Exploratory Data Analysis on Robots.txt

Vajinder Kaur

2025-03-24

Load the Required Libraries

1. dplyr : For Data Manipulation
2. ggplot2 : Data Visualization
3. lubridate : For working with date and time.
4. tidyr : Data Cleaning
5. knitr : For including Graphs in the output file

```
suppressPackageStartupMessages({  # Load required libraries
  library(dplyr)
  library(ggplot2)
  library(lubridate)
  library(tidyr)
  library(knitr)
  library(kableExtra)
})
```

Creating Functions

load__ data

This function reads the given csv file and later converts the Wayback Timestamps into standard time and date format. Additionally, it creates a column called `Blocked_All_Bots`, which is binary variable, containing `Yes` if the domain blocks all the bots in addition to certain bots for that timestamp and `No` if it doesn't block all bots but only certain ones.

extract__ai__blocks

This function goes through the column "Blocked__Crawlers and identify specific AI-based crawlers, such as GPTBot, ChatGPT-User, OAI-SearchBot, Google-Extended, PerplexityBot, anthropic-ai, ClaudeBot, and Claude-Web. Further, it generates single row per AI crawler.

plot__crawler__frequency

This Function creates a frequency bar graph showing how many times a certain AI crawler was blocked by Group A/B Publishers. Saves the plot as per the given format by user.

plot__heatmap

This function creates a Heatmap of How Publishers blocked AI Crawlers over the time.

plot__facet__heatmap

This function creates a heatmap depicting how different publishers blocked certain AI crawlers over the time.This specifically contains Crawlers.

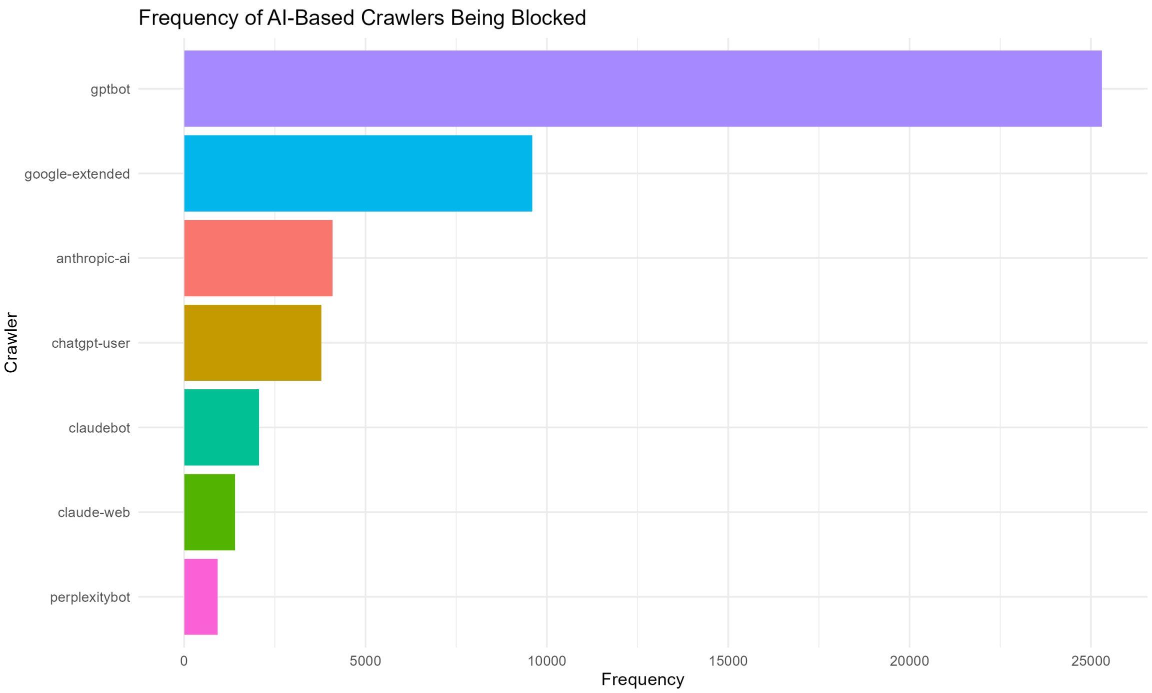
```r
load_data <- function(file) {
  data <- read.csv(file)
  data <- data %>%  # Add a column indicating if all bots were blocked
    mutate(Blocked_All_Bots = ifelse(grepl("\\*", Blocked_Crawlers), "Yes", "No"),
           Timestamp = as.POSIXct(as.character(Timestamp),
                                  format = "%Y%m%d%H%M%S", tz = "UTC"))
  return(data) }

extract_ai_blocks <- function(data, ai_crawlers) {
  data %>% filter(grepl(paste(ai_crawlers, collapse = "|"), Blocked_Crawlers)) %>%
    mutate(Blocked_Crawlers = strsplit(as.character(Blocked_Crawlers), ", ")) %>%
    unnest(Blocked_Crawlers) %>% filter(Blocked_Crawlers %in% ai_crawlers)}

plot_crawler_frequency <- function(block_data, filename) {
  freq <- block_data %>% count(Blocked_Crawlers)
  p <- ggplot(freq, aes(x = reorder(Blocked_Crawlers, n),
    y = n, fill = Blocked_Crawlers)) + geom_bar(stat = "identity") +
    coord_flip() +labs(title = "Frequency of AI-Based Crawlers Being Blocked",
    x = "Crawler", y = "Frequency") + theme_minimal() + theme(legend.position = "none")
  ggsave(filename, plot = p, width = 10, height = 6, dpi = 300, bg = "white") }

plot_heatmap <- function(data, filename) {
  p <- ggplot(data, aes(x = YearMonth, y = Domain, fill = Count)) +
    geom_tile(color = "white") + scale_fill_gradient(low = "white", high = "red") +
    labs(title = "Heatmap of Blocked Crawlers Over Time",
         x = "Month", y = "Domain", fill = "Block Count") +
    theme_minimal() + theme(axis.text.x = element_text(angle = 45, hjust = 1))
   # Save the heatmap with the specified filename
  ggsave(filename, plot = p, width = 12, height = 8, dpi = 300, bg = "white") }

plot_facet_heatmap <- function(data, filename = "blocked_crawlers_domains.png") {
  p <- ggplot(data, aes(x = YearMonth, y = Domain, fill = Count)) +
    geom_tile(color = "white") + # Color scale for intensity
    scale_fill_viridis_c(option = "plasma", direction = -1) +
    facet_wrap(~ Blocked_Crawlers, scales = "free_y")+ # Separate heatmap/crawler
    labs(title = "Blocked Crawlers Across Domains Over Time",
         x = "Month", y = "Domain", fill = "Block Count") +
    theme_minimal() + theme(axis.text.x = element_text(angle = 45, hjust = 1),
          strip.text = element_text(size = 7, face = "bold"))
  ggsave(filename, plot = p, width = 12, height = 8, dpi = 300, bg = "white") }
```

Frequency Graphs

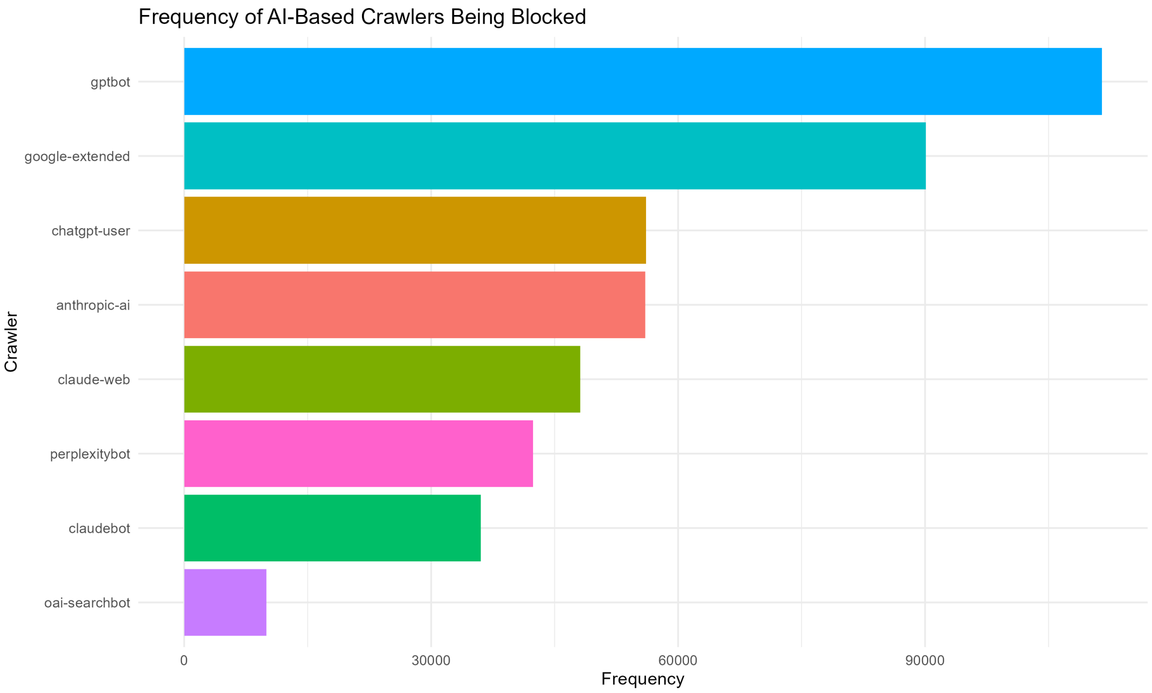
Clearly `GptBot` is blocked the most of the time in both of the Groups. `Google-Extended` is blocked by Group B Publishers almost as much as `GptBot`. Additionally, Group B publishers are blocking AI crawlers more than Group A in general as well.

```r
ai_crawlers <- c("gptbot", "chatgpt-user", "claudebot", # Define AI crawlers
   "perplexitybot", "google-extended", "anthropic-ai", "claude-web", "oai-searchbot")

data1 <- load_data("UpdatedGroup1.csv")  # Process Group 1
ai_blocks1 <- extract_ai_blocks(data1, ai_crawlers)
plot_crawler_frequency(ai_blocks1, "Frequency_GroupA.png")
```

```r
data2 <- load_data("UpdatedGroup2.csv") # Process Group 2
ai_blocks2 <- extract_ai_blocks(data2, ai_crawlers)
plot_crawler_frequency(ai_blocks2, "Frequency_GroupB.png")

include_graphics("Frequency_GroupA.png")
```



```r
include_graphics("Frequency_GroupB.png")
```



Getting Data ready for the Group A heatmap

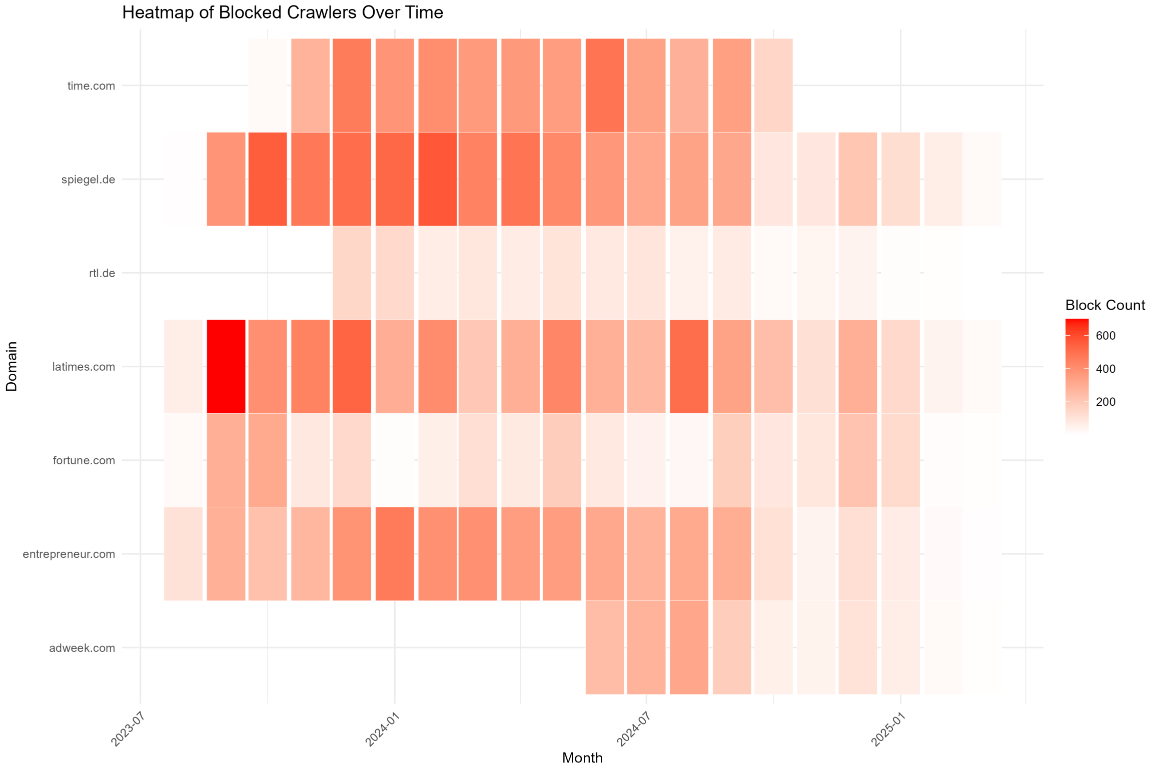
Filter out Domains, YearMonth (created using Timestamp) and AI Crawlers in Blocked_Crawlers to have the data ready for Heatmaps.

```
ai_blocks1$Timestamp <- ymd_hms(ai_blocks1$Timestamp)
ai_blocks1$YearMonth <- format(ai_blocks1$Timestamp, "%Y-%m")

# Count occurrences per domain, per crawler, per month
ai_blocks1_summary <- ai_blocks1 %>%
  group_by(YearMonth, Blocked_Crawlers, Domain) %>%
  summarise(Count = n(), .groups = "drop")
#Making sure about YM Format
ai_blocks1_summary$YearMonth <- as.Date(paste0(ai_blocks1_summary$YearMonth, "-01"))
```

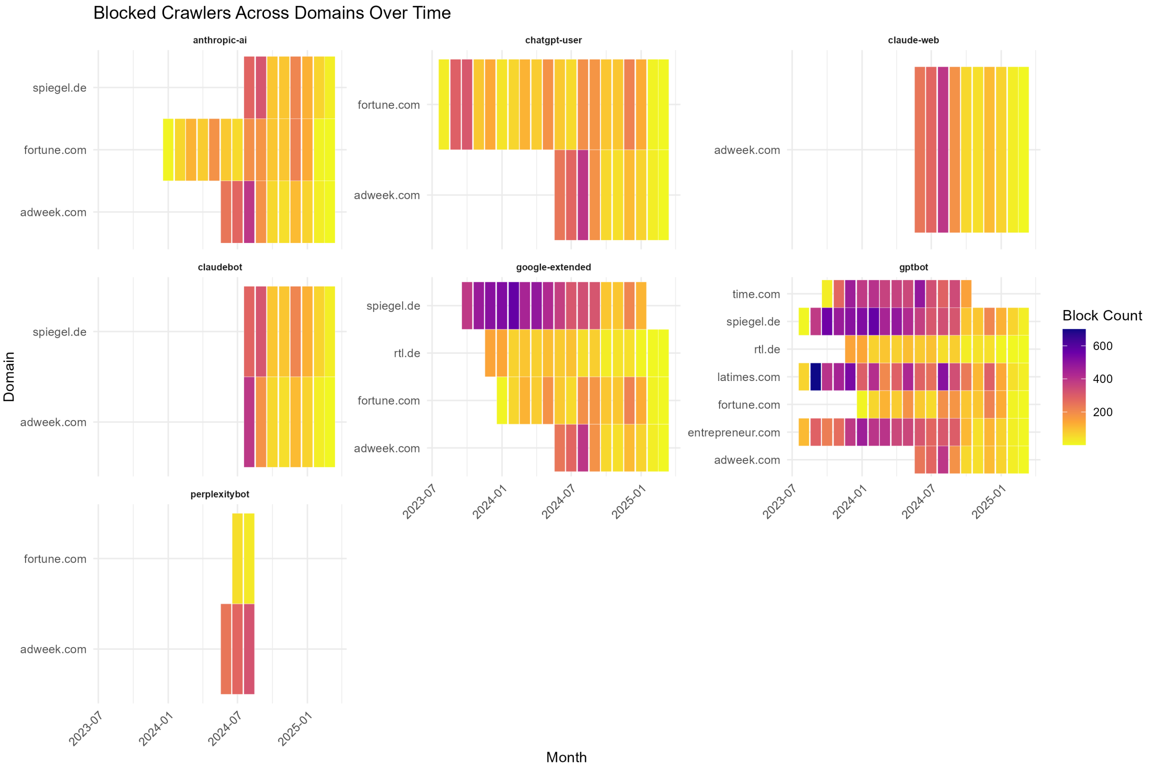
Group A Heatmaps

Out of 10 Publishers in Group A, only 7 specifically blocks out AI Crawlers. Other three (Blavity.com, independent.co.uk, prisa.com) usually block all the bots. We can see some of the Publishers such as rtl.de, adweek.com, time.com) started blocking AI Crawlers later than others. GptBot is mostly guarded against while perplexitybot is barely blocked as it was blocked for couple of months around July 2024. But common pattern is as the time passed, Blocking against Bots actually lessened which is surprising.

```
plot_heatmap(ai_blocks1_summary, "GroupA_Heatmap_DomainTime.png")
plot_facet_heatmap(ai_blocks1_summary, "GroupA_Heatmap_DomainTimeBots.png")
include_graphics("GroupA_Heatmap_DomainTime.png")
```



```r
include_graphics("GroupA_Heatmap_DomainTimeBots.png")
```



Getting Data ready for the Group B heatmap

Filter out Domains, YearMonth (created using Timestamp) and AI Crawlers in Blocked_Crawlers to have the data ready for Heatmaps.

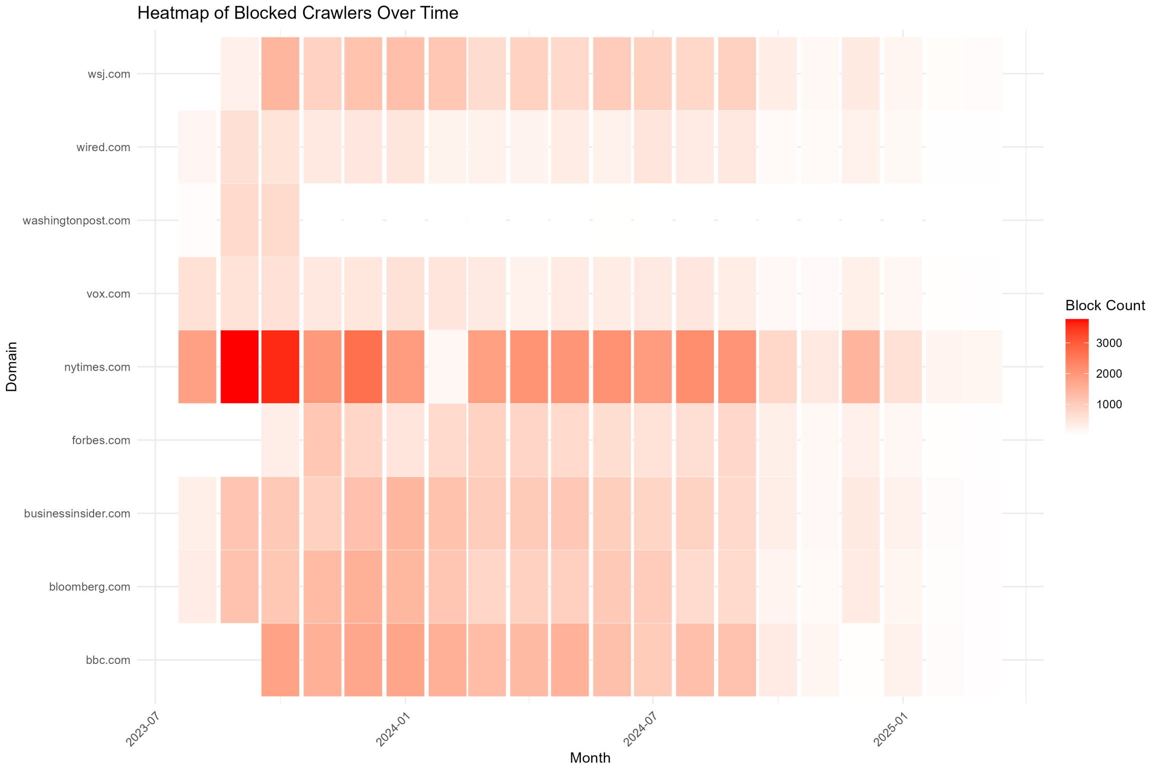
```r
ai_blocks2$Timestamp <- ymd_hms(ai_blocks2$Timestamp)
ai_blocks2$YearMonth <- format(ai_blocks2$Timestamp, "%Y-%m")

ai_blocks2_summary <- ai_blocks2 %>%
  group_by(YearMonth, Blocked_Crawlers, Domain) %>%
  summarise(Count = n(), .groups = "drop")
ai_blocks2_summary$YearMonth <- as.Date(paste0(ai_blocks2_summary$YearMonth, "-01"))
```

Group A Heatmap

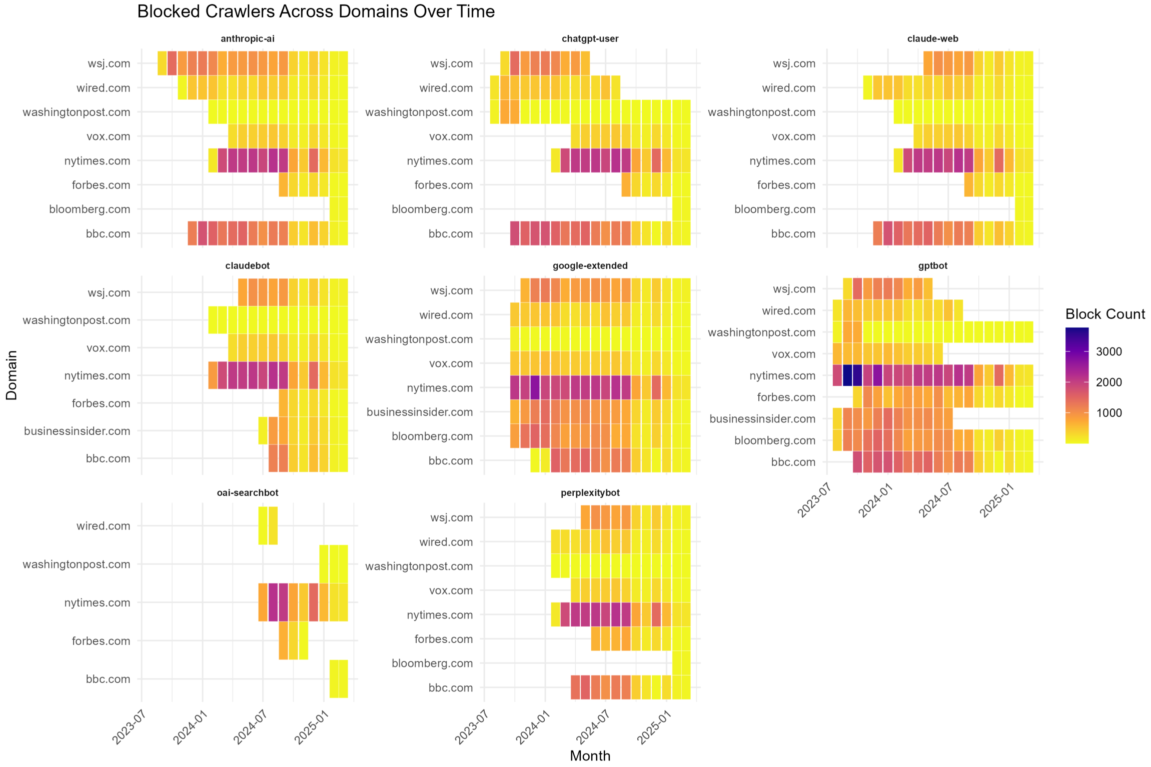
All 9 Publishers in Group B (of which data could be collected) guard against most of the AI crawlers. Most of them started blocking these Crawlers quite early as well. Compared to the Group A where the maximum blocks are 600, Group B blocked Crawlers more (3000 max). But surprising thing was for washingtonpost.com to block less of AI Crawlers in the later years. We will investigate it further.

Second heatmap shows how every bot is blocked by most of the Publishers as compared to just one or two Publishers Blocking some of the bots in Group A. It leads to possibility that Group B publishers are more guarded against AI than Group A.

```r
plot_heatmap(ai_blocks2_summary, "GroupB_Heatmap_DomainTime.png")
plot_facet_heatmap(ai_blocks2_summary, "GroupB_Heatmap_DomainTimeBots.png" )
include_graphics("GroupB_Heatmap_DomainTime.png")
```



```r
include_graphics("GroupB_Heatmap_DomainTimeBots.png")
```



Why Washingtonpost.com having less blockage in the later years?

As spectulated, Washingtonpost.com guard against not only AI bots like but all poossible crawlers like archive.org_bot, gptbot, yandexbot, ia_archiver, web-archive-net.com.bot, claude-web, applebot-extended, imagesiftbot, awariosmartbot, meta-externalagent, anthropic-ai, linkarchiver, archivebot, arquivo-web-crawler, europarchive.org, nicecrawler, magpie-crawler, diffbot, ccbot, dataforseobot, omgili, ia_archiver-web.archive.org, perplexitybot, ahrefsbot, omgilibot, awariorssbot, twitterbot, semrushbot, claudebot, heritrix, meta-externalfetcher, bytespider, chatgpt-user, oai-searchbot, facebookbot, google-extended, amazonbot, primalbot.

But due to filtering just couple of AI Bots, we got the bad results in above heatmap.

```r
washington_data <- data2 %>%
  filter(Domain == "washingtonpost.com")

unique_blocked_crawlers <- washington_data %>%
  select(Blocked_Crawlers) %>%
  distinct()# Get unique values from the blocked_crawlers column

unique_blocked_crawlers_split <- unique_blocked_crawlers %>%
  separate_rows(Blocked_Crawlers, sep = ", ") %>%
  distinct()  # Keep only distinct values

multi_col_table <- matrix(unique_blocked_crawlers_split$Blocked_Crawlers,
                          ncol = 2, byrow = FALSE) %>% as.data.frame()

# Render the table and keep it on the same page
multi_col_table %>%  kbl(booktabs = TRUE, col.names = c("Blocked Crawlers 1",
        "Blocked Crawlers 2")) %>% kable_styling(latex_options = "hold_position")
```

Blocked Crawlers 1	Blocked Crawlers 2
*	amazonbot
twitterbot	yandexbot
semrushbot	applebot-extended
ahrefsbot	meta-externalagent
gptbot	meta-externalfetcher
chatgpt-user	archive.org_bot
google-extended	ia_archiver
ccbot	web-archive-net.com.bot
claude-web	imagesiftbot
awariosmartbot	linkarchiver
anthropic-ai	archivebot
magpie-crawler	arquivo-web-crawler
dataforseobot	europarchive.org
omgili	nicecrawler
perplexitybot	diffbot
omgilibot	ia_archiver-web.archive.org
claudebot	awariorssbot
bytespider	heritrix
facebookbot	oai-searchbot
peer39_crawler/1.0	primalbot