DSP505 – FINAL PROJECT

POPULAR WEBSITES DATA ANALYSIS

Submitted By: Raghav Borikar – M24DS010

Submitted To: Prof. Gagan Raj Gupta

INDEX of CONTENTS

INTRODUCTION

This project focuses on cleaning, analyzing, and visualizing a dataset that provides insights into the top 50 websites across 191 countries. The dataset includes a mix of categorical and continuous variables, capturing various aspects of website performance, such as global and country-specific rankings, user engagement metrics, social media influence, trustworthiness, and hosting details.

The primary goal of this project is to uncover meaningful patterns and trends in website performance at a global and regional level. The analysis involves handling missing values, correcting data inconsistencies, and transforming the data into a usable format. Exploratory Data Analysis (EDA) is conducted to identify correlations, trends, and anomalies, followed by creating dashboards that present key insights in an interactive and visually engaging format.

By leveraging tools such as Python, SQL, and dashboarding platforms (Power BI), this project aims to analyse various websites on their regional & global parameters as well as based on their global functioning & identity.

The final deliverables include clean and well-structured data, comprehensive insights from analysis, and intuitive dashboards, making this project a valuable resource for understanding the global and regional dynamics of website performance.

OBJECTIVES

- Perform Data Cleaning on the Dataset using Pandas.
- Perform exploratory data analysis (EDA) to identify patterns, trends, and correlations by using Matplotlib, Seaborn & Plotly.
- Use SQL queries to aggregate, filter, and rank data.
- Create interactive dashboards to visualize key metrics.(Using PowerBI)

DATASET

The dataset used for this project was sourced from Kaggle and provides comprehensive information about the top 50 websites in 191 countries. It consists of approximately 9,500 rows and 31 columns, capturing data from around 3,401 distinct websites. Some websites appear as region-specific instances, reflecting how their popularity and performance vary by country.

The dataset includes a mix of categorical and continuous variables, providing a rich foundation for analysis. Here's a detailed breakdown of the data:

1. Dataset Size

- **Rows**: ~9,500 (representing website-country combinations).
- **Columns**: 31 (covering various metrics and attributes).
- **Unique Websites**: ~3,400 (with some websites appearing in multiple countries).

2. Key Features

The dataset contains the following types of information:

- **Website Identification and Ranking**:
    - **Website Name**: The domain or website name.
    - **Country Rank**: The ranking of the website in a specific country based on traffic.
    - **Global Rank**: The global ranking of the website in terms of traffic and reach.
- **Traffic Metrics**:
    - **Average Daily Visitors**: The average number of daily visitors to the website.
    - **Daily Pageviews Per User**: The average number of pages viewed by a user daily.

- **Monthly Pageviews**: Total pageviews accumulated over a month.
- **Reach Day**: A measure of the website's reach on a daily basis.

- **Trust and Safety**:
  - **Trustworthiness**: A rating or score indicating the perceived reliability of the website.
  - **Child Safety**: A score reflecting how safe the website is for younger audiences.
  - **Privacy**: Indicators related to the website's user data handling and security.

- **Social Media Influence**:
  - **Facebook Likes**: The number of likes the website has received on Facebook.
  - **Google Pluses**: Interactions or votes from Google users.
  - **LinkedIn Mentions**: Socials interactions on LinkedIn
- **Hosting and Infrastructure**:
  - **Hosted By**: Information about the hosting provider.
  - **Subnetworks**: The network infrastructure supporting the website.
  - **Hosted Location**: The geographic location of the hosting server.

- **Administrative and Technical Details**:
  - **Registrant**: The entity or individual who registered the website.
  - **Registrar**: The organization responsible for the website's registration.

- **Status**: Simply OK or Not OK (utilized differently in the project)

3. Data Types

- **Categorical Variables**:
  - Columns like Trustworthiness, Privacy, Child Safety, Status.
- **Continuous Variables**:
  - Metrics such as Average Daily Visitors, Monthly Pageviews, Reach Day.

5. Missing and Inconsistent Data

- Some columns contain missing or incomplete values, particularly in metrics like Trustworthiness or hosting details.
- There are inconsistencies in how certain metrics are recorded, such as zeros in place of missing values or variations in data types (e.g., numeric columns containing strings).

6. Dataset Challenges

- **Handling Missing Values**: Addressing gaps in critical columns like visitor counts and safety scores.
- **Standardizing Metrics**: Ensuring consistency in numerical data.
- **Filtering Redundant Entries**: Managing instances of duplicated websites with different regional attributes.

This dataset provides an excellent opportunity to explore the dynamics of website performance globally and locally. It offers insights into how various factors like traffic, trust, safety, and hosting influence a website's ranking and reach, making it a rich source for both technical and business analysis.

DATA CLEANING

In this project, the dataset presented several challenges related to missing values, inconsistencies, and data type mismatches. Rather than removing the data or dropping rows with missing values, I adopted a strategy that focused on preserving as much information as possible while ensuring the data was cleaned and prepared for meaningful analysis. Below is a detailed description of the cleaning process that was implemented to address these challenges:

1. Handling Missing Values

- **Strategy**: Instead of simply removing rows with missing values, I chose to fill the missing data with either website-specific or country-specific values to maintain the integrity of the dataset and prevent unnecessary data loss.
- **Website-Specific Replacement**: For websites that had multiple country-specific instances (e.g., Google.com appearing in many countries), I replaced the missing values with the **mean** or **median** of the existing values for that specific website. This approach helped preserve website-specific trends, particularly for metrics like **Average Daily Visitors** that could be impacted by regional differences.
    - Example: If google.com had missing data for **Average Daily Visitors** in a specific country, the missing value was replaced with the mean value of **Average Daily Visitors** for google.com across all countries where data was available.
- **Country-Specific Replacement**: In cases where no website-specific data was available (or if only one regional instance of the website existed), I replaced the missing values with the **median** or **mean** values for that particular **country**. This general approach ensured that country-level trends were accounted for without introducing biases from specific website data.

- Example: If the data for **Trustworthiness** was missing for a website in India, the value was filled with the median Trustworthiness value for all websites in India.

- **Replacing Null Values with Zero**: In certain columns where missing values implied a **lack of interaction** or **absence** (e.g., **Facebook Likes**, **Google Pluses**), I opted to replace NaN values with 0. This approach was particularly useful in cases where a 0 represented meaningful data (e.g., a website that had no social media presence).
    - Example: For columns like **Facebook Likes** and **Google Pluses**, missing values were replaced with 0 because they indicated the absence of social media interactions for some websites.

2. Dealing with Data Type Inconsistencies

- The dataset contained several **data type mismatches**, which could cause issues during analysis and visualization. To resolve this, I carefully checked each column and converted the data types where necessary:
    - **Numeric Columns**: Some numeric columns were erroneously imported as strings due to the presence of commas or special characters (e.g., Avg Daily Visitors, Pageviews). These were cleaned by removing any non-numeric characters (like commas or spaces) and converting them to the appropriate numeric data types (int64 or float64).

3. Handling Subnetworks Column

- **Problem**: The **Subnetworks** column contained lists of subnetwork identifiers (e.g., ['subnet1', 'subnet2', 'subnet3']), which was difficult to work with for numerical analysis.
- **Solution**: To convert this information into a more useful form, I transformed the **Subnetworks** column into a **count** of subnetworks per website. This involved:
    - Parsing the list of subnetworks and counting the number of items in each list (i.e., how many subnetworks were associated with each website).

- This transformation allowed for easier aggregation and analysis, as it converted a non-numeric column into a numeric one suitable for further statistical analysis.
- Example: A website with three subnetworks was recorded as 3, while a website with no subnetworks had a value of 0. Later, all the 0 values were converted to 1, as it is obvious that the website must definitely be on some subnetwork (at least 1).

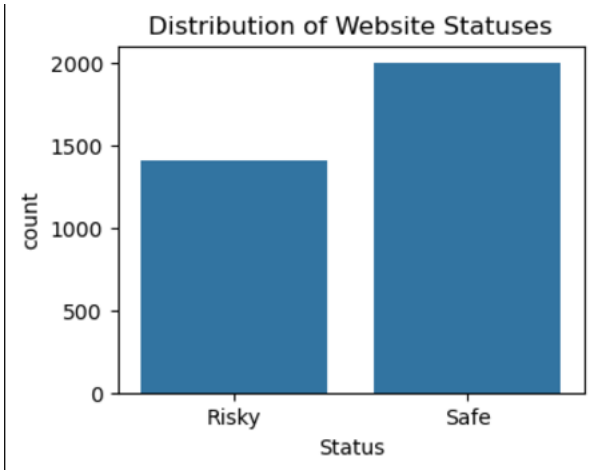
5. Redefining Status Attribute

The change introduced a new Status column to categorize websites as either "Safe" or "Risky" based on key indicators. Websites with missing critical information such as unknown location, registrar, or hosting provider were marked as "Risky." Additionally, websites with poor or unknown ratings in both child safety and privacy were also classified as "Risky." This was done to streamline the dataset and highlight websites that may pose trust or security concerns. The objective was to provide a quick and effective way to identify and analyze problematic websites, making the dataset more insightful for further exploration and analysis.

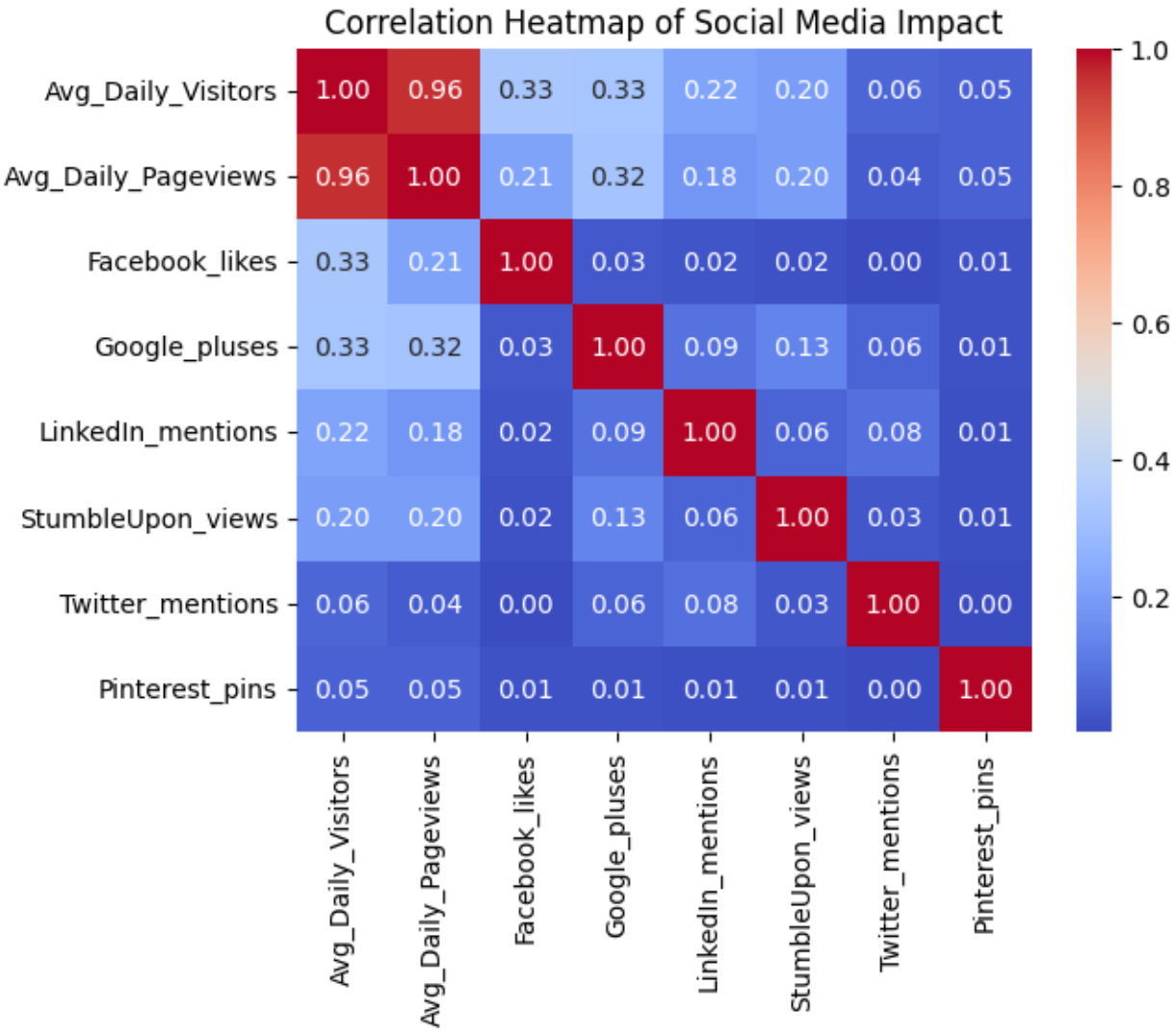
EXPLORATORY DATA ANALYSIS

The exploratory data analysis phase involved investigating various aspects of the dataset to uncover trends, relationships, and patterns. Key visualizations and analyses performed include:

- **Distribution of Statuses:**
  A bar graph was created to visualize the distribution of website statuses, categorized as "risky" or "safe." This provided insights into the safety ratings of websites globally. According to our new definition of Statuses, close to 2000 websites were classified as Safe while about 1400 were risky.



- **Correlation Matrix based on social media:**
  A correlation matrix of all social media features was generated to identify relationships between metrics, such as daily visitors, pageviews & social media presence & interaction. This helped uncover significant correlations and dependencies among variables.



This correlation matrix reveals that only facebook likes & google pluses tend to have low-to-moderate correlation with the Visitors & Pageviews. (Correlation follows the following custom defined scheme)

0.0-0.2 – Low Correlation

0.2-0.4 – Low to Moderate Correlation

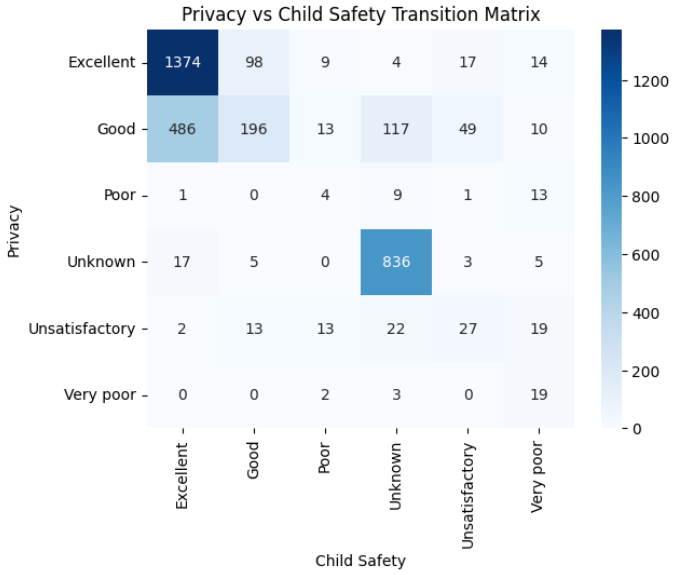
0.4-0.6 – Moderation Correlation

0.6-0.8 – Moderate to High Correlation

0.8-1.0 – High Correlation

- **A Bubble Chart of Log(Avg. Daily Visitors) vs Log(Avg. Daily Pageviews):**
  A bubble chart of logs of aforementioned 2 attributes tell us that the logs follow a linear relationship, this can be justified by the fact both Avg. Daily Visitors & Avg. Daily Pageviews follow an exponentially decreasing curve as their rank goes down in any country. Also, it gives us a convenience of using simple Linear Regression or SVM (Regressor) if we want to work on predictive model covering these 2 attributes.

- **A Transtition Matrix of Privacy vs Child Safety:**
  This shows us that, except the websites whose child safety & privacy status is unknown, about 2154 websites have both of these in either good or excellent category, the only concern that remains is of those 836 odd websites, whose details are unknown.



DIMENSIONALITY REDUCTION & MODEL BUILDING

Model Selection

The dataset contains various features related to social media activity, reach statistics, and website attributes. The goal is to address two distinct tasks:

1. **Predicting the traffic rank of websites** (a continuous variable) using a regression model.
2. **Classifying websites as "risky" or "safe"** (a categorical variable) using a classification model.

Based on the problem requirements and data structure:

- **Random Forest Regressor** was selected for the traffic rank prediction task due to its ability to handle non-linear relationships and high-dimensional data effectively. Another Reason to use this was the fact that contrary to the belief that Traffic Rank strongly correlates with the Pageviews or Visitors, it was observed that it had a negligible correlation coefficient against both these values, which prompted me to use Ensemble methods to reach the solution to this problem.
- **K-Nearest Neighbors (KNN) Classifier** was chosen for the classification task due to its simplicity and interpretability in handling proximity-based decision-making.

Feature Selection and Dimensionality Reduction

Principal Component Analysis (PCA):

To manage the high dimensionality and reduce noise, PCA was applied to two groups of features:

1. **Social Media Features (6 Columns):**
   Metrics such as Facebook likes, Twitter mentions, Google Plus interactions, and LinkedIn mentions were standardized and reduced to **4 principal components**, retaining 90% of the variance.

   - **Mathematical Formulation of PCA:**
     PCA projects data onto new axes (principal components) that maximize variance while minimizing redundancy. Z=XW where Z is the transformed feature matrix, X is the original feature matrix, and W is the matrix of principal components (eigenvectors).

Data Standardization

$$X_{\text{standardized}} = \frac{X - \mu}{\sigma}$$

Covariance Matrix Computation

$$\Sigma = \frac{1}{n-1} X_{\text{standardized}}^{\top} X_{\text{standardized}}$$

Eigenvalue Decomposition of the Covariance Matrix Σ

$$\Sigma v_i = \lambda_i v_i$$

Projection of Data

$$Z = X_{\text{standardized}} W_k$$

Reconstruction of Data (Optional)

$$X_{\text{reconstructed}} = Z W_k^\top$$

Variance Explained by v_i

$$\text{Variance explained by } v_i = \frac{\lambda_i}{\sum_{j=1}^{d} \lambda_j}$$

2. **Reach Statistics (5 Columns):**
   Features such as daily reach and pageviews were similarly standardized and reduced to **1 principal component** while retaining 93% variance.

The transformed components (4 for social media and 1 for reach statistics) were added back to the dataset, replacing the original columns.

Random Forest Regressor for Traffic Rank Prediction

Features Considered:

- Principal Components of Social Media Activity (SM_PC1 to SM_PC4).
- Reach Statistics Principal Component (RS_PC1).
- Average Daily Visitors and Average Daily Pageviews.
- Trustworthiness, Child Safety, Privacy, Subnetwork Count.

Model Description:

The **Random Forest Regressor** is an ensemble model that uses multiple decision trees to improve prediction accuracy and reduce overfitting. Each tree operates on a bootstrap sample of the data, and predictions are averaged to produce the final output. Default parameters of SKLEARN model were used.

- **Mathematical Formulation:**

Decision Tree Prediction:

$$\hat{y}_i = \frac{1}{|L_j|} \sum_{x_k \in L_j} y_k$$

Aggregated Random Forest Prediction:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} \hat{y}^{(t)}$$

Mean Squared Error (Loss Function):

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Results:

- **Validation Accuracy:** 93.23 %
- **Test Accuracy:** 94.40 %

The model effectively captures the relationships between features and traffic rank, making it a reliable predictor.

```
Model trained successfully!
Validation Score: 0.9323
Test Score: 0.9440
```

K-Nearest Neighbors Classifier for Risk Status (with K=3)

Features Considered:

- Principal Components of Social Media Activity (SM_PC1 to SM_PC4).
- Reach Statistics Principal Component (RS_PC1).
- Average Daily Visitors and Average Daily Pageviews.
- Trustworthiness, Child Safety, Privacy, Subnetwork Count.

Model Description:

The **K-Nearest Neighbors (KNN) Classifier** determines the class of a website based on the majority label among its k-nearest neighbors in the feature space. Euclidean distance was used to compute similarity between data points.

- **Mathematical Formulation:**

Distance Metric

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

Prediction for Classification

$$\text{Class}(x_q) = \arg\max_c \sum_{i=1}^{k} I(y_i = c)$$

where $I(y_i = c)$ is the indicator function:

$$I(y_i = c) = \begin{cases} 1, & \text{if } y_i = c \\ 0, & \text{otherwise} \end{cases}$$

Results:

- **Validation Accuracy:** ~79%
- **Test Accuracy:** ~77%

While the classifier achieves a moderate accuracy, further tuning (e.g., varying k or feature engineering) could improve performance.

```
Model trained successfully!
Validation Accuracy: 0.7955
Classification Report (Validation):
              precision    recall  f1-score   support

           0       0.77      0.70      0.73       213
           1       0.81      0.86      0.83       320

    accuracy                           0.80       533
   macro avg       0.79      0.78      0.78       533
weighted avg       0.79      0.80      0.79       533

Test Accuracy: 0.7736
Classification Report (Test):
              precision    recall  f1-score   support

           0       0.72      0.72      0.72       120
           1       0.81      0.81      0.81       176

    accuracy                           0.77       296
   macro avg       0.77      0.77      0.77       296
weighted avg       0.77      0.77      0.77       296
```

Conclusion

By applying PCA for dimensionality reduction and employing suitable models for regression and classification, the project demonstrates the effectiveness of machine learning in predicting website performance and assessing risks. The Random Forest Regressor provides excellent predictions for traffic rank, while the KNN Classifier offers a reasonable starting point for risk classification, with scope for further optimization.

SOME FUN WITH MySQL

To enhance the analysis, the dataset was imported into MySQL, enabling structured exploration of the data. MySQL allowed for efficient querying, aggregation, and filtering of the dataset, providing insights into various aspects such as website rankings, performance metrics, and geographical patterns.

Using SQL, the data was segmented and ranked to uncover regional and global trends. Relationships between different variables, such as hosting locations, safety ratings, and traffic metrics, were also explored. This integration added a dynamic layer to the analysis, enabling deeper insights and supporting informed decision-making.

Starting with some easy ones,

1. To get websites originating from India

```
1 •  use dsp505;
2 •  SELECT Website, Location
3    FROM websites_data
4    WHERE Location = 'India';
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Website	Location
flipkart.com	India
onlinesbi.com	India
irctc.co.in	India
billdesk.com	India
uidai.gov.in	India

2. Most popular websites by PageViews (greater than 1 Billion)

```
6 •  SELECT Website, Avg_Daily_Pageviews
7    FROM unique_websites
8    WHERE Avg_Daily_Pageviews > 1000000000;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Website	Avg_Daily_Pageviews
google.com	4192159833
youtube.com	2679159025
facebook.com	1082985733

3. Most Popular websites on Social Media

```
10 •  SELECT Website, Facebook_likes
11    FROM unique_websites
12    ORDER BY Facebook_likes DESC
13    LIMIT 10;
14
15
16
17
```

Result Grid | Filter Rows: | Export:

Website	Facebook_likes
facebook.com	5870000
whatsapp.com	166000
blogspot.kr	94200
google.com	94200
blogspot.ru	94200
blogspot.co.id	94200
blogspot.qa	94200
blogspot.ch	93700
blogspot.sn	93700
blogspot.rs	93700

Now, some interesting ones,

4. Listing Registrars based on descending order of the number of websites it hosts.

```
19 •  SELECT Registrar, COUNT(*) AS Website_Count
20    FROM unique_websites
21    GROUP BY Registrar
22    ORDER BY Website_Count DESC
23    LIMIT 5;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Registrar	Website_Count
Unknown	809
GODADDY.COM, LLC	361
NETWORK SOLUTIONS, LLC.	155
ENOM, INC.	127
MARKMONITOR INC.	97

5. Gathering websites having highest Pageviews per visitor

```
25 •  SELECT Website, (Avg_Daily_Pageviews / Avg_Daily_Visitors) AS Pageviews_Per_Visitor
26    FROM unique_websites
27    WHERE Avg_Daily_Visitors > 0  -- Avoid division by zero
28    ORDER BY Pageviews_Per_Visitor DESC
29    LIMIT 10;
30
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

Website	Pageviews_Per_Visitor
inbet.cc	69.00430570505921
neobux.com	38.50003139459626
buyma.com	31.100005542717295
weshare.mu	30.006191950464395
lightrabbit.co.uk	30.00451467268623
ss.lv	25.40010205600323
mangapanda.com	23.800069930502694
rst.ua	23.70029627873904
asb.by	23.700043351155475
etorrent.co.kr	23.000022426912295

6. This is more of an experiment to learn some complex queries, this query returns the websites which have the subnetworks count in double digits & have more views than the median in their respective countries.

```sql
WITH RankedData AS (
    SELECT
        Location,
        Avg_Daily_Pageviews,
        ROW_NUMBER() OVER (PARTITION BY Location ORDER BY Avg_Daily_Pageviews) AS RowNum,
        COUNT(*) OVER (PARTITION BY Location) AS TotalRows
    FROM unique_websites
)
SELECT Website, Location, Avg_Daily_Pageviews, subnetwork_count
FROM unique_websites AS w
WHERE subnetwork_count > 9
  AND Avg_Daily_Pageviews > (
      SELECT AVG(Avg_Daily_Pageviews)
      FROM RankedData
      WHERE Location = w.Location
        AND (RowNum = FLOOR((TotalRows + 1) / 2) OR RowNum = CEIL((TotalRows + 1) / 2))
  );
```

Website	Location	Avg_Daily_Pageviews	subnetwork_count
xvideos.com	Netherlands	146500484	17
xnxx.com	Netherlands	50749572	17
google.cd	United States	1369115	16
noaa.gov	United States	1640887	10
emag.ro	Romania	2989492	16
msmtrakk10a.com	France	441501	16
eksisozluk.com	Turkey	8717214	10
24h.com.vn	Vietnam	3492013	11

This resulted in some Pornographic websites appearing at the top, as they seemed popular but also maintained a high subnetwork count. I dug deeper into it to get it's reasons.

Content Distribution and Mirroring:

Porn websites often utilize **content delivery networks (CDNs)** or maintain multiple mirror sites to ensure accessibility across regions and networks. This creates a large number of subnetworks to distribute their content efficiently and circumvent local restrictions or bans.

Avoidance of Regulatory Measures:

To evade content restrictions, many porn websites operate on multiple subnetworks, domains, and hosting providers. This allows them to continue functioning even if some of their networks or domains are blocked by authorities or ISPs (Internet Service Providers).

Affiliate Marketing and Third-Party Partnerships:

These websites often participate in extensive affiliate marketing programs, partnering with third-party sites that redirect traffic to their platforms. This increases the number of associated subnetworks as affiliate programs are hosted on separate infrastructure.

Legal and Operational Challenges:

In response to legal challenges, some porn websites fragment their infrastructure into multiple subnetworks to distribute liability and avoid legal scrutiny. This distributed setup makes it harder for authorities to target or shut down their operations completely.

Monetization Strategies:

Many porn websites monetize through ads, pop-ups, and content redirection to partner sites. These monetization strategies often involve integration with multiple external networks and hosting providers, increasing the overall subnetwork count.

User Anonymity and Privacy:

Due to the nature of their content, these websites also prioritize user anonymity and privacy. Using multiple subnetworks and geographically dispersed servers enhances their ability to obscure user data and provide secure access.

Dashboarding in PowerBI

The project culminated in the creation of an interactive and insightful dashboard to visualize key metrics and findings. The dashboard highlights several aspects of the dataset, enabling users to explore the data in a visually engaging and dynamic manner. Below are the key components of the dashboard:

- **Scatter Plot:**
- Average of Daily Pageviews per User by Country Rank. This scatter plot visualizes the relationship between country rank and average daily pageviews per user, providing insights into how website engagement varies by rank.

- **Treemap:**
- Count of Websites by Privacy Rating, A treemap was used to display the distribution of privacy ratings among websites. Categories like "Excellent," "Good," and "Unknown" showcase the dominance of certain ratings and highlight potential privacy concerns.

- **Pie Chart**
- Count of Websites by Child Safety Rating. The pie chart represents the distribution of websites based on child safety ratings, including "Excellent," "Good," "Unsatisfactory," and others. This gives a quick overview of website safety levels.

- **Map**

  Count of Websites by Location and Status, An interactive world map shows the distribution of websites marked as "Safe" or "Risky" based on their geographical hosting locations. It provides a global perspective on website safety.

- **Tabular Data**

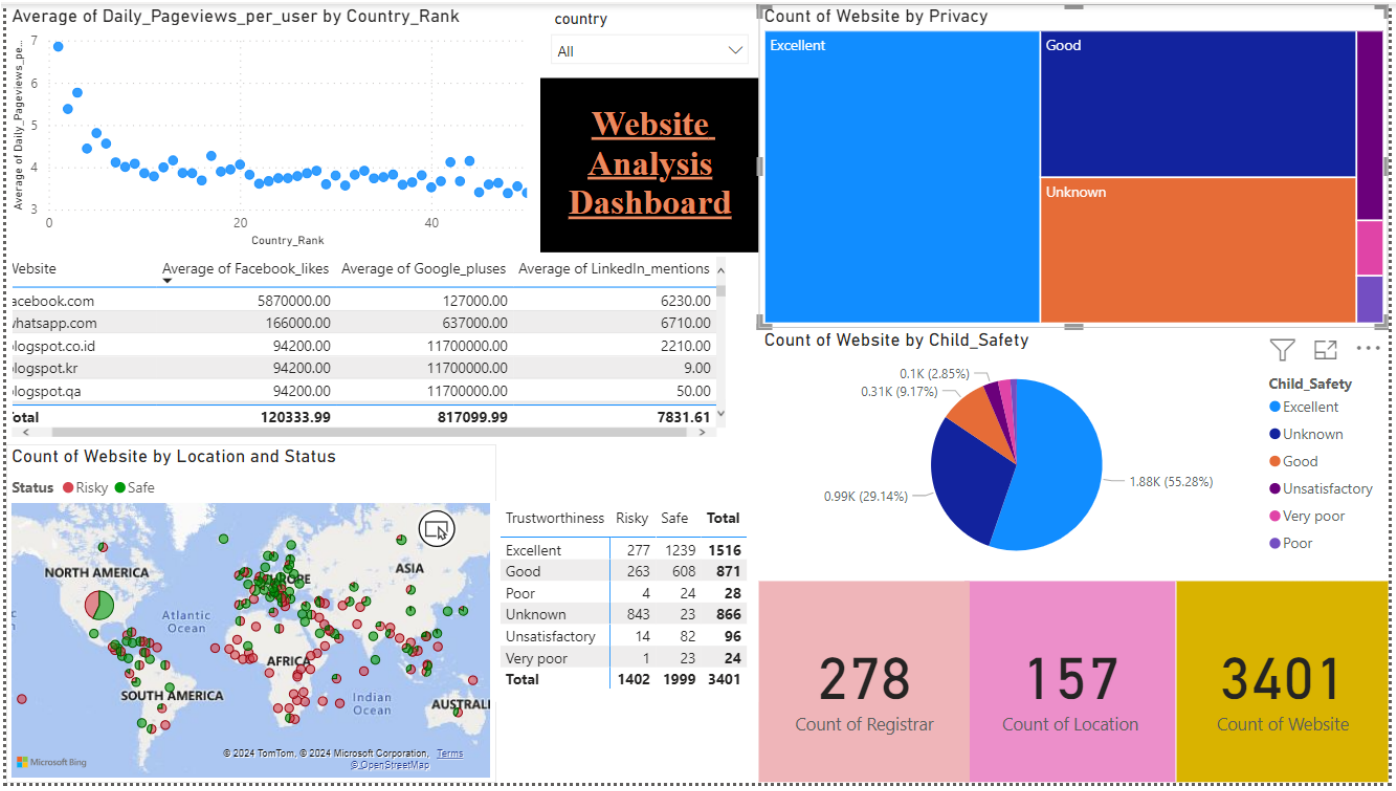
  A table summarizes the averages of social media metrics, such as Facebook likes, Google pluses, and LinkedIn mentions, for each website. This highlights the websites' popularity on social platforms.

- **KPIs**

  Key performance indicators (KPIs) are displayed to provide quick metrics:

  - Total count of unique websites: 3,401
  - Number of distinct hosting locations: 157
  - Number of registrars: 278

This dashboard enables stakeholders to interact with the data and draw actionable insights, offering a holistic view of website performance, safety, and engagement across different dimensions.


Average of Daily_Pageviews_per_user by Country_Rank
Average of Daily_Pageviews_pe...
Country_Rank
country
All
Website Analysis Dashboard
Count of Website by Privacy
Excellent
Good
Unknown
Vebsite
Average of Facebook_likes
Average of Google_pluses
Average of LinkedIn_mentions
acebook.com 5870000.00 127000.00 6230.00
/hatsapp.com 166000.00 637000.00 6710.00
logspot.co.id 94200.00 11700000.00 2210.00
logspot.kr 94200.00 11700000.00 9.00
logspot.qa 94200.00 11700000.00 50.00
otal 120333.99 817099.99 7831.61
Count of Website by Child_Safety
0.1K (2.85%)
0.31K (9.17%)
0.99K (29.14%)
1.88K (55.28%)
Child_Safety
Excellent
Unknown
Good
Unsatisfactory
Very poor
Poor
Count of Website by Location and Status
Status Risky Safe
NORTH AMERICA
EUROPE
ASIA
Atlantic Ocean
AFRICA
SOUTH AMERICA
Indian Ocean
AUSTRALI
© 2024 TomTom, © 2024 Microsoft Corporation, Terms
© OpenStreetMap
Trustworthiness Risky Safe Total
Excellent 277 1239 1516
Good 263 608 871
Poor 4 24 28
Unknown 843 23 866
Unsatisfactory 14 82 96
Very poor 1 23 24
Total 1402 1999 3401
278
Count of Registrar
157
Count of Location
3401
Count of Website

SOME DERIVED INSIGHTS

- USA happens to be the epicentre of the world of internet as a plurality of the websites **(about 1641 out of 3401)** are hosted in that country, but one must not be fooled by sheer numbers as many of them also fall in the category of risky (704) websites based on our analysis.

- All the websites originating from India (5) fall in the category of Safe Companies.

- Not all websites have a wide reaching or global social media presence across platforms as I came across many companies that were not active socially, it can mean 2 things, first is the lack of accessibility of Social Media sites in some nooks & corners of the world, while it can also mean the website is not engaging much which indirectly means that it is doing so to remain under the hood (it may mean potential fraud as well).

- As per our analysis, suspicious or Pornographic websites tend to be hosted on more subnetworks for reasons mentioned in the analysis.

- Websites with lower country ranks (higher positions) tend to have significantly higher daily pageviews per user, indicating a strong exponential correlation between rank and engagement.

- A majority of websites are rated "Excellent" or "Good" in terms of privacy, but a significant portion remains in the "Unknown" category, suggesting gaps in available privacy data.

- Over **55%** of websites are rated "Excellent" for child safety, but a considerable number fall under “Unknown"(29 %) category highlighting potential safety concerns. Unsatisfactory,"(2.85 %) "Poor,"(1.21 %) or "Very Poor,"(2.35 %) make up quite less of the total count.

- The global map shows that the distribution of "Safe" websites far outweighs "Risky" ones in most regions. However, some countries host a higher percentage of risky websites, warranting further investigation. To go in detail, only about **16** websites originating from the Region of Africa are safe, while the rest lie in risky category. This can be driven by the fact that African countries tend to have much weak digital safety infrastructure & also loopholes in the digital laws exist which makes it a breeding ground for Risky & unsafe websites.

- The table of social media reveals that google+ & facebook likes drive the website's popularity across countries having correlation coefficient of 0.33 & 0.32. Rest of the social media sites are bound by their less reach as compared to above listed platforms or by the different purpose of their existenct.

- With **157 distinct hosting locations** and **278 registrars**, the dataset highlights a diverse ecosystem of website management and hosting, distributed across various countries.

SUGGESTIONS & ACTIONS

Actionable Insights to Improve Website Traffic Rank

Based on the analysis and model results, the following insights can help websites improve their traffic rank:

1. Leverage Social Media Effectively

The analysis shows a moderate correlation between social media metrics and website visits & pageviews. Websites with higher engagement on platforms like Facebook & Google plus tend to perform better in Pageviews.

- **Action Steps:**
  - Invest in targeted social media campaigns to boost interactions (e.g., likes, shares, and mentions).
  - Focus on creating shareable content to increase visibility and drive traffic.

2. Enhance User Engagement

Metrics like **Average Daily Visitors** and **Daily Pageviews per User** significantly impact the traffic rank. Engaging content and user-friendly navigation can boost these metrics.

- **Action Steps:**
  - Regularly update the website with fresh, high-quality, and relevant content to encourage repeat visits.
  - Optimize website navigation to improve the user experience and reduce bounce rates.
  - Introduce interactive elements such as quizzes, polls, or videos to engage users for longer durations.

3. Ensure Trustworthiness and Safety

Websites with high **Trustworthiness** and **Child Safety Ratings** rank better and attract more traffic. A website's credibility directly affects its perception and traffic.

- **Action Steps:**
  - Obtain SSL certifications and prominently display trust badges.
  - Regularly monitor and address user complaints or negative reviews.
  - Comply with data privacy regulations to build user confidence.

4. Expand Reach Through Targeted Marketing

The **Reach Statistics** PCA component highlights the importance of reaching a wider audience. Increasing visibility across regions and demographics can significantly enhance traffic.

- **Action Steps:**
  - Implement SEO strategies to rank higher in search results and target niche keywords.
  - Use targeted advertisements on search engines and social media platforms to attract specific user segments.
  - Analyze competitor strategies to identify opportunities for expanding into untapped markets.

5. Focus on Quality Hosting and Infrastructure

Reliable hosting and optimized infrastructure can improve both website speed and uptime, contributing positively to the user experience and traffic rank.

- **Action Steps:**
  - Use the subnetworks optimally to maintain uptime while keeping the website up to date.
  - Use Content Delivery Networks (CDNs) to ensure fast loading times globally.

By implementing these insights, websites can systematically enhance their visibility, user engagement, and credibility, leading to improved traffic ranks over time.

DOMAIN KNOWLEDGE GAINED

- Had a comprehensive outing with pandas figuring out new ways to clean & transform data, sometimes using measures of central tendency or sometimes using grouped means.
- Brushed up on the Data Visualization Skills encompassing Matplotlib, Seaborn & Plotly & also learned about some new plots like Sankey, Network Graphs & how they are implemented.
- Some Practice over Dimensionality Reduction using PCA, & Model Building using KNN & Ensemble Methods.
- Learned how to connect a Notebook to a MySQL server & transfer a dataframe to a database without explicit schema definition.
- Practiced some Basic & Advanced SQL Queries ranging from simple WHERE clause to PARTITION & ORDER statements.
- Revisited PowerBI to build a Dashboard that covers some key attributes from the Dataset using various interactive & intriguing visuals.