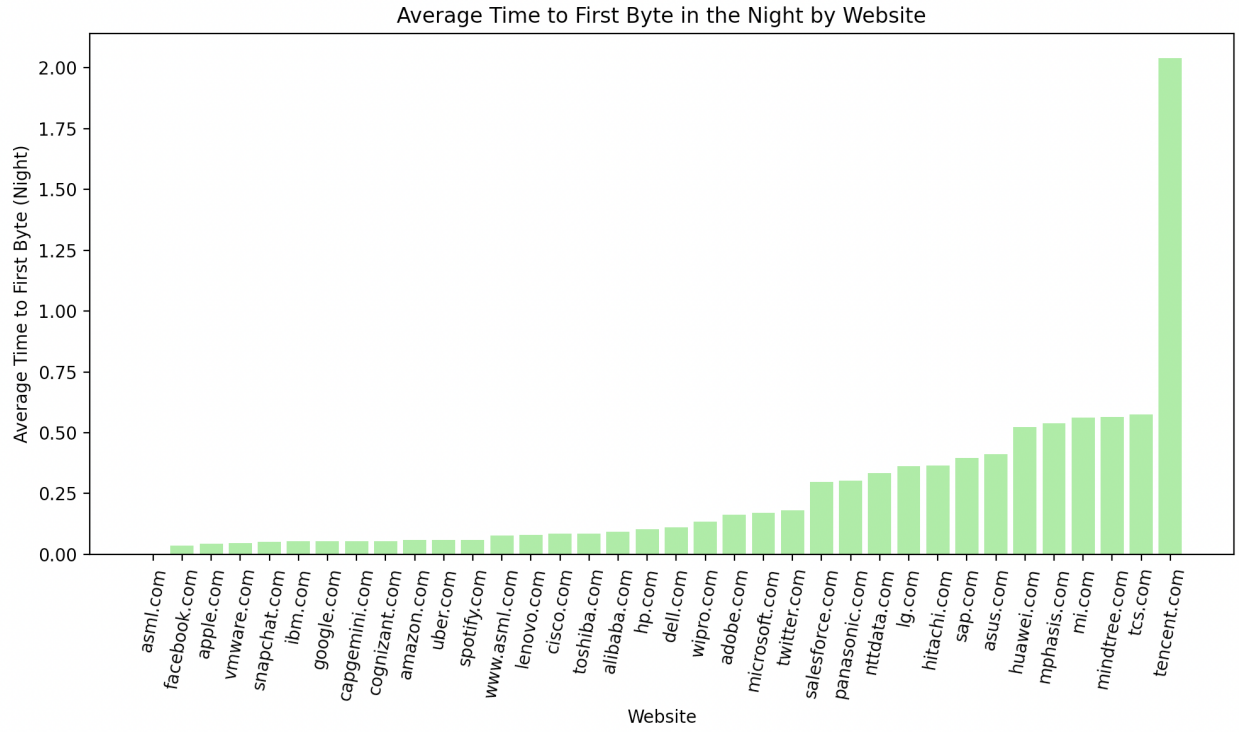
Morning:



Evening:

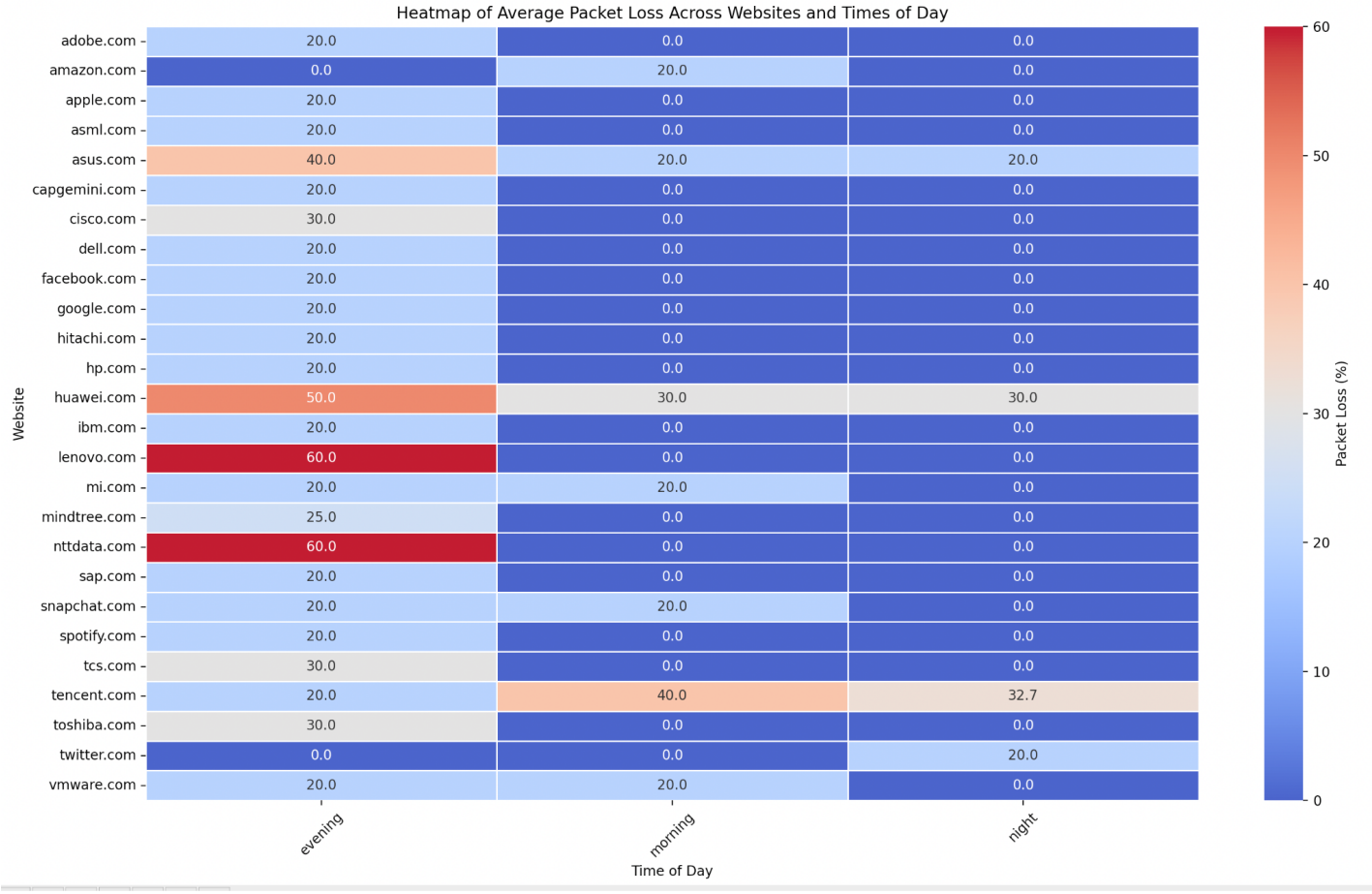


Night:



Similar to the Round Trip Time, the Time to First byte for US Companies is typically much faster than that of non-US companies. The ones taking the most time are Tencent, TCS Mindtree, and Mi which are all non-US Companies. Companies like Facebook.com, Amazon.com, and Google.com which have to cater to a huge variety of requests are able to maintain their ranking in the top 10 across different times of the day.

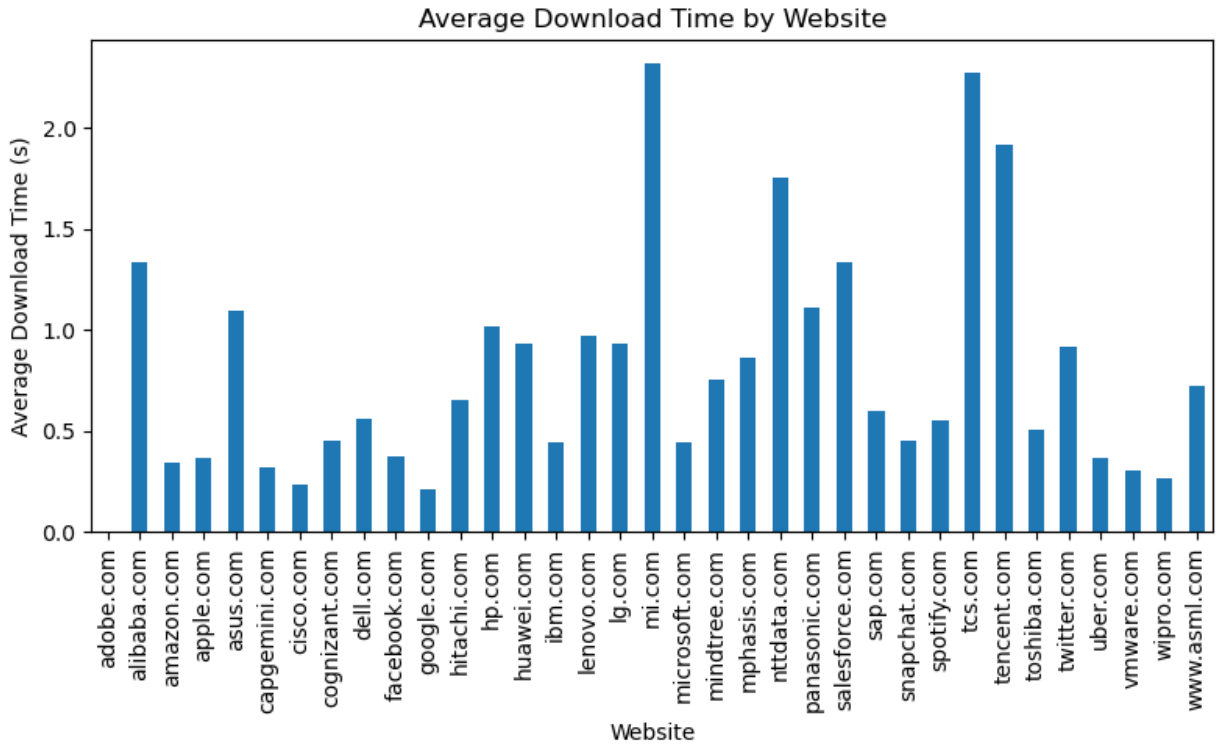
III. Heatmap Showing Average Packet Loss for each website Based on the time of the day



The data shows that out of the 35 websites I analyzed, 26 of those suffered some packet loss during the ping execution. If we look at the websites with the highest average packet loss percentage, all of them are non-US-based companies like Asus(Taiwanese), Huawei(Chinese), Lenovo(Chinese), nttdata(Japanese), and Tencent (Chinese). However, there are US-based companies also that suffered packet loss. The lowest number of packet losses was at night which could be because of the reduction in the traffic at night and the highest number of packet losses occurred during the evening when we can assume that the traffic is highest.

IV. Change in Average Page download time by company:

Here is a visualization of the average Page download time by company:



Please disregard adobe since the server of adobe was not responding to any wget requests and always timed out.

The following data shows the percentage change in the average time taken to download a webpage for the 35 different websites. A positive change reflects that the average time taken to download the webpage using wget increased and a negative change reflects that the average time taken to download the webpage reduced.

For the analysis, I have assumed that the usage increases as the day progresses. The change from morning to evening was mainly positive for companies such as Alibaba, Facebook. This is understandable because these are social media and e-commerce websites whose usage increases as the day passes. However, there were some cases such as Amazon.com, Google.com, and Snapchat.com whose average page load time reduced rather than increasing with the increase in usage. This might be because of better resource scaling and optimization adopted by large companies. Companies might be allocating resources more efficiently especially during the day when they know that the traffic is going to increase.

We have a similar case for change from evening to night where Amazon.com, and uber.com have increased page load time because of an increase in traffic. There are also cases where the average time reduces. This could be because of a lot of reasons, such as a decrease in traffic but we cannot conclude anything specifically.

TimeOfDay	Morning_to_Evening_%	Evening_to_Night_%
Website		
Website	-18.9	40.8
adobe.com	NaN	NaN
alibaba.com	15.7	-21.3
amazon.com	-7.3	16.2
apple.com	-13.9	-2.4
asus.com	1.5	4.1
capgemini.com	4.9	-4.0
cisco.com	-15.8	7.5
cognizant.com	7.4	17.7
dell.com	-9.1	-2.9
facebook.com	49.5	-35.2
google.com	-4.8	11.3
hitachi.com	-6.7	7.5
hp.com	13.2	-6.8
huawei.com	-20.6	13.9
ibm.com	-8.3	1.2
lenovo.com	13.3	-0.7
lg.com	-8.1	40.2
mi.com	-4.8	86.7
microsoft.com	-19.8	3.7
mindtree.com	-3.8	-3.6
mphasis.com	-13.0	12.7
nttdata.com	2.3	73.2
panasonic.com	2.8	5.2
salesforce.com	-16.3	60.5
sap.com	-9.3	35.8
snapchat.com	-7.4	0.8
spotify.com	-8.4	-0.4
tcs.com	-20.8	44.6
tencent.com	-8.7	-9.3
toshiba.com	-2.2	0.1
twitter.com	-0.1	5.6
uber.com	-9.6	16.1
vmware.com	2.1	2.5
wipro.com	-25.8	30.2
www.asml.com	-20.2	11.4

One interesting thing that I found from the results is that the websites with the highest change in average page load time from evening to night are:

Website	Morning_to_Evening	Evening_to_Night
mi.com	-4.8%	76.7%
nttdata.com	2.3%	73.2%
salesforce.com	-16.3%	60.5%
tcs.com	-20.8%	44.6%
lg.com	-8.1%	40.2%

Except Salesforce, all the other websites are non-US and the increase in average page load would be because of an increase in traffic due to timezone differences. The recordings have been done per the EST Timezone and the increase in average page load time would be because even though it is night in the US, it is still day in the Asian countries. Websites like mi.com, nttdata.com, tcs.com, and lg.com which are based out of Asian countries and this drastic increase could be attributed to geographical and timezone differences. This could very well be because of some other reasons, but this is my analysis based on speculation and the evident differences. The sudden increase in time taken by salesforce.com is an outlier and could be because of several reasons.

V. Correlation Coefficient calculation for websites for different parameters

Now, I wanted to find the numeric correlation between the time of the day and the parameters we recorded earlier for each website. This will allow me to understand how strong or weak the correlation between the two factors is based on the data I collected. The picture on the left shows the correlation between the time of the day (Morning, evening, night) and the time taken for page download. The picture on the right shows the correlation between the time of the day and the Time to First Byte(TTFB). The values are within the range of -1 to 1 where a value closer to 1 represents a strong correlation and a number closer to -1 represents a weak correlation.

	Company	Correlation
0	nttdata.com	0.446031
1	mi.com	0.383725
2	sap.com	0.348791
3	asus.com	0.231349
4	salesforce.com	0.209100
5	lg.com	0.204028
6	panasonic.com	0.203001
7	cognizant.com	0.195215
8	tcs.com	0.177935
9	amazon.com	0.153451
10	twitter.com	0.140031
11	vmware.com	0.120377
12	google.com	0.118306
13	uber.com	0.084440
14	lenovo.com	0.079137
15	hp.com	0.060734
16	capgemini.com	0.020640
17	hitachi.com	0.008133
18	facebook.com	-0.023130
19	mphasis.com	-0.028337
20	toshiba.com	-0.040373
21	wipro.com	-0.054450
22	alibaba.com	-0.086139
23	ibm.com	-0.088051
24	snapchat.com	-0.091628
25	huawei.com	-0.126969
26	spotify.com	-0.136096
27	dell.com	-0.153899
28	www.asml.com	-0.186420
29	cisco.com	-0.189115
30	apple.com	-0.195260
31	tencent.com	-0.212259
32	microsoft.com	-0.260575
33	mindtree.com	-0.293641
34	adobe.com	NaN

	Company	Correlation
0	tencent.com	0.276696
1	salesforce.com	0.185237
2	adobe.com	0.142939
3	nttdata.com	0.126289
4	wipro.com	0.076477
5	mi.com	0.009674
6	mphasis.com	-0.032137
7	mindtree.com	-0.035087
8	tcs.com	-0.035457
9	hp.com	-0.038221
10	lg.com	-0.039480
11	capgemini.com	-0.041269
12	huawei.com	-0.044239
13	vmware.com	-0.053394
14	sap.com	-0.060069
15	facebook.com	-0.061488
16	hitachi.com	-0.069794
17	dell.com	-0.071506
18	asus.com	-0.078386
19	cisco.com	-0.122224
20	snapchat.com	-0.140547
21	www.asml.com	-0.147314
22	toshiba.com	-0.149072
23	cognizant.com	-0.154590
24	panasonic.com	-0.156217
25	alibaba.com	-0.163711
26	microsoft.com	-0.187231
27	google.com	-0.191440
28	uber.com	-0.191631
29	spotify.com	-0.236237
30	lenovo.com	-0.244401
31	amazon.com	-0.271446
32	twitter.com	-0.277282
33	ibm.com	-0.318362
34	apple.com	-0.378601
35	asml.com	NaN

Correlation between Time of day and page download time(left) and correlation between Time of day and Time to First Byte(right).

From the above correlation values for page download time, we can see that there is a positive correlation against the Time of the day for almost half of the websites. However, for the Time to First Byte, only 5 of the 35 websites have a positive correlation coefficient which represents that the Time to First Byte might not be correlated with the various times of the day.