Documentation for Web Data Analysis

Somesh Kumar

1. Introduction

This project analyzes web traffic data from a website. The goal is to explore various metrics such as visits, page views, time spent on pages, bounces, and exits. The analysis includes statistical summaries, correlation analysis, and visualizations like scatter plots, bar plots, and heatmaps to help understand how different web metrics are related to each other.

---

2. Data Loading and Initial Exploration

The first step in the analysis is to load the dataset. The dataset is stored as a CSV file, which is read into a DataFrame using the **Pandas** library. Once the data is loaded, we inspect the first few rows to understand the structure of the dataset (e.g., column names, data types, and the first few entries).

We also check the data types of each column to ensure that they match the expected type (e.g., numeric, string). For example, we expect columns like `Visits`, `Bounces`, and `Timeinpage` to be numeric, while columns like `Sourcegroup` and `Continent` might be categorical.

Additionally, we check for any missing values (`NaN`) in the dataset. This is important because missing values can affect the accuracy of our analysis. If missing data is found, it will need to be addressed before performing further analysis.

---

3. Descriptive Statistics

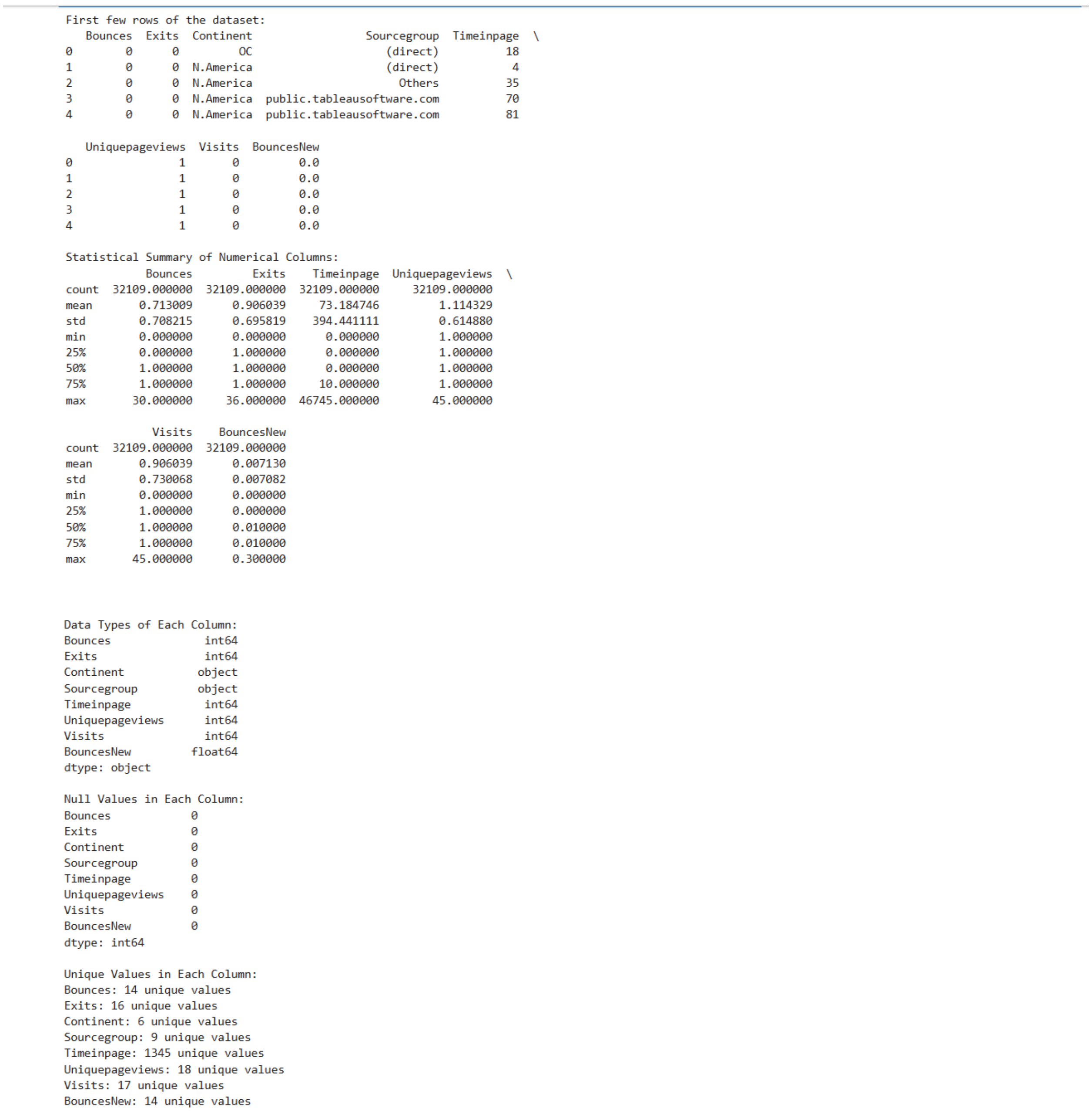
To get a better understanding of the data, we calculate summary statistics for all numerical columns. This includes:

- **Mean**: The average value of each numeric column.
- **Median**: The middle value of each numeric column when sorted.
- **Standard Deviation**: A measure of how spread out the values are around the mean.
- **Minimum and Maximum**: The smallest and largest values in each numeric column.

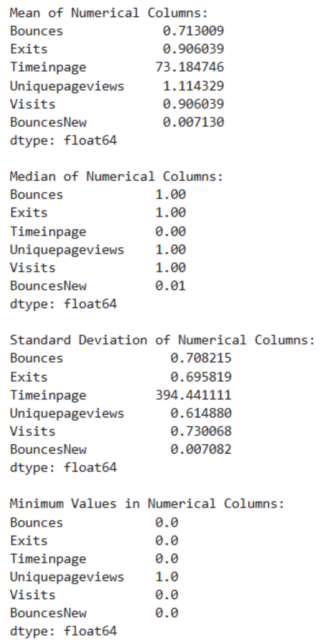
We also calculate **skewness**, which tells us if the data is symmetrically distributed or if it's skewed in one direction (positive skew means data is skewed right, negative skew means data is skewed left).

These statistical summaries provide a foundation for understanding the general behavior of each variable.

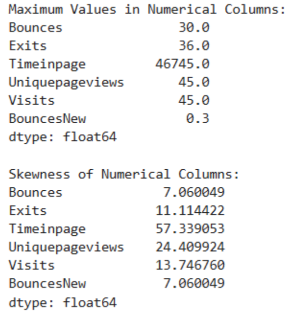
Result



```
First few rows of the dataset:
   Bounces  Exits  Continent                Sourcegroup  Timeinpage  \
0        0      0         OC                   (direct)          18
1        0      0  N.America                   (direct)           4
2        0      0  N.America                     Others          35
3        0      0  N.America  public.tableausoftware.com          70
4        0      0  N.America  public.tableausoftware.com          81

   Uniquepageviews  Visits  BouncesNew
0                1       0         0.0
1                1       0         0.0
2                1       0         0.0
3                1       0         0.0
4                1       0         0.0

Statistical Summary of Numerical Columns:
            Bounces         Exits    Timeinpage  Uniquepageviews  \
count  32109.000000  32109.000000  32109.000000     32109.000000
mean       0.713009      0.906039     73.184746         1.114329
std        0.708215      0.695819    394.441111         0.614880
min        0.000000      0.000000      0.000000         1.000000
25%        0.000000      1.000000      0.000000         1.000000
50%        1.000000      1.000000      0.000000         1.000000
75%        1.000000      1.000000     10.000000         1.000000
max       30.000000     36.000000  46745.000000        45.000000

             Visits    BouncesNew
count  32109.000000  32109.000000
mean       0.906039      0.007130
std        0.730068      0.007082
min        0.000000      0.000000
25%        1.000000      0.000000
50%        1.000000      0.010000
75%        1.000000      0.010000
max       45.000000      0.300000



Data Types of Each Column:
Bounces              int64
Exits                int64
Continent           object
Sourcegroup         object
Timeinpage           int64
Uniquepageviews      int64
Visits               int64
BouncesNew         float64
dtype: object

Null Values in Each Column:
Bounces            0
Exits              0
Continent          0
Sourcegroup        0
Timeinpage         0
Uniquepageviews    0
Visits             0
BouncesNew         0
dtype: int64

Unique Values in Each Column:
Bounces: 14 unique values
Exits: 16 unique values
Continent: 6 unique values
Sourcegroup: 9 unique values
Timeinpage: 1345 unique values
Uniquepageviews: 18 unique values
Visits: 17 unique values
BouncesNew: 14 unique values
```



```
Mean of Numerical Columns:
Bounces             0.713009
Exits               0.906039
Timeinpage         73.184746
Uniquepageviews     1.114329
Visits              0.906039
BouncesNew          0.007130
dtype: float64

Median of Numerical Columns:
Bounces            1.00
Exits              1.00
Timeinpage         0.00
Uniquepageviews    1.00
Visits             1.00
BouncesNew         0.01
dtype: float64

Standard Deviation of Numerical Columns:
Bounces              0.708215
Exits                0.695819
Timeinpage         394.441111
Uniquepageviews      0.614880
Visits               0.730068
BouncesNew           0.007082
dtype: float64

Minimum Values in Numerical Columns:
Bounces            0.0
Exits              0.0
Timeinpage         0.0
Uniquepageviews    1.0
Visits             0.0
BouncesNew         0.0
dtype: float64
```



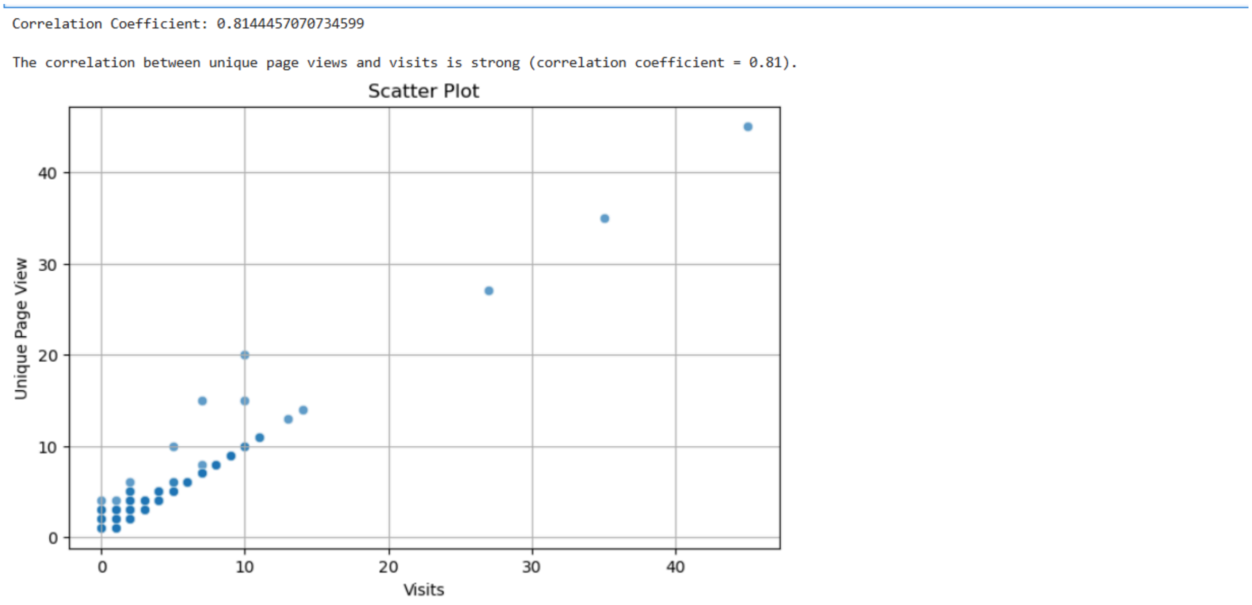
```
Maximum Values in Numerical Columns:
Bounces               30.0
Exits                 36.0
Timeinpage         46745.0
Uniquepageviews       45.0
Visits                45.0
BouncesNew             0.3
dtype: float64

Skewness of Numerical Columns:
Bounces             7.060049
Exits              11.114422
Timeinpage         57.339053
Uniquepageviews    24.409924
Visits             13.746760
BouncesNew          7.060049
dtype: float64
```

4. Correlation Analysis

Correlation analysis helps us understand how strongly two or more variables are related. In this case, we look at the relationship between `Uniquepageviews` and `Visits`. The **correlation coefficient** tells us if these variables are strongly, moderately, or weakly related. A positive correlation means that as one variable increases, the other also increases. A negative correlation means that as one variable increases, the other decreases.

We visualize this relationship through a **scatter plot**, which displays the data points for `Visits` and `Uniquepageviews` on a two-dimensional plane. This helps in identifying any linear relationship between the variables.

Result



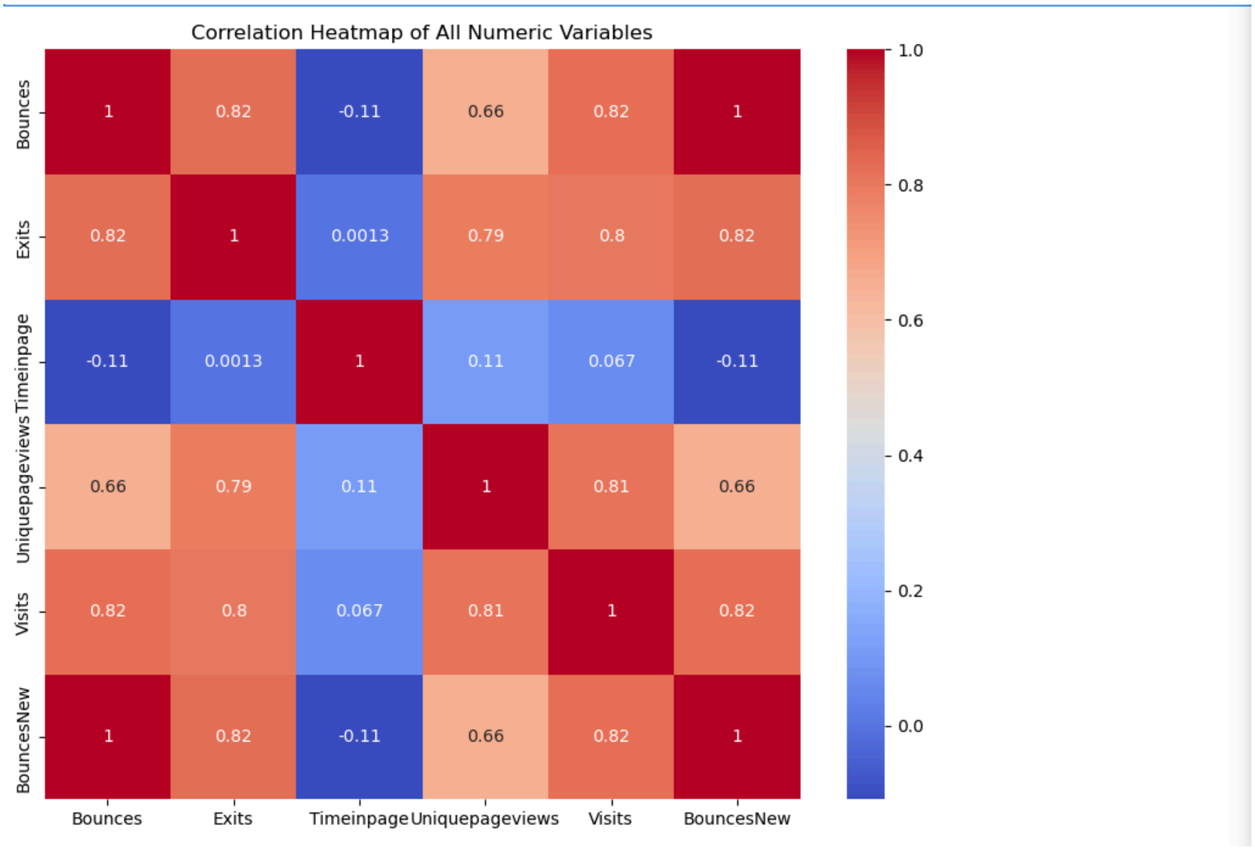
```
Correlation Coefficient: 0.8144457070734599

The correlation between unique page views and visits is strong (correlation coefficient = 0.81).
```

5. Heatmap for Correlation Between All Numeric Variables

A **correlation heatmap** is generated to show the correlation between all numeric columns in the dataset. Each cell in the heatmap represents the correlation between two variables, with colors indicating the strength of the correlation (red for high positive correlation, blue for high negative correlation, and white for no correlation).

This heatmap provides a visual summary of how different variables interact with each other. For instance, if `Timeinpage` and `Exits` have a strong negative correlation, it might suggest that users who spend more time on the page are less likely to exit quickly.

Result



	Bounces	Exits	Timeinpage	Uniquepageviews	Visits	BouncesNew
Bounces	1	0.82	-0.11	0.66	0.82	1
Exits	0.82	1	0.0013	0.79	0.8	0.82
Timeinpage	-0.11	0.0013	1	0.11	0.067	-0.11
Uniquepageviews	0.66	0.79	0.11	1	0.81	0.66
Visits	0.82	0.8	0.067	0.81	1	0.82
BouncesNew	1	0.82	-0.11	0.66	0.82	1

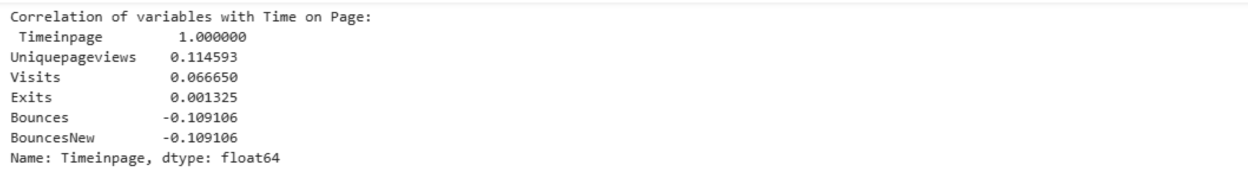
6. Time on Page Analysis

Next, we analyze the `Timeinpage` column to understand how it correlates with other variables. First, we convert the `Timeinpage` column to numeric values (in case it contains any non-numeric entries). After conversion, we check the correlation of `Timeinpage` with other variables in the dataset, such as `Visits`, `Exits`, `Bounces`, and others.

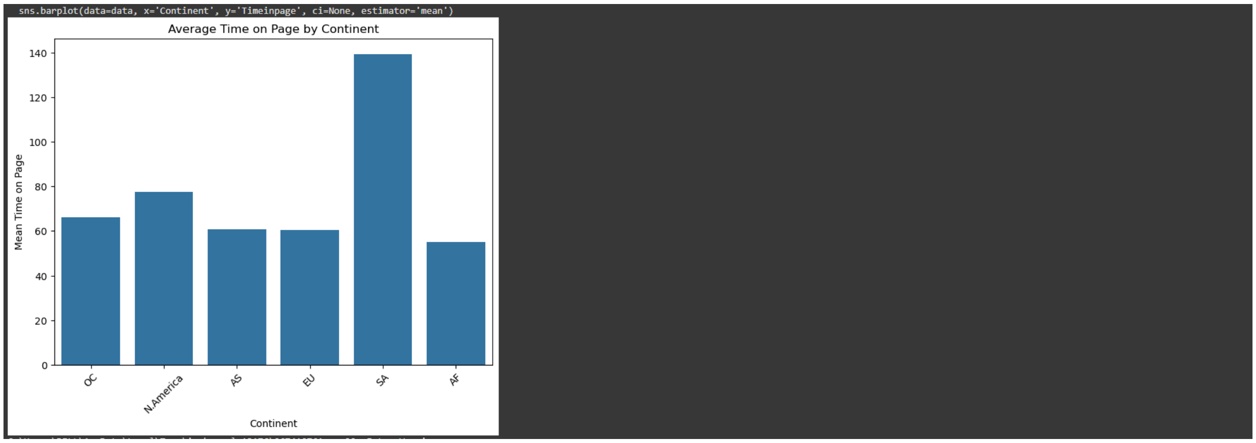
We also examine how the average `Timeinpage` varies across different categories, such as by **Continent** or **Sourcegroup**, by creating **bar plots**. These plots show the mean time spent on pages for each category, which can help identify geographic or source-related trends.

Additionally, we create scatter plots to examine the relationship between `Timeinpage` and other continuous variables (like `Visits` and `Exits`). This allows us to visually explore whether there’s a pattern between how much time a user spends on a page and their likelihood of exiting or how many pages they visit.

Result

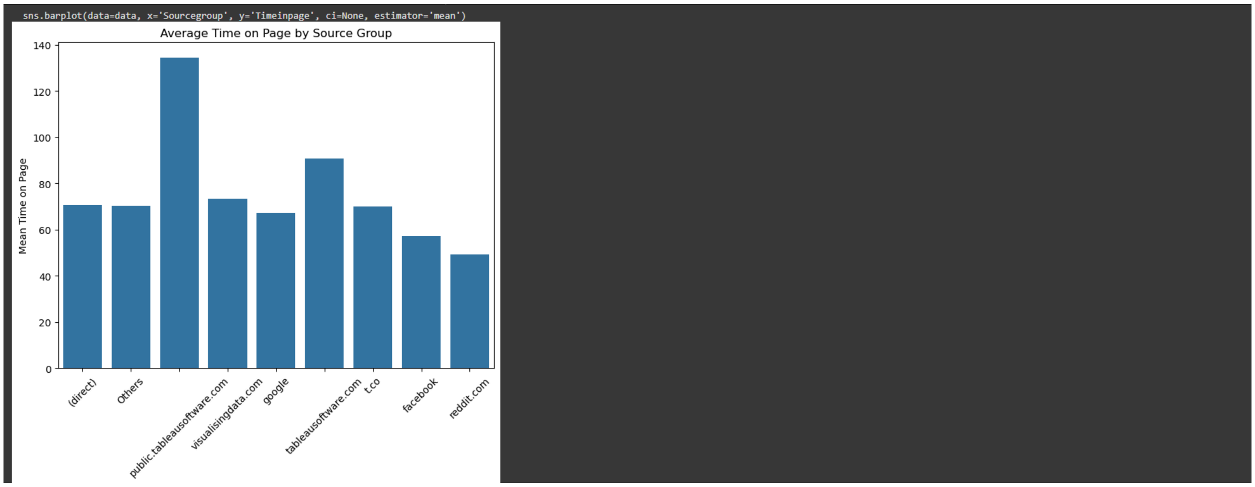


```
Correlation of variables with Time on Page:
 Timeinpage          1.000000
Uniquepageviews     0.114593
Visits              0.066650
Exits               0.001325
Bounces            -0.109106
BouncesNew         -0.109106
Name: Timeinpage, dtype: float64
```



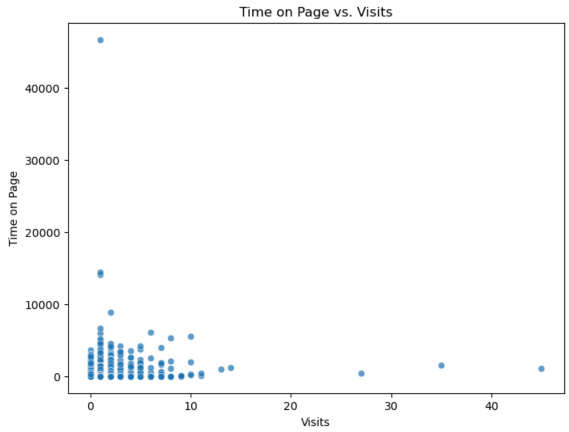
```
sns.barplot(data=data, x='Continent', y='Timeinpage', ci=None, estimator='mean')
```

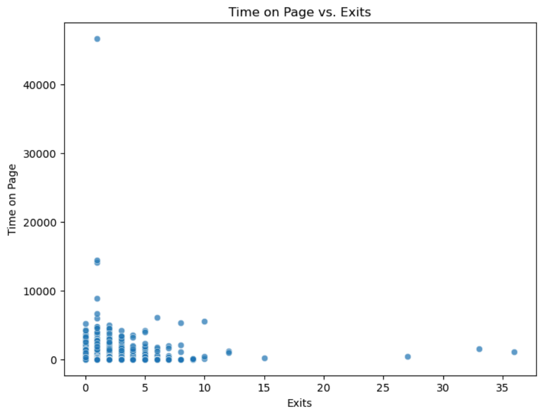
Average Time on Page by Continent



```
sns.barplot(data=data, x='Sourcegroup', y='Timeinpage', ci=None, estimator='mean')
```

Average Time on Page by Source Group



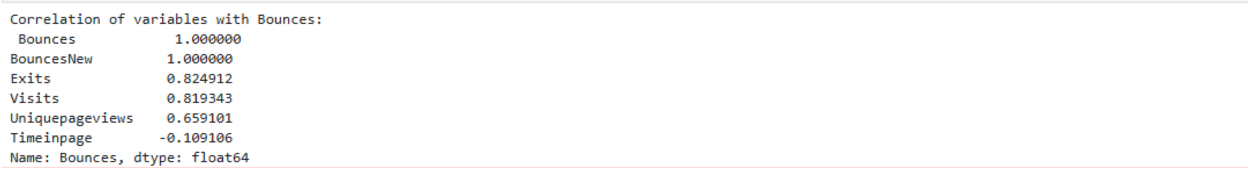


7. Bounces Analysis

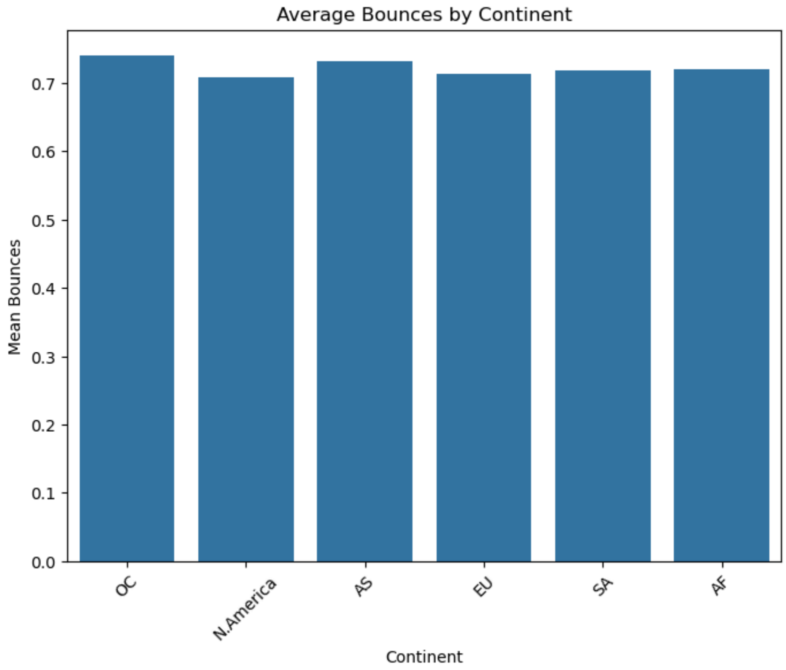
The `Bounces` column refers to users who land on a page and leave without interacting further. We analyze how `Bounces` correlate with other variables, such as `Timeinpage` and `Visits`. A positive correlation between `Bounces` and `Exits` might suggest that users who exit the page are more likely to bounce.

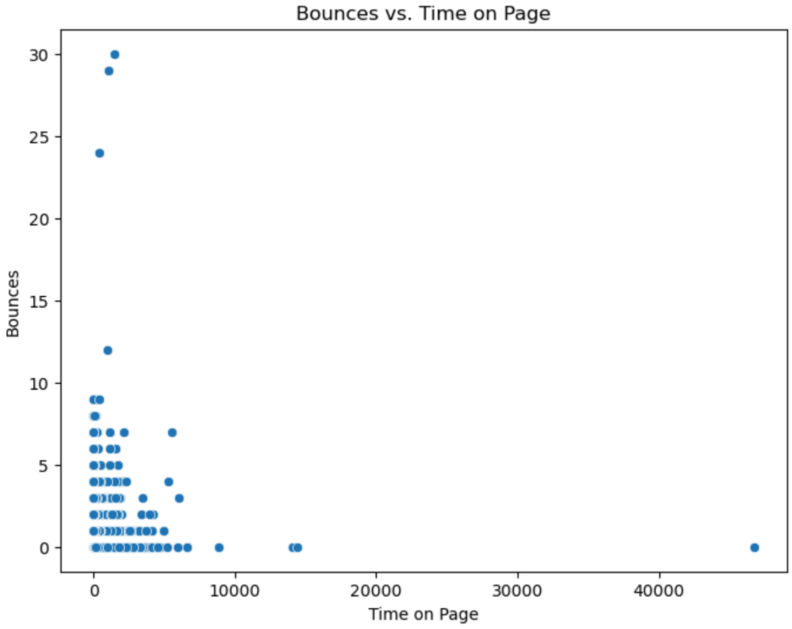
We visualize this relationship with **bar plots** (showing average bounces by `Continent` and `Sourcegroup`) and a **scatter plot** (showing bounces vs. time spent on the page). These visualizations help understand if there are specific regions or sources where users tend to bounce more frequently.

Result



```
Correlation of variables with Bounces:
 Bounces            1.000000
BouncesNew         1.000000
Exits              0.824912
Visits             0.819343
Uniquepageviews    0.659101
Timeinpage        -0.109106
Name: Bounces, dtype: float64
```





8. Conclusion

This analysis provides a deep dive into the behavior of website visitors. By calculating correlations, examining descriptive statistics, and visualizing relationships between key metrics, we can gain insights into how different aspects of web traffic are connected.

For example, understanding the correlation between `Visits` and `Uniquepageviews` can help us determine how traffic patterns relate to user engagement. Similarly, by analyzing `Timeinpage`, `Bounces`, and `Exits`, we can uncover patterns related to user retention and abandonment.

These insights can guide website optimizations and inform business decisions on where to focus marketing or improve user experience.