Coding Challenge

Read the data from excel sheet

```
import pandas as pd
```

```
data= pd.read_excel('C:\\Users\\chand\\Downloads\\WebdataAnalysis.xlsx')
data
```

	Bounces	Exits	Continent	Sourcegroup	Timeinpage	Uniquepageviews	Visits	BouncesNew
0	0	0	OC	(direct)	18	1	0	0.00
1	0	0	N.America	(direct)	4	1	0	0.00
2	0	0	N.America	Others	35	1	0	0.00
3	0	0	N.America	public.tableausoftware.com	70	1	0	0.00
4	0	0	N.America	public.tableausoftware.com	81	1	0	0.00
...	...	...	...	...	...	...	...	...
32104	1	1	N.America	public.tableausoftware.com	12	2	2	0.01
32105	2	2	N.America	(direct)	0	2	2	0.02
32106	2	2	N.America	(direct)	0	2	2	0.02
32107	2	2	N.America	(direct)	0	2	2	0.02
32108	2	2	N.America	google	0	2	2	0.02

32109 rows × 8 columns

The team wants to analyze each variable of the data collected through data summarization to get a basic understanding of the dataset and to prepare for further analysis.

```
data.describe()
```

	Bounces	Exits	Timeinpage	Uniquepageviews	Visits	BouncesNew
count	32109.000000	32109.000000	32109.000000	32109.000000	32109.000000	32109.000000
mean	0.713009	0.906039	73.184746	1.114329	0.906039	0.007130
std	0.708215	0.695819	394.441111	0.614880	0.730068	0.007082
min	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000
25%	0.000000	1.000000	0.000000	1.000000	1.000000	0.000000
50%	1.000000	1.000000	0.000000	1.000000	1.000000	0.010000
75%	1.000000	1.000000	10.000000	1.000000	1.000000	0.010000
max	30.000000	36.000000	46745.000000	45.000000	45.000000	0.300000

```
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
```

- Describe() method gives the basic information like mean, maximum, mimimun, count, standard deviation of each column
- Info() method give information about datypes of the column

As mentioned earlier, a unique page view represents the number of sessions during which that page was viewed one or more times. A visit counts all instances, no matter how many times the same visitor may have been to your site. So the team needs to know whether the unique page view value depends on visits.

```
#coefficient between Uniquepageviews and Visits
corr_coef = data['Uniquepageviews'].corr(data['Visits'])

print(corr_coef)
```

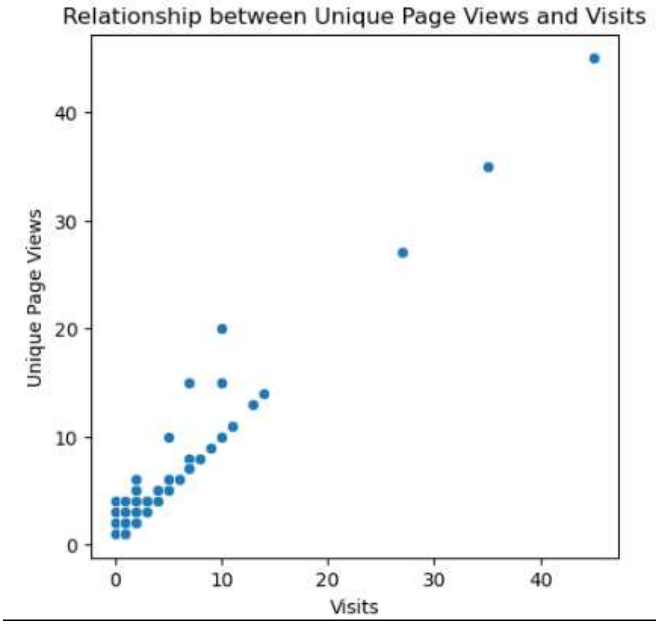
```
0.8144457070734599
```

Here value is almost nearer to 1. So uniquepageviews highly depends on visits.

To represent visually we have used scatterplot here

A scatter plot is a type of data visualization that displays the relationship between two continuous variables. It is a useful tool for identifying patterns, trends, and correlations in data.

```
# Create a scatter plot to visualize the relationship
import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(5, 5))
sns.scatterplot(x='Visits', y='Uniquepageviews', data=data)
plt.xlabel('Visits')
plt.ylabel('Unique Page Views')
plt.title('Relationship between Unique Page Views and Visits')
plt.show()
```



Correlation matrix:

- It gives the dependencies of the column with the another column

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

	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

Find out the probable factors from the dataset, which could affect the exits. Exit Page Analysis is usually required to get an idea about why a user leaves the website for a session and moves on to another one. Please keep in mind that exits should not be confused with bounces.

Here we need to find the probable factors that depend on the Exits.

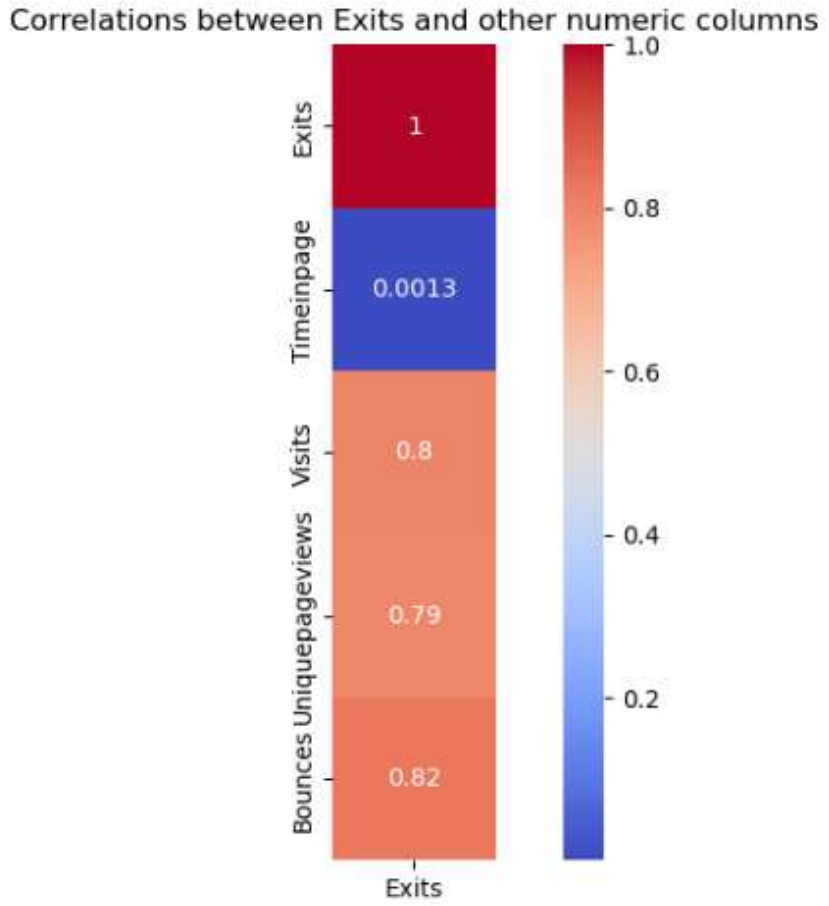
From the below data we can understand that visits, uniquepageviews, bounces depends on the exits

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

```
Exits has a strong positive relationship with: Visits
Exits has a strong positive relationship with: Uniquepageviews
Exits has a strong positive relationship with: Bounces
```

Visual representation using heat map:

```
plt.figure(figsize=(10, 6))
sns.heatmap(Exits.to_frame(), annot=True, cmap='coolwarm', square=True)
plt.title('Correlations between Exits and other numeric columns')
plt.show()
```



Every site wants to increase the time on page for a visitor. This increases the chances of the visitor understanding the site content better and hence there are more chances of a transaction taking place. Find the variables which possibly have an effect on the time on page.

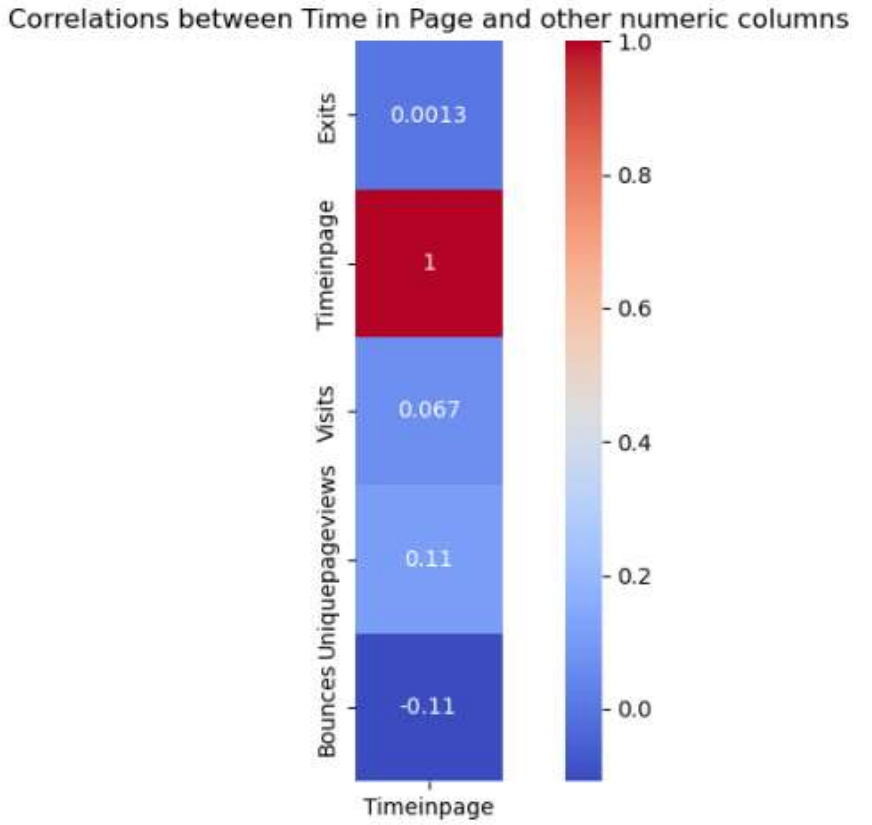
Here we need to find the probable factors that depend on the Time in page.

```
correlation_matrix = data[['Exits','Timeinpage', 'Visits', 'Uniquepageviews', 'Bounces']].corr()
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 affecting Time in page")
```

```
Nothing is affecting Time in page
```

Visual representation using heat map:

```
plt.figure(figsize=(10, 6))
sns.heatmap(Time_in_Page.to_frame(), annot=True, cmap='coolwarm', square=True)
plt.title('Correlations between Time in Page and other numeric columns')
plt.show()
```



A high bounce rate is a cause of alarm for websites which depend on visitor engagement. Help the team in determining the factors that are impacting the bounce.

Here we need to find the probable factors that depend on the Bounces.

From the below data we can understand that exits, visits, uniquepageviews depends on the Bounces

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

```
Bounces has a strong positive relationship with: Exits
Bounces has a strong positive relationship with: Visits
Bounces has a strong positive relationship with: Uniquepageviews
```

Visual representation using heat map:

```
plt.figure(figsize=(10, 6))
sns.heatmap(Bounce.to_frame(), annot=True, cmap='coolwarm', square=True)
plt.title('Correlations between Bounces and other numeric columns')
plt.show()
```