# INTERNET USAGE

**A Project Report**

***Submitted by***

**Priyanshi (Roll No -202401100300185)**

***in partial fulfilment for the award of the degree of***
**Bachelor of Computer Science and Engineering (Artificial Intelligence),**

**KIET Group of Institutions, Ghaziabad**
**Affiliated to Dr. A.P.J. Abdul Kalam Technical University, Lucknow**

## *1. Introduction*

As internet usage becomes increasingly integral to daily life, understanding user behavior through data-driven methods is essential. This project focuses on analyzing internet usage patterns to uncover trends and relationships between user activity metrics such as daily usage hours, site categories visited, and sessions per day. By utilizing a dataset containing detailed behavioral information, the goal is to develop insights that can support digital product development, user segmentation, or future predictive modeling in areas like digital engagement or online behavior forecasting.

## *2. Problem Statement*

To analyze and understand patterns in internet usage behavior using key activity metrics such as daily usage hours, number of site categories visited, and sessions per day. The goal is to uncover meaningful insights that can help digital platforms optimize user engagement strategies, support behavior-based segmentation, and lay the groundwork for predictive modeling in areas such as user retention or digital well-being.

## *3. Objectives*

- Preprocess the dataset for analysis.
- Perform exploratory data analysis (EDA).
- Generate visualizations to highlight patterns and relationships.
- Derive insights based on statistical summaries and graphical representations.

## *4. Methodology*

**Data Collection:** A CSV file containing internet usage statistics is uploaded for analysis.

**Data Preprocessing:**

- Handling missing values using mean/mode imputation.
- Converting data types where necessary.
- Creating new derived columns if needed.

**Exploratory Data Analysis (EDA):**

- Descriptive statistics and data summaries.
- Univariate and bivariate analysis using visualizations.

**Visualization:**

- Bar charts, line graphs, heatmaps, and histograms to present insights effectively.

## *5. Data Preprocessing*

The dataset is cleaned and prepared using the following steps:

- Identifying and filling missing values.
- Converting date/time fields and numeric columns to appropriate formats.
- Removing irrelevant or duplicate entries.

## *6. Analysis Implementation*

Python libraries like pandas, matplotlib, and seaborn are used for analysis and visualization. Jupyter Notebook or Google Colab is used as the coding environment.

## *7. Evaluation Metrics*

As this is an analysis and visualization project, key outputs include:

- Completeness and clarity of visualizations.
- Insights extracted from the dataset.
- Quality and informativeness of statistical summaries.

## *8. Results and Analysis*

- Summarized trends and outliers were identified.
- Visualizations revealed key patterns (e.g., usage spikes, demographic correlations).
- The analysis highlighted valuable insights relevant to internet usage behavior.

## *9. Conclusion*

The project successfully demonstrates how Python can be leveraged to clean, analyze, and visualize internet usage data. The results can inform strategic decisions, marketing plans, or further academic research in digital behavior.

## *10. References*


- pandas documentation
- matplotlib and seaborn documentation
- Official Python documentation

# Code:

```
#importing the libraries

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns


# Load the dataset

df = pd.read_csv('/content/internet_usage.csv')


# Displaying the basic information

print("First 5 rows of the dataset:")

print(df.head())


print("\nDataset Info:")

print(df.info())


print("\nSummary Statistics:")

print(df.describe())


# Check for missing values

print("\nMissing Values:")
```

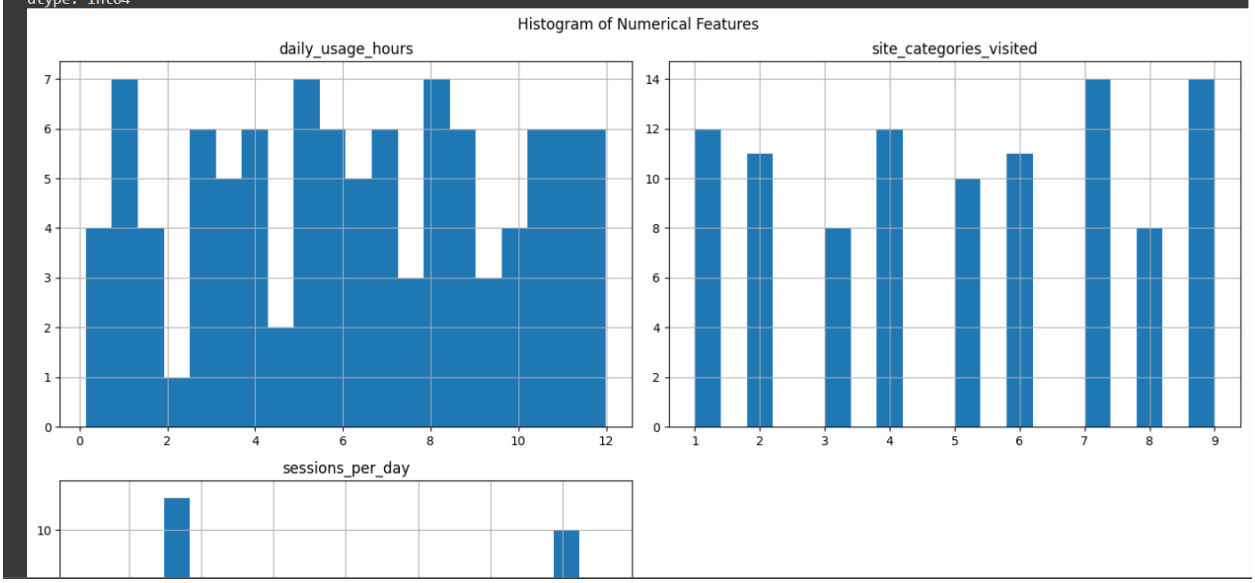
```
print(df.isnull().sum())


# Visualizations


# 1. Distribution of numerical features
df.hist(bins=20, figsize=(14, 10))
plt.suptitle('Histogram of Numerical Features')
plt.tight_layout()
plt.show()


# 2. Correlation Heatmap
plt.figure(figsize=(10, 8))
sns.heatmap(df.corr(numeric_only=True), annot=True, cmap='coolwarm',
fmt='.2f')
plt.title('Correlation Heatmap')
plt.show()


# 3. Line plot of internet usage over time (if applicable)
if 'Year' in df.columns and 'Internet Users' in df.columns:
    plt.figure(figsize=(10, 6))
    sns.lineplot(data=df, x='Year', y='Internet Users')
    plt.title('Internet Users Over Time')
```

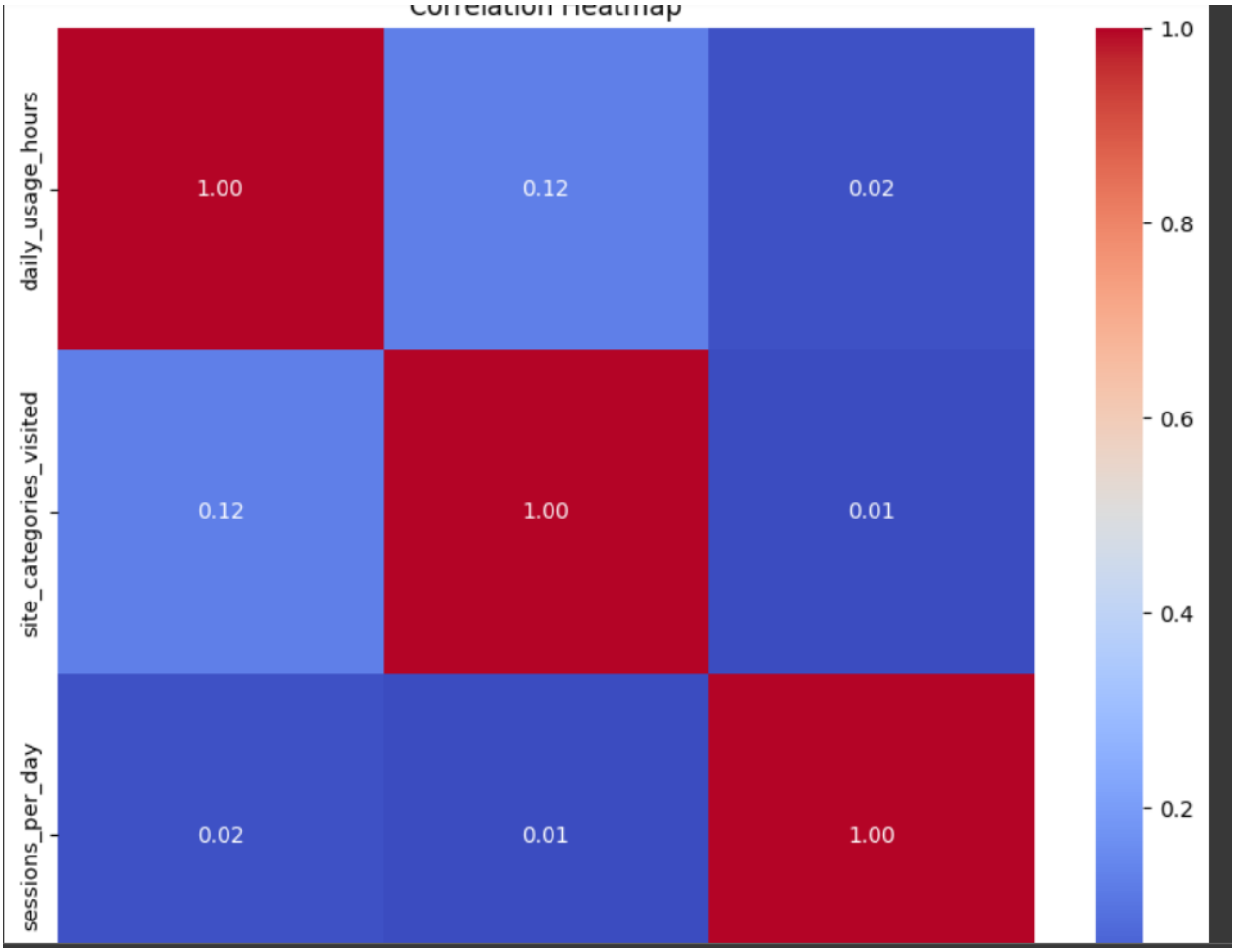
```
    plt.xlabel('Year')
    plt.ylabel('Internet Users')
    plt.grid(True)
    plt.show()
else:
    print("\nColumns 'Year' and/or 'Internet Users' not found for time series plot.")
```

## Output/Result:

```
First 5 rows of the dataset:
   daily_usage_hours  site_categories_visited  sessions_per_day
0           9.884957                        2                13
1           1.023220                        9                 1
2          10.394205                        9                 3
3           5.990237                        6                16
4           3.558451                        4                 4

Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 3 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   daily_usage_hours        100 non-null    float64
 1   site_categories_visited  100 non-null    int64
 2   sessions_per_day         100 non-null    int64
dtypes: float64(1), int64(2)
memory usage: 2.5 KB
None

Summary Statistics:
       daily_usage_hours  site_categories_visited  sessions_per_day
count         100.000000               100.000000        100.000000
mean            6.298375                 5.100000         10.870000
std             3.448911                 2.653376          5.799086
min             0.143016                 1.000000          1.000000
25%             3.494349                 3.000000          5.000000
50%             6.169502                 5.000000         11.500000
75%             9.069780                 7.000000         16.000000
max            11.988594                 9.000000         19.000000

Missing Values:
daily_usage_hours          0
site_categories_visited    0
```

```
Missing Values:
daily_usage_hours          0
site_categories_visited    0
sessions_per_day           0
dtype: int64
```



## 5. References/Credits:

### 1. Dataset Source:

https://archive.ics.uci.edu/dataset/126/internet+usage+data

### 2. Libraries Used:

- **Pandas**
  Used for data manipulation and analysis, including reading the CSV file, preprocessing data, and exploring features

- **NumPy**
  Supports numerical operations, which are essential for certain calculations and transformations.

- **Scikit-learn**
  Includes tools for machine learning models like Logistic Regression, as well as clustering algorithms like KMeans. It also provides utilities for splitting data, evaluating models (accuracy, precision, recall), and generating confusion matrices.

- **Seaborn**
  Facilitates advanced data visualization, particularly for creating heatmaps of confusion matrices and scatterplots.

- **Matplotlib**
  Used for plotting graphs and visualizing clusters or metrics in combination with Seaborn.