PYTHON - CODING CHALLENGE
Web Data Analysis

–Nidya Thirshala M

The coding file is uploaded as:Python Coding Challenge.ipynb

Domain: Web

Dataset Description:
The variables in the dataset are defined here for better understanding:

Attribute	Description
Bounces	It represents the percentage of visitors who enter the site and "bounce" (leave the site) rather than continuing to view other pages within the same site.
Continent	It shows the continent from which the site has been accessed.
Source group	It shows how the visitor has accessed the site.
Time on page	It shows how long the user has spent on that particular page of the website.
Unique pageview	It represents the number of sessions during which that page was viewed one or more times.
Visits	A visit counts all visitors, no matter how many times the same visitor may have been to your site.

Tasks:

1.Provide Basic Summarization about the Dataset

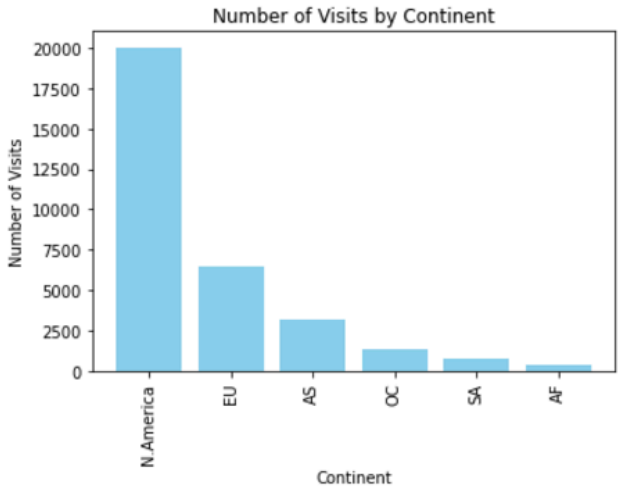
Data summarization can be done by both pandas and matplotlib.Matplotlib gives visualized data which will be easy to understand.

```python
#Using Pandas
import pandas as pd
data = pd.read_csv("project.xlsx - InternetCaseStudy.csv")
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32109 entries, 0 to 32108
Data columns (total 8 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   Bounces          32109 non-null  int64  
 1   Exits            32109 non-null  int64  
 2   Continent        32109 non-null  object 
 3   Sourcegroup      32109 non-null  object 
 4   Timeinpage       32109 non-null  int64  
 5   Uniquepageviews  32109 non-null  int64  
 6   Visits           32109 non-null  int64  
 7   BouncesNew       32109 non-null  float64
dtypes: float64(1), int64(5), object(2)
memory usage: 2.0+ MB
```

```python
#using matplotlib
import matplotlib.pyplot as plt
continent_counts = data['Continent'].value_counts()

plt.bar(continent_counts.index, continent_counts.values, color='skyblue')
plt.xlabel('Continent')
plt.ylabel('Number of Visits')
plt.title('Number of Visits by Continent')
plt.xticks(rotation=90)
plt.show()
```



I have attached only two samples of the given data .The visualized data is called Bar plot.

2.The team needs to know whether the unique page view value depends on visits.

First to find the dependency between visits and unique page views.We must know about their actual definition.

Visits: Counts all instances of page visits, regardless of whether they are from the same session or the same user.

Unique pageview :If a user visits the same page multiple times within a single session, it is counted as one unique page view for that session.

The actual question asked:Does the **unique page view value depend** on the **visits**

```
#using pandas
# Calculate correlation between 'Visits' and 'Unique pageview'
correlation = data['Visits'].corr(data['Uniquepageviews'])
print(f"Correlation between Visits and Unique Pageview: {correlation}")
```

```
Correlation between Visits and Unique Pageview: 0.8144457070734604
```

Here those two variables has perfect positive correlation as it is closest to +1

Correlation: Correlation is a statistical concept that describes the relationship or association between two variables.

Types of Correlation:

- **Positive correlation**: When one variable increases, the other also tends to increase.
- **Negative correlation**: When one variable increases, the other tends to decrease.
- **No correlation**: No apparent relationship between the two variables.

Range: The correlation coefficient typically ranges from **-1 to +1**:

- **+1**: Perfect positive correlation (both variables increase together).
- **0**: No correlation (no linear relationship).

- **-1**: Perfect negative correlation (one variable increases as the other decreases).

3.Find out the probable factors from the dataset, which could affect the exits.

Exits: Exits refer to when a user leaves a website after viewing one or more pages, but not necessarily the first page they visited. It's the last page a user views during their session before leaving the site.

There will definitely be a reason why the user is exiting the website so to find the possible factors that could be affecting we can use **Correlation.**

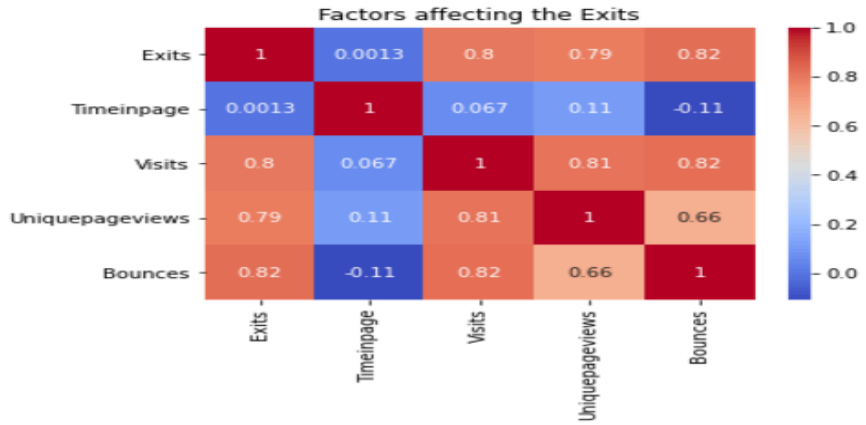
```
# Correlation analysis between Exits and other variables
correlation_matrix = data[['Exits','Timeinpage', 'Visits', 'Uniquepageviews', 'Bounces']].corr()
print(correlation_matrix)

for column in correlation_matrix.columns:
    if column != 'Exits':
        if correlation_matrix.loc['Exits', column] >0.5  and correlation_matrix.loc['Exits', column]<=1 :
            print(f"Exits has a strong positive relationship with: {column}")
```

```
                   Exits  Timeinpage    Visits  Uniquepageviews   Bounces
Exits           1.000000    0.001325  0.800979         0.791129  0.824912
Timeinpage      0.001325    1.000000  0.066650         0.114593 -0.109106
Visits          0.800979    0.066650  1.000000         0.814446  0.819343
Uniquepageviews 0.791129    0.114593  0.814446         1.000000  0.659101
Bounces         0.824912   -0.109106  0.819343         0.659101  1.000000
Exits has a strong positive relationship with: Visits
Exits has a strong positive relationship with: Uniquepageviews
Exits has a strong positive relationship with: Bounces
```

Here I have created a if statement to check for the strong positive correlation of some variables that affects the exiting status

Heatmap:



A heatmap is a **data visualization** technique that uses color to represent the values of a matrix or table. It's particularly useful for visualizing **correlations.**

4. Find the variables which possibly have an effect on the time on page.

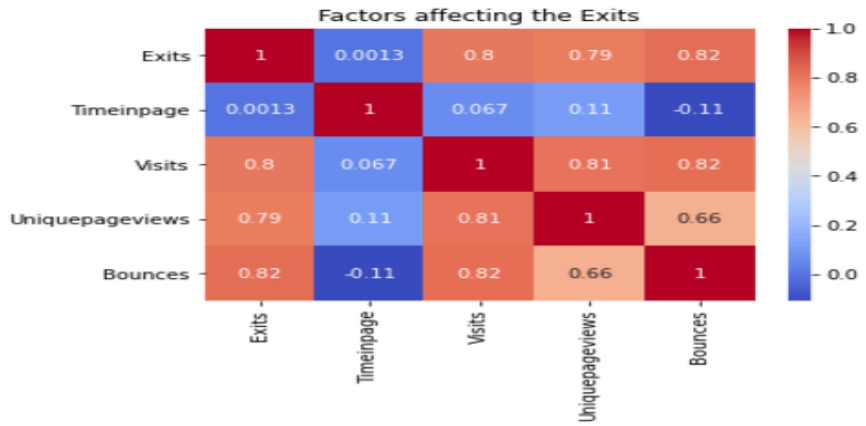
The time utilized for a user on using a website is relatively important on the content engagement of the page.

```
correlation_matrix = data[['Exits','Timeinpage', 'Visits', 'Uniquepageviews', 'Bounces']].corr()
print(correlation_matrix)

for column in correlation_matrix.columns:
    if column != 'Timeinpage':
        if correlation_matrix.loc['Timeinpage', column] >0.6:
            print(f"Timeinpage has a strong positive relationship with: {column}")
    else:
            print("Nothing is much affecting the time on page")
```

```
                   Exits  Timeinpage    Visits  Uniquepageviews   Bounces
Exits           1.000000    0.001325  0.800979         0.791129  0.824912
Timeinpage      0.001325    1.000000  0.066650         0.114593 -0.109106
Visits          0.800979    0.066650  1.000000         0.814446  0.819343
Uniquepageviews 0.791129    0.114593  0.814446         1.000000  0.659101
Bounces         0.824912   -0.109106  0.819343         0.659101  1.000000
Nothing is much affecting the time on page
```

When we notice the **timeinpage** row there is not much value**>0.5** which means it is a negligible correlation between all the variables.



When we see the correlation map the **timeonpage** column has only the **cool tones(blue)** which denotes **lesser impact.**

5.Help the team in determining the factors that are impacting the bounce.

A high bounce rate typically indicates that visitors are leaving the website quickly without interacting with additional pages.

```
# Calculate the correlation matrix
correlation_matrix = data[['Exits', 'Timeinpage', 'Visits', 'Uniquepageviews', 'Bounces']].corr()
print(correlation_matrix)

for column in correlation_matrix.columns:
    if column != 'Bounces':
        correlation_value = correlation_matrix.loc['Bounces', column]
        if correlation_value > 0.6:
            print(f"Bounces has a strong positive relationship with: {column}")
```

```
                   Exits  Timeinpage    Visits  Uniquepageviews   Bounces
Exits           1.000000    0.001325  0.800979         0.791129  0.824912
Timeinpage      0.001325    1.000000  0.066650         0.114593 -0.109106
Visits          0.800979    0.066650  1.000000         0.814446  0.819343
Uniquepageviews 0.791129    0.114593  0.814446         1.000000  0.659101
Bounces         0.824912   -0.109106  0.819343         0.659101  1.000000
Bounces has a strong positive relationship with: Exits
Bounces has a strong positive relationship with: Visits
Bounces has a strong positive relationship with: Uniquepageviews
```

The bounces have a strong relationship between three variables which are **Exits,Visits,Unique Pageviews.**