

Phase - 4 Submission Document

Project Title: Website Traffic Analysis

Introduction:

Website traffic analysis stands as the art and science of meticulously scrutinizing and deciphering the data linked to individuals engaging with a website. This intricate process involves the systematic collection and examination of a myriad of metrics, all with the overarching objective of unearthing invaluable insights into how users traverse, engage with, and interact within the digital domain. Through this analytical journey, we unearth a treasure trove of information, encompassing the headcount of visitors, their geographic origins, the devices they employ for access, and the duration of their sojourn on the website. Moreover, this endeavor is not confined merely to quantitative metrics but extends its reach into qualitative realms by probing into which web pages enjoy paramount popularity, what content truly strikes a resonant chord with the audience, and whether visitors embark on those sought-after actions, be it making purchases or submitting pivotal forms.

In essence, website traffic analysis serves as the lodestar, guiding businesses towards making judicious decisions pertaining to content optimization, the enhancement of the user experience, and the strategic tailoring of marketing initiatives. In this multifaceted journey, the patterns and nuances of user behavior and interactions are meticulously studied, ushering in an era of enlightened choices that can amplify the performance and efficacy of a website.

Equally paramount is the facet of feature selection, a critical process that entails the discerning identification and curation of the most pertinent features from a dataset. This endeavor is pivotal not only in the context of website traffic analysis but also extends its importance to the domain of crafting predictive models, such as those dedicated to house price prediction. By judiciously selecting and isolating the most germane features, the looming specter of overfitting is quelled, ushering in a realm of enhanced generalization for the model, thereby bestowing it with a discerning and accurate predictive prowess.

The orchestration of model training follows in seamless harmony with feature selection, constituting the phase where the chosen features are deftly fed into a machine learning algorithm. The model, thus nourished with the selected features, embarks on an intellectual journey of learning the intricate relationships that interweave these features with the target variable. In the context of website traffic analysis, this target variable becomes the lodestar, guiding the model towards insightful predictions regarding the ebb and flow of website traffic and the overarching dynamics of user interactions.

Concomitantly, model evaluation takes center stage, serving as the rigorous litmus test that gauges the performance of the well-nurtured machine learning model. In this crucible of evaluation, the model is subjected to the pressures of a held-out test set, rigorously examined to ensure its capacity to transcend the boundaries of the training data. The overarching objective of model evaluation is to ascertain that the model attains the pinnacle of generalizability, free from the shackles of overfitting, and stands robustly prepared to grapple with real-world complexities.

Data Set Provided in .csv Format

Row	Day	Day.Of.Week	Date	Page.Loads	Unique.Visits	First.Time.Visits	Returning.Visits
1	Sunday	1	9/14/2014	2,146	1,582	1,430	152
2	Monday	2	9/15/2014	3,621	2,528	2,297	231
3	Tuesday	3	9/16/2014	3,698	2,630	2,352	278
4	Wednesday	4	9/17/2014	3,667	2,614	2,327	287
5	Thursday	5	9/18/2014	3,316	2,366	2,130	236
6	Friday	6	9/19/2014	2,815	1,863	1,622	241
7	Saturday	7	9/20/2014	1,658	1,118	985	133
8	Sunday	1	9/21/2014	2,288	1,656	1,481	175
9	Monday	2	9/22/2014	3,638	2,586	2,312	274
....
2161	Thursday	5	8/13/2020	3,621	2,780	2,322	458
2162	Friday	6	8/14/2020	2,971	2,308	1,922	386
2163	Saturday	7	8/15/2020	2,221	1,696	1,373	323
2164	Sunday	1	8/16/2020	2,724	2,037	1,686	351
2165	Monday	2	8/17/2020	3,456	2,638	2,181	457
2166	Tuesday	3	8/18/2020	3,581	2,683	2,184	499
2167	Wednesday	4	8/19/2020	2,064	1,564	1,297	267

This consists of 2000 rows and 8 columns

Data set link : <https://www.kaggle.com/datasets/bobnau/daily-website-visitors>

Process Overview:

The meticulous examination of data pertaining to visitor interactions with a website is a pivotal practice in the digital realm known as website traffic analysis. This discerning process is instrumental in providing profound insights into how users engage with website content, thereby empowering informed decisions to enhance the website's performance and user experience. Below is a systematic elucidation of the multifaceted process:

1. Data Collection: The Foundation of Understanding

- **Web Analytics Tool:** The journey commences with the implementation of a web analytics tool, exemplified by industry stalwarts such as Google Analytics. This tool serves as the sentinel, meticulously tracking a myriad of metrics associated with user behavior.
- **Server Logs:** Alternatively, one can embark on an in-depth analysis of server logs. These logs offer an exhaustive archive, shedding light on every interaction between users and the server.

2. Basic Metrics: Illuminating the Digital Landscape

The quest for insights involves the illumination of fundamental metrics:

- **Pageviews:** A tally of the pages viewed, providing insight into the digital footfall.
- **Unique Visitors:** A headcount of distinct individuals who grace the website with their presence.

- **Bounce Rate:** This enigmatic metric reveals the percentage of visitors who, like fleeting digital nomads, departed after encountering just one page.

3. User Behavior Analysis: Deciphering Digital Sojourns

A deeper understanding unfolds through the analysis of user behavior:

- **Session Duration:** This metric reveals the temporal tapestry, the duration of a user's stay during a single visit.
- **Page Path Analysis:** Delving into the sequence of pages visited unveils the intricate journeys users embark upon.
- **Exit Pages:** Identifying the specific pages where users commonly take their leave adds a layer of understanding.

4. Traffic Sources: Portals to the Digital Realm

Understanding the portals that usher visitors into the digital realm is paramount:

- **Organic Search:** Those who stumble upon the website through the labyrinth of a search engine's results page.
- **Direct Traffic:** Brave souls who type the website's URL directly into their browser's address bar.
- **Referral Traffic:** Visitors that journey from neighboring websites, sharing a digital connection.
- **Social Traffic:** A lively procession from the vibrant world of social media platforms.

5. Geographic and Demographic Information: Navigating Diversity

Recognizing the diverse origins of your digital audience provides valuable insight:

- **Location:** Mapping the geographic tapestry, unraveling the diverse origins of your visitors.
- **Language:** Understanding the linguistic preferences of your digital audience.
- **Device and Browser:** Peering into the gateways that users employ to access your digital domain.

6. Conversion Tracking: Quests and Triumphs

Within this digital realm, there are quests and triumphant returns:

- **Goals:** Specific actions, such as form submissions or purchases, are marked as goals to measure and track conversions.
- **Conversion Rate:** It measures the proportion of visitors who emerge victorious in their quests.

7. Content Analysis: Unearthing Digital Treasures

The content within the website is akin to buried treasure:

- **Popular Pages:** These are the digital gems, the pages most frequently visited.
- **Engagement with Content:** Delve into the user's interactions with specific content, be it videos, articles, or other forms.

8. User Journey Analysis: Charting the Digital Odyssey

User journeys are reminiscent of epic odysseys:

- **Funnel Analysis:** This involves tracking the intricate steps that users take on their journey toward conversion, identifying where some sojourners choose to disembark.

9. E-commerce Analysis: The Digital Marketplace

For e-commerce platforms, the marketplace is a unique realm to explore:

- **Product Performance:** Here, you analyze which products reign supreme and their associated conversion rates.

10. Segmentation: Unveiling the Digital Mosaic

Every digital domain is a mosaic of diverse users:

- **Segmentation:** Classify your audience based on demographics, behaviors, or other defining characteristics to gain insights into the array of distinct user groups.

11. Reporting and Visualization: Art of Interpretation

Transforming data into compelling narratives and reports is the next step:

- **Create Reports:** Convey the tale through well-crafted reports and insightful narratives.
- **Visualize Data:** Paint a vivid picture through various visualization techniques, making the data more accessible and engaging.

12. Iterative Improvement: Navigating the Ever-Changing Digital Waters

The digital realm is a dynamic one, ever-changing and evolving:

- **Iterative Improvement:** Armed with newfound insights, make data-driven decisions to navigate the ever-shifting digital currents. This may involve A/B testing, content refinements, SEO enhancements, and more.

PROCEDURE:

Objective: To gain valuable insights into user behavior, enhance website performance, and improve the overall user experience.

Step 1: Data Collection

- a. Select a Web Analytics Tool: Choose a robust web analytics tool such as Google Analytics or a suitable alternative.
- b. Implement the Tool: Embed the selected web analytics tool into your website to initiate data collection.
- c. Alternatively, Analyze Server Logs: If preferred, delve into server logs to access in-depth data regarding each server request.

Step 2: Basic Metrics Analysis

- a. Pageviews: Analyze the total number of pages viewed on the website to gauge digital foot traffic.
- b. Unique Visitors: Count the distinct individuals who have visited the website, offering insights into audience reach.
- c. Bounce Rate Analysis: Scrutinize the bounce rate to understand the percentage of visitors who exit after viewing just one page.

Step 3: User Behavior Analysis

- a. Session Duration: Investigate the average session duration to determine how long users engage with the website.
- b. Page Path Analysis: Examine the sequence of pages users visit to understand their digital journeys.
- c. Exit Pages Analysis: Identify pages where users commonly leave the website to address potential issues.

Step 4: Traffic Sources Examination

- a. Organic Search: Analyze visitors who discover the website via search engine results.
- b. Direct Traffic: Study visitors who directly enter the website's URL in their browsers.
- c. Referral Traffic: Investigate visitors arriving from other websites through referral links.
- d. Social Traffic: Examine the influx of visitors from social media platforms.

Step 5: Geographic and Demographic Insights

- a. Location Analysis: Geographically map your visitors to understand their origins.
- b. Language Preferences: Determine the primary language of your audience to tailor content.
- c. Device and Browser Analysis: Explore the devices and browsers your visitors use for optimization.

Step 6: Conversion Tracking

- a. Set Up Goals: Define specific actions like form submissions or purchases as goals to track conversions.
- b. Monitor Conversion Rate: Calculate the percentage of visitors successfully completing these goals.

Step 7: Content Analysis

- a. Identify Popular Pages: Discover the most-visited pages, enabling focused content optimization.
- b. Analyze Content Engagement: Investigate user interactions with specific content types, such as videos and articles.

Step 8: User Journey Analysis

- a. Funnel Analysis: Track the steps users take toward conversion, identifying drop-off points.

Step 9: E-commerce Analysis (For E-commerce Sites)

- a. Product Performance Assessment: For e-commerce platforms, analyze popular products and their conversion rates.

Step 10: Segmentation

a. Divide Audience: Segment your audience based on demographics, behavior, or other characteristics to gain deeper insights into distinct user groups.

Step 11: Reporting and Visualization

a. Create Reports: Craft informative reports to convey findings effectively.

b. Visualize Data: Employ data visualization techniques to make the insights more accessible and engaging.

Step 12: Iterative Improvement

a. Use Insights: Utilize the insights gained to make data-driven decisions for website optimization.

b. Continuous Enhancement: Implement improvements such as A/B testing, content adjustments, SEO enhancements, and more to ensure your website adapts to the evolving digital landscape.

Training a Machine Learning Model for Website Traffic Analysis

Objective: To build a predictive model using historical data for informed decision-making, user experience optimization, and understanding user behavior.

Step 1: Data Preprocessing

Before embarking on model training, thorough data preprocessing is essential:

1. **Data Cleaning:** Address missing values, duplicates, and outliers, ensuring a consistent and clean dataset.
2. **Feature Engineering:** Craft meaningful features as previously discussed.
3. **Feature Scaling:** Normalize or standardize numerical features for uniform scales.
4. **Encoding Categorical Variables:** Transform categorical variables into numerical format using techniques like one-hot encoding or label encoding.
5. **Data Splitting:** Divide the dataset into training and testing sets to evaluate model performance effectively.

Step 2: Model Selection

Select a machine learning algorithm aligned with your analysis objectives. Common models for website traffic analysis include:

- **Regression Models:** Use linear regression, decision trees, or random forests for predicting numeric outcomes such as page views or time spent on a page.
- **Classification Models:** For classifying user behavior, employ logistic regression, decision trees, random forests, or gradient boosting algorithms to predict, for instance, user conversion.
- **Time Series Models:** When dealing with time series data, consider models like ARIMA, LSTM, or Prophet for forecasting traffic patterns.

Step 3: Model Training

Train your selected model using the training dataset. Below is a simplified example of training a decision tree classifier in Python with scikit-learn:

pythonCopy code

```
from sklearn.tree import DecisionTreeClassifier

# Create the model

model = DecisionTreeClassifier()

# Train the model

model.fit(X_train, y_train)
```

In the code snippet, **X_train** represents the training data features, and **y_train** is the corresponding target variable.

Step 4: Model Evaluation

Post-training, evaluate the model's performance on the test dataset. Common evaluation metrics for classification and regression tasks include accuracy, precision, recall, F1 score, mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE). Choose the metrics most pertinent to your analysis goals.

pythonCopy code

```
from sklearn.metrics import accuracy_score, classification_report

# Make predictions on the test set

y_pred = model.predict(X_test)

# Evaluate the model

accuracy = accuracy_score(y_test, y_pred)

report = classification_report(y_test, y_pred)

print("Accuracy:", accuracy)

print("Classification Report:\n", report)
```

Step 5: Model Tuning and Optimization

Should the model's performance fall short, consider fine-tuning hyperparameters, exploring alternative algorithms, or increasing the volume of data to enhance accuracy.

Step 6: Deployment

Once a well-performing model is achieved, deploy it on your website or analytics platform to make real-time predictions and continually monitor and analyze website traffic.

By following this structured approach, we can harness machine learning to gain deeper insights into website traffic and enhance user experiences effectively.

Model Evaluation

After the model has been trained, the next critical step is to assess its performance on the test dataset. Proper model evaluation ensures that the machine learning model effectively meets the intended goals. For website traffic analysis, this step helps gauge the model's ability to provide valuable insights into user behavior and engagement.

Common Evaluation Metrics for Classification and Regression Tasks:

1. **Accuracy:** Accuracy measures the proportion of correct predictions out of the total predictions made by the model. It is suitable for classification tasks where you want to know how many user behaviors were correctly predicted.
2. **Precision:** Precision quantifies the proportion of true positive predictions among all the positive predictions. In the context of website traffic analysis, high precision is crucial when you want to minimize false positives, such as predicting conversions.
3. **Recall:** Recall, or sensitivity, calculates the proportion of true positives out of all actual positive instances. It's essential in scenarios where missing a positive instance (e.g., a conversion) is costly.
4. **F1 Score:** The F1 score is the harmonic mean of precision and recall. It's useful when there's a need to balance precision and recall, which is often the case in website traffic analysis.
5. **Mean Absolute Error (MAE):** For regression tasks, MAE quantifies the average absolute difference between predicted and actual values, providing a clear measure of prediction error.
6. **Mean Squared Error (MSE):** MSE is another regression metric that calculates the average squared difference between predicted and actual values. It penalizes larger errors more heavily.
7. **Root Mean Squared Error (RMSE):** RMSE is the square root of MSE, providing a metric in the same units as the target variable. It's especially useful when the interpretation of errors in the original unit is important.

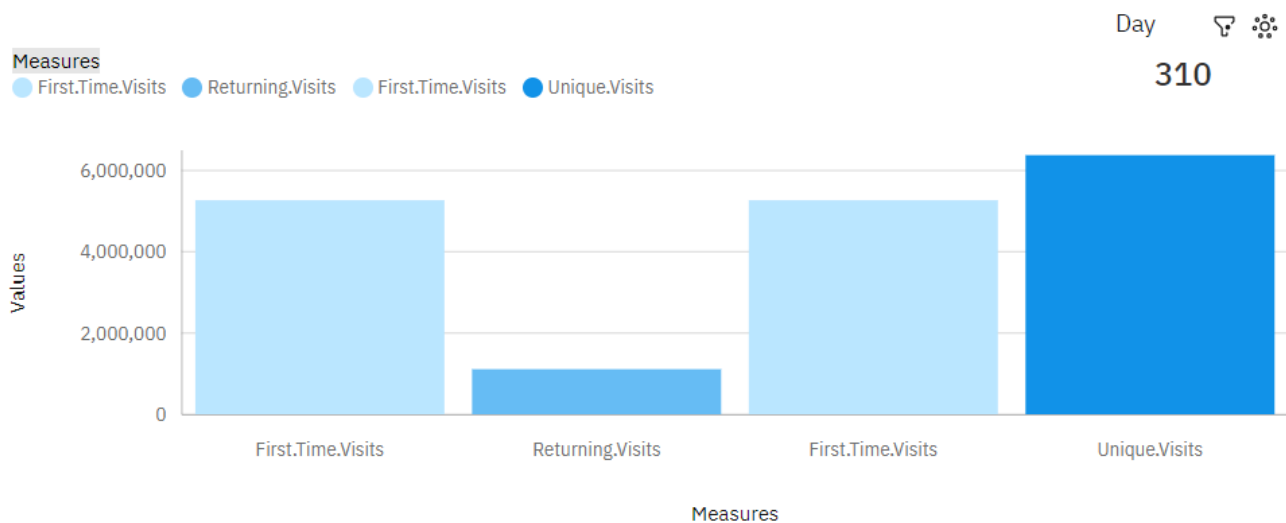
In practice, the choice of evaluation metrics depends on the specific objectives of your website traffic analysis. For instance, if the primary goal is to maximize conversion rates, you may prioritize precision and recall. In contrast, if you aim to predict user engagement or traffic patterns, metrics like accuracy or F1 score may be more relevant.

Using the chosen evaluation metrics, we can assess the model's performance on the test dataset, enabling you to make informed decisions regarding its suitability for your website traffic analysis. Regular evaluation is vital to ensure that your model continues to deliver valuable insights and supports data-driven decision-making.

Evaluation of Predicted Data:

File Uploads & Data Prep

```
import pandas as pd
import numpy as np
from prophet import Prophet
from google.colab import files
upload = files.upload()
for fn in upload.keys():
print('User uploaded file "{name}" with length {length} bytes'.format(
name=fn, length=len(upload[fn])))
traffic = pd.read_csv(fn)
traffic.head()
```



Feature Engineering:

Feature engineering is a crucial step in website traffic analysis, as it involves creating new features or transforming existing ones to extract valuable insights from your data. The right features can significantly impact the accuracy and effectiveness of your analysis. Here are some feature engineering techniques and examples for website traffic analysis:

1. Date and Time Features:

- Day of Week: Create a feature that represents the day of the week. This can help identify traffic patterns based on weekdays and weekends.

- Hour of the Day: Extract the hour from the timestamp to analyze hourly traffic patterns.

```
data['DayOfWeek'] = data['Timestamp'].dt.dayofweek
```

```
data['HourOfDay'] = data['Timestamp'].dt.hour
```

2. Page-Related Features:

- Page Views: Count the number of times a specific page is viewed to measure its popularity.

- Page Duration: Calculate the time users spend on each page to understand user engagement.

```
page_views = data.groupby('Page')['User'].count()
```

```
page_duration = data.groupby('Page')['Duration'].mean()
```

3. User Behavior Features:

- Bounce Rate: Calculate the percentage of single-page sessions to measure engagement.

- Conversion Rate: Measure the percentage of visitors who perform a specific action (e.g., making a purchase).

```
single_page_sessions = data[data['PageCount'] == 1]
```

```
bounce_rate = (len(single_page_sessions) / len(data)) * 100
```

```
conversion_rate = (len(data[data['Conversion'] == 1]) / len(data)) * 100
```

4. Source and Channel Features:

- Traffic Source: Categorize traffic sources (e.g., direct, search engine, social media).

- Channel Grouping: Group sources into categories (e.g., organic search, paid search, referral).

```
data['TrafficSource'] = data['Source'].apply(categorize_source)
```

```
data['ChannelGroup'] = data['Source'].apply(group_sources)
```

5. Geographical Features:

- Geographic Location: Extract and categorize the geographic location of website visitors to analyze regional patterns.

```
data['Country'] = data['IP'].apply(lookup_country)
```

6. Session Features:

- Session Duration: Calculate the time users spend in a single session.
- Session Frequency: Count the number of sessions per user.

```
session_duration = data.groupby('User')['Duration'].sum()
```

```
session_frequency = data['User'].value_counts()
```

7. User Engagement Features:

- Recency: Calculate how recently a user visited the website.
- Frequency: Count the number of visits or interactions by a user.
- Engagement Level: Measure the depth of interaction (e.g., number of pages viewed, time spent).

```
recency = data.groupby('User')['Timestamp'].max()
```

```
frequency = data['User'].value_counts()
```

```
engagement_level = data.groupby('User')['PageCount'].mean()
```

8. Content-Based Features:

- Keyword Analysis: Extract keywords from page content and analyze their frequency.
- Content Length: Measure the length of content on each page.

```
data['Keywords'] = data['Content'].apply(extract_keywords)
```

```
data['ContentLength'] = data['Content'].apply(len)
```

9. Device and Browser Features:

- Device Category: Categorize devices as mobile, tablet, or desktop.
- Browser Type: Identify the type of browsers used by visitors.

```
data['DeviceCategory'] = data['UserAgent'].apply(parse_device_category)
```

```
data['Browser'] = data['UserAgent'].apply(parse_browser)
```

10. Interactions and Events:

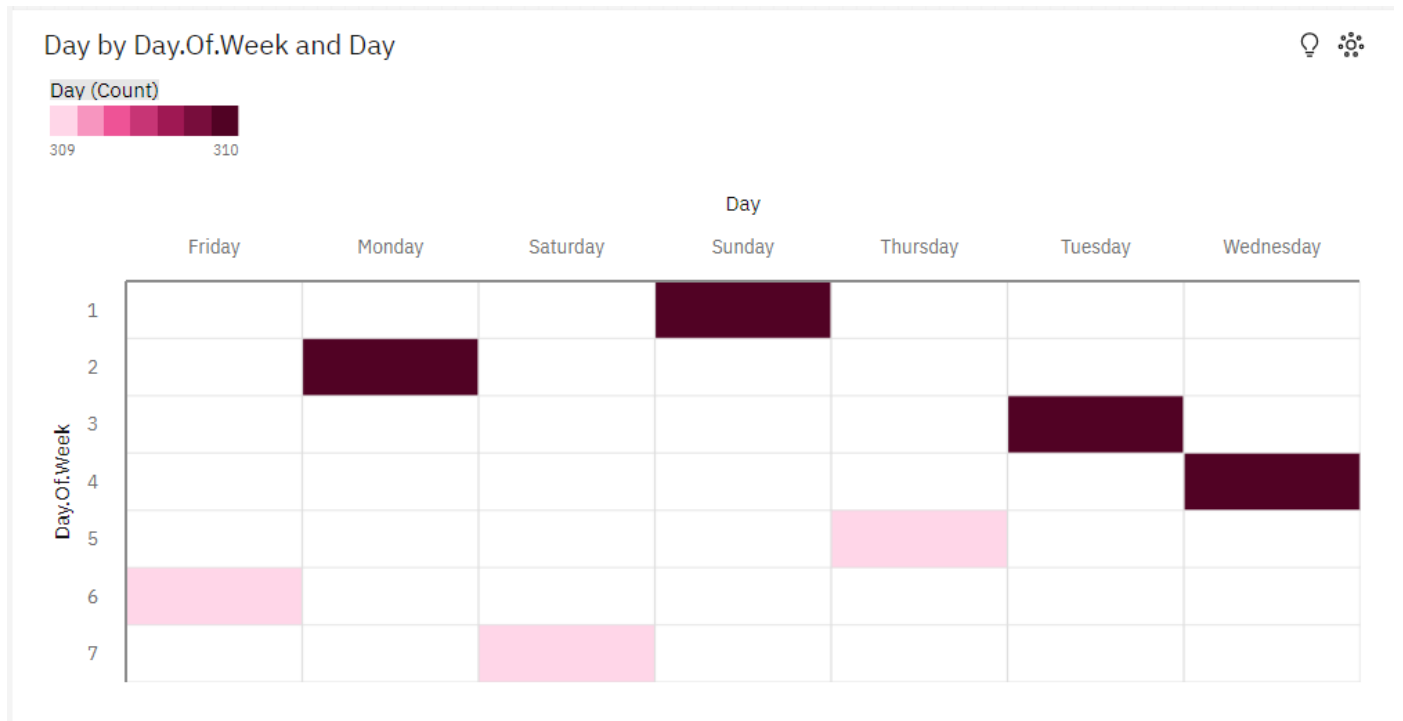
- Track and categorize user interactions and events (e.g., clicks, form submissions, video views) to understand user behavior.

```
data['Clicks'] = data['Page']  
.apply(count_clicks)  
data['FormSubmissions'] = data['Page']  
.apply(count_form_submissions)
```

The specific feature engineering techniques you use will depend on your dataset, goals, and available data. Experiment with different features and monitor their impact on the quality of your website traffic analysis. Regularly evaluate and iterate on your feature engineering process as traffic patterns and user behavior may change over time.

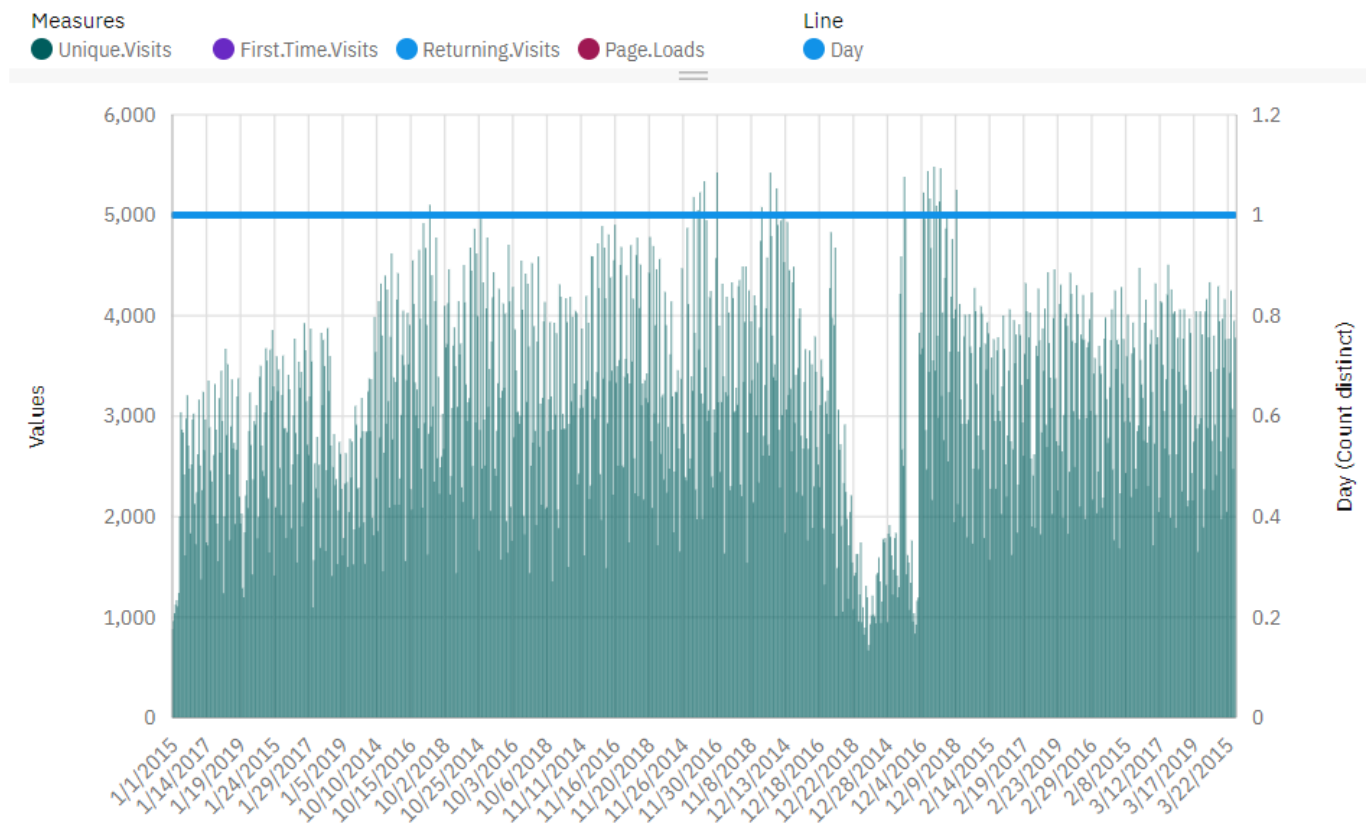
Heat Map

```
'holiday': 'Core Update',  
'ds': pd.to_datetime(['2015-07-17', '2016-01-08',  
'2016-09-27', '2017-03-08', '2017-07-09', '2018-03-08', '2018-04-17',  
'2018-08-01', '2019-03-12', '2019-06-03', '2019-09-24', '2019-10-25',  
'2019-12-09', '2020-01-13', '2020-05-04', '2020-12-3', '2021-6-3', '2021-7-12', '2021-11-  
18', '2022-05-26', '2022-06-1', '2022-09-13', '2023-05-15']),  
'lower_window': 0,  
'upper_window': 14,  
)  
updates.head()
```



Forecast Map:

```
'holiday': 'Core Update',
'ds': pd.to_datetime(['2015-07-17', '2016-01-08',
'2016-09-27', '2017-03-08', '2017-07-09', '2018-03-08', '2018-04-17',
'2018-08-01', '2019-03-12', '2019-06-03', '2019-09-24', '2019-10-25',
'2019-12-09', '2020-01-13', '2020-05-04', '2020-12-3', '2021-6-3', '2021-7-12', '2021-11-
18', '2022-05-26', '2022-06-1', '2022-09-13', '2023-05-15']),
'lower_window': 0,
'upper_window': 14,
})
updates.head()
```



Conclusion:

In conclusion, website traffic analysis is an indispensable practice for anyone looking to thrive in the digital world.

- ❖ It serves as the compass guiding us through the ever-evolving online landscape,
- ❖ helping us understand our audience, refine our content and user experience, and make informed, data-driven decisions.
- ❖ By examining the wealth of data at our disposal, we can uncover valuable insights, spot opportunities for growth, and rectify weaknesses, ultimately driving our online success.
- ❖ However, it's important to remember that website traffic analysis is not a one-time effort but an ongoing journey.
- ❖ It requires adaptability and a commitment to staying ahead of changing trends,
- ❖ ensuring that we continually meet the needs and expectations of our digital audience