Team Members
Vaneesha S Kumar (221166)
Nikhil Gupta (220708)
Nischay Patel (220721)

CGS616 - Assignment 1A

OVERVIEW

The objective of the assignment was to explore the nature and extent of biases in search results. To do this expansive searches were conducted on different search engines and the results were compared to identify any biases or differences in the search results.

METHODOLOGY

To observe the biases in search engines, a methodical approach of data collection, data analysis and data observation was followed.

DATA COLLECTION

The analysis was conducted for 3 topics: Gun Ownership, LGBTQ+ rights and Abortion. For each topic a list of 10 search query phrases were prepared. Each search query phrase was then used to obtain the search results for 4 search engines: Google, Bing, Brave and DuckDuckGo. The search results for each topic were stored in separate csv files, the links of which are attached below.

TOOLS AND TECHNIQUES

1. Web Scraping: Automated scripts were used to collect search results, extracting titles and links using selenium and beautifulsoup.
2. Sentiment Analysis: For each topic, polarity and subjectivity analyses were conducted to assess the tone and objectivity of the content using the textblob library.
3. Domain Analysis: For each topic, the frequency of top-level domains (TLDs) was analysed to identify the types of sources prioritized by each search engine. A word cloud of domain types was generated for each search engine to assess the variety and frequency of domains generated. A heatmap was generated for all the links to visualise the distribution of links across domains.

4. Geographical Bias: For each topic, the frequency of regional domain types such as .au, .uk, .ca, .ie etc. were visualised by generating bar plots. The results were then analysed to observe if there was bias of a search engine towards any one particular region.
5. Text Analysis: For each topic, the frequency of key terms were visualised by generating a word cloud for each search engine. Additionally, key 2-word and 3-word terms were analysed through N-gram analysis.
6. Location Analysis: A separate csv per topic was generated for location analysis. The locations analysed were US, UK and South Africa. The above mentioned analyses were carried out for Google search in these three locations and biases resulting from difference in location were observed.

TOPIC 1: ABORTION

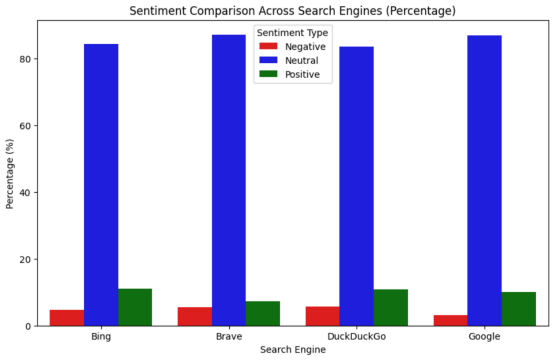
Generated Search Results stored in csv file: Search_results_3

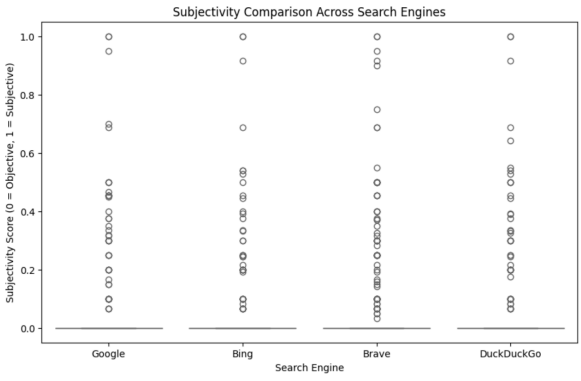
Location based Search Results stored in csv file: Google_results_3

Query phrases used:
"Abortion laws",
"Should it be legal to abort",
"Early abortion",
"Reproductive rights activism",
"Public opinion and abortion laws",
"Abortion bans and women's health",
"Abortion health risks",
"Pro-life vs pro-choice debate",
"Religious views on abortion",
"Abortion access during COVID-19"

DESCRIPTION OF SEARCH RESULTS

1. Sentiment Analysis - Polarity and Subjectivity:





Google:

Sentiment: Mostly neutral, with a minimal amount of positive and negative sentiments. This indicates a balanced presentation of information, with an emphasis on factual content.

Subjectivity: Primarily objective, which means that Google is usually giving more information that is less opinionated and more factual.

Bing

Sentiment: As with Google, Bing has a high rate of neutral sentiment, but slightly more positive content than negative. This is indicative of a focus on presenting a balanced perspective with a little positive framing.

Subjectivity: Mainly objective, indicating that Bing too prioritizes fact reporting over subjective opinions.

Brave

Sentiment: Exhibits a very strong neutral sentiment, with little positive and negative content. This suggests that there is a concentration on presenting plain information with little bias.

Subjectivity: Extremely objective, supporting the notion that Brave places emphasis on accurate facts rather than subjective stories.

DuckDuckGo

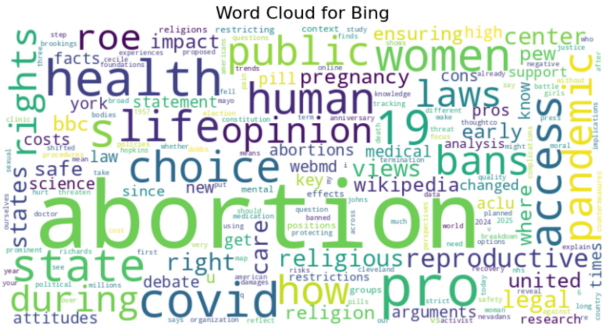
Sentiment: Similar to the others, DuckDuckGo is highly neutral with very little positive and negative content. This indicates a similar style of delivering balanced information.

Subjectivity: Primarily objective, reflecting a preference for factual content, consistent with its reputation for privacy-oriented, unbiased search results.

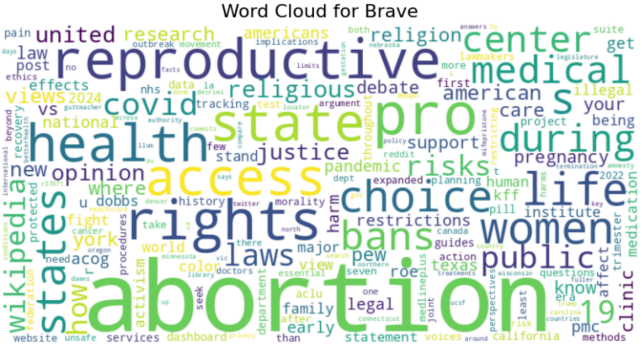
2. Text Analysis using Word Clouds:

Word Cloud for Google

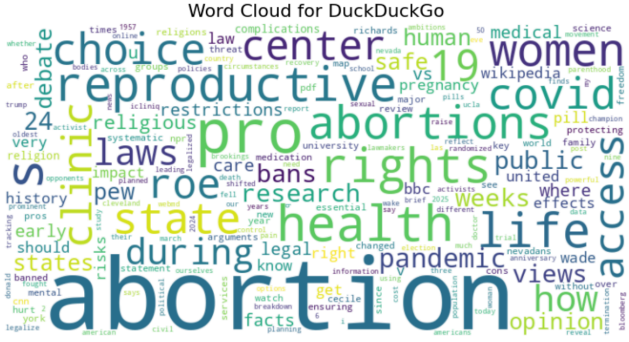
Word Cloud for Bing



Word Cloud for Brave



Word Cloud for DuckDuckGo



Google

Focus: The word cloud highlights terms like "abortion," "rights," "health," and "legal," indicating a strong emphasis on legal and health aspects of abortion.

Diversity: Words like "religion," "pandemic," and "women" suggest a broad coverage of topics, including religious views and the impact of COVID-19.

Bing

Focus: Prominent words include "abortion," "choice," "health," and "public," reflecting a focus on public opinion and health-related issues.

Diversity: Terms like "pandemic," "religious," and "laws" indicate coverage of diverse perspectives, including religious and legal aspects.

Brave

Focus: Key terms such as "abortion," "reproductive," "rights," and "state" suggest a focus on reproductive rights and state-level legislation.

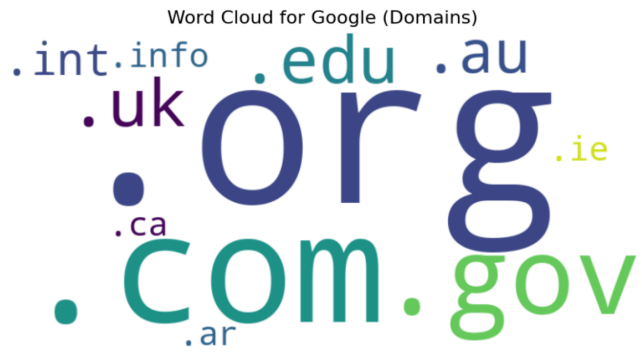
Diversity: Words like "activism," "pandemic," and "religious" show an interest in activism and the broader societal context.

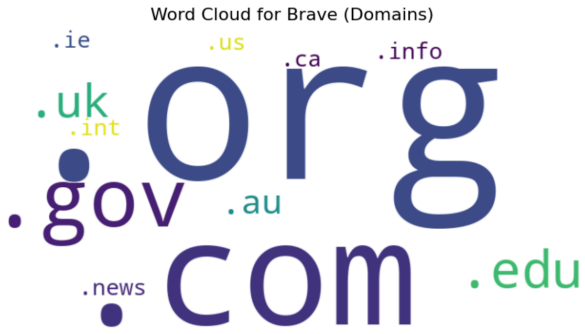
DuckDuckGo

Focus: The word cloud emphasizes "abortion," "rights," "health," and "state," indicating a focus on rights and health implications.

Diversity: Terms like "pandemic," "religious," and "access" suggest a comprehensive approach, covering access issues and religious perspectives.

3. Domain Type Analysis using Word Clouds:





Google

Prominent Domains: The word cloud shows a strong presence of .org, .com, and .edu domains, indicating a mix of nonprofit, commercial, and educational sources.

Diversity: Includes domains like .gov and international domains such as .uk and .au, suggesting a wide range of governmental and international perspectives.

Bing

Prominent Domains: Similar to Google, Bing features .org, .com, and .edu prominently, reflecting a balanced mix of nonprofit, commercial, and educational content.

Diversity: The presence of .gov and .uk domains indicates a focus on governmental and UK-based sources, providing varied viewpoints.

Brave

Prominent Domains: The word cloud highlights .org and .com domains, suggesting a strong emphasis on nonprofit and commercial content.

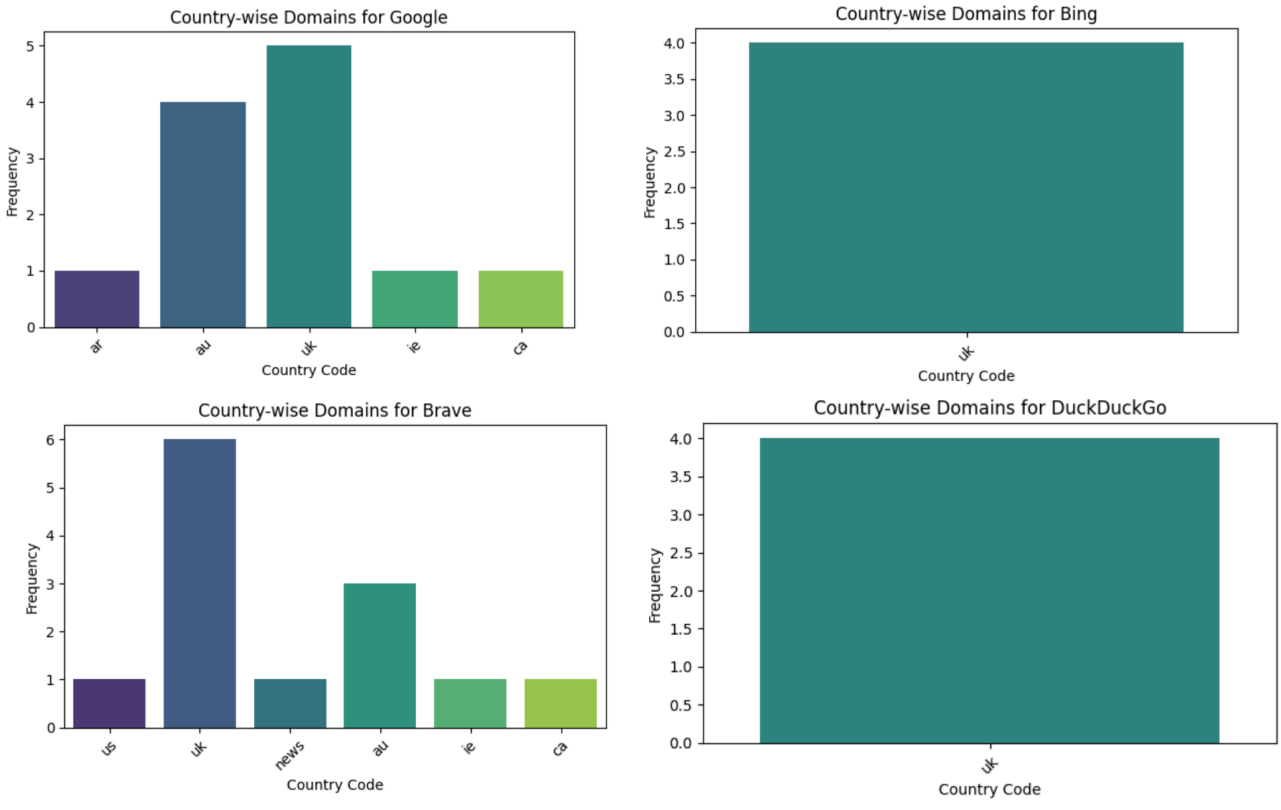
Diversity: Includes .gov, .edu, and international domains like .uk and .ca, indicating a broad spectrum of sources, including governmental and educational.

DuckDuckGo

Prominent Domains: Features .org, .com, and .edu domains, showing a mix of nonprofit, commercial, and educational content.

Diversity: The inclusion of .gov and .uk domains suggests a focus on governmental and UK-based perspectives, offering a diverse range of information.

4. Geographical Bias Analysis:



Google

Diversity: Google shows a variety of country domains, with a notable presence from the UK (.uk), Australia (.au), and Canada (.ca), indicating a broad international perspective.

Focus: The inclusion of domains like .ie (Ireland) and .ar (Argentina) suggests an interest in diverse regional viewpoints.

Bing

Concentration: Bing predominantly features UK domains (.uk), indicating a strong focus on content from the United Kingdom.

Limited Diversity: The lack of other country domains suggests a narrower regional focus compared to other search engines.

Brave

Diversity: Brave displays a wide range of country domains, with significant representation from the UK (.uk) and Australia (.au), reflecting a broad international scope.

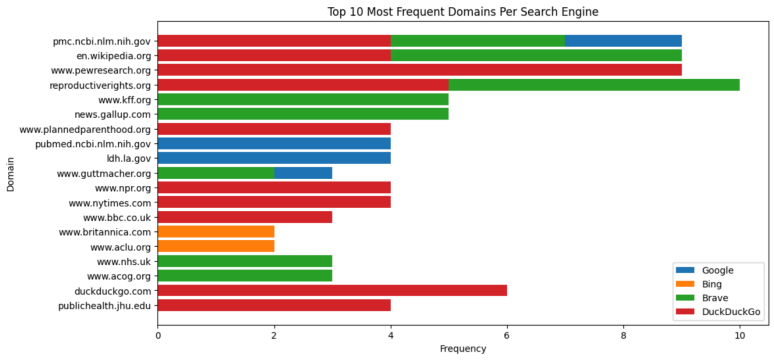
Additional Focus: The presence of .us (United States) and .ca (Canada) domains indicates an interest in North American perspectives.

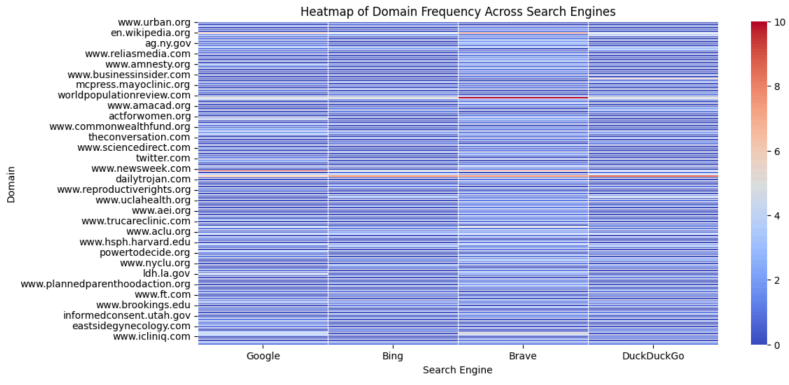
DuckDuckGo

Concentration: Similar to Bing, DuckDuckGo primarily features UK domains (.uk), suggesting a strong emphasis on UK-based content.

Limited Diversity: The focus on a single region may indicate a more concentrated regional approach.

5. Analysis of Domain Diversity:





Google

Prominent Domains: Features domains like pubmed.ncbi.nlm.nih.gov and ldh.la.gov, indicating a focus on authoritative and governmental sources.

Diversity: Includes a mix of educational (.edu) and nonprofit (.org) domains, suggesting a balanced approach to providing reliable information.

Bing

Prominent Domains: Highlights www.britannica.com and www.aclu.org, reflecting a focus on encyclopedic and civil rights perspectives.

Diversity: Shows a preference for well-established and reputable sources, emphasizing comprehensive and factual content.

Brave

Prominent Domains: Features www.kff.org and www.nhs.uk, indicating a focus on health-related and UK-based content.

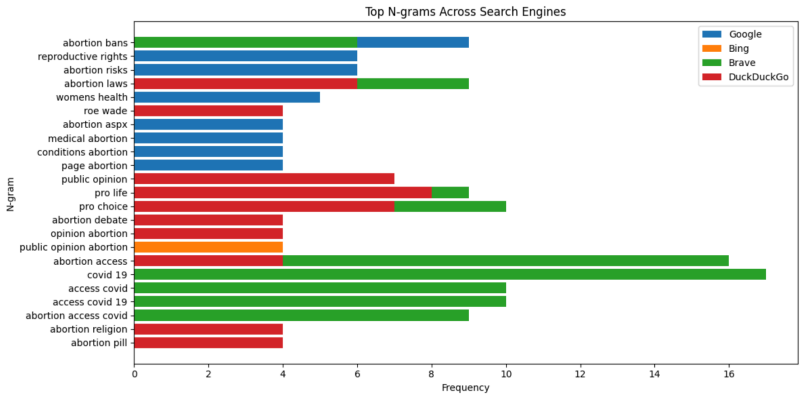
Diversity: Includes a variety of nonprofit and governmental domains, suggesting a commitment to diverse and authoritative information.

DuckDuckGo

Prominent Domains: Highlights en.wikipedia.org and www.nytimes.com, reflecting a focus on widely recognized and mainstream sources.

Diversity: Shows a strong presence of media and educational domains, indicating a broad approach to covering different perspectives.

6. N-Gram Analysis:



Google

Focus: Emphasizes terms like "abortion bans," and "medical abortion," indicating a focus on legal and medical aspects.

Diversity: Includes terms like "reproductive rights," indicating a broader view on rights and health.

Bing

Focus: Highlights "opinion abortion" and "public opinion abortion," reflecting a strong emphasis on public sentiment and debate.

Brave

Focus: Dominates with n-grams like "abortion access" and "access covid," suggesting a focus on accessibility and the impact of COVID-19.

Features "pro life" and "pro choice" prominently, reflecting a focus on the ideological debate.

DuckDuckGo

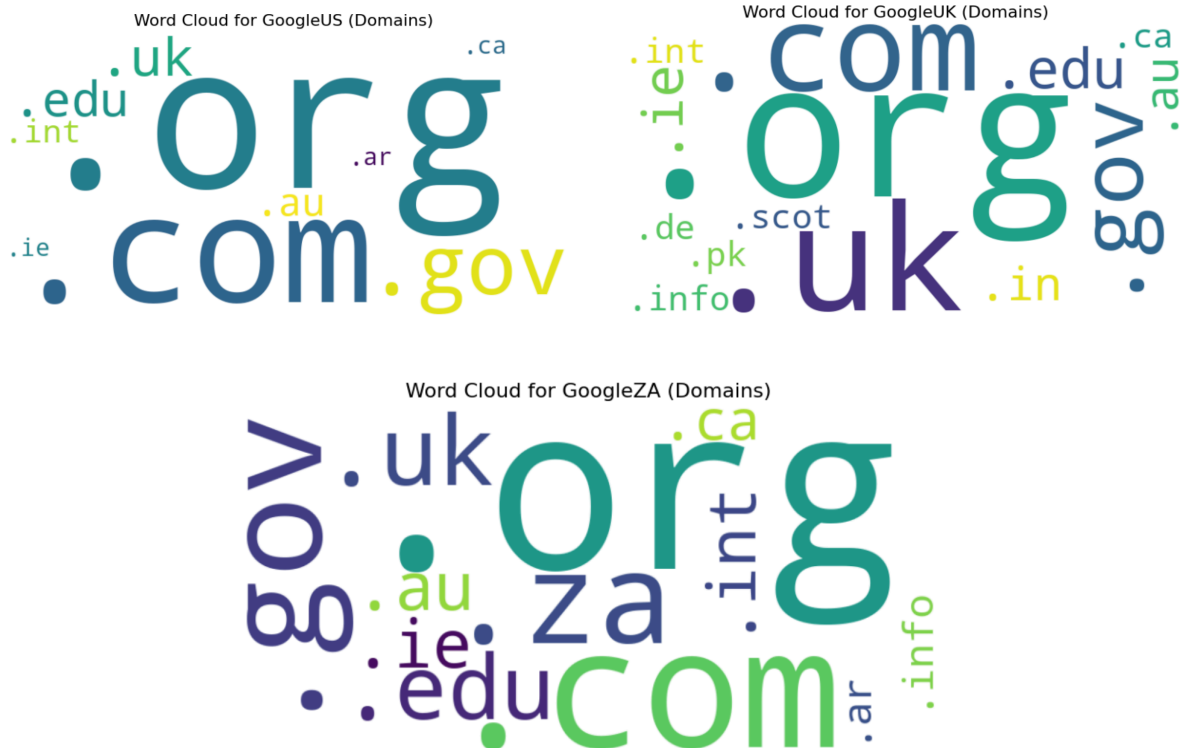
Focus: Features "pro life" and "pro choice" prominently, reflecting a focus on the ideological debate.

Diversity: Covers "abortion pill" and "abortion religion," indicating a range of topics from medical to religious perspectives.

Includes n-grams related to "public opinion," suggesting an interest in societal perspectives.

7. Bias based on Location (Search Engine Used: Google):

A) Domain Analysis using Word Cloud:



Google US

Diversity: Includes international domains like .uk and .au, suggesting a broad range of perspectives.

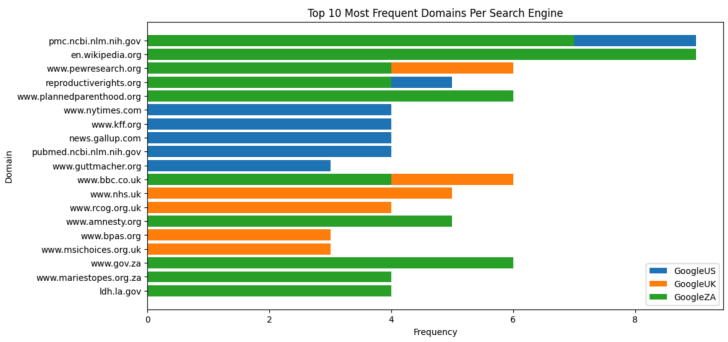
Google UK

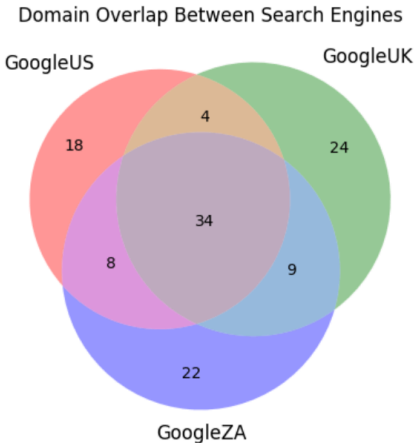
Diversity: Features UK-specific domains like .uk and .scot, indicating a focus on local content alongside international sources.

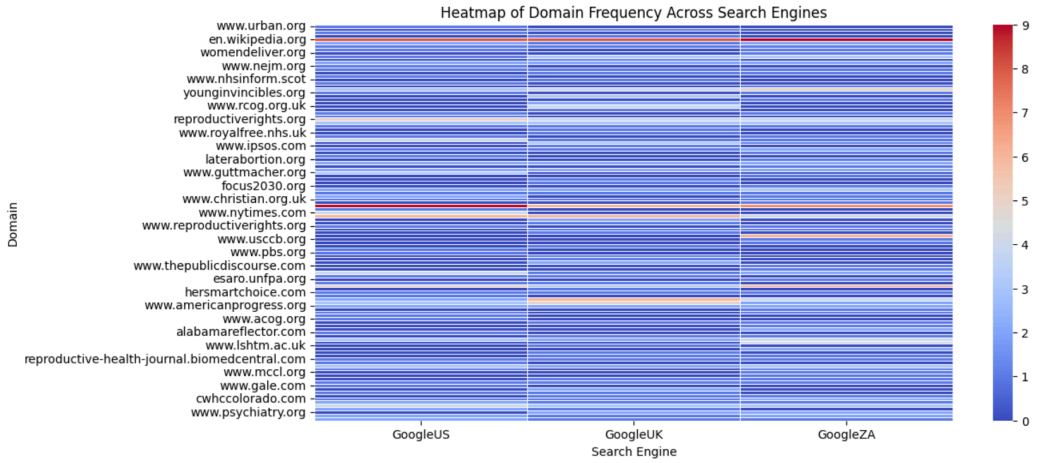
Google ZA

Diversity: Includes .za for South Africa, reflecting a focus on local content, along with international domains like .uk and .au.

B) Analysing Domain Diversity:







Google US

Prominent Domains: Features domains like pmc.ncbi.nlm.nih.gov, en.wikipedia.org, and www.nytimes.com, indicating a focus on authoritative, educational, and media sources.
Diversity: Shows a mix of nonprofit and governmental domains, reflecting a broad range of perspectives on abortion-related topics.

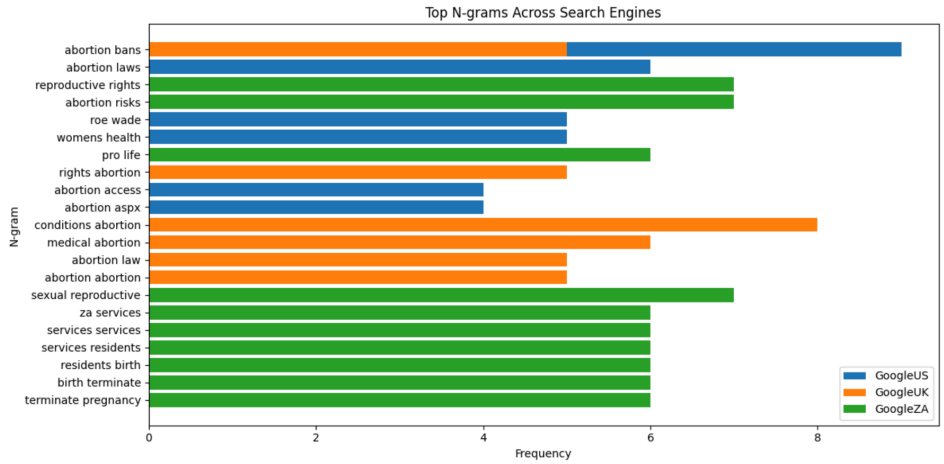
Google UK

Prominent Domains: Highlights www.bbc.co.uk, www.nhs.uk, and www.rcog.org.uk, suggesting a focus on UK-specific media and health organizations.
Diversity: Includes a variety of nonprofit and governmental sources, emphasizing local content alongside international perspectives.

Google ZA

Prominent Domains: Features www.gov.za, www.mariestopes.org.za, and www.amnesty.org, indicating a focus on governmental and nonprofit organizations within South Africa.
Diversity: Reflects a strong emphasis on local content, with a mix of international sources to provide a comprehensive view.

C) N-Gram Analysis:



Google US

Abortion Bans: This topic is highly prevalent, indicating a strong focus on legislative aspects and restrictions.
Abortion Laws: Significant interest, reflecting ongoing legal debates and changes.
Abortion Access: Concerns about accessibility, possibly influenced by recent legal changes.

Google UK

Medical Abortion: High frequency, indicating a focus on medical procedures and options.

Abortion Law: Legal aspects are a major concern, reflecting the UK's regulatory environment.

Google ZA

Reproductive Rights: Dominant topic, highlighting activism and rights in the South African context.

Services and Access: High frequency of terms related to services, indicating a focus on availability and support.

Pro-Life: The presence of this term suggests active engagement in the pro-life vs. pro-choice debate.

Sexual Reproductive Health: Emphasizes a comprehensive approach to reproductive health.

TOPIC 2: LGBTQ+

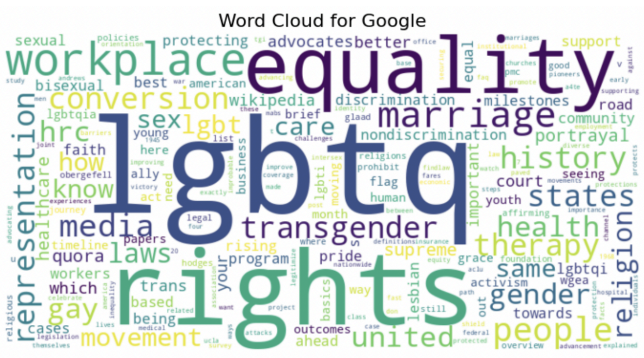
Generated Search Results stored in csv file: Google_Results_2

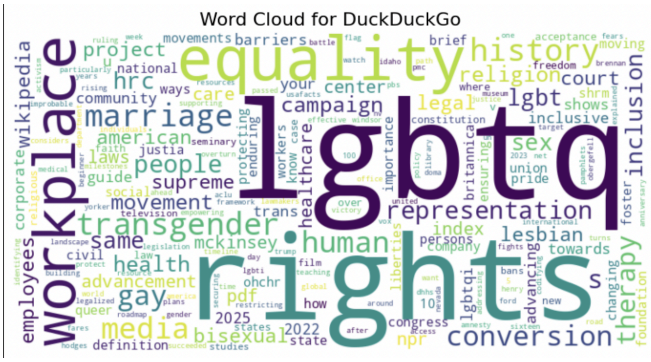
Location based Search Results stored in csv file: Search_Results_2

Query phrases used:
"Importance of LGBTQ+ rights",
"Legal protections for LGBTQ+ individuals",
"History of LGBTQ+ movements",
"LGBTQ+ equality in the workplace",
"Same-sex marriage legal battles",
"Transgender rights and healthcare access",
"Religious views on LGBTQ+ rights",
"LGBTQ+ representation in media",
"Conversion therapy bans",
"LGBTQ+ workplace equality"

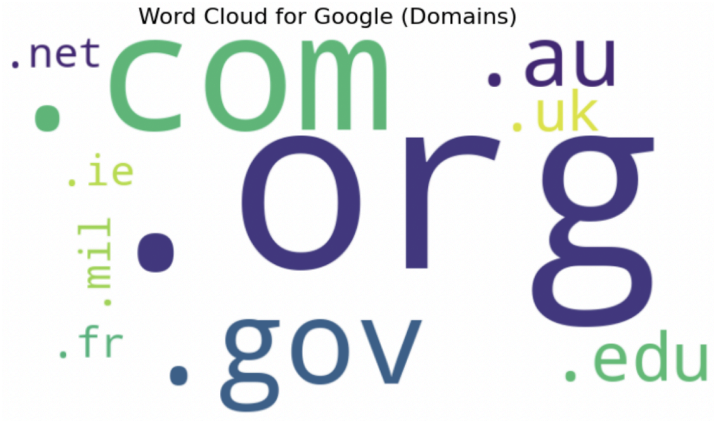
DESCRIPTION OF SEARCH RESULTS

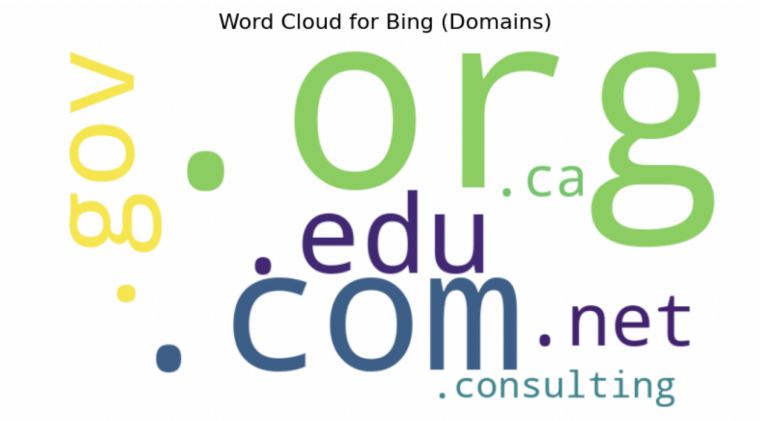
1. Text Analysis using Word Clouds:

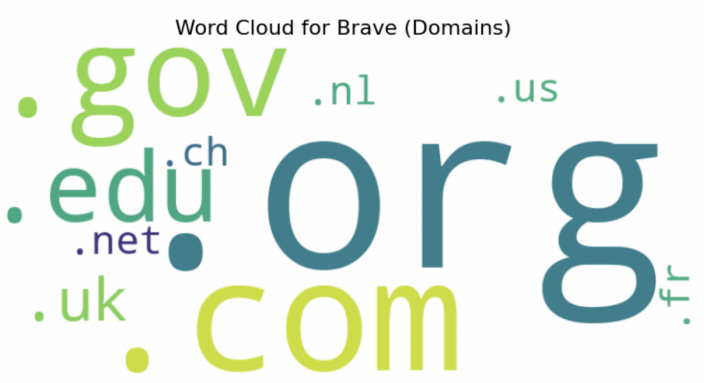


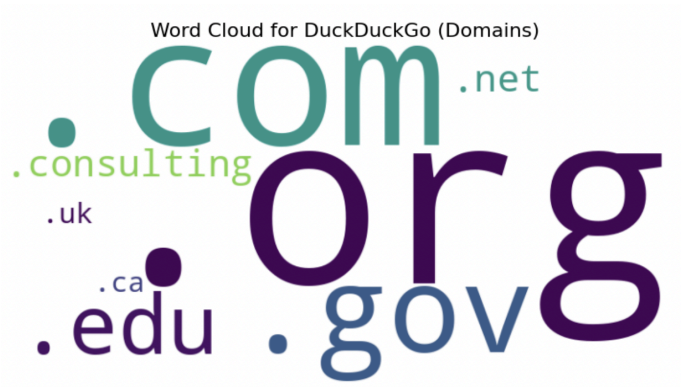


2. Domain Type Analysis using Word Clouds:









Google

- Prominent .com and .org domains indicate a balance between commercial and non-profit sources.
- Domains like .uk, .au, and .fr domains indicating international reach
- Presence of .gov and .edu suggests a mix of official and academic information.

Bing

- Strong emphasis on .org and.edu domains, highlighting non-profit and academic perspectives.
- Frequent use of .gov indicates a reliance on official government information.
- .consulting domain indicates involvement of consulting firms

Brave

- Domains like .nl, .us, .uk, .ch, and .fr suggest a broad international perspective.
- Emphasis on .edu and .gov reflects a focus on educational and official sources.

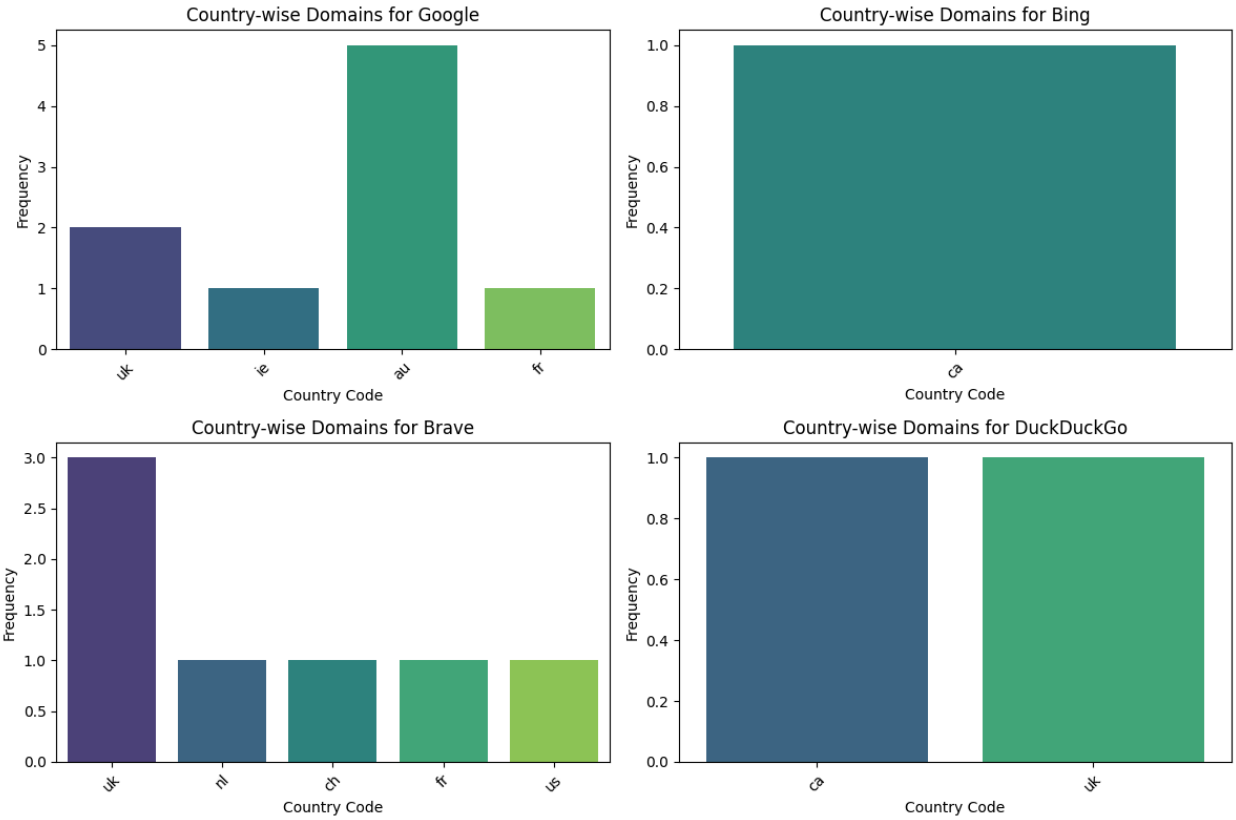
DuckDuckGo

- Prominent domains like .com, .org, and .edu indicate a mix of commercial, non-profit, and educational sources.

- Presence of .gov and .consulting suggests reliance on official and expert insights.

- .consulting domain indicates involvement of consulting firms

3. Geographical Bias Analysis:



Google

- High frequency of .au domains indicates a strong focus on Australian sources.

- Presence of .uk, .ie, france suggests additional emphasis on **UK** and **Irish** perspectives.

Bing

- Predominantly Canadian sources, as indicated by the high frequency of .ca domains
- The lack of other country domains suggests a bias toward **canadian** domains

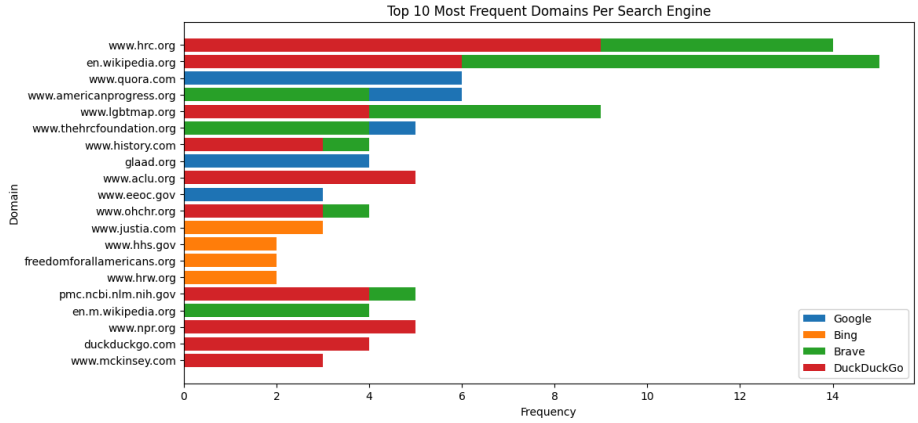
Brave

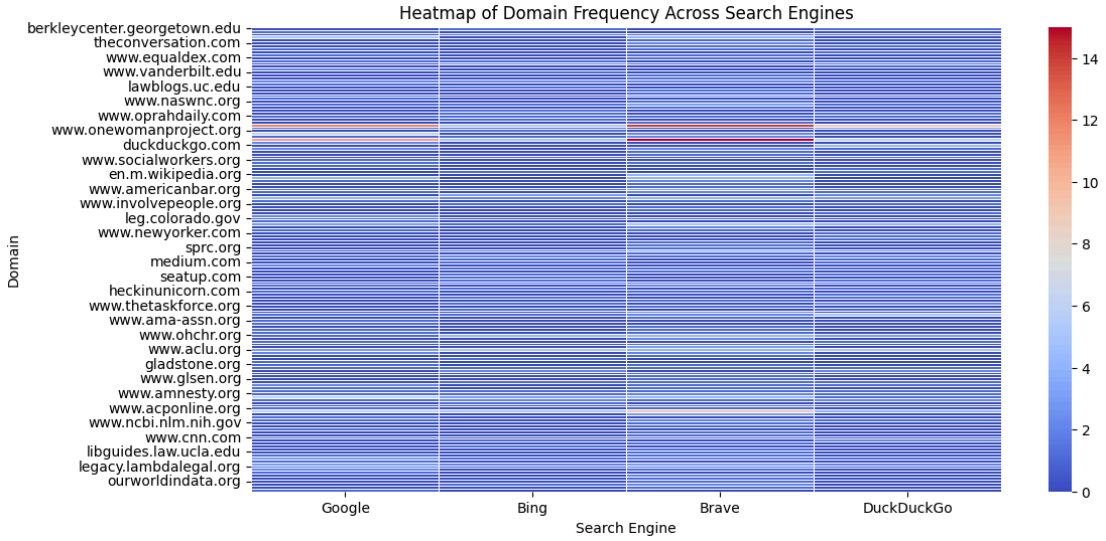
- Significant presence of .uk domains, highlighting a focus on UK-based content.
- Inclusion of .nl, .ch, and .fr domains suggests an emphasis on **French** and **European perspective.**

DuckDuckGo

- Balanced focus on Canadian and UK sources, with equal representation of .ca and .uk domains.
- Biased towards 2 domains.

4. Analysis of Domain Diversity:





Google

- Google shows a wide range of domains, indicating a broad spectrum of perspectives on LGBTQ+ topics.

- Surprisingly, Google shows a good amount of results from Quora (a questionable site in terms of fact-based content). It usually contains human subjective opinions on many topics

- Notable domains include Wikipedia and HRC, suggesting a mix of general information and advocacy-focused content. The search results balance between educational, governmental, and advocacy sites.

Bing

- Bing results are more concentrated on specific domains like Justia, hrw, HHS and freedomforallamericans.
- Compared to Google, Bing shows less domain diversity, possibly reflecting a narrower range of POV

Brave

- Brave frequently features domains like Wikipedia and LGBTMap, highlighting both informational and advocacy content.
- The repeated appearance of certain domains suggests a reliance on trusted sources for LGBTQ+ information.
- The inclusion of educational domains indicates a focus on providing comprehensive background information.

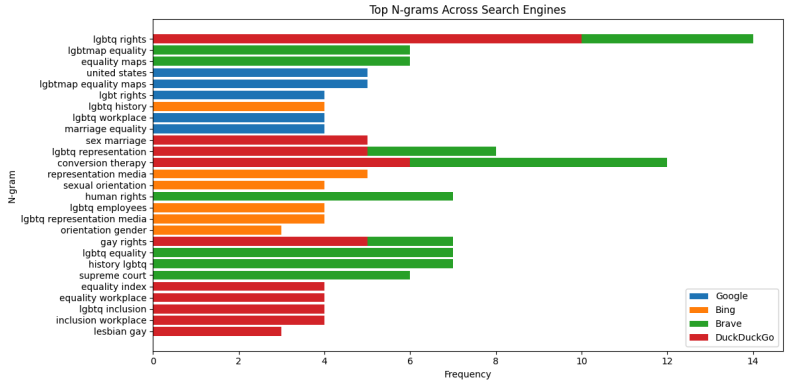
DuckDuckGo

- DuckDuckGo prominently features advocacy sites like HRC and ACLU, emphasizing rights and policy discussions.
- It covers a good variety of sites and duckduckgo itself is represented in good numbers.

General Observations

- Across all search engines, domains like Wikipedia and HRC are common, showing their central role in providing LGBTQ+ information.
- Varied Emphasis: Each search engine has a unique emphasis, from legal and health (Bing) to advocacy and education (DuckDuckGo and Brave).

5. N-Gram Analysis:



Google

- Feature “united states” to a certain extent while no other search engine featured this.
- Focuses on the general aspects like “Marriage equality”, “lgbtq workplace” and “lgbtq rights”

Bing

- Focus on Legal and Social Orientation N-grams like "orientation gender", “sexual orientation” and "LGBTQ employees" suggest a focus on legal definitions and workplace issues.
- The presence of "representation media" indicates an interest in how LGBTQ+ individuals are portrayed in media. Additionally, history of LGBTQ is largely focused

Brave

- The variety of N-grams, including "equality maps" and "LGBTQ representation," suggests a broad coverage of topics.
- Dominant N-grams such as "LGBTQ equality", “gay ights” and "conversion therapy" emphasize themes of equality and conversion rights.

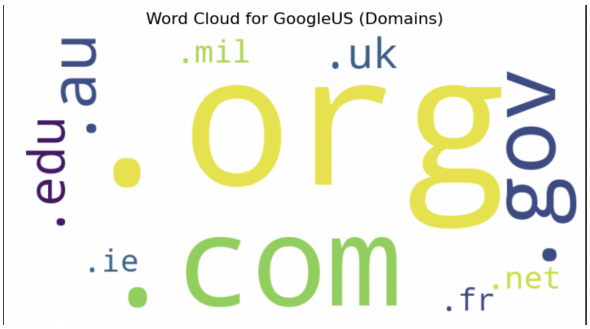
DuckDuckGo

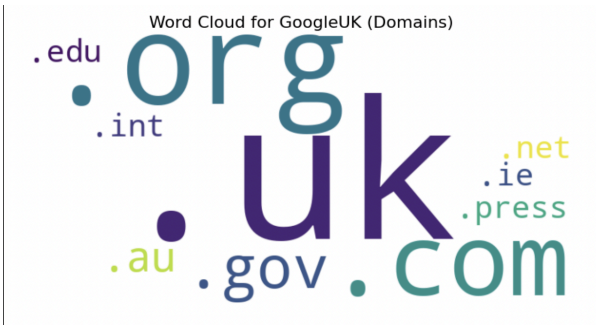
- “sex marriage”, “conversion”, “equality index”, “lesbian gay”, “equality workplace”, “lgbtq inclusion” represents an oriented approach towards inclusion and equality

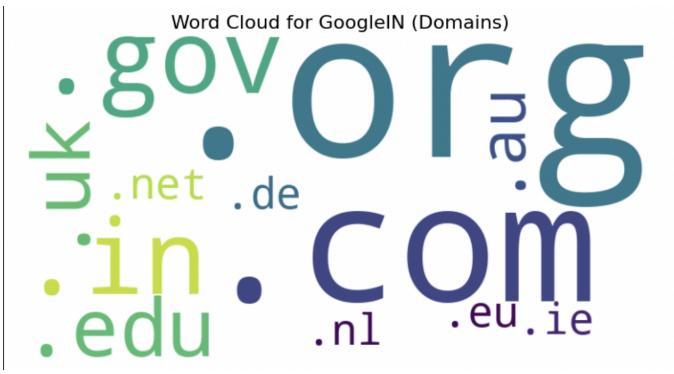
Brave and Duckduckgo covers a wide variety of topics with respect to google and bing

6. Bias based on Location (Search Engine Used: Google):

A) Domain Analysis using Word Cloud:







Google US

- .mil (military organisation), .au(australia), .fr(france) were represented in the word cloud

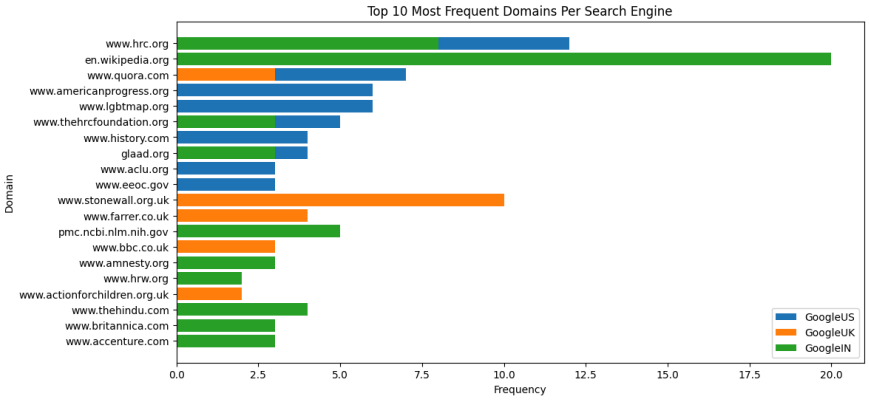
Google UK

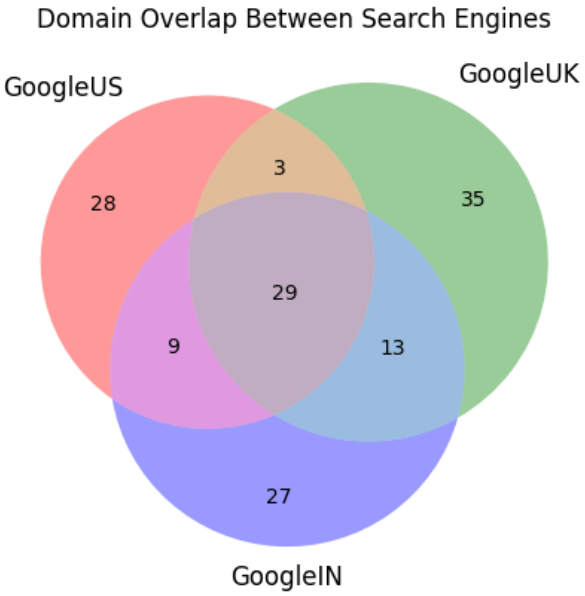
- .press, .int were some different domain comparing other location that showed in UK location

Google IN

- Despite the location being India, it shows domains from **uk, au, nl** which indicates that searches through indian location tend to redirect to other countries domains

B) Analysing Domain Diversity:





Google US

- .Unique Domains: 28 domains, mostly U.S.-centric
- Key Themes: Advocacy (e.g., www.hrc.org, www.lgbtmap.org), historical movements (www.history.com).

Google UK

- Unique Domains: 35 domains, U.K.-specific
- Key Themes: Legal protections (www.stonewall.org.uk, www.farrer.co.uk), community support.

Google IN

- Key Themes: Education (en.wikipedia.org, www.britannica.com), mainstream media (www.thehindu.com).
- More Focused on General awareness, cultural context, international perspectives.

TOPIC 3: GUN OWNERSHIP

Generated Search Results stored in csv file: Search_Results_1

Location based Search Results stored in csv file: Google_Results_1

Query phrases used:
"Does gun ownership increase violence?",
"Should civilians be allowed to own assault weapons?"
"Gun ownership: Right or privilege?",
"The link between gun ownership and mass shootings",
"Is gun control unconstitutional?",
"Do more guns make societies safer?",
"Concealed carry permits",
"Impact of gun ownership on crime rates",
"Gun rights advocacy groups",
"International perspectives on gun ownership"

DESCRIPTION OF SEARCH RESULTS

1. Sentiment Analysis:

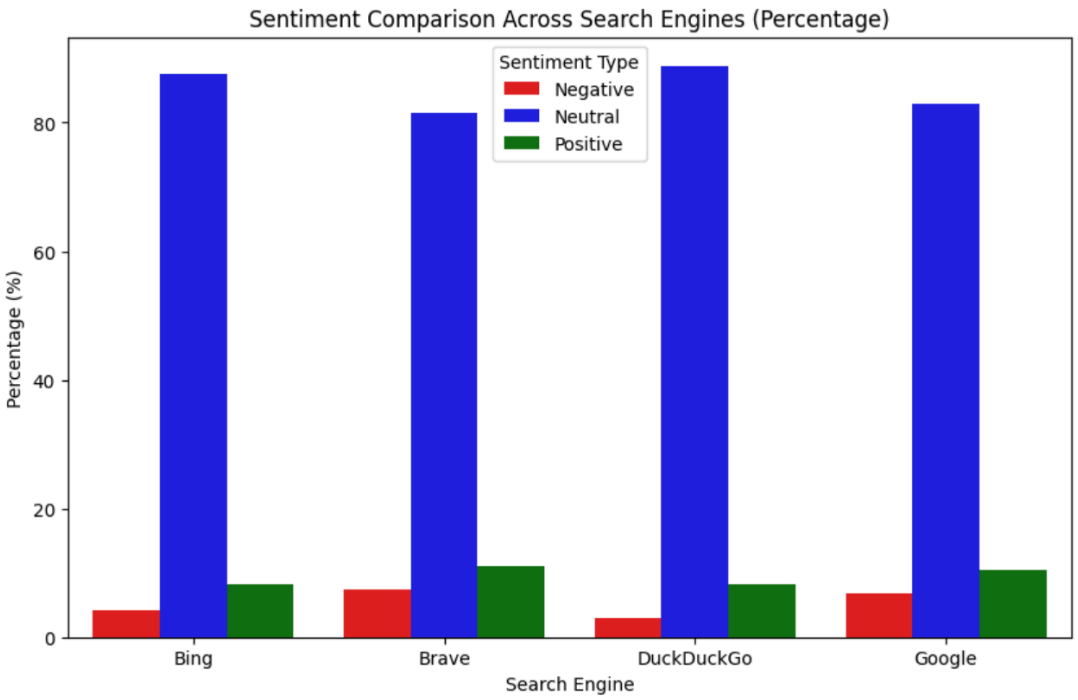
From the figure given below, we can draw the following conclusions:

Google and Brave: Majority search results have a neutral tone suggesting that the search results are mostly fact based and not opinionated. However, we can observe that both the browsers give the least neutral search results amongst the 4 search engines. They tend to produce more positive results than negative results suggesting that Google and Brave search results may lean towards the positive aspects of a search query phrase rather than the negative aspects of it.

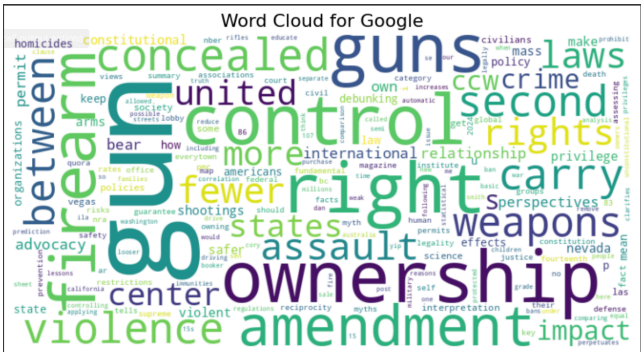
For Example:
Query Phrase: Does gun ownership increase violence?
Search Result: Fewer Guns Mean Fewer Gun Homicides | NBER

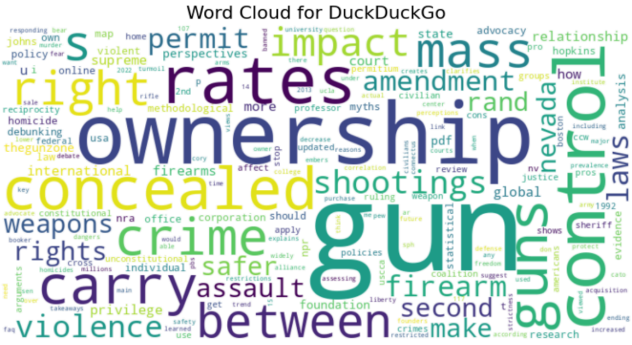
Query Phrase: Do more guns make societies safer?
Search Result: Gun Ownership Provides Effective Self-Defense (From Gun Control, P 142...

As we can observe from this example, the search result tends to follow the tone of the search query.



2. Text Analysis Through Word Clouds:





Key Observations:

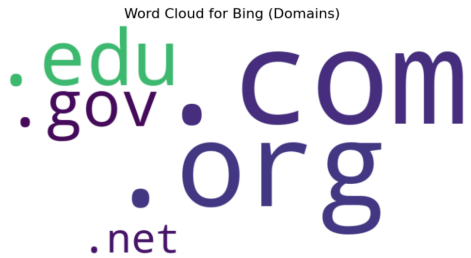
All the 4 search engines highlight terms like “control”, “rights”, “ownership”, “violence”. This suggests a balanced approach, covering both legal and societal aspects of gun ownership.

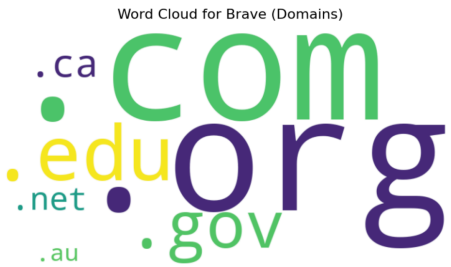
Bing word cloud’s relative emphasis on “nevada” is due to the fact that in the search results generated for the query “concealed carry permits”, every single search result generated was focused on Las Vegas Nevada, even though the query was not.

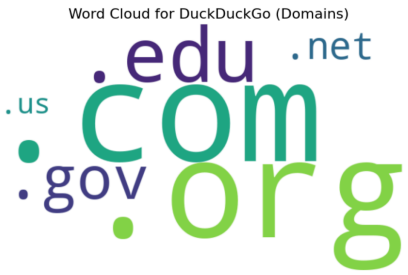
In Brave’s word cloud the relatively large emphasis on “concealed” shows that for the same query as mentioned for Bing (“concealed carry permits”) Brave does not bias towards any particular region, rather gives generalised search results for a large number of regions.

3. Domain Analysis through Word Clouds:

- All 4 search engines have a strong preference for commercial, non-profit and educational





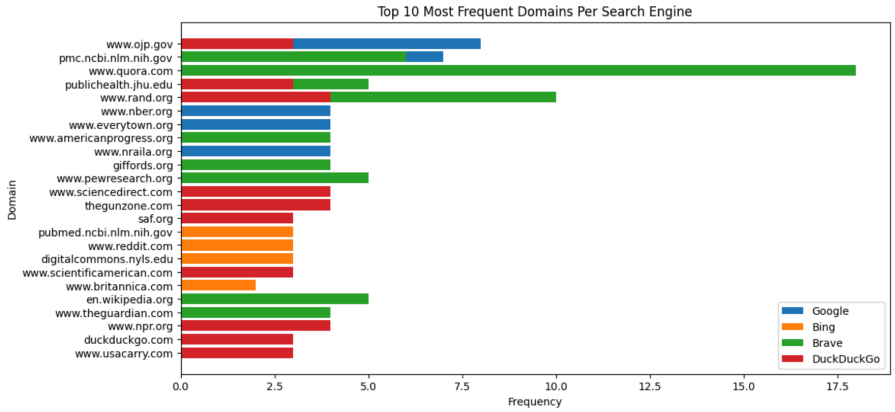


domain types.

- Brave also has some regional domain types which suggests that Brave chooses from a diverse range of sources.

4. Analysing Domain Diversity:

Google: From the above plot, we can see that Google tends to display government and



non-profit sources frequently.

Bing: From the above plot, we can see that Bing does not favour any one domain. That is, it equally shows government, commercial and educational sources. However, Bing has a disproportionate number of search results of reddit.com. This suggests that Bing values diverse perspectives and real-world experiences shared by users, potentially offering a more grassroots view on topics like gun ownership. This approach can provide insights into public opinion and discussions, complementing the more traditional sources like government and educational sites.

Brave: Brave's algorithm may prioritize community-driven content and user-generated insights, as indicated by the frequent appearance of quora.com. This suggests that Brave values diverse perspectives and personal experiences shared by individuals.

Having said that, the frequent appearance of reddit.com in Bing and quora.com in Brave could also have some downsides:

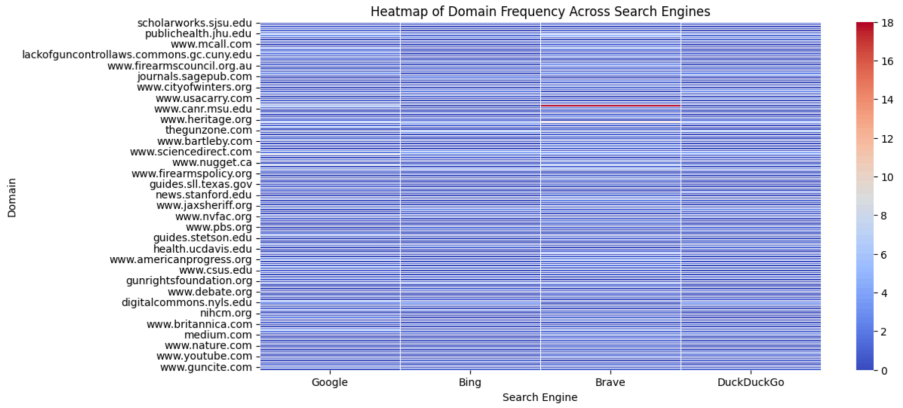
User-generated content can sometimes include inaccurate or misleading information, as it may not be fact-checked or verified.

Answers can reflect personal biases, which might skew the information presented.

The quality of content can vary widely, making it challenging for users to discern reliable information.

Popular opinions may dominate, potentially reinforcing existing biases and limiting exposure to diverse perspectives.

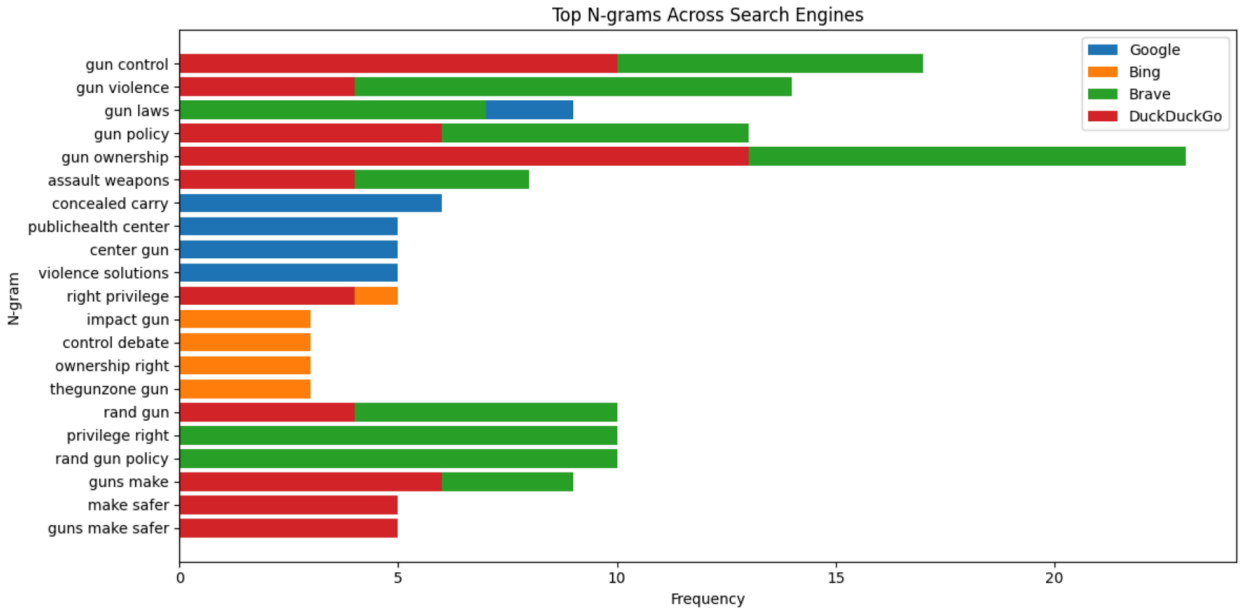
From the heatmap we can make the following observations:



Google, Bing and DuckDuckGo have a fairly uniform distribution of domains which suggests that they do not have bias towards a particular domain. However, we can clearly see that .org, .com, .edu are the predominant domain types across all search engines.

Brave has a clear bias towards one particular domain www.canr.msu.edu, which could be due to the algorithmic bias of SEO.

5. N-gram Analysis:



Google: Emphasizes "concealed carry," "public health center," and "violence solutions," which indicate a broader approach including public health and safety solutions.

Bing: Has a more balanced distribution but shows interest in "impact gun" and "control debate," suggesting a focus on the effects and discussions around gun control.

Brave: Shows a strong emphasis on "gun ownership" and "rand gun," indicating a potential focus on ownership rights and specific gun policies.

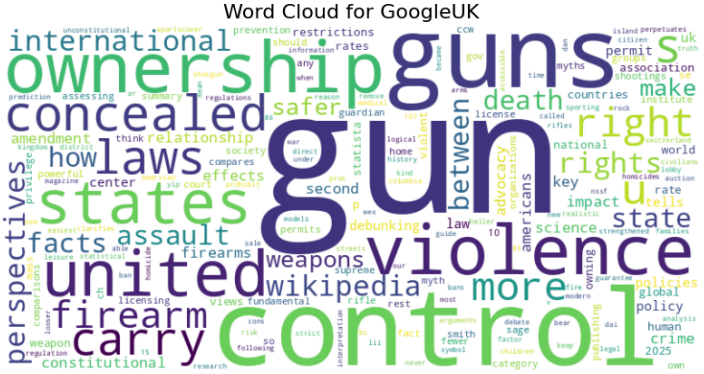
DuckDuckGo: Highlights terms like "gun control," "gun policy," and "assault weapons," suggesting a focus on regulatory aspects.

6. Bias based on Location (Search Engine Used: Google):

A) Text Analysis Through Word Clouds:

Word Cloud for GoogleUS

Word Cloud for GoogleUK



Word Cloud for GoogleZA

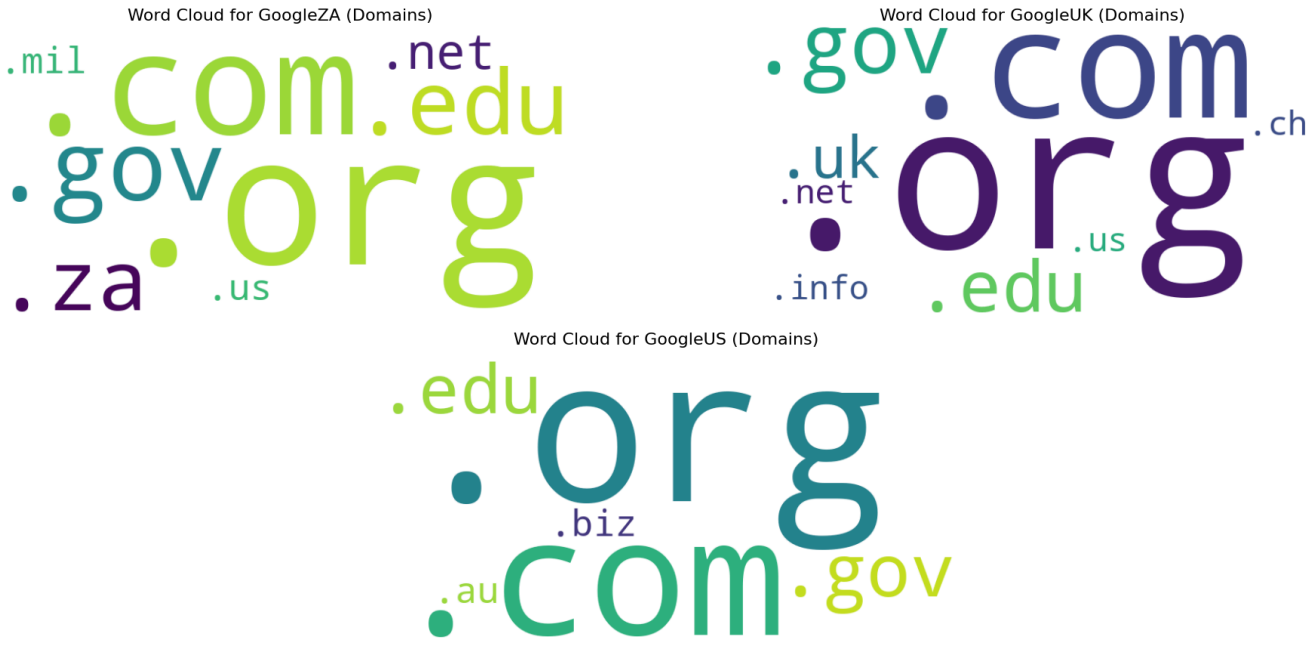
Observations:

The analysis of the word clouds reveals notable differences in search result emphasis across regions.

- In the United States and South Africa, terms like *'violence'*, *'firearm'*, *'rights'*, and *'amendment'* are more prominent, indicating a focus on legal and social issues surrounding gun ownership. Conversely, in the United Kingdom, terms such as *'perspectives'*, *'facts'*, and *'Wikipedia'* appear more frequently, suggesting a more informational or neutral tone in search results. This highlights a potential bias in searches from the US and South Africa compared to the UK, where there seems to be a greater emphasis on public perspectives and general information.

- The word clouds show the frequent occurrence of regional identifiers such as *'United'* and *'States'* in both US and UK searches, while terms like *'South Africa'* prominently appear in searches from South Africa, reflecting a regional focus in search behavior.

B) Domain Analysis through Word Clouds:



The word clouds of domain names across the three regions (UK, US, and SA) provide insight into the sources of information highlighted in search results. Here's an analysis:

GoogleUK Word Cloud:

Dominance of `.org` and `.com` suggests that non-profit organizations and commercial entities are primary sources of information.

The presence of `.gov` and `.edu` indicates contributions from government and educational institutions, ensuring authoritative content.
The inclusion of `.uk` aligns with regional relevance, signifying localized sources.

GoogleUS Word Cloud:

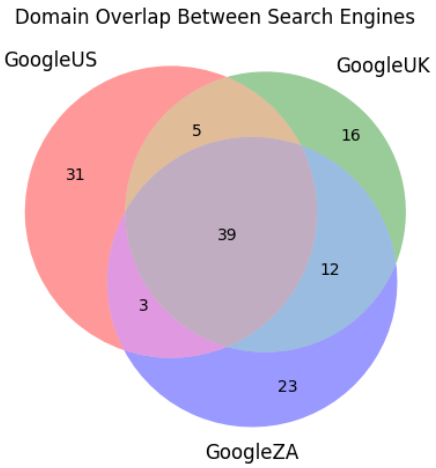
Similar to GoogleUK, `.org` and `.com` dominate, reflecting a mix of non-profit and commercial sources.
A significant presence of `.edu` and `.gov` domains suggests strong representation from academic and governmental perspectives.
The addition of `.biz` highlights some business-oriented results, while `.au` indicates occasional cross-regional results, possibly globalized searches.

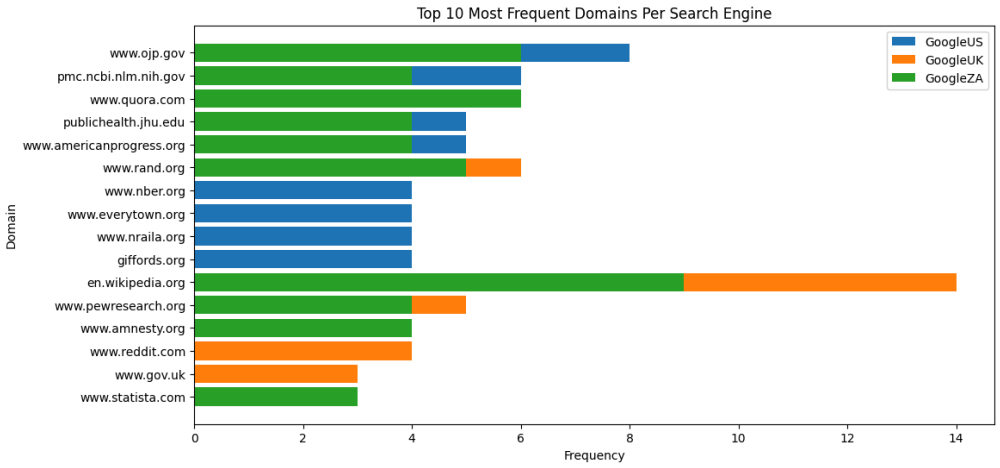
GoogleSA Word Cloud:

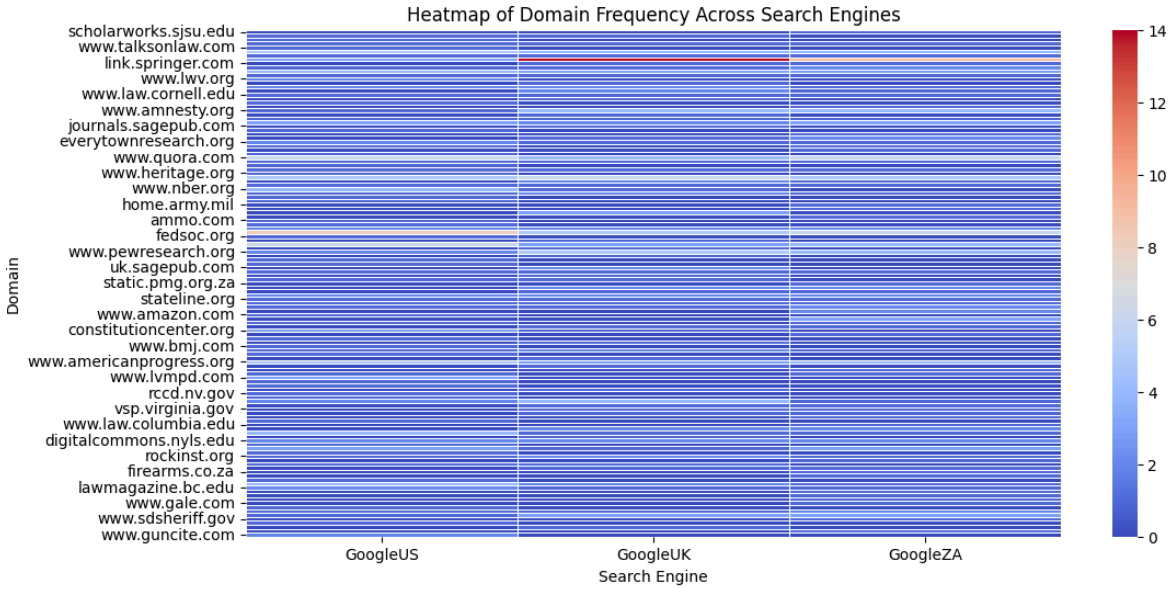
Like the others, `.org` and `.com` dominate, indicating reliance on international non-profits and commercial sources.
The presence of `.gov` and `.edu` ensures access to governmental and academic content but may not be as prominent as in the US and UK clouds.
Country-specific domains such as `.za` (if present) would reinforce localized content in the South African context, though it is not explicit in the cloud.

The analysis indicates varying levels of regional and global bias, with a general tendency toward prioritizing widely accepted international sources.

C) Analysing Domain Diversity:







Google US

Government and Research Domains: High frequency of domains like www.ojp.gov and www.nber.org indicates a focus on official and research-based perspectives.
Gun Rights Advocacy: Domains such as www.nraila.org and giffords.org suggest active engagement in advocacy and policy discussions.
Crime and Safety: Interest in the impact of gun ownership on crime rates and societal safety is evident.

Google UK

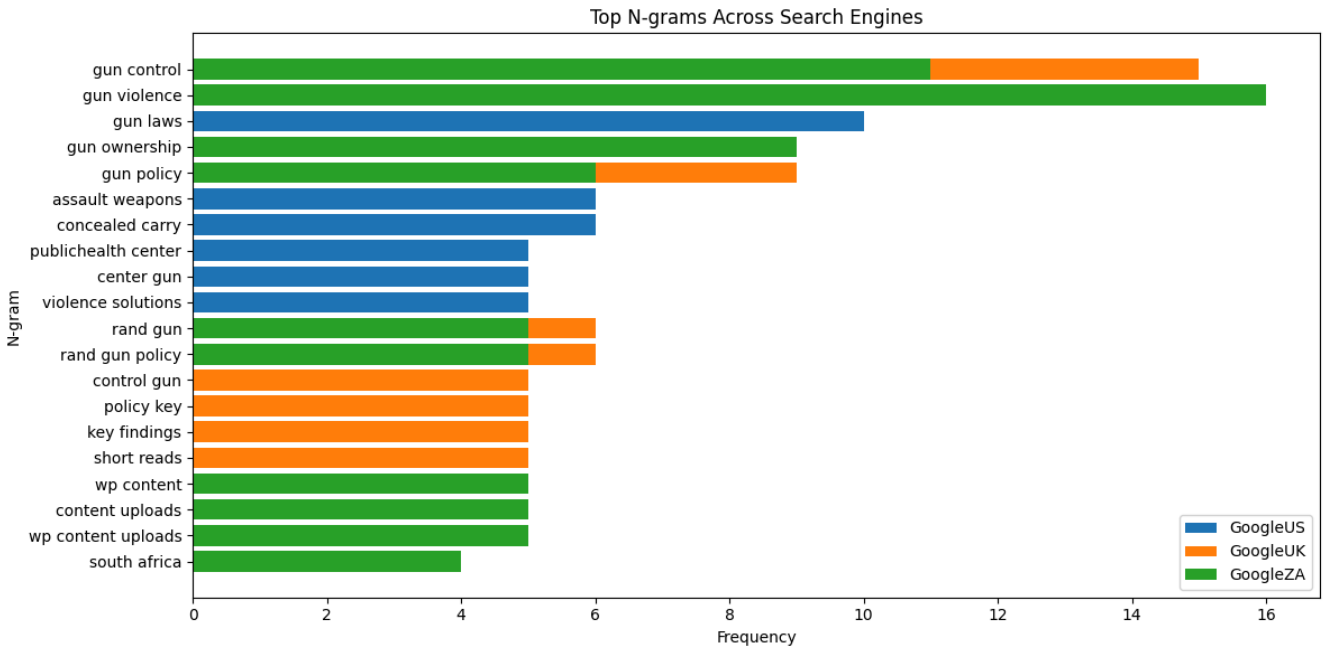
International Perspectives: The presence of domains like www.amnesty.org and www.gov.uk highlights a focus on international views and regulatory frameworks.
Public Opinion and Research: Domains such as www.pewresearch.org indicate interest in public opinion and statistical analysis.
Legal and Constitutional Aspects: Emphasis on whether gun control is constitutional reflects ongoing legal debates.

Google ZA

Health and Safety: Domains like publichealth.jhu.edu suggest a focus on the health implications of gun ownership.

Crime and Violence: Interest in the link between gun ownership and violence, as seen in domains like www.americanprogress.org.
Educational and Informative: High frequency of educational domains indicates a focus on informing the public about gun-related issues.

D) N-Gram Analysis:



Google US

Gun Control and Violence: High frequency of terms like "gun control" and "gun violence" indicates a focus on regulatory measures and societal impacts.
Gun Laws and Ownership: Interest in legal aspects and the debate over gun ownership as a right or privilege.
Concealed Carry: Emphasis on permits and regulations related to carrying firearms.

Google UK

Policy and Key Findings: Focus on policy discussions and research findings, reflecting a more analytical approach.
International Perspectives: Interest in how gun ownership is viewed globally, indicating a broader perspective.
Violence Solutions: Emphasis on finding solutions to gun-related violence.

Google ZA

Gun Control and Violence: Similar to the US, there's a strong focus on control measures and violence.

Public Health: Interest in the health implications of gun ownership, as seen in terms like "public health center."
South Africa Specific: Unique focus on local context and implications, as indicated by terms like "South Africa."

POSSIBLE CAUSES FOR BIAS IN SEARCH ENGINES:

Editorial Judgments: Search engines make editorial choices about which data to collect and how to present it, leading to systematic biases in the ranking and indexing processes.
Popularity Metrics: Algorithms often prioritize popularity-based factors (e.g., Google's PageRank), which favor well-known or economically powerful websites, reinforcing existing power structures.
Majority Interests: Search engines tune their algorithms to cater to majority interests, often marginalizing minority viewpoints and less popular content.
Structural Preferences: Choices such as indexing only parts of large documents or excluding certain web pages based on technical criteria can introduce biases.
Neutralized, Fact-Based Opinions: Some search engines deliberately avoid incorporating diverse public opinions to maintain a perception of neutrality. This approach can suppress controversial or alternative viewpoints, leading to a bias toward sanitized, non-confrontational content

WAYS TO MINIMIZE BIAS:

Improved Transparency:
Require search engines to disclose details about their ranking algorithms and practices. This can empower users to critically assess biases.
Mandating Changes:
Implement regulations to give marginalized websites better visibility. For example:
-Introduce "randomized rank promotion" for less prominent sites.
-Include corrective information or alternative viewpoints alongside contentious results.
Algorithmic Evolution:
Continually refine algorithms to adapt to changing user needs and reduce reliance on outdated or overly simplistic ranking factors

CODE FILES

CGS616_1A.ipynb.

CGS616 - Assignment 1B

OVERVIEW

To simulate the behavioral analysis of the Online shoppers purchasing intention dataset, we developed the Drift diffusion model from scratch due to the constraints in implementing existing libraries. We modelled the population intention for the people who purchased the product and identified the mean and standard deviation for their Response times. To enhance the fitting of DDM on the data we estimated the parameters (Drift rate, Decision boundary and Gaussian noise) using linear estimation and trial calling. Additionally, we analysed the variation of drift rates for a few specific “traffic types” and identified the traffic medium which ended with the quickest purchase.

METHODOLOGY

Exploratory Data Analysis:

The Online shoppers purchasing intention dataset has very few missing values and all features of the dataset are relevant to the purchasing intention based on inference. We visualised the behaviour of features with respect to the “Revenue” using various plots.

Feature Selection:

Based on the visualisations, we selected the following feature which favoured [Revenue=True] [exit_rates, Page_value, Bounce_rate, traffic_type , visitor_type]

Feature scaling and Drift rates estimation:

We used a linear regression to predict the drift rate (A_t) based on the selected features.
Eqn: (A_t) = A1*exit_rates + A2*page_value + A3*bounce_rate + A4*traffic_type + A5*visitor_type + intercept

The model was trained on the dataset we used, and the learned weights *(A1, A2, A3, A4 and A5)* and intercept were extracted. The predicted drift rates for each time step were computed and displayed for inspection, and optionally, added to the original data frame for further analysis.

Drift Diffusion Model Simulation:

We wrote the code using the fundamental equation *dx= Adt+ cdW* to simulate decision-making processes, where the drift is dynamically estimated based on factors like *exit_rates, Page_value, Bounce_rate, traffic_type , visitor_type*. The model simulates multiple trials, updating evidence over time with noise, and tracks whether the decision reaches a threshold for purchase or not. This generates the simulated graph for all the rows by changing the num_trials parameter.

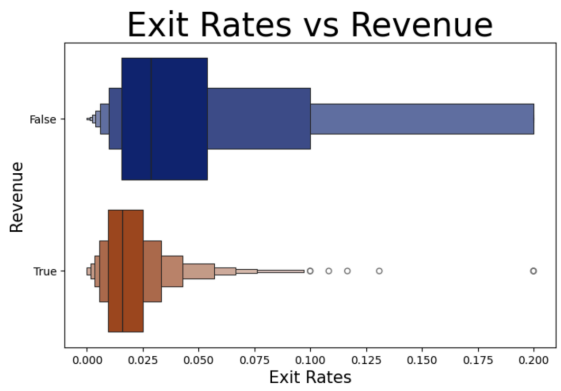
We computed mean and standard deviation of the reaction time and plotted Frequency vs Reaction Time Distribution.

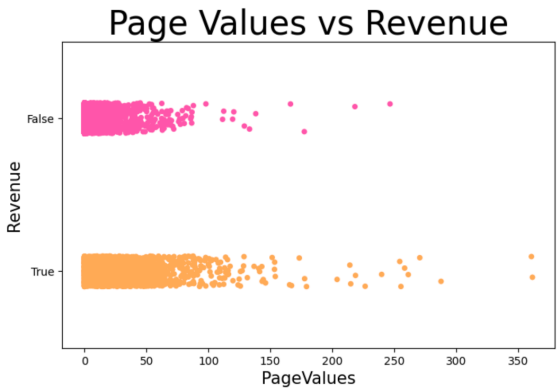
Traffic Type analysis:

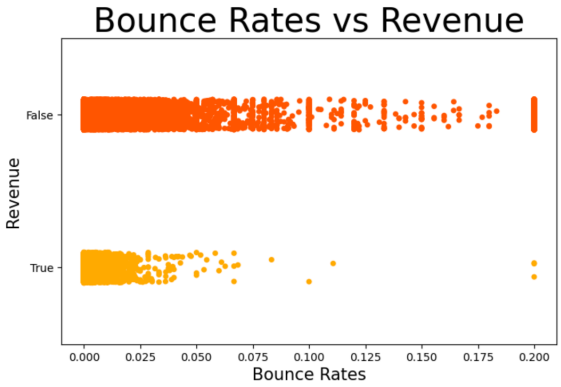
Using the simulation and linear regression, we analysed drift rates for all 20 traffic types and observed the traffic type with the quickest purchase.

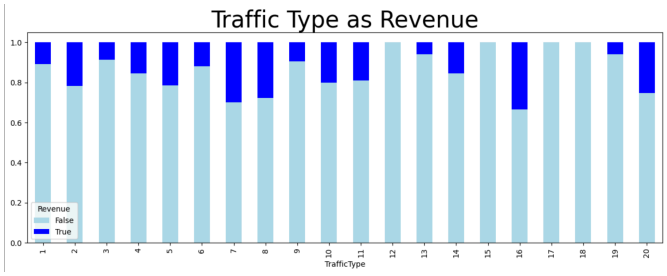
RESULTS AND DISCUSSION

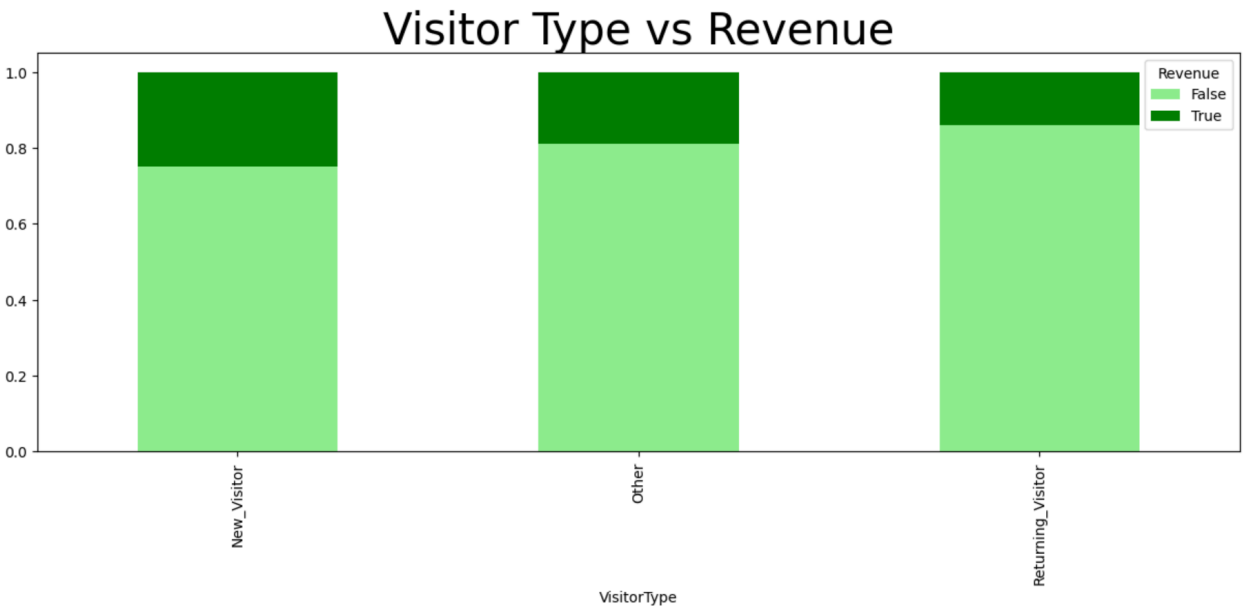
Feature Selection:











All features of the dataset were plotted wrt revenue to observe which features the target variable (Revenue) depended on. After a thorough analysis, it was observed that ExitRates, BounceRates, PageValues, TrafficType and VisitorType were the required features.

RATIONALE FOR PARAMETER FINALIZATION

Decision Boundary:

The Decision Boundary was set to 1 for positive boundary and -0.8 for negative boundary.

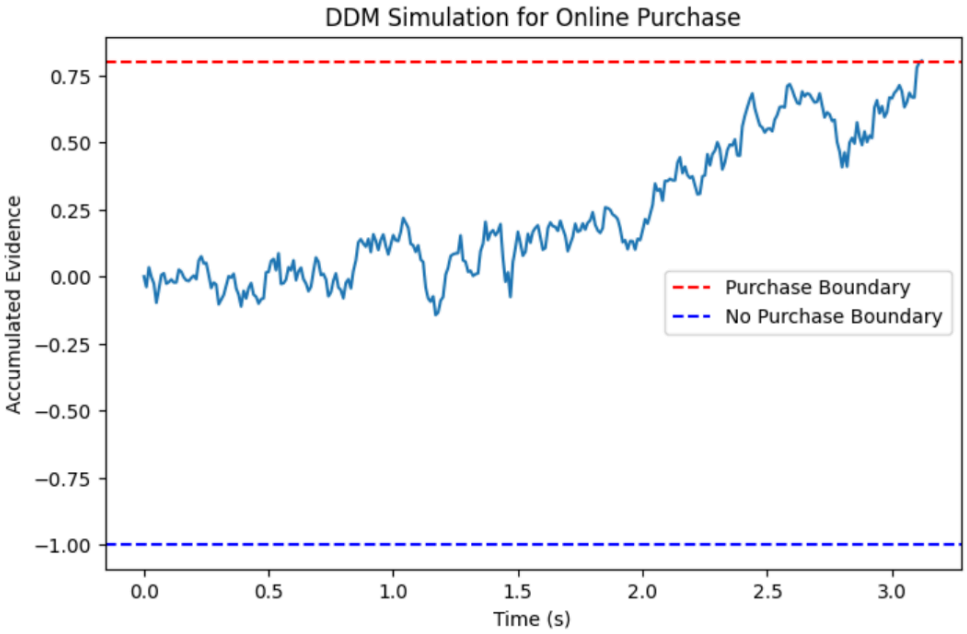
- The linear regression model learns a function for A_t (drift rate) based on historical data.
- Since only 6.6% of users purchased, the model is mostly trained on non-purchasers (93.4%).
- As a result, the learned drift rates A_t will be biased towards predicting no-purchase.
- Most samples will have a negative or small A_t, meaning the accumulated evidence will more often go toward −B (no purchase).
- Fewer users will have high A_t, making it harder to reach +B (purchase).
- This means purchases will be predicted less frequently, even if a user has high "purchase-intent" features.
- For most users, evidence accumulation will move quickly to −B (no purchase), meaning short reaction times.
- For the small number of purchasing users, evidence accumulation will take longer to reach +B, leading to longer reaction times for purchases.

To ensure that the drift diffusion model models this trend, the decision boundary was adjusted to the above mentioned values.

Starting Point:
Given that only 6.6% of users purchased the product, an asymmetric starting point (e.g., closer to the no-purchase boundary) might lead to premature terminations.
By setting it at 0, users have a fair chance to accumulate evidence in either direction without an artificial bias toward no-purchase.

Drift Rate:
A linear regression model was trained on features like *exit_rates, Page_value, Bounce_rate, traffic_type , visitor_type* to predict the drift rate for each user.
This means A_t is a function of real-world behavioral factors, rather than being arbitrarily chosen.

MEAN AND SD OF REACTION TIME FOR THE PURCHASE (REVENUE = TRUE)



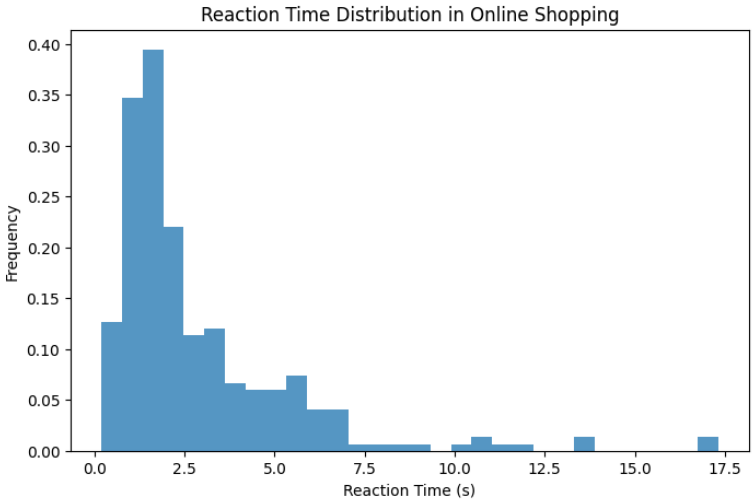
This is an example graph for evidence accumulation for one user. The reaction time is around 3s.
The mean of reaction times was found to be **4.292699161425577s.**
The standard deviation of reaction times was found to be **3.539405856235836s.**

The figure below shows the distribution of reaction times of all users. The distribution follows the expected trends as it is initially **right skewed** with a long tail at the end.

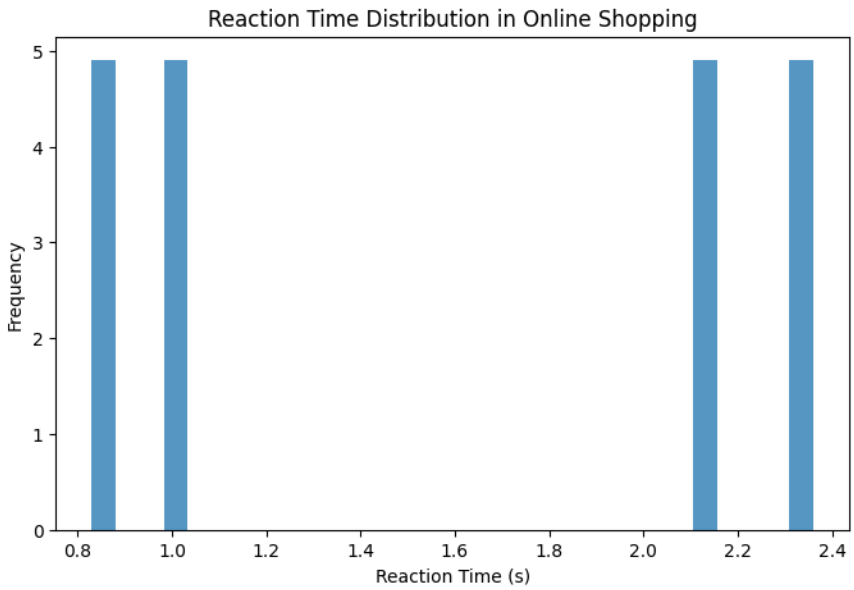
TRAFFIC TYPE ANALYSIS

Using the simulation and linear estimation, we analysed drift rates for all 20 traffic types and found out the traffic type with the quickest purchase (reaction time = 0.19).

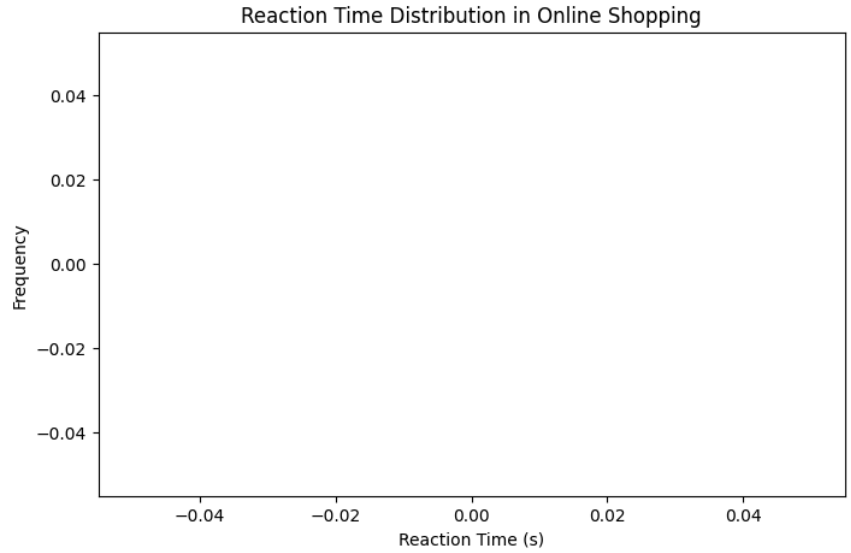
- We obtain some roughly right skewed graphs for reaction time where there is enough data for plotting the graph.
- Some traffic types have very less or no data for purchase. For traffic types with less data points, we obtain a comparatively very high or very low value of reaction time due to fewer and extreme values.



Minimum reaction time for traffic type 1: 0.19s (lowest of all)
Mean drift rate for traffic type 1: 3.580963740458015



Minimum reaction time for traffic type 9: 0.8300000000000001 (very high value due to less data)
Mean drift rate for traffic type 9: 1.9874999999999998



No Purchase in traffic type 12 (no data).

LIMITATIONS

Limited Data for Each Traffic Type:

Due to the limited data available, the response time graph lacks smoothness, resulting in a rough right-skewed plot.

Implementing the DDM Model from Scratch:

As the Hddm library is deprecated and the PyDDM library did not offer a model compatible with our dataset, we had to define and implement the model ourselves.

REFERENCES

[1]“Real-time prediction of online shoppers’ purchasing intention using multilayer perceptron and LSTM recurrent neural networks | Semantic Scholar.” Accessed: Jan. 26, 2025. [Online]. Available: https://www.semanticscholar.org/paper/Real-time-prediction-of-online-shoppers%E2%80%99-purchasing-Sakar-Polat/747e098f85ca2d20afd6313b11242c0c427e6fb3

[2]C. E. Myers, A. Interian, and A. A. Moustafa, “A practical introduction to using the drift diffusion model of decision-making in cognitive psychology, neuroscience, and health sciences,” Front. Psychol., vol. 13, Dec. 2022, doi: 10.3389/fpsyg.2022.1039172.

[3]S. Lalvani and A. Katsaggelos, “Crowdsourcing with the drift diffusion model of decision making,” Sci. Rep., vol. 14, no. 1, p. 11311, May 2024, doi: 10.1038/s41598-024-61687-y.

[4]“PyDDM Cookbook — PyDDM 0.8.1 documentation.” Accessed: Jan. 26, 2025. [Online]. Available: https://pyddm.readthedocs.io/en/latest/cookbook/index.html#model-components

CODE FILES:

CGS616_1B.ipynb

CONTRIBUTIONS:

We all worked equally on all aspects of the Assignment 1A and 1B, including coding, report writing and analysis of the results

Particularly for Assignment 1A, we analyzed and interpreted data individually on the following topics:

Vaneesha S. Kumar: “Abortion Rights”

Nischay Patel: “LGBTQ+ Rights”

Nikhil Gupta: “Gun Ownership”