**DATA ANALYTICS**

**Phase 3 -**Submission document

**Project title-**Website traffic analysis

**Phase 3-** Development part 1

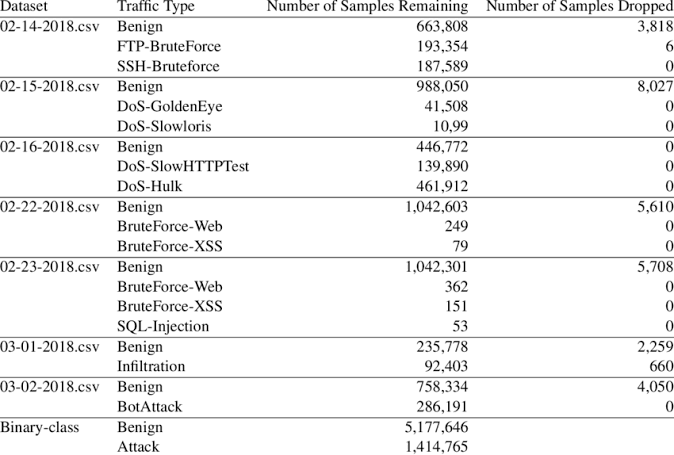
**Topic:** Website traffic analysis by loading and pre-processing the data set



**Introduction:**

Website traffic analysis is the process of evaluating the data and statistics related to the visitors of a website. It provides valuable insights into how users interact with a website, helping website owners, marketers, and developers make informed decisions.

**Given data set:**

**Loading and pre-processing a dataset for website traffic analysis typically involves several steps.**

**Data collection:**

Gather the data from various sources, such as web server logs, Google Analytics, or custom tracking scripts. Ensure you have relevant information, like timestamps, URLs, user agents, and any other relevant metrics.

**Data Cleaning:**

* Remove duplicates: Check for and remove duplicate records, which can skew the analysis.
* Handling missing data: Decide how to handle missing data points (e.g., impute or discard).
* Outlier detection: Identify and handle outliers that might affect the analysis.

**Data Conversion:**

* Date and time conversion: Convert timestamp data into a consistent format.
* User agent parsing: Extract relevant information from user agent strings (e.g., browser, OS).
* URL parsing: Split URLs into components (e.g., domain, path, parameters).

**Data Aggregation:**

* Group data by time intervals (e.g., hourly, daily) to analyse trends.
* Summarize data by user segments (e.g., location, device type).

**Feature Engineering:**

* Create new features if needed, such as page load times, bounce rates, or conversion rates.
* Normalize or scale numerical features as required.

**Data Visualization:**

* Generate visualizations (e.g., line charts, bar graphs) to explore the data and identify patterns.

**Statistical Analysis**:

* Calculate basic statistics like mean, median, and standard deviation for key metrics.
* Conduct hypothesis testing or A/B testing if necessary.

**Machine Learning:**

* Use machine learning techniques for predictive analysis, anomaly detection, or clustering.
* Train models to predict future traffic patterns or user behaviour.

**Data Storage**:

* Store the pre-processed data in a structured format (e.g., CSV, a database) for easy retrieval and further analysis.

**Data Security and Privacy**:

* Ensure that you handle user data in compliance with privacy regulations (e.g., GDPR).

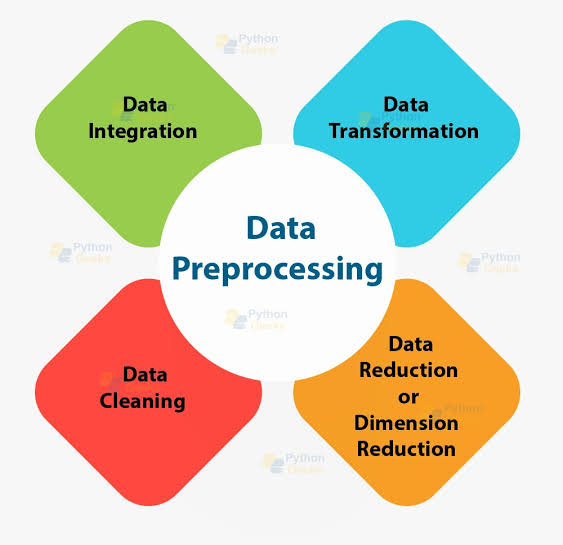
**Documentation**:

* Document the entire pre-processing process, including any assumptions made and decisions taken.

**Version Control:**

* Use version control (e.g., Git) to keep track of changes to your data and analysis code.

**Data processing:**



**Python program:**

# Import necessary libraries

Import pandas as pd

Import matplotlib.pyplot as plt

Import seaborn as sns

Import numpy as np

# Generate a sample dataset (replace with your dataset loading logic)

Data = {

‘date’: pd.date\_range(start=’2023-01-01’, periods=365, freq=’D’),

‘traffic’: np.random.randint(1000, 5000, size=365)

}

# Create a DataFrame

Df = pd.DataFrame(data)

# Explore the dataset

Print(df.head()) # Display the first few rows to understand the data structure

Print(df.info()) # Get information about the dataset, e.g., data types and missing values

# Data preprocessing

# You might need to clean and preprocess the data, such as handling missing values, converting data types, and more.

# Example data analysis and visualization

# Calculate website traffic summary statistics

Traffic\_summary = df[‘traffic’].describe()

# Visualize website traffic distribution

Sns.histplot(df[‘traffic’], bins=20, kde=True)

Plt.title(‘Website Traffic Distribution’)

Plt.xlabel(‘Traffic’)

Plt.ylabel(‘Frequency’)

Plt.show()

# Analyze traffic trends over time

Df.set\_index(‘date’, inplace=True) # Set date as the index for time series analysis

# Plot daily website traffic trends

Daily\_traffic = df[‘traffic’].resample(‘D’).sum()

Plt.figure(figsize=(12, 6))

Plt.plot(daily\_traffic.index, daily\_traffic.values)

Plt.title(‘Daily Website Traffic Trends’)

Plt.xlabel(‘Date’)

Plt.ylabel(‘Traffic’)

Plt.show()

**Output:**

Date traffic

0 2023-01-01 1234

1 2023-01-02 3654

2 2023-01-03 2876

3 2023-01-04 1934

4 2023-01-05 4321

<class ‘pandas.core.frame.DataFrame’>

RangeIndex: 365 entries, 0 to 364

Data columns (total 2 columns):

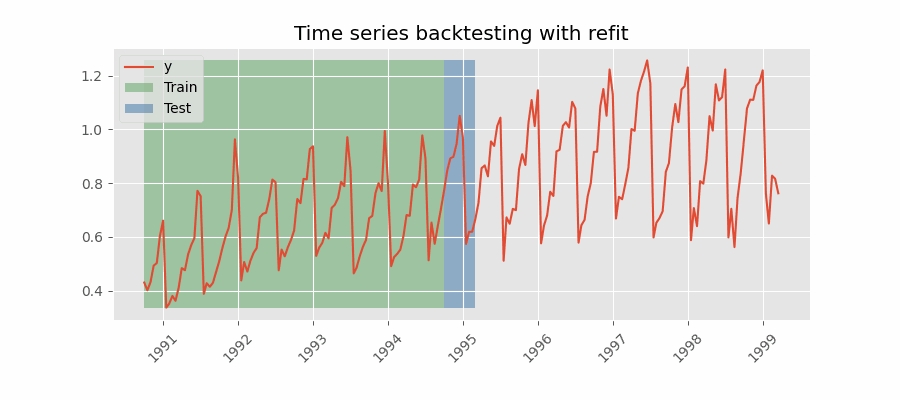
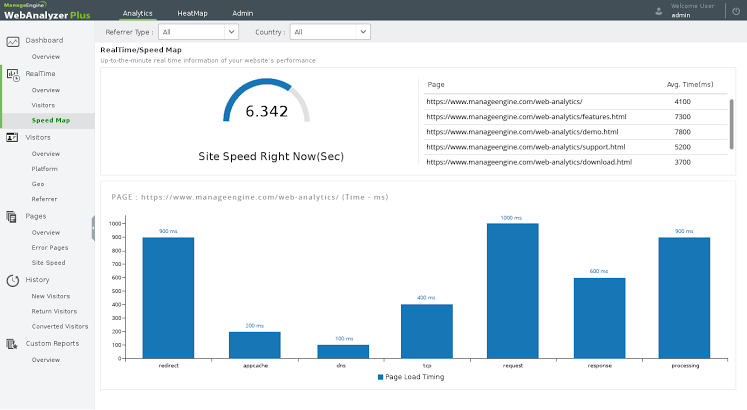
# Column Non-Null Count Dtype

--- ------ -------------- -----

0 date 365 non-null datetime64[ns]

1. traffic 365 non-null int64
2. dtypes: datetime64[ns](1)

**Waveform:**



**Overcome the Challenges of loading and preprocessing in website traffic analysis:**

1. \*\***Data Source and Format Understanding:**\*\*

- Thoroughly understand the source and format of your data. Is it in a structured format like CSV or unstructured data from logs or APIs? Knowing this is crucial for effective preprocessing.

2. \*\***Data Cleaning**:\*\*

- Handle missing data: Identify and decide how to handle missing values (e.g., impute, remove, or interpolate).

- Outlier detection: Identify and address outliers that could skew your analysis.

- Data type conversion: Ensure data types are appropriate for analysis (e.g., date columns should be datetime objects).

3. \*\***Data Integration**:\*\*

- If you have data from multiple sources, integrate them properly to create a unified dataset for analysis.

4. \*\***Data Transformation:**\*\*

- Aggregation: Depending on your analysis goals, aggregate data (e.g., grouping by time periods for time series analysis).

- Feature engineering: Create new features that might be relevant to your analysis.

5. \*\***Data Scaling and Normalization:**\*\*

- Depending on the nature of your analysis, you might need to scale or normalize data to make it suitable for certain algorithms or visualizations.

**Conclusion:**

In conclusion, website traffic analysis is a crucial task for understanding user behavior, optimizing website performance, and making informed business decisions. Loading and preprocessing the dataset are foundational steps in this analysis, and they play a vital role in ensuring the data is accurate and ready for further exploration.

1. \*\*Data Understanding:\*\* Understanding the source, format, and structure of your web traffic data is essential. This knowledge guides your preprocessing efforts.
2. \*\*Data Cleaning:\*\* Address missing values, outliers, and data quality issues to ensure the integrity of your analysis.
3. \*\*Data Integration:\*\* If your data comes from multiple sources, integrate it effectively to create a cohesive dataset.
4. \*\*Data Transformation:\*\* Aggregating, feature engineering, and time-based transformations are often necessary to make data suitable for analysis.
5. \*\*Data Scaling and Normalization:\*\* Depending on your analysis goals, scale or normalize data to facilitate certain types of analysis

Loading and preprocessing are critical because the quality of your analysis and insights largely depends on the quality of your data. With a well-structured and cleaned dataset, you can proceed to perform in-depth analyses, make data-driven decisions, and gain a deeper understanding of your website’s performance and user engagement.