shopping-dataset-analysis

September 3, 2024

[1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

[2]:

```
# 1. Basic data cleaning and exploration:
# Loading Data from google drive.

df1=pd.read_csv("/content/drive/MyDrive/Data sets/shopping.csv")
```

[2]:

```

```

[3]:

```
df1.head()
```

[3]:

```
   Administrative  Administrative_Duration  Informational  \
0               0                      0.0              0
1               0                      0.0              0
2               0                      0.0              0
3               0                      0.0              0
4               0                      0.0              0

   Informational_Duration  ProductRelated  ProductRelated_Duration  \
0                     0.0               1                 0.000000
1                     0.0               2                64.000000
2                     0.0               1                 0.000000
3                     0.0               2                 2.666667
4                     0.0              10               627.500000

   BounceRates  ExitRates  PageValues  SpecialDay Month  OperatingSystems  \
0         0.20       0.20         0.0         0.0   Feb                 1
1         0.00       0.10         0.0         0.0   Feb                 2
2         0.20       0.20         0.0         0.0   Feb                 4
3         0.05       0.14         0.0         0.0   Feb                 3
4         0.02       0.05         0.0         0.0   Feb                 3

   Browser  Region  TrafficType        VisitorType  Weekend  Revenue
0        1       1            1  Returning_Visitor    False    False
```

```
1         2        1             2  Returning_Visitor    False    False
2         1        9             3  Returning_Visitor    False    False
3         2        2             4  Returning_Visitor    False    False
4         3        1             4  Returning_Visitor     True    False
```

[4]:
```
df1.describe()
```

[4]:
```
       Administrative  Administrative_Duration  Informational  \
count    12330.000000             12330.000000   12330.000000   
mean         2.315166                80.818611       0.503569   
std          3.321784               176.779107       1.270156   
min          0.000000                 0.000000       0.000000   
25%          0.000000                 0.000000       0.000000   
50%          1.000000                 7.500000       0.000000   
75%          4.000000                93.256250       0.000000   
max         27.000000              3398.750000      24.000000   

       Informational_Duration  ProductRelated  ProductRelated_Duration  \
count            12330.000000    12330.000000             12330.000000   
mean                34.472398       31.731468              1194.746220   
std                140.749294       44.475503              1913.669288   
min                  0.000000        0.000000                 0.000000   
25%                  0.000000        7.000000               184.137500   
50%                  0.000000       18.000000               598.936905   
75%                  0.000000       38.000000              1464.157214   
max               2549.375000      705.000000             63973.522230   

        BounceRates     ExitRates    PageValues    SpecialDay  \
count  12330.000000  12330.000000  12330.000000  12330.000000   
mean       0.022191      0.043073      5.889258      0.061427   
std        0.048488      0.048597     18.568437      0.198917   
min        0.000000      0.000000      0.000000      0.000000   
25%        0.000000      0.014286      0.000000      0.000000   
50%        0.003112      0.025156      0.000000      0.000000   
75%        0.016813      0.050000      0.000000      0.000000   
max        0.200000      0.200000    361.763742      1.000000   

       OperatingSystems       Browser        Region   TrafficType  
count      12330.000000  12330.000000  12330.000000  12330.000000  
mean           2.124006      2.357097      3.147364      4.069586  
std            0.911325      1.717277      2.401591      4.025169  
min            1.000000      1.000000      1.000000      1.000000  
25%            2.000000      2.000000      1.000000      2.000000  
50%            2.000000      2.000000      3.000000      2.000000  
75%            3.000000      2.000000      4.000000      4.000000  
max            8.000000     13.000000      9.000000     20.000000  
```

[5]:
```
df1.shape
```

[5]:
```
(12330, 18)
```

[6]:
```
df1.isnull().sum()
```

[6]:
```
Administrative             0
Administrative_Duration    0
Informational              0
Informational_Duration     0
ProductRelated             0
ProductRelated_Duration    0
BounceRates                0
ExitRates                  0
PageValues                 0
SpecialDay                 0
Month                      0
OperatingSystems           0
Browser                    0
Region                     0
TrafficType                0
VisitorType                0
Weekend                    0
Revenue                    0
dtype: int64
```

[7]:
```
df1.duplicated().sum()
```

[7]:
```
125
```

[8]:
```
df1.sample(10)
```

[8]:
```
       Administrative  Administrative_Duration  Informational  \
1129                1                 4.000000              0
4320                6               200.333333              0
10764               5               107.125000              0
6847                0                 0.000000              0
6950                9               124.426667              0
1135                0                 0.000000              0
8466                1                27.900000              0
2290                2                29.333333              0
10087               9               181.275000              2
4447                0                 0.000000              0

       Informational_Duration  ProductRelated  ProductRelated_Duration  \
1129                      0.0              37              1096.750000
4320                      0.0               9               154.000000
```

```
10764                    0.0              34           916.000000
6847                     0.0              24           985.000000
6950                     0.0              19           720.676667
1135                     0.0               2            36.500000
8466                     0.0              64          3875.427778
2290                     0.0              98          2680.828571
10087                   65.0             120          3283.715359
4447                     0.0               8           459.000000

       BounceRates  ExitRates  PageValues  SpecialDay Month  OperatingSystems  \
1129      0.000000   0.020175    0.000000         0.0   Mar                 2
4320      0.000000   0.030769    0.000000         0.0   May                 1
10764     0.000000   0.002857    0.000000         0.0   Dec                 2
6847      0.070833   0.087500    0.000000         0.0  June                 3
6950      0.000000   0.018933    0.000000         0.0   Oct                 2
1135      0.000000   0.033333    0.000000         0.0   Mar                 3
8466      0.000000   0.019583    5.698585         0.0   Nov                 2
2290      0.009184   0.014160   16.048607         0.0   May                 3
10087     0.004651   0.024543    2.442153         0.0   Nov                 1
4447      0.000000   0.028571   26.980000         0.0   May                 2

       Browser  Region  TrafficType        VisitorType  Weekend  Revenue
1129         2       6           10  Returning_Visitor    False    False
4320         1       1            1  Returning_Visitor    False    False
10764        2       2            1  Returning_Visitor    False    False
6847         2       6           13  Returning_Visitor    False    False
6950         2       7            4  Returning_Visitor    False    False
1135         2       1            1  Returning_Visitor    False    False
8466         2       1            2  Returning_Visitor    False    False
2290         2       3           13  Returning_Visitor    False     True
10087        2       1            2  Returning_Visitor    False    False
4447         4       3            2  Returning_Visitor    False     True
```

[9]:

```
class_distribution = df1['Revenue'].value_counts()
print(class_distribution)
```

```
Revenue
False    10422
True      1908
Name: count, dtype: int64
```

[10]:

```
class_distribution = df1['Revenue'].value_counts(normalize=True) * 100
print(class_distribution)
```

```
Revenue
False    84.525547
True     15.474453
Name: proportion, dtype: float64
```

[11]:
```
page_cnt=df1.
 ↪groupby('Revenue')[['Administrative','Informational','ProductRelated']].
 ↪mean().reset_index()
page_cnt.loc[page_cnt['Revenue'] == True,'Revenue'] = 'Buyers'
page_cnt.loc[page_cnt['Revenue'] == False,'Revenue'] = 'Non-Buyers'
page_cnt
```

```
<ipython-input-11-fdd6e0a45714>:2: FutureWarning: Setting an item of
incompatible dtype is deprecated and will raise in a future error of pandas.
Value 'Buyers' has dtype incompatible with bool, please explicitly cast to a
compatible dtype first.
  page_cnt.loc[page_cnt['Revenue'] == True,'Revenue'] = 'Buyers'
```

[11]:
```
      Revenue  Administrative  Informational  ProductRelated
0  Non-Buyers        2.117732       0.451833       28.714642
1      Buyers        3.393606       0.786164       48.210168
```

[12]:
```
# creating df for page Duration calc..

page_duration=df1.
 ↪groupby('Revenue')[['Administrative_Duration','Informational_Duration','ProductRelated_Dura
 ↪mean().reset_index()
page_duration.loc[page_duration['Revenue'] == True,'Revenue'] = 'Buyers'
page_duration.loc[page_duration['Revenue'] == False,'Revenue'] = 'Non-Buyers'
page_duration
```

```
<ipython-input-12-b2cde7266cbc>:4: FutureWarning: Setting an item of
incompatible dtype is deprecated and will raise in a future error of pandas.
Value 'Buyers' has dtype incompatible with bool, please explicitly cast to a
compatible dtype first.
  page_duration.loc[page_duration['Revenue'] == True,'Revenue'] = 'Buyers'
```

[12]:
```
      Revenue  Administrative_Duration  Informational_Duration  \
0  Non-Buyers                73.740111               30.236237
1      Buyers               119.483244               57.611427

   ProductRelated_Duration
0              1069.987809
1              1876.209615
```

Calculation of mean of “visited_all_page_categories”

[13]:
```
df1['visted_all_page_categories'] = (df1['Administrative'] > 0) &␣
 ↪(df1['Informational'] > 0) & (df1['ProductRelated'] > 0)

print(df1.groupby('visted_all_page_categories')['Revenue'].mean())
```

```
visted_all_page_categories
```

```
False    0.135983
True     0.242732
Name: Revenue, dtype: float64
```

[14]:

```
# MOM customers buyers Vs NON buyers
monthly_cust = df1.groupby('Revenue')['Month'].value_counts().reset_index()
monthly_cust.loc[monthly_cust['Revenue'] == True,'Revenue'] = 'Buyers'
monthly_cust.loc[monthly_cust['Revenue'] == False,'Revenue'] = 'Non-Buyers'
monthly_cust
```

```
<ipython-input-14-23dc61626f21>:3: FutureWarning: Setting an item of
incompatible dtype is deprecated and will raise in a future error of pandas.
Value 'Buyers' has dtype incompatible with bool, please explicitly cast to a
compatible dtype first.
  monthly_cust.loc[monthly_cust['Revenue'] == True,'Revenue'] = 'Buyers'
```

[14]:

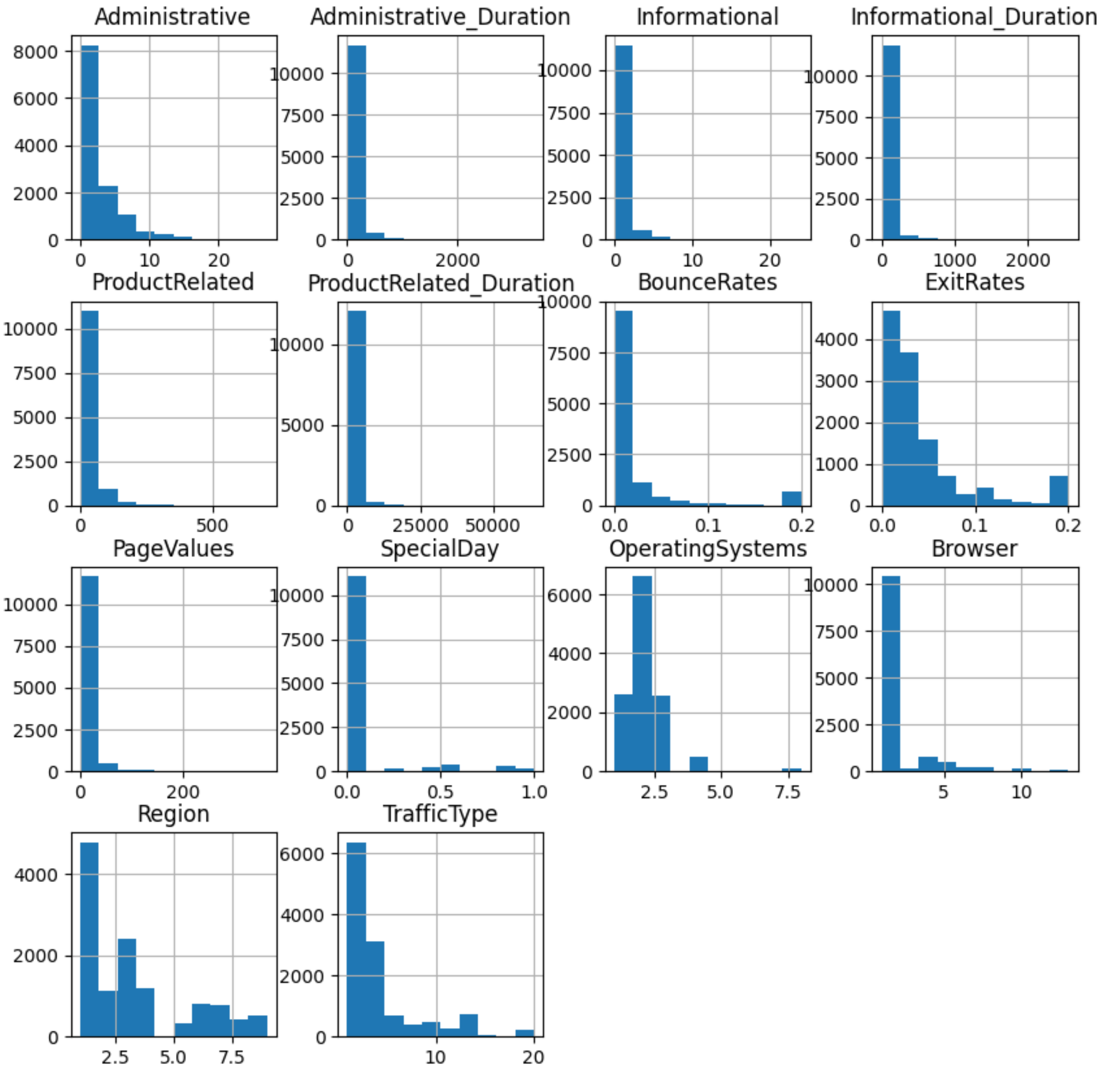
```
       Revenue Month  count
0   Non-Buyers   May   2999
1   Non-Buyers   Nov   2238
2   Non-Buyers   Mar   1715
3   Non-Buyers   Dec   1511
4   Non-Buyers   Oct    434
5   Non-Buyers   Jul    366
6   Non-Buyers   Sep    362
7   Non-Buyers   Aug    357
8   Non-Buyers  June    259
9   Non-Buyers   Feb    181
10      Buyers   Nov    760
11      Buyers   May    365
12      Buyers   Dec    216
13      Buyers   Mar    192
14      Buyers   Oct    115
15      Buyers   Sep     86
16      Buyers   Aug     76
17      Buyers   Jul     66
18      Buyers  June     29
19      Buyers   Feb      3
```

[15]:

```
df1.hist(figsize=(10,10))
```

[15]:

```
array([[<Axes: title={'center': 'Administrative'}>,
        <Axes: title={'center': 'Administrative_Duration'}>,
        <Axes: title={'center': 'Informational'}>,
        <Axes: title={'center': 'Informational_Duration'}>],
       [<Axes: title={'center': 'ProductRelated'}>,
        <Axes: title={'center': 'ProductRelated_Duration'}>,
        <Axes: title={'center': 'BounceRates'}>,
        <Axes: title={'center': 'ExitRates'}>],
       [<Axes: title={'center': 'PageValues'}>,
        <Axes: title={'center': 'SpecialDay'}>,
        <Axes: title={'center': 'OperatingSystems'}>,
        <Axes: title={'center': 'Browser'}>],
       [<Axes: title={'center': 'Region'}>,
        <Axes: title={'center': 'TrafficType'}>, <Axes: >, <Axes: >]],
      dtype=object)
```

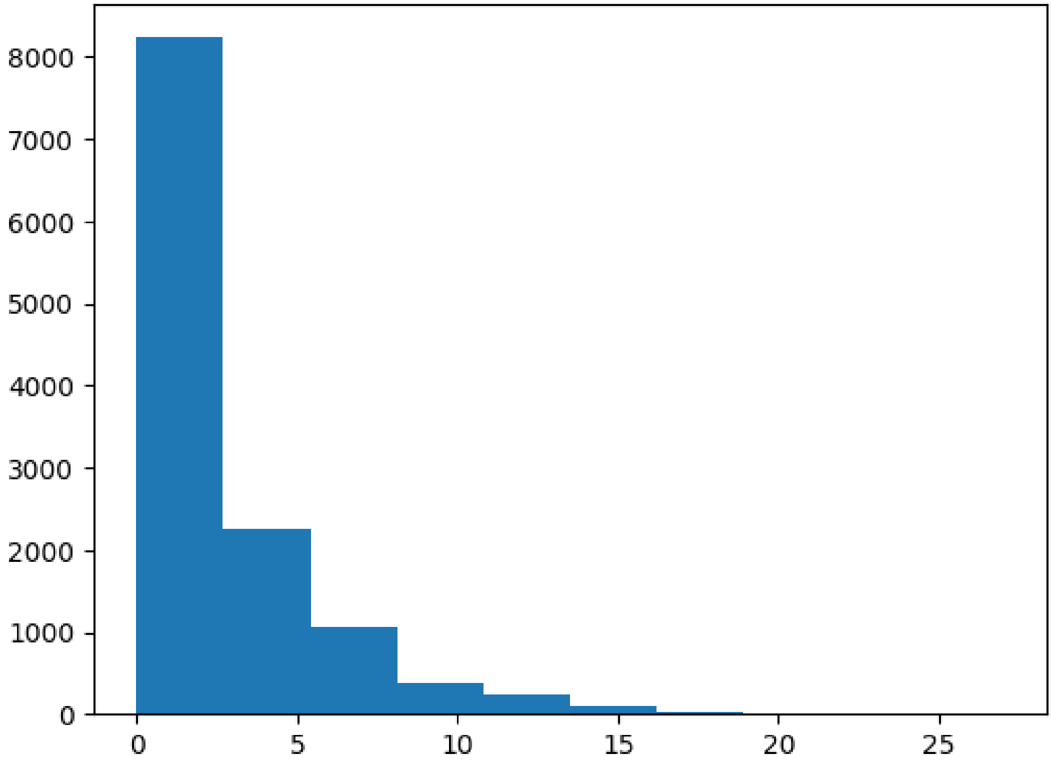


[15]:

0.1 Q.1. Univariate Analysis:Plot histograms or box plots for each numerical feature to identify outliers and distribution shapes.

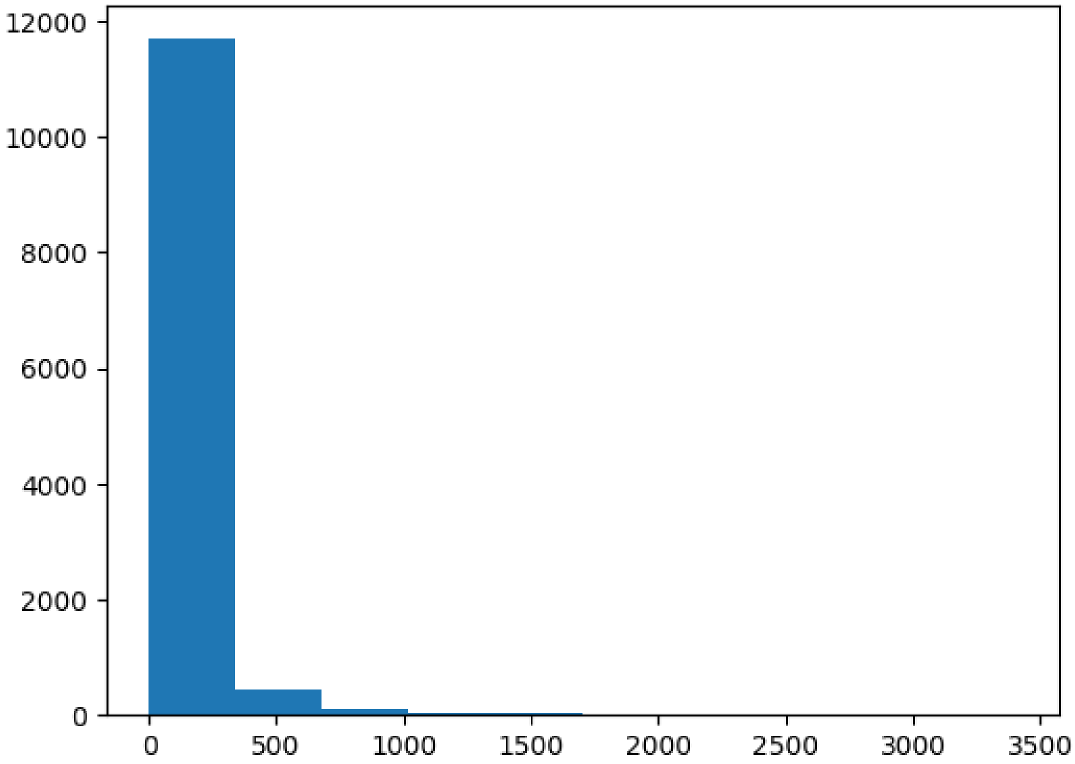
```
[16]: plt.hist(df1['Administrative'])
      #sns.histplot(df1['Administrative'])
```

```
[16]: (array([8.236e+03, 2.255e+03, 1.057e+03, 3.780e+02, 2.470e+02, 1.060e+02,
              2.800e+01, 1.000e+01, 1.100e+01, 2.000e+00]),
       array([ 0. ,  2.7,  5.4,  8.1, 10.8, 13.5, 16.2, 18.9, 21.6, 24.3, 27. ]),
       <BarContainer object of 10 artists>)
```



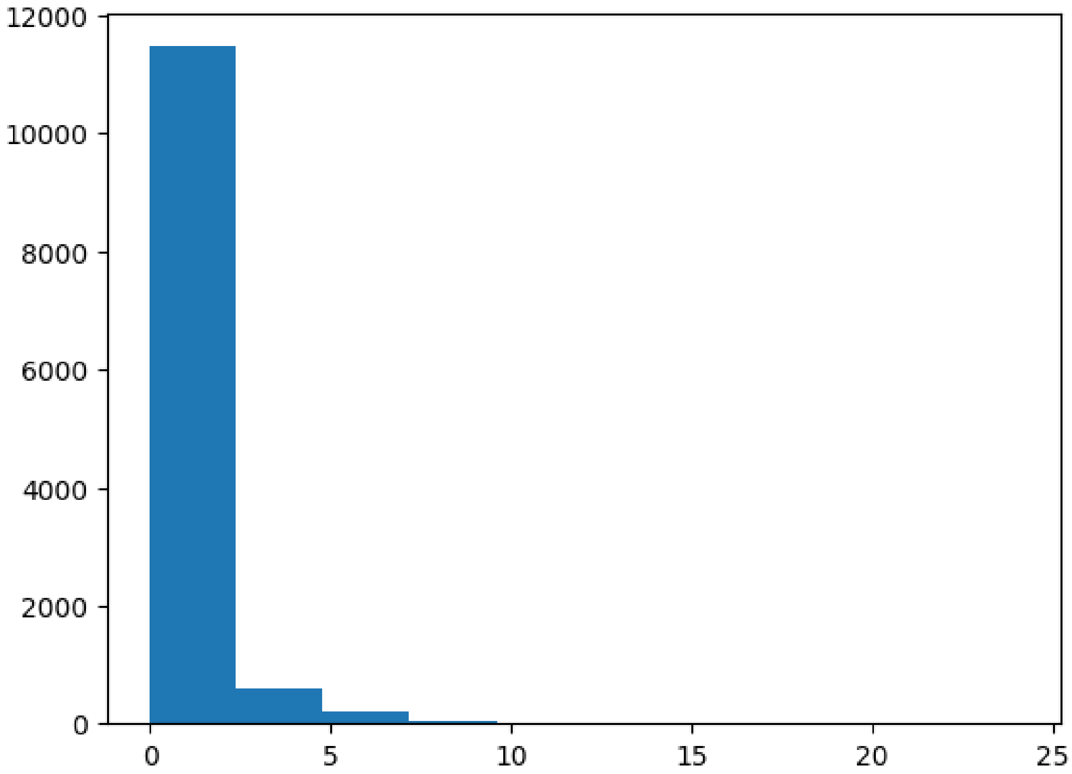
```
[17]: plt.hist(df1['Administrative_Duration'])
      #sns.histplot(df1['Administrative_Duration'])
```

```
[17]: (array([1.1687e+04, 4.5600e+02, 1.0800e+02, 3.7000e+01, 2.7000e+01,
              6.0000e+00, 4.0000e+00, 3.0000e+00, 1.0000e+00, 1.0000e+00]),
       array([   0.   ,  339.875,  679.75 , 1019.625, 1359.5  , 1699.375,
              2039.25 , 2379.125, 2719.   , 3058.875, 3398.75 ]),
       <BarContainer object of 10 artists>)
```



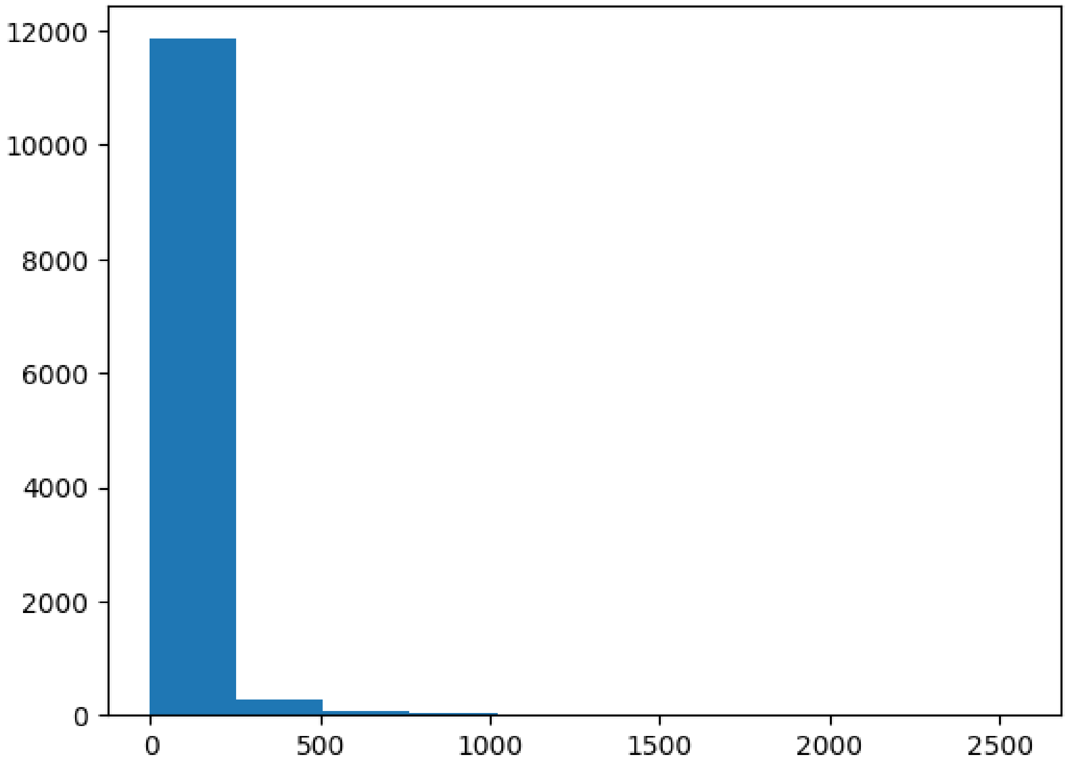
```
[18]: plt.hist(df1['Informational'])
      #sns.histplot(df1['Informational'])
```

```
[18]: (array([1.1468e+04, 6.0200e+02, 2.1300e+02, 2.9000e+01, 8.0000e+00,
              8.0000e+00, 1.0000e+00, 0.0000e+00, 0.0000e+00, 1.0000e+00]),
       array([ 0.  ,  2.4,  4.8,  7.2,  9.6, 12. , 14.4, 16.8, 19.2, 21.6, 24. ]),
       <BarContainer object of 10 artists>)
```



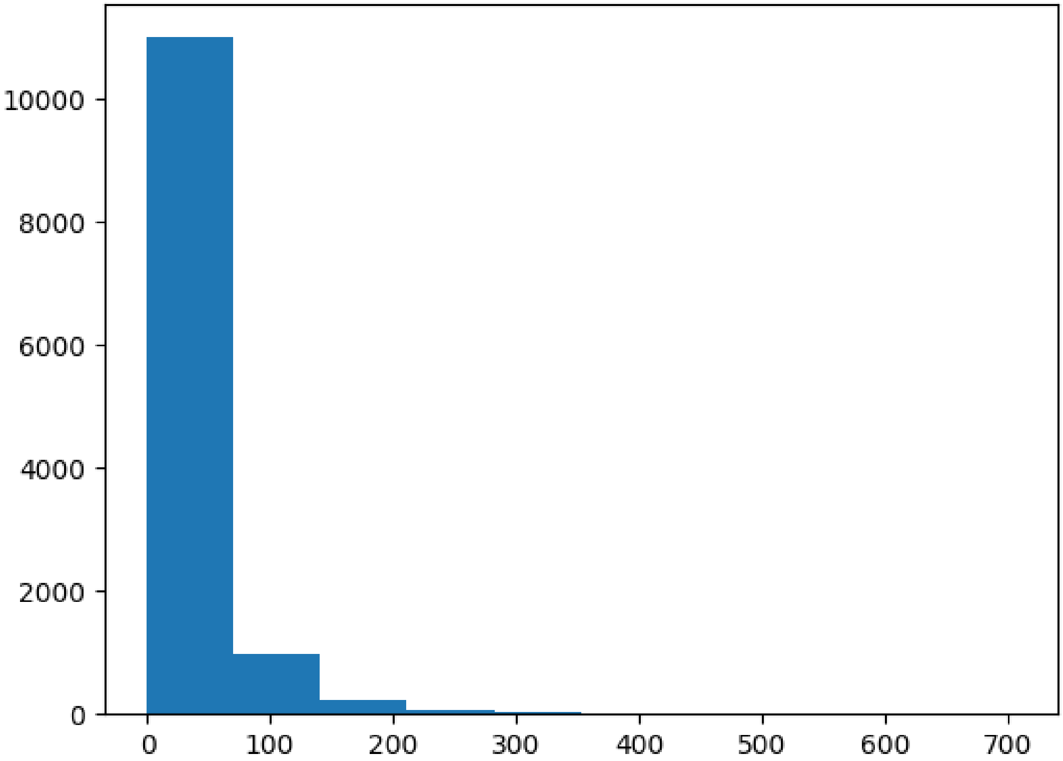
[19]:
```
plt.hist(df1['Informational_Duration'])
#sns.histplot(df1['Informational_Duration'])
```

[19]:
```
(array([1.1858e+04, 2.7700e+02, 8.4000e+01, 5.2000e+01, 2.6000e+01,
        1.5000e+01, 1.0000e+01, 2.0000e+00, 5.0000e+00, 1.0000e+00]),
 array([   0.    ,  254.9375,  509.875 ,  764.8125, 1019.75  , 1274.6875,
        1529.625 , 1784.5625, 2039.5   , 2294.4375, 2549.375 ]),
 <BarContainer object of 10 artists>)
```



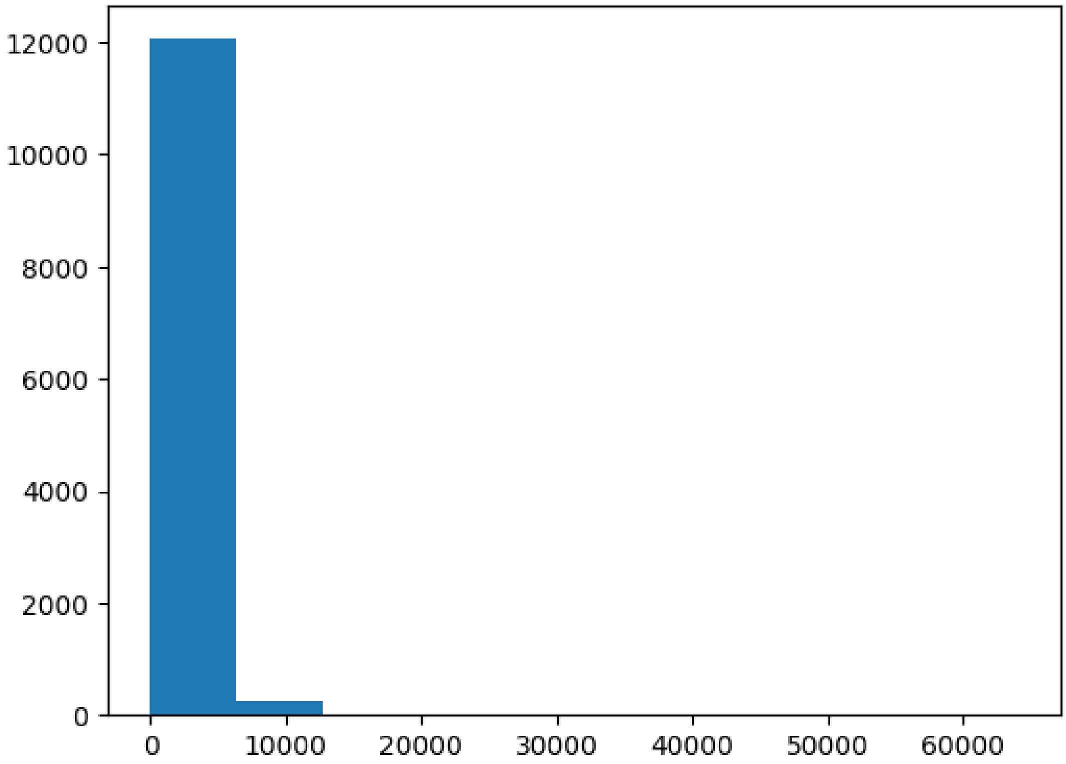
```
[20]: plt.hist(df1['ProductRelated'])
```

```
[20]: (array([1.1006e+04, 9.6800e+02, 2.1900e+02, 7.5000e+01, 3.3000e+01,
              1.4000e+01, 8.0000e+00, 4.0000e+00, 1.0000e+00, 2.0000e+00]),
       array([  0. ,  70.5, 141. , 211.5, 282. , 352.5, 423. , 493.5, 564. ,
              634.5, 705. ]),
       <BarContainer object of 10 artists>)
```



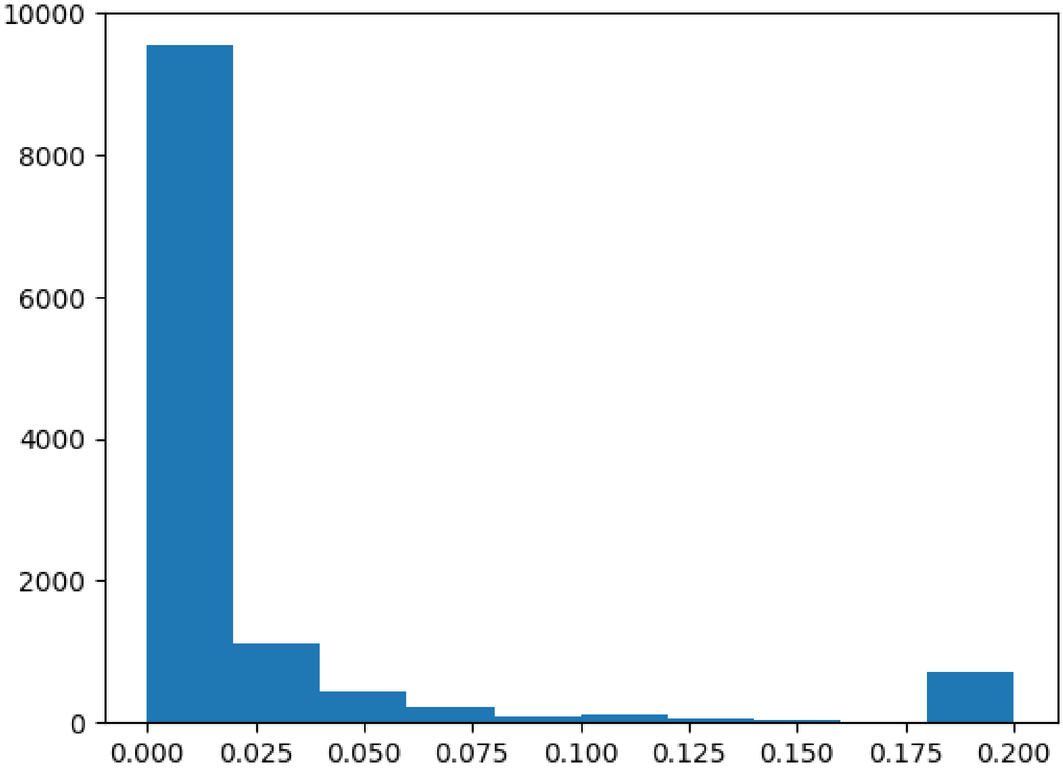
```
[21]: plt.hist(df1['ProductRelated_Duration'])
```

```
[21]: (array([1.206e+04, 2.340e+02, 2.600e+01, 6.000e+00, 2.000e+00, 0.000e+00,
              1.000e+00, 0.000e+00, 0.000e+00, 1.000e+00]),
       array([    0.      ,  6397.352223, 12794.704446, 19192.056669,
              25589.408892, 31986.761115, 38384.113338, 44781.465561,
              51178.817784, 57576.170007, 63973.52223 ]),
       <BarContainer object of 10 artists>)
```



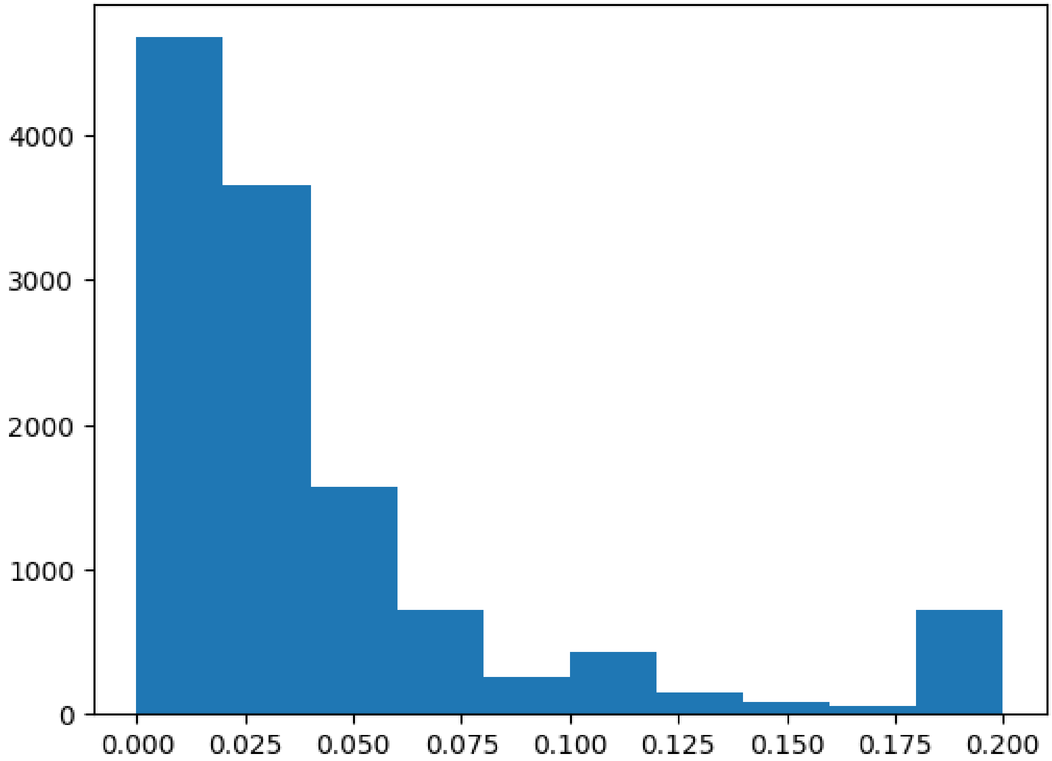
```
[22]: plt.hist(df1['BounceRates'])
```

```
[22]: (array([9542., 1119.,  449.,  225.,   81.,  118.,   51.,   27.,   12.,
                706.]),
       array([0.  , 0.02, 0.04, 0.06, 0.08, 0.1 , 0.12, 0.14, 0.16, 0.18, 0.2 ]),
       <BarContainer object of 10 artists>)
```



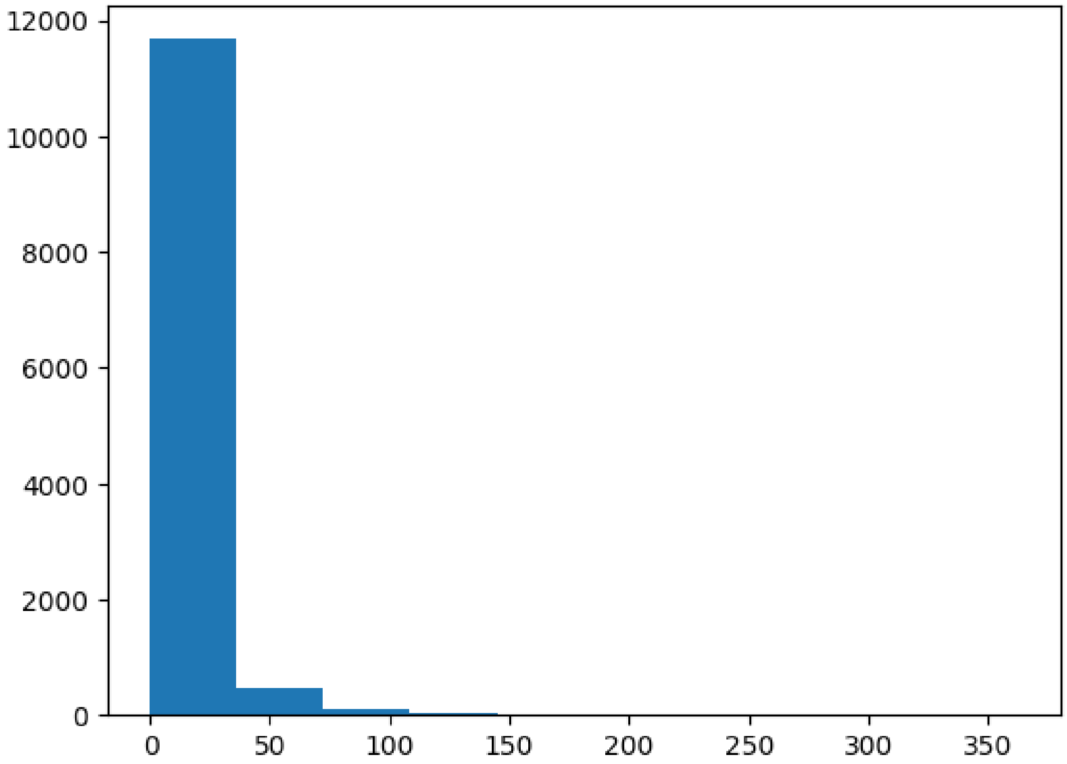
```
[23]: plt.hist(df1['ExitRates'])
```

```
[23]: (array([4675., 3662., 1570.,  716.,  257.,  431.,  148.,   90.,   57.,
               724.]),
       array([0.  , 0.02, 0.04, 0.06, 0.08, 0.1 , 0.12, 0.14, 0.16, 0.18, 0.2 ]),
       <BarContainer object of 10 artists>)
```



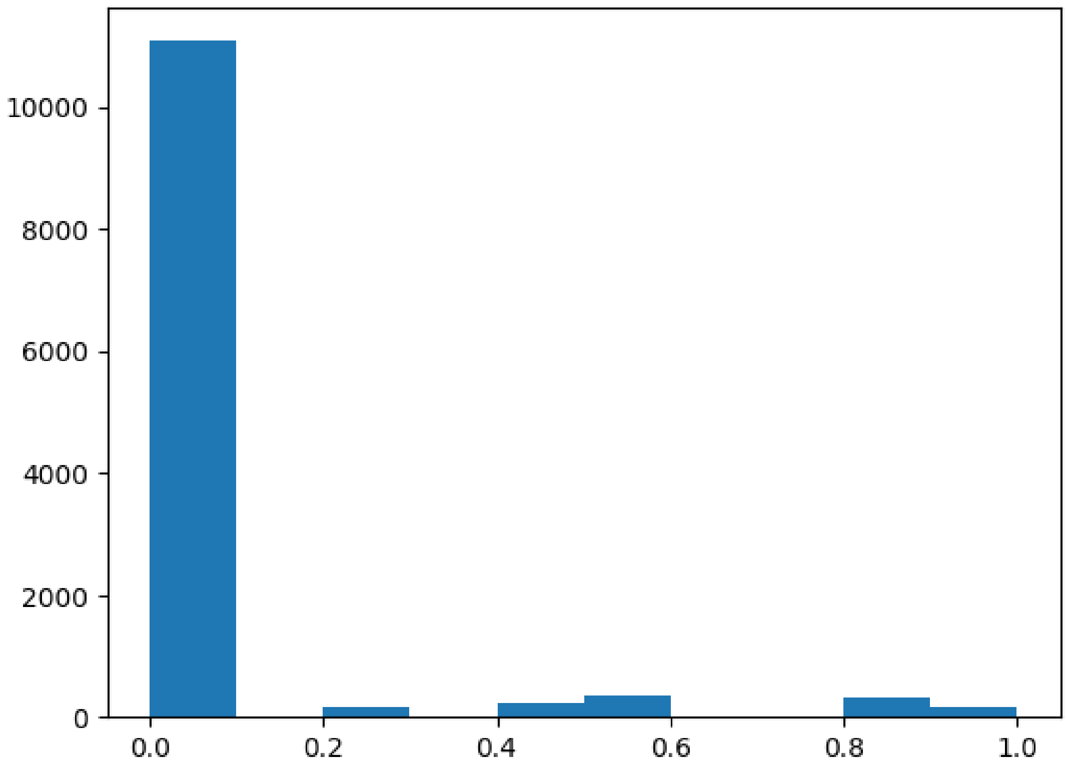
```
[24]: plt.hist(df1['PageValues'])
```

```
[24]: (array([1.1676e+04, 4.7300e+02, 1.1400e+02, 3.9000e+01, 1.2000e+01,
              3.0000e+00, 5.0000e+00, 6.0000e+00, 0.0000e+00, 2.0000e+00]),
       array([  0.        ,  36.17637419,  72.35274838, 108.52912257,
              144.70549676, 180.88187095, 217.05824514, 253.23461933,
              289.41099352, 325.58736771, 361.7637419 ]),
       <BarContainer object of 10 artists>)
```



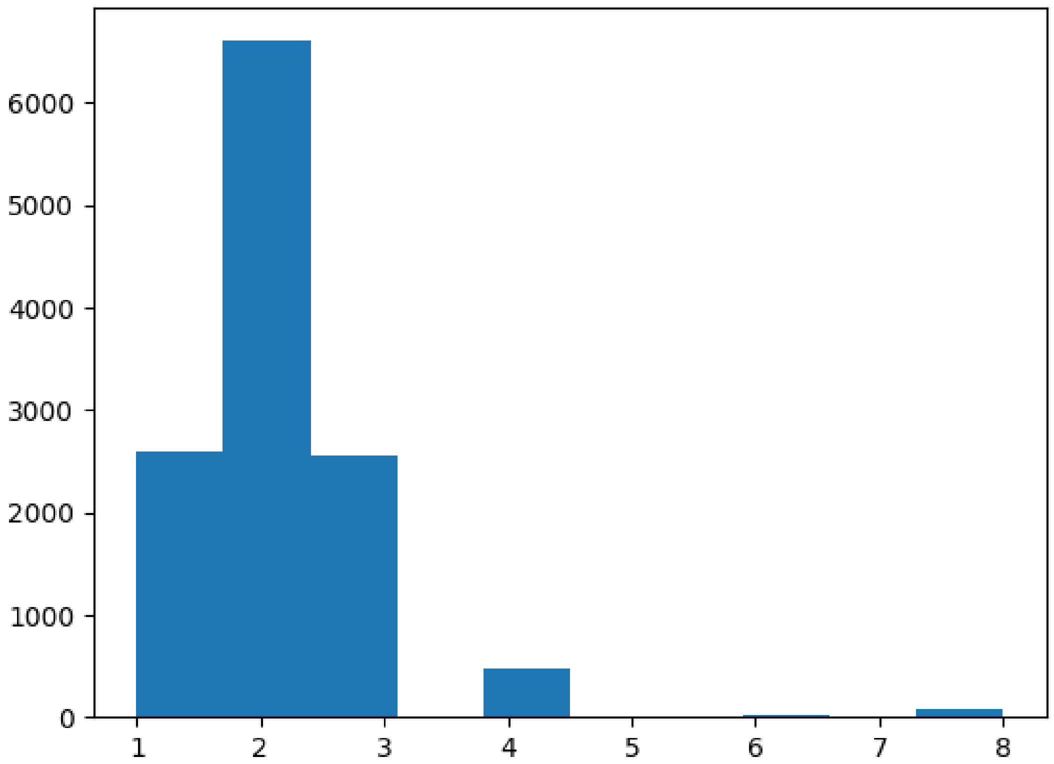
```
[25]: plt.hist(df1['SpecialDay'])
```

```
[25]: (array([11079.,     0.,   178.,     0.,   243.,   351.,     0.,     0.,
                325.,   154.]),
       array([0. , 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1. ]),
       <BarContainer object of 10 artists>)
```



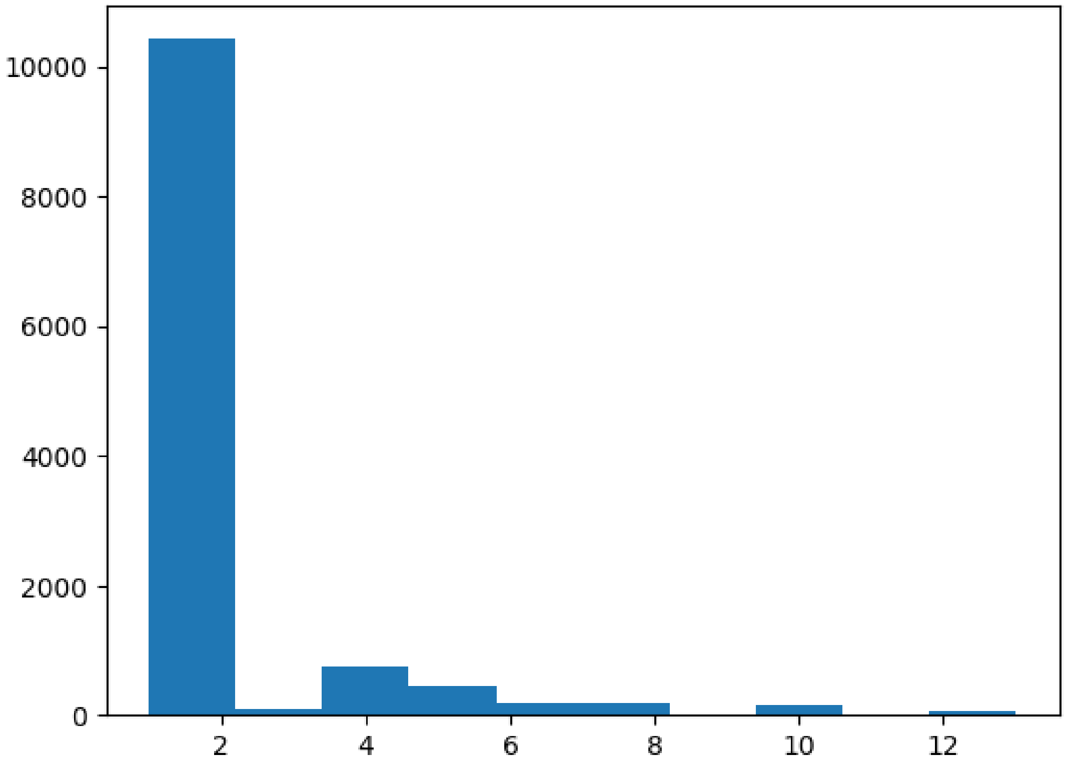
```
[26]: plt.hist(df1['OperatingSystems'])
```

```
[26]: (array([2.585e+03, 6.601e+03, 2.555e+03, 0.000e+00, 4.780e+02, 6.000e+00,
              0.000e+00, 1.900e+01, 7.000e+00, 7.900e+01]),
       array([1. , 1.7, 2.4, 3.1, 3.8, 4.5, 5.2, 5.9, 6.6, 7.3, 8. ]),
       <BarContainer object of 10 artists>)
```



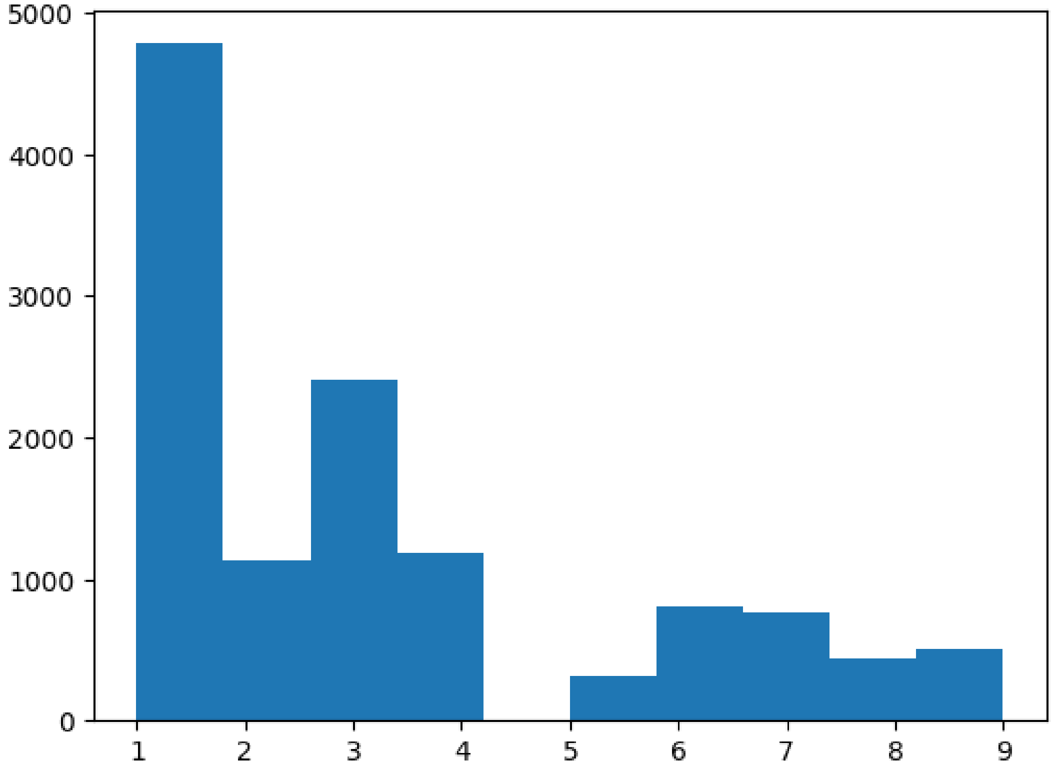
```
[27]: plt.hist(df1['Browser'])
```

```
[27]: (array([1.0423e+04, 1.0500e+02, 7.3600e+02, 4.6700e+02, 1.7400e+02,
              1.8400e+02, 1.0000e+00, 1.6300e+02, 6.0000e+00, 7.1000e+01]),
       array([ 1. ,  2.2,  3.4,  4.6,  5.8,  7. ,  8.2,  9.4, 10.6, 11.8, 13. ]),
       <BarContainer object of 10 artists>)
```



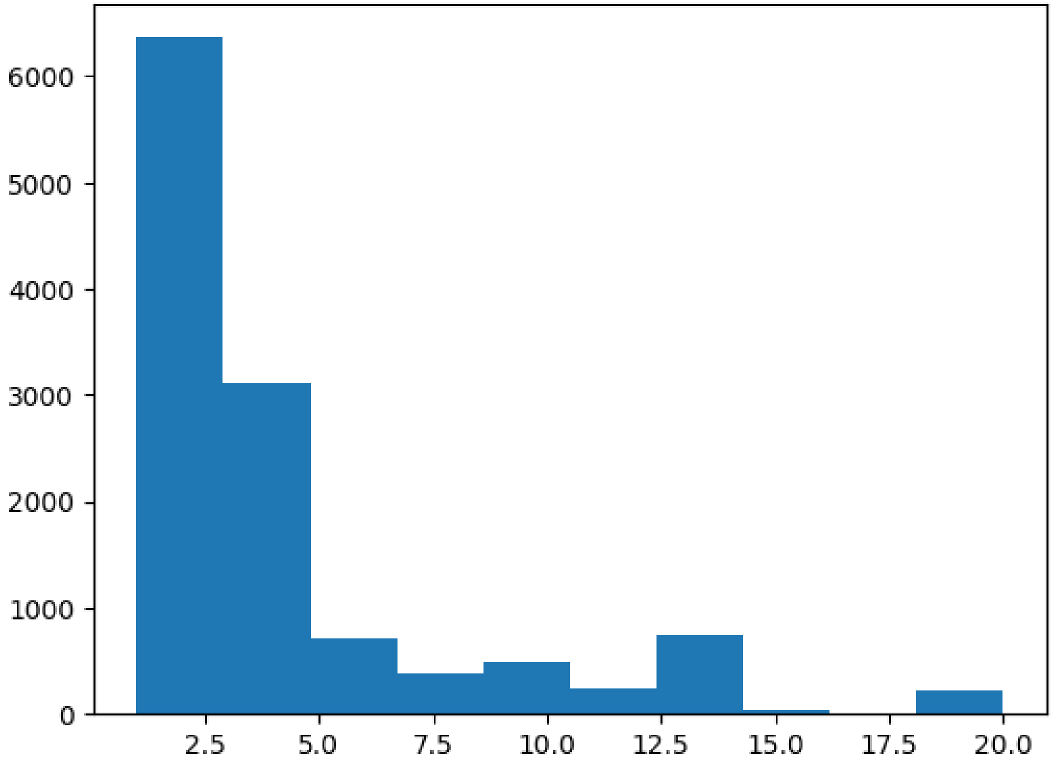
```
[28]: plt.hist(df1['Region'])
```

```
[28]: (array([4780., 1136., 2403., 1182.,    0.,  318.,  805.,  761.,  434.,
               511.]),
       array([1. , 1.8, 2.6, 3.4, 4.2, 5. , 5.8, 6.6, 7.4, 8.2, 9. ]),
       <BarContainer object of 10 artists>)
```



[29]:
```
plt.hist(df1['TrafficType'])
```

[29]:
```
(array([6364., 3121.,  704.,  383.,  492.,  248.,  751.,   41.,   11.,
         215.]),
 array([ 1. ,  2.9,  4.8,  6.7,  8.6, 10.5, 12.4, 14.3, 16.2, 18.1, 20. ]),
 <BarContainer object of 10 artists>)
```



1 *Q.2. Correlation Analysis: Calculate correlations between numerical features to identify potential relationships.*

[30]:
```
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12330 entries, 0 to 12329
Data columns (total 19 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Administrative           12330 non-null  int64
 1   Administrative_Duration  12330 non-null  float64
 2   Informational            12330 non-null  int64
 3   Informational_Duration   12330 non-null  float64
 4   ProductRelated           12330 non-null  int64
 5   ProductRelated_Duration  12330 non-null  float64
 6   BounceRates              12330 non-null  float64
 7   ExitRates                12330 non-null  float64
 8   PageValues               12330 non-null  float64
 9   SpecialDay               12330 non-null  float64
 10  Month                    12330 non-null  object
 11  OperatingSystems            12330 non-null  int64
 12  Browser                     12330 non-null  int64
 13  Region                      12330 non-null  int64
 14  TrafficType                 12330 non-null  int64
 15  VisitorType                 12330 non-null  object
 16  Weekend                     12330 non-null  bool
 17  Revenue                     12330 non-null  bool
 18  visted_all_page_categories  12330 non-null  bool
dtypes: bool(3), float64(7), int64(7), object(2)
memory usage: 1.5+ MB
```

[31]:
```
numeric_df = df1.select_dtypes(include=['number'])
```

[32]:
```
correlation_matrix = numeric_df.corr()
correlation_matrix
```

[32]:
```
                         Administrative  Administrative_Duration  \
Administrative                 1.000000                 0.601583
Administrative_Duration        0.601583                 1.000000
Informational                  0.376850                 0.302710
Informational_Duration         0.255848                 0.238031
ProductRelated                 0.431119                 0.289087
ProductRelated_Duration        0.373939                 0.355422
BounceRates                   -0.223563                -0.144170
ExitRates                     -0.316483                -0.205798
PageValues                     0.098990                 0.067608
SpecialDay                    -0.094778                -0.073304
OperatingSystems              -0.006347                -0.007343
Browser                       -0.025035                -0.015392
Region                        -0.005487                -0.005561
TrafficType                   -0.033561                -0.014376

                         Informational  Informational_Duration  \
Administrative                0.376850                0.255848
Administrative_Duration       0.302710                0.238031
Informational                 1.000000                0.618955
Informational_Duration        0.618955                1.000000
ProductRelated                0.374164                0.280046
ProductRelated_Duration       0.387505                0.347364
BounceRates                  -0.116114               -0.074067
ExitRates                    -0.163666               -0.105276
PageValues                    0.048632                0.030861
SpecialDay                   -0.048219               -0.030577
OperatingSystems             -0.009527               -0.009579
Browser                      -0.038235               -0.019285
Region                       -0.029169               -0.027144
TrafficType                  -0.034491               -0.024675
```

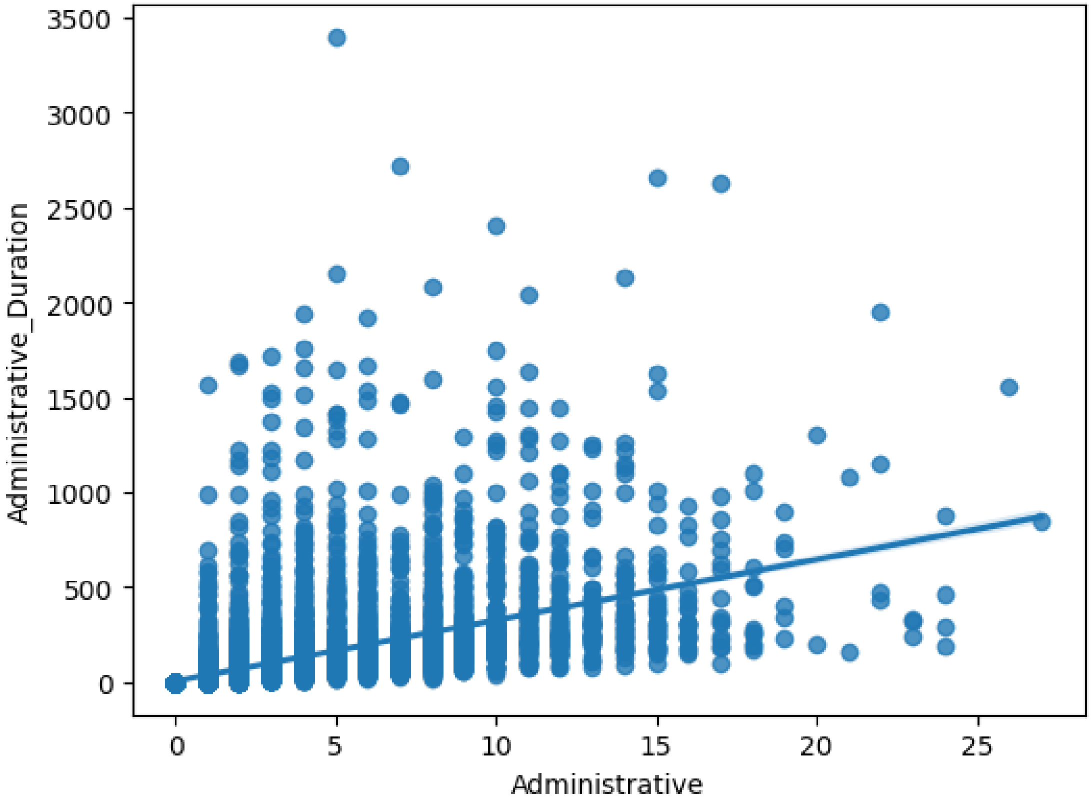
```
                         ProductRelated  ProductRelated_Duration  BounceRates  \
Administrative                 0.431119                 0.373939    -0.223563
Administrative_Duration        0.289087                 0.355422    -0.144170
Informational                  0.374164                 0.387505    -0.116114
Informational_Duration         0.280046                 0.347364    -0.074067
ProductRelated                 1.000000                 0.860927    -0.204578
ProductRelated_Duration        0.860927                 1.000000    -0.184541
BounceRates                   -0.204578                -0.184541     1.000000
ExitRates                     -0.292526                -0.251984     0.913004
PageValues                     0.056282                 0.052823    -0.119386
SpecialDay                    -0.023958                -0.036380     0.072702
OperatingSystems               0.004290                 0.002976     0.023823
Browser                       -0.013146                -0.007380    -0.015772
Region                        -0.038122                -0.033091    -0.006485
TrafficType                   -0.043064                -0.036377     0.078286

                         ExitRates  PageValues  SpecialDay  OperatingSystems  \
Administrative           -0.316483    0.098990   -0.094778         -0.006347
Administrative_Duration  -0.205798    0.067608   -0.073304         -0.007343
Informational            -0.163666    0.048632   -0.048219         -0.009527
Informational_Duration   -0.105276    0.030861   -0.030577         -0.009579
ProductRelated           -0.292526    0.056282   -0.023958          0.004290
ProductRelated_Duration  -0.251984    0.052823   -0.036380          0.002976
BounceRates               0.913004   -0.119386    0.072702          0.023823
ExitRates                 1.000000   -0.174498    0.102242          0.014567
PageValues               -0.174498    1.000000   -0.063541          0.018508
SpecialDay                0.102242   -0.063541    1.000000          0.012652
OperatingSystems          0.014567    0.018508    0.012652          1.000000
Browser                  -0.004442    0.045592    0.003499          0.223013
Region                   -0.008907    0.011315   -0.016098          0.076775
TrafficType               0.078616    0.012532    0.052301          0.189154

                          Browser    Region  TrafficType
Administrative          -0.025035 -0.005487    -0.033561
Administrative_Duration -0.015392 -0.005561    -0.014376
Informational           -0.038235 -0.029169    -0.034491
Informational_Duration  -0.019285 -0.027144    -0.024675
ProductRelated          -0.013146 -0.038122    -0.043064
ProductRelated_Duration -0.007380 -0.033091    -0.036377
BounceRates             -0.015772 -0.006485     0.078286
ExitRates               -0.004442 -0.008907     0.078616
PageValues               0.045592  0.011315     0.012532
SpecialDay               0.003499 -0.016098     0.052301
OperatingSystems         0.223013  0.076775     0.189154
Browser                  1.000000  0.097393     0.111938
Region                   0.097393  1.000000     0.047520
```

```
TrafficType              0.111938  0.047520     1.000000
```

Strong correlation b/w

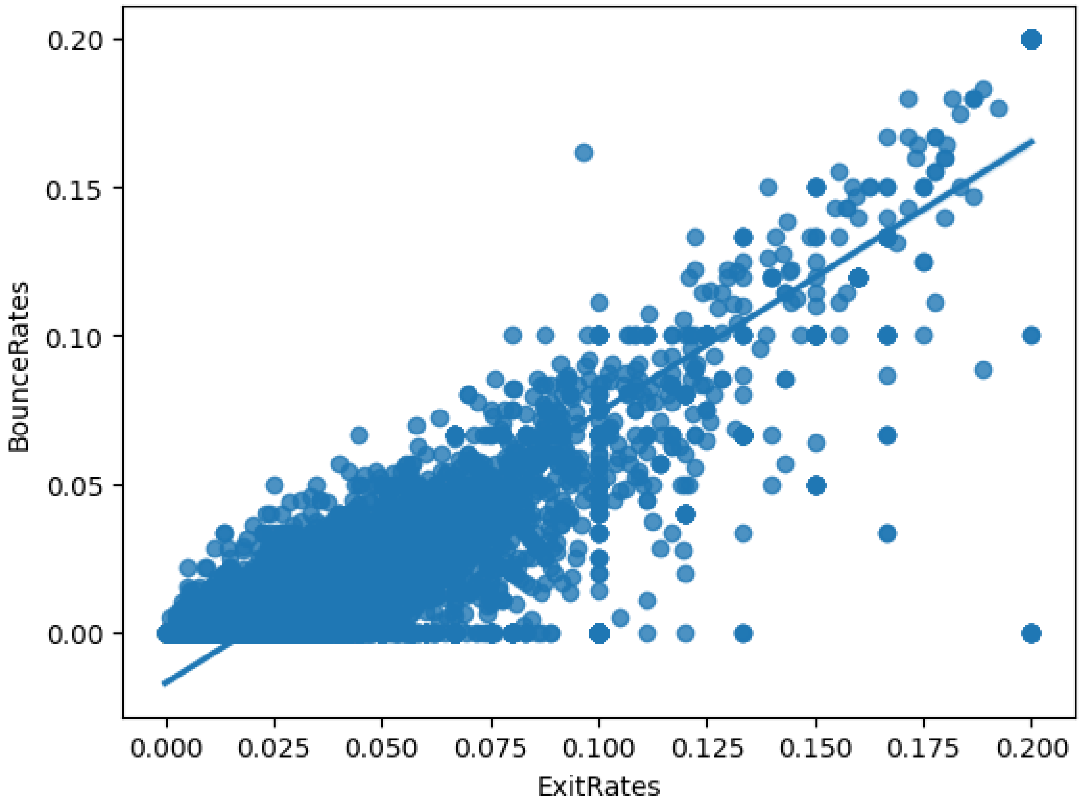
```
[33]: sns.regplot(x='Administrative', y='Administrative_Duration', data=df1)
```

```
[33]: <Axes: xlabel='Administrative', ylabel='Administrative_Duration'>
```



```
[34]: sns.regplot(x='ExitRates', y='BounceRates', data=df1)
```

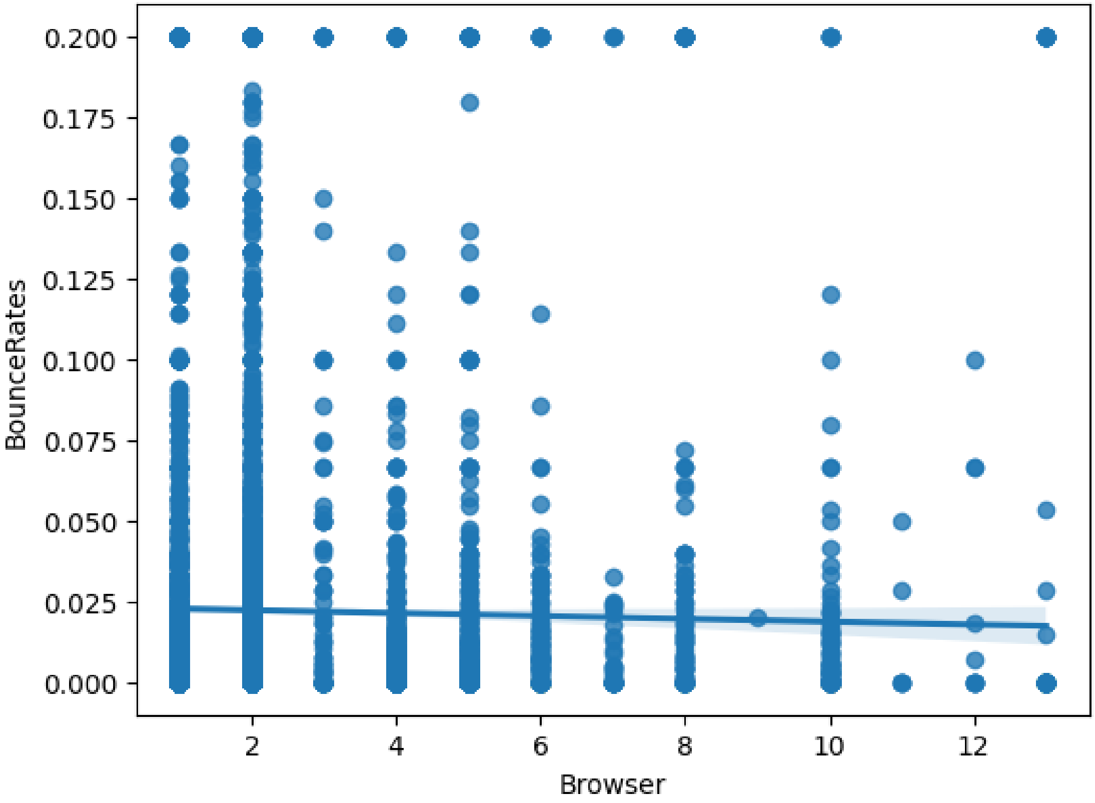
```
[34]: <Axes: xlabel='ExitRates', ylabel='BounceRates'>
```



2 Conclusion: There is Strong positive correlation B/W ExitRates & Bounce Rates

```
[35]: sns.regplot(x='Browser', y='BounceRates', data=df1)
```

```
[35]: <Axes: xlabel='Browser', ylabel='BounceRates'>
```

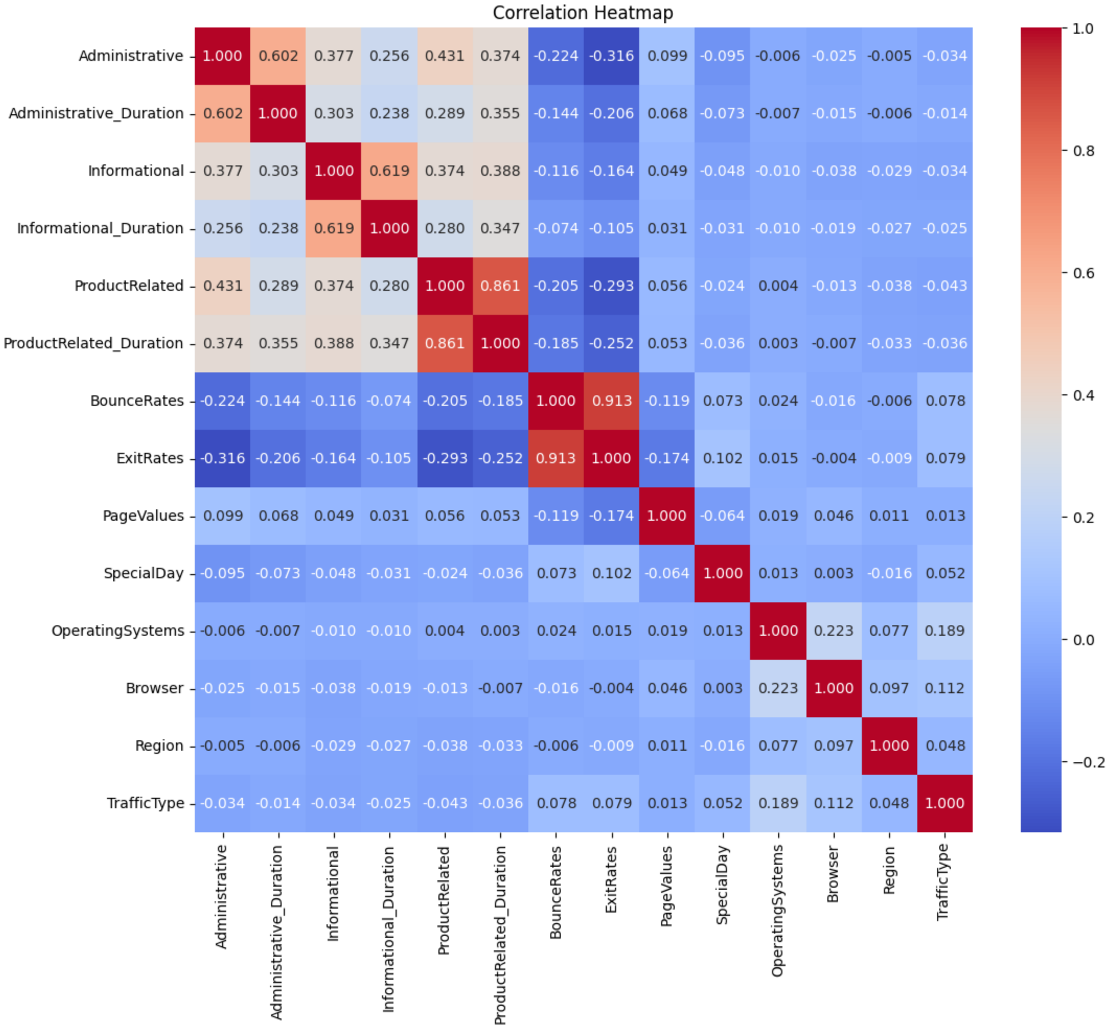


3 Conclusion: There is Negative Correlation B/W Browser & BounceRates.

4 Q.3. Visualizations: Use scatter plots, pair plots, or heatmaps to visualize relationships between numerical features.

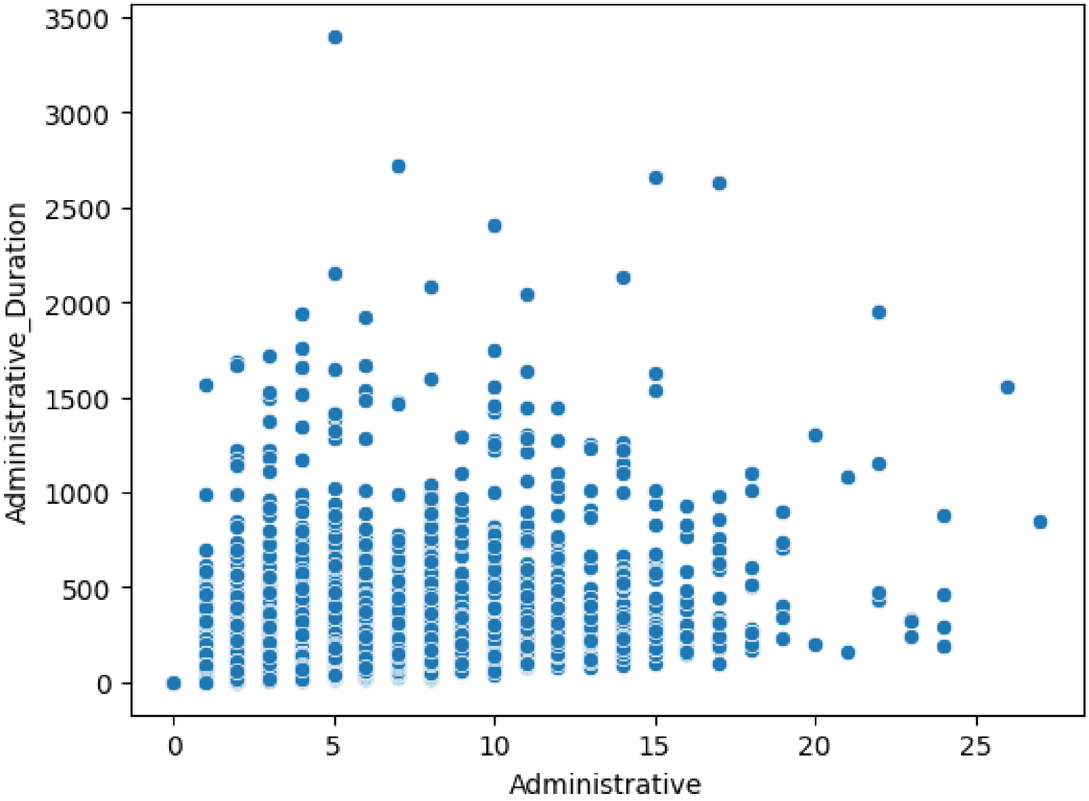
[36]:

```python
correlation_matrix
plt.figure(figsize=(12, 10))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm',fmt='.3f')
plt.title('Correlation Heatmap')
plt.show()
```



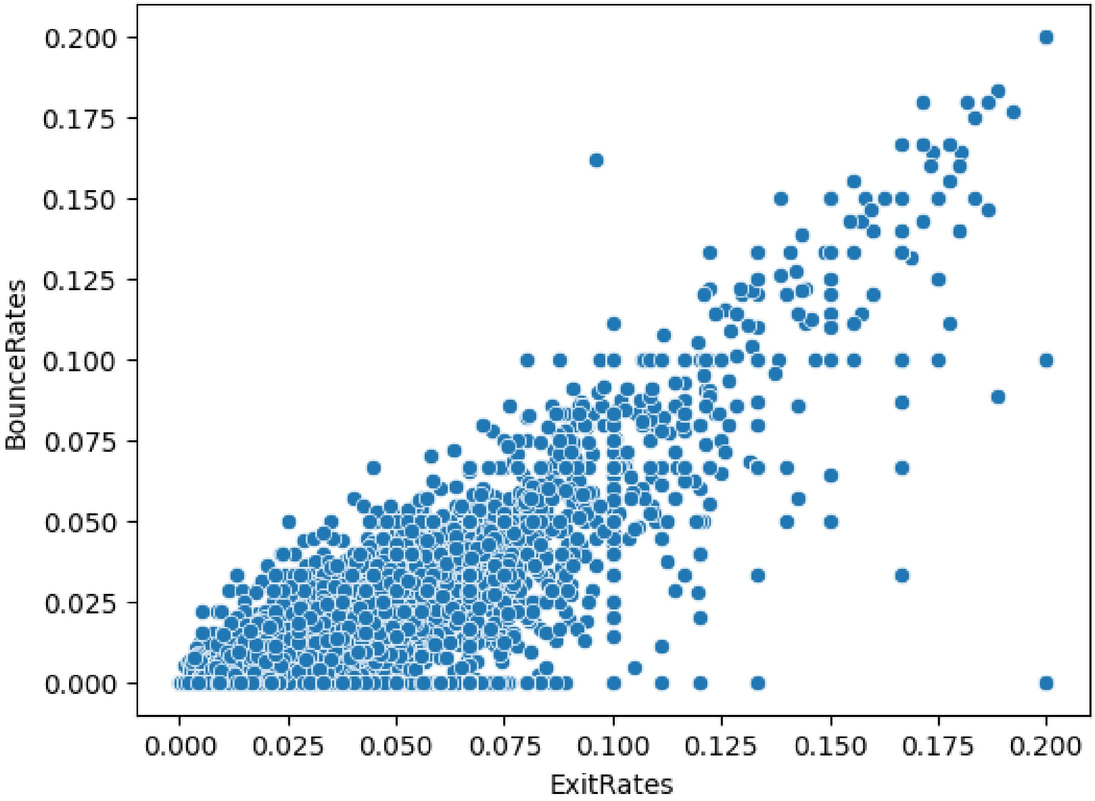
[37]:
```
sns.scatterplot(x='Administrative', y='Administrative_Duration', data=df1)
```

[37]:
```
<Axes: xlabel='Administrative', ylabel='Administrative_Duration'>
```



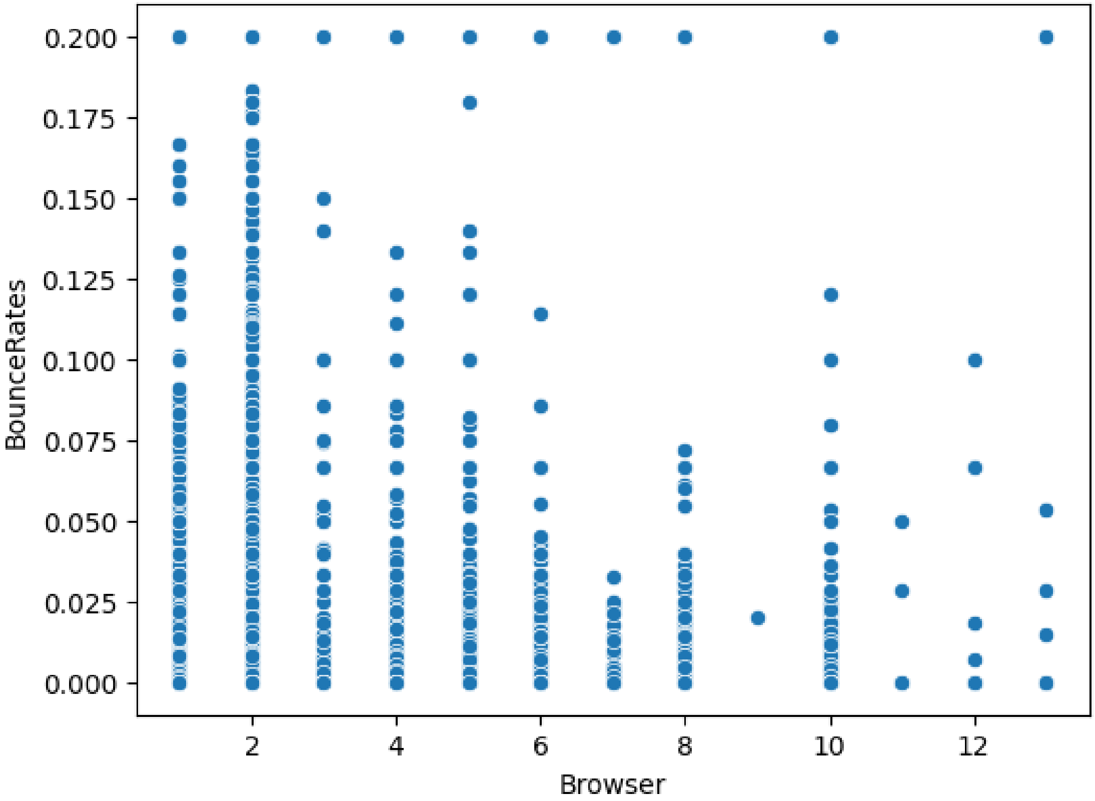
```
[38]: sns.scatterplot(x='ExitRates', y='BounceRates', data=df1)
```

```
[38]: <Axes: xlabel='ExitRates', ylabel='BounceRates'>
```



```
[39]: sns.scatterplot(x='Browser', y='BounceRates', data=df1)
```

```
[39]: <Axes: xlabel='Browser', ylabel='BounceRates'>
```



5 Q.4. Class Distribution: Check the distribution of the target variable ('Revenue') to understand class balance.

[40]:
```
df1.head()
```

[40]:
```
   Administrative  Administrative_Duration  Informational  \
0               0                      0.0              0
1               0                      0.0              0
2               0                      0.0              0
3               0                      0.0              0
4               0                      0.0              0

   Informational_Duration  ProductRelated  ProductRelated_Duration  \
0                     0.0               1                 0.000000
1                     0.0               2                64.000000
2                     0.0               1                 0.000000
3                     0.0               2                 2.666667
4                     0.0              10               627.500000

   BounceRates  ExitRates  PageValues  SpecialDay Month  OperatingSystems  \
0          0.20       0.20         0.0         0.0   Feb                   1
1          0.00       0.10         0.0         0.0   Feb                   2
2          0.20       0.20         0.0         0.0   Feb                   4
3          0.05       0.14         0.0         0.0   Feb                   3
4          0.02       0.05         0.0         0.0   Feb                   3

   Browser  Region  TrafficType        VisitorType  Weekend  Revenue  \
0        1       1            1  Returning_Visitor    False    False
1        2       1            2  Returning_Visitor    False    False
2        1       9            3  Returning_Visitor    False    False
3        2       2            4  Returning_Visitor    False    False
4        3       1            4  Returning_Visitor     True    False

   visted_all_page_categories
0                       False
1                       False
2                       False
3                       False
4                       False
```

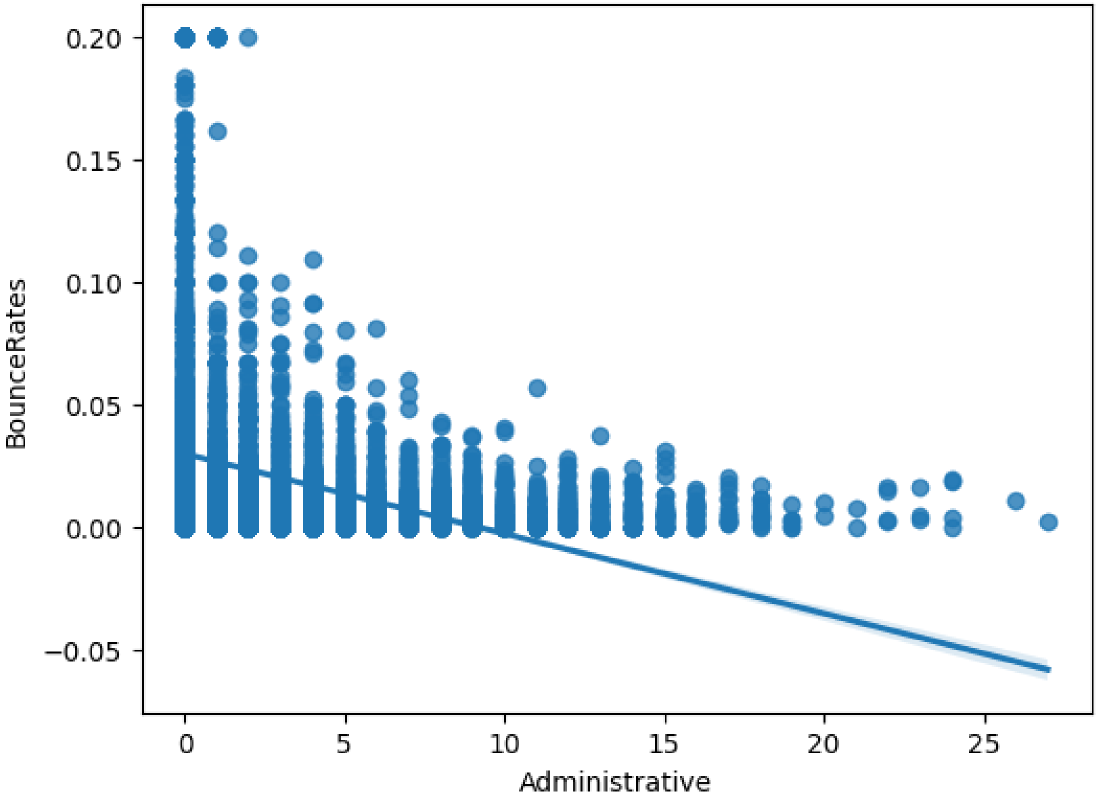
[41]:
```
df1.groupby('Revenue').size()
```

[41]:
```
Revenue
False    10422
True      1908
dtype: int64
```

6 Q.5. Summarize page views, durations, and bounce/exit rates for each page category.

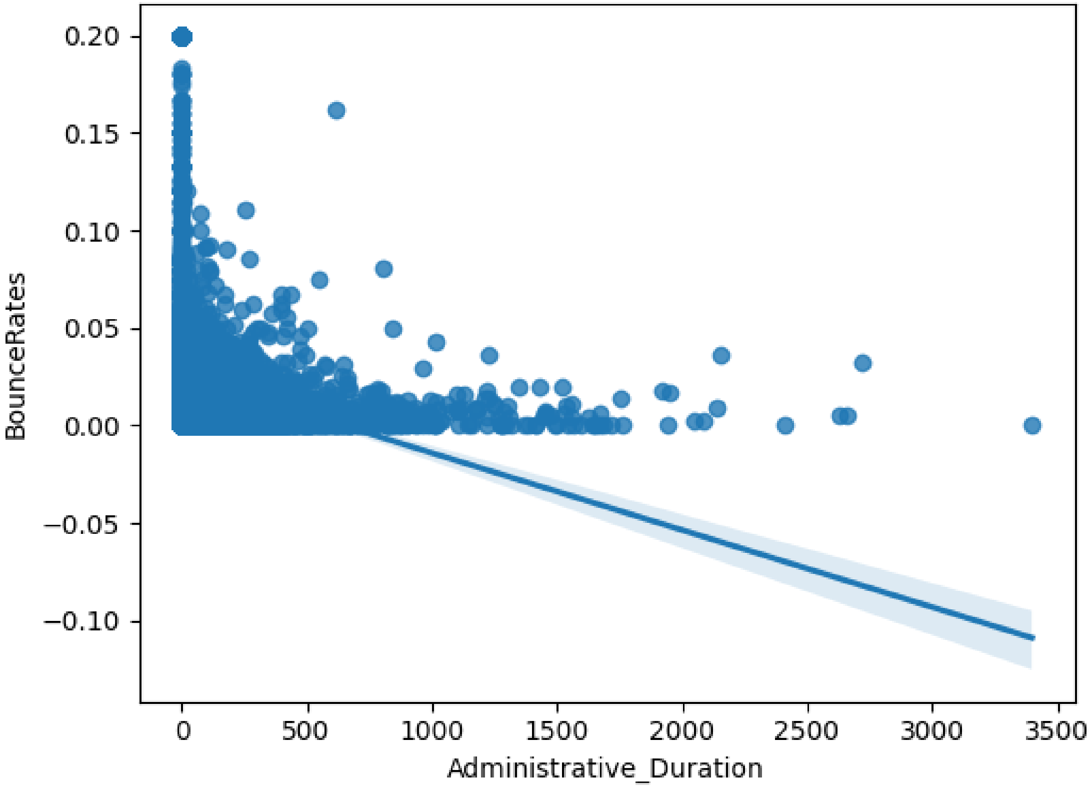
[42]:
```
sns.regplot(x='Administrative', y='BounceRates', data=df1)
#
```

[42]:
```
<Axes: xlabel='Administrative', ylabel='BounceRates'>
```



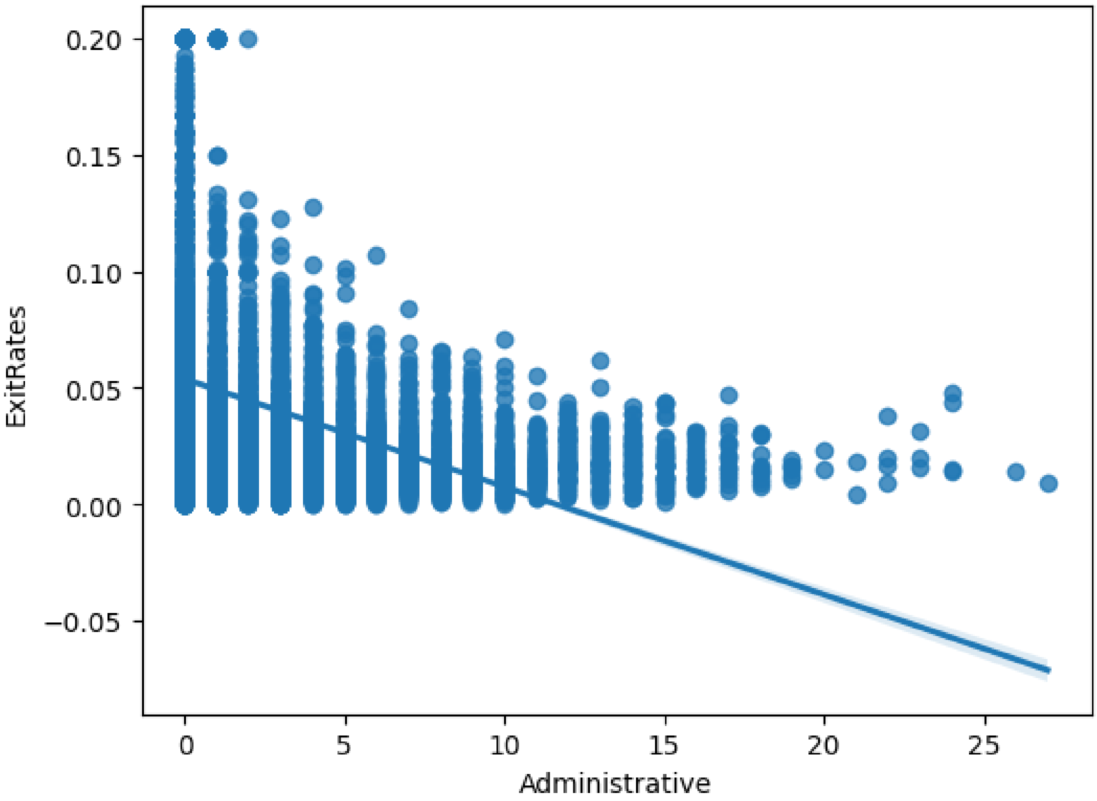
```
[43]: sns.regplot(x='Administrative_Duration', y='BounceRates', data=df1)
```

```
[43]: <Axes: xlabel='Administrative_Duration', ylabel='BounceRates'>
```



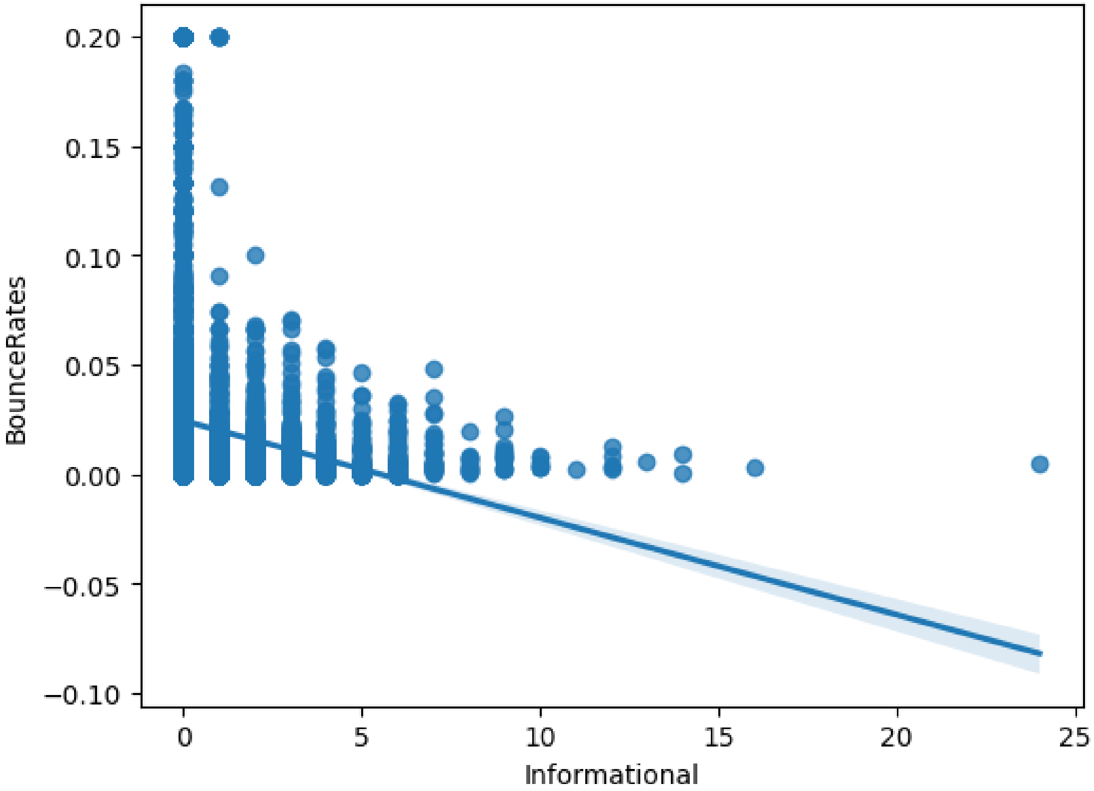
```
[44]: sns.regplot(x='Administrative', y='ExitRates', data=df1)
```

```
[44]: <Axes: xlabel='Administrative', ylabel='ExitRates'>
```



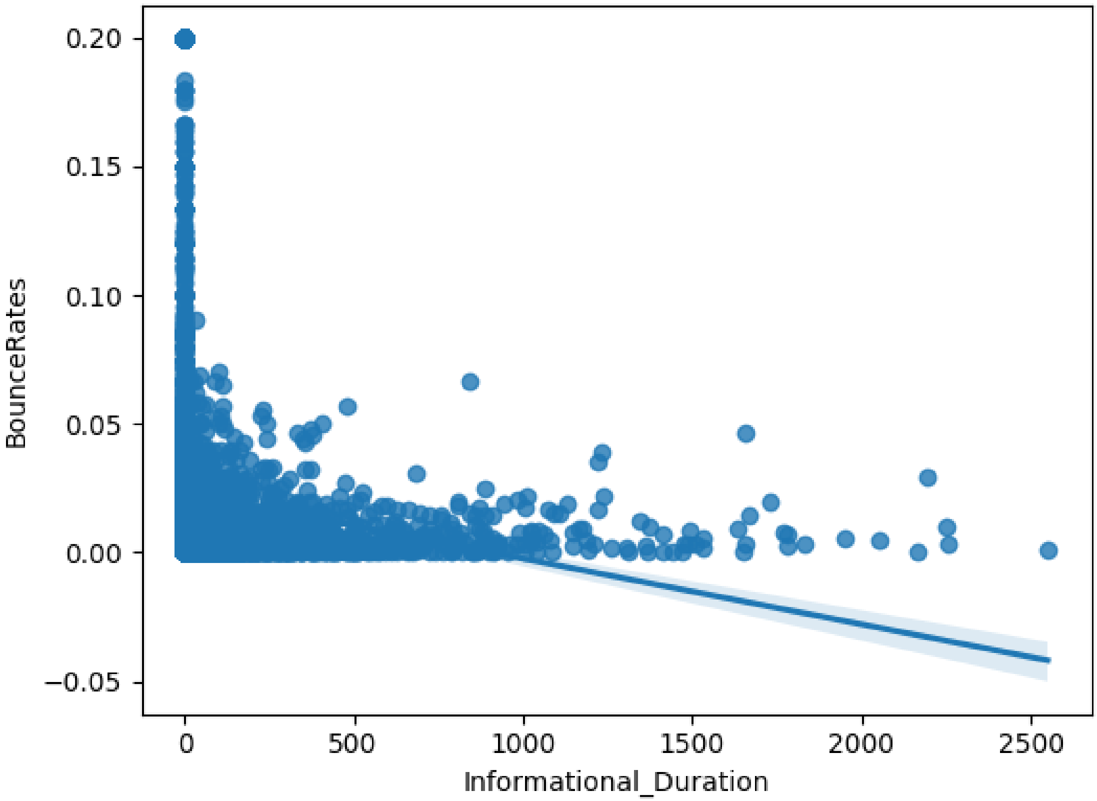
```
[45]: sns.regplot(x='Informational', y='BounceRates', data=df1)
```

```
[45]: <Axes: xlabel='Informational', ylabel='BounceRates'>
```



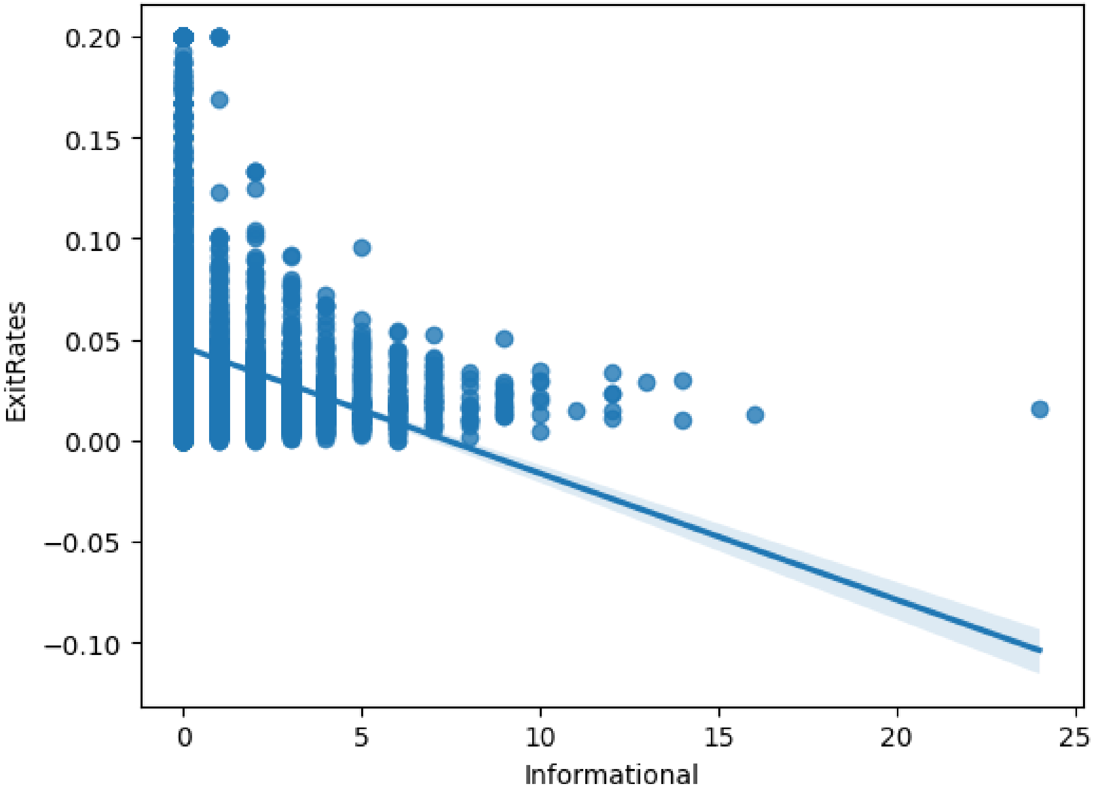
```
[46]: sns.regplot(x='Informational_Duration', y='BounceRates', data=df1)
```

```
[46]: <Axes: xlabel='Informational_Duration', ylabel='BounceRates'>
```



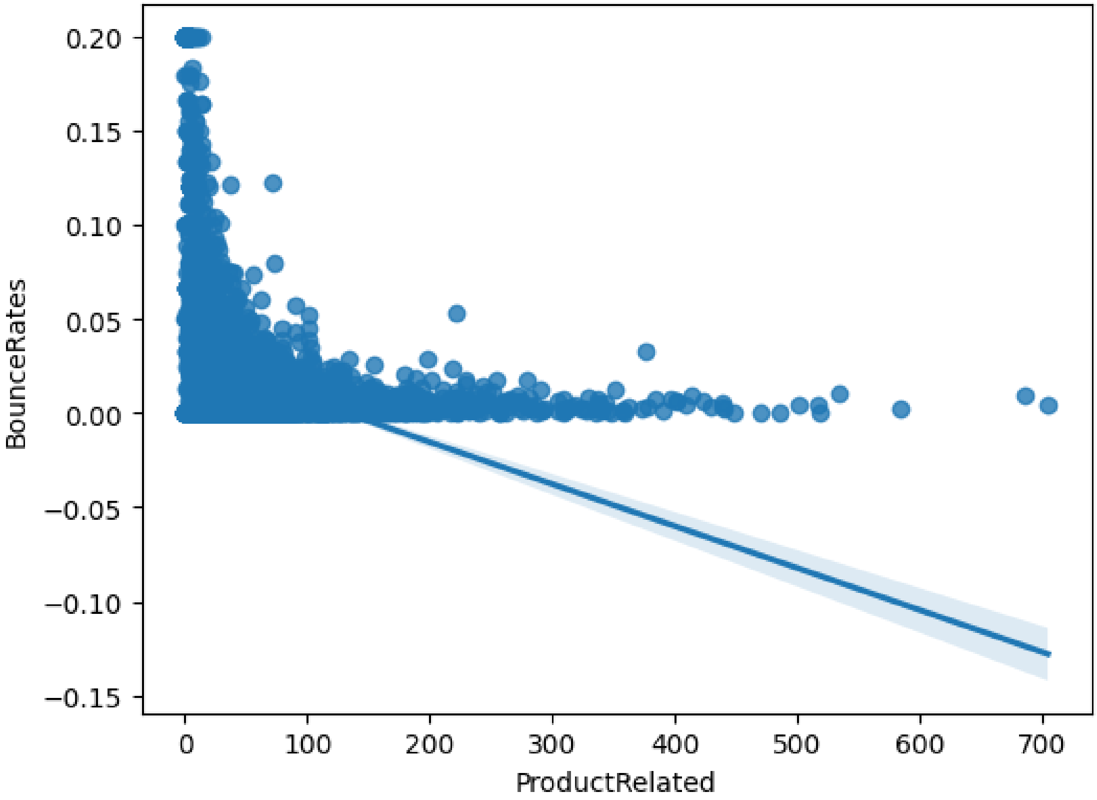
[47]:
```
sns.regplot(x='Informational', y='ExitRates', data=df1)
```

[47]:
```
<Axes: xlabel='Informational', ylabel='ExitRates'>
```



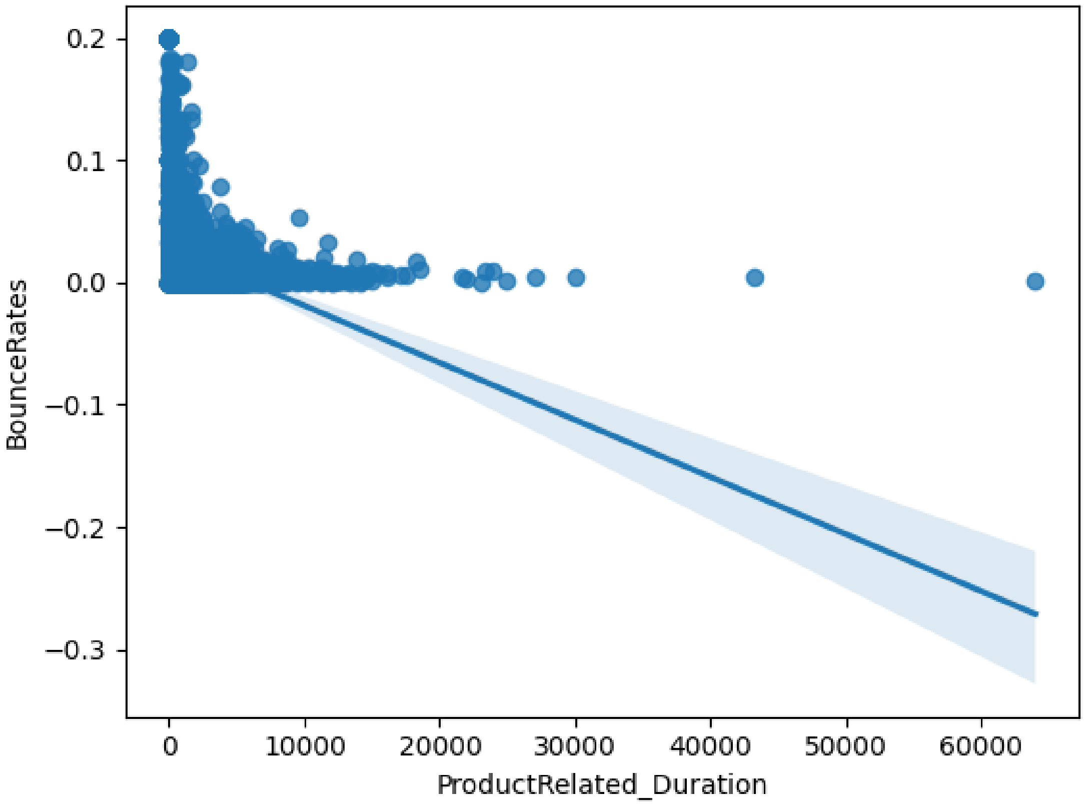
```
[48]: sns.regplot(x='ProductRelated', y='BounceRates', data=df1)
```

```
[48]: <Axes: xlabel='ProductRelated', ylabel='BounceRates'>
```



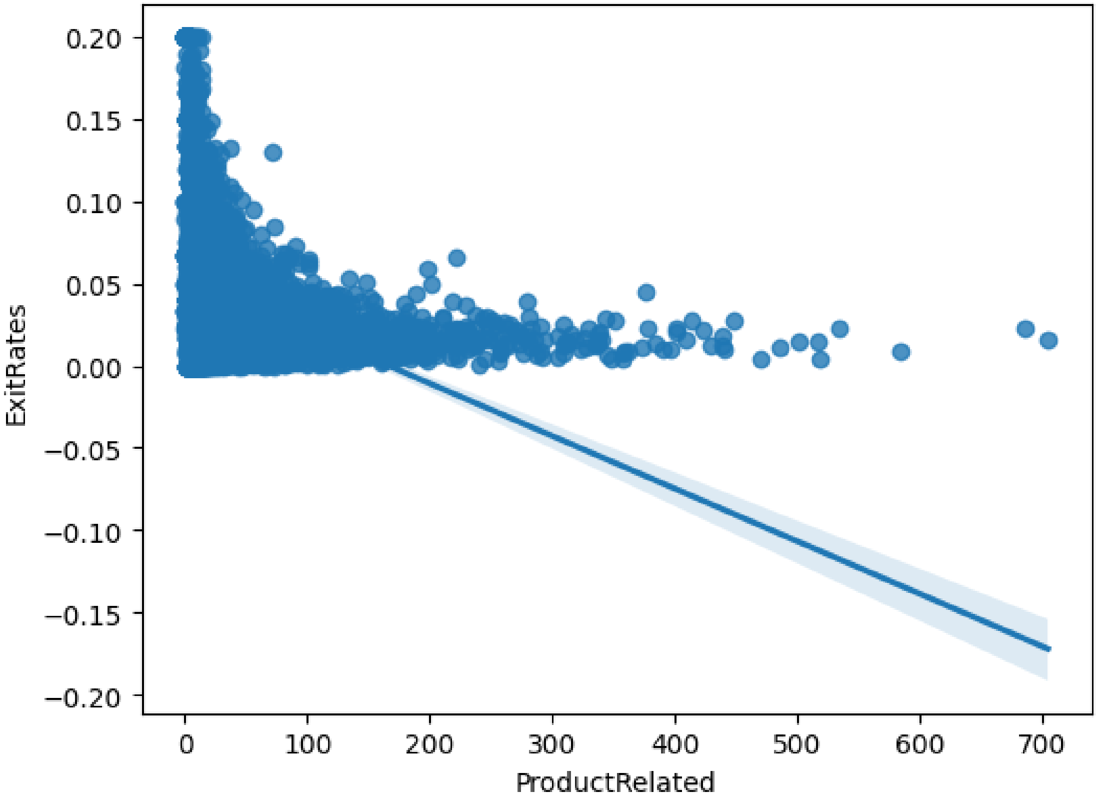
```
[49]: sns.regplot(x='ProductRelated_Duration', y='BounceRates', data=df1)
```

```
[49]: <Axes: xlabel='ProductRelated_Duration', ylabel='BounceRates'>
```



```
[50]: sns.regplot(x='ProductRelated', y='ExitRates', data=df1)
```

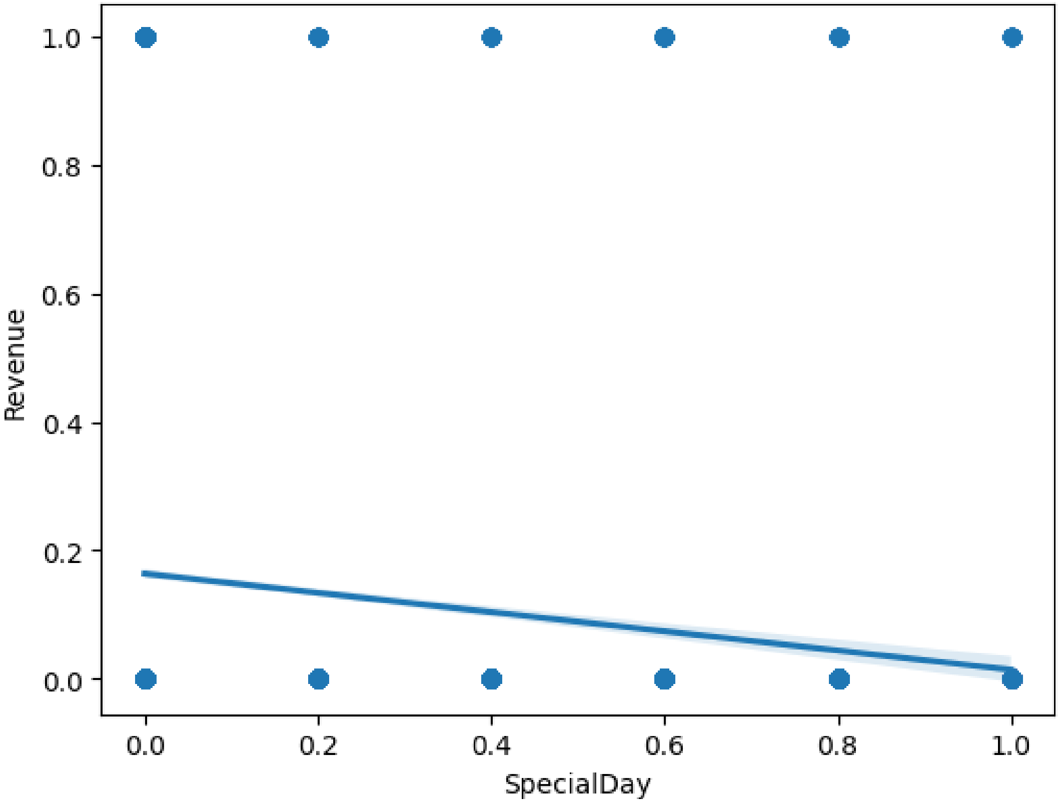
```
[50]: <Axes: xlabel='ProductRelated', ylabel='ExitRates'>
```



7 Q.6. Analyze SpecialDay distribution and its correlation with Revenue.

[51]:
```
sns.regplot(x='SpecialDay', y='Revenue', data=df1)
```

[51]:
```
<Axes: xlabel='SpecialDay', ylabel='Revenue'>
```

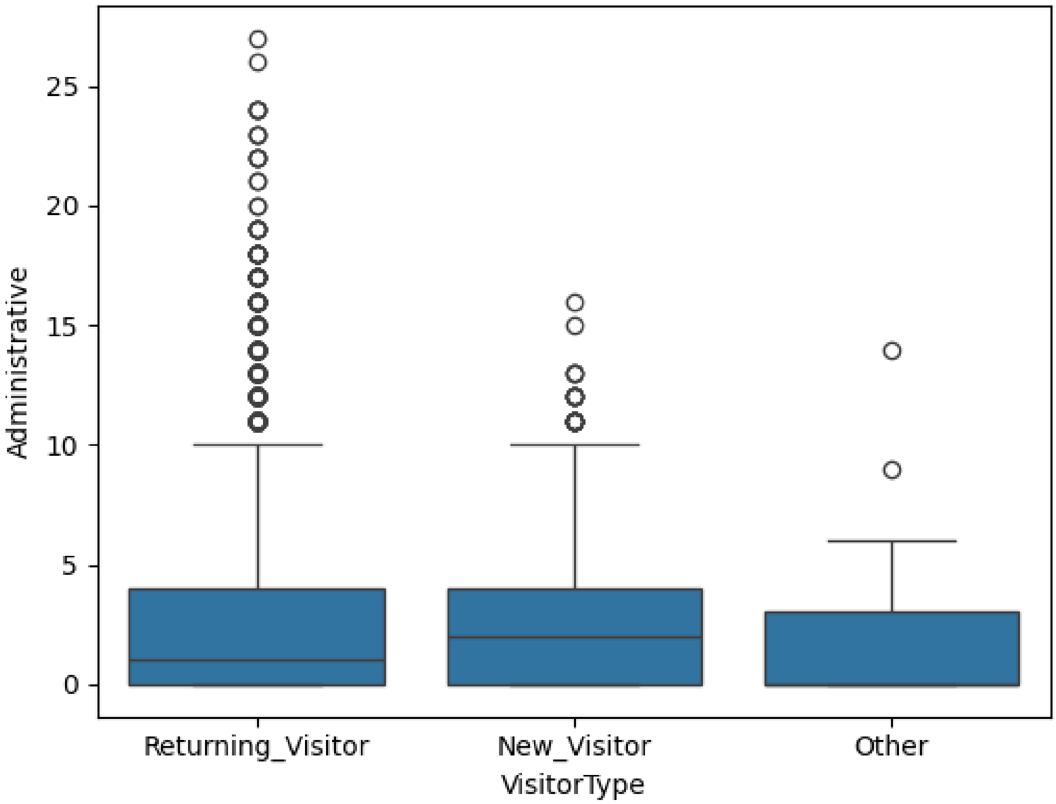


Conclusion: There is Negative Correlation B/W SpecialDay & Revenue

8 Q.7. Generate a binary feature indicating whether the user visited all three page categories.

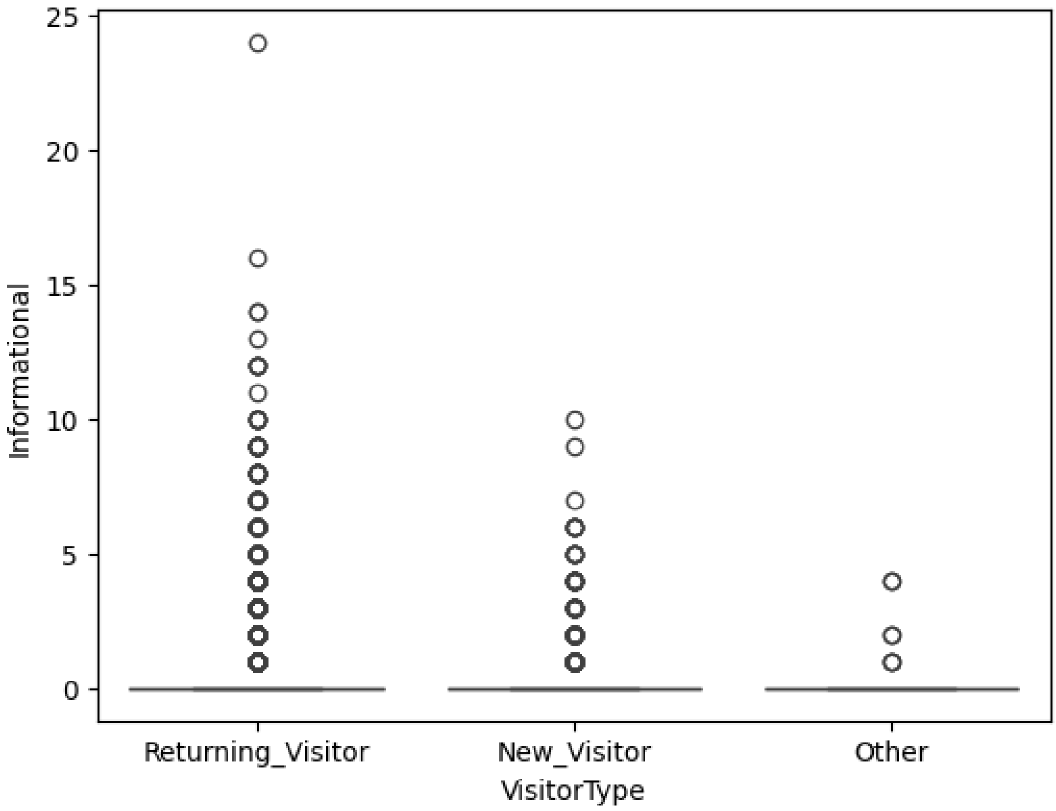
[52]:
```
sns.boxplot(x='VisitorType', y='Administrative', data=df1)
```

[52]:
```
<Axes: xlabel='VisitorType', ylabel='Administrative'>
```



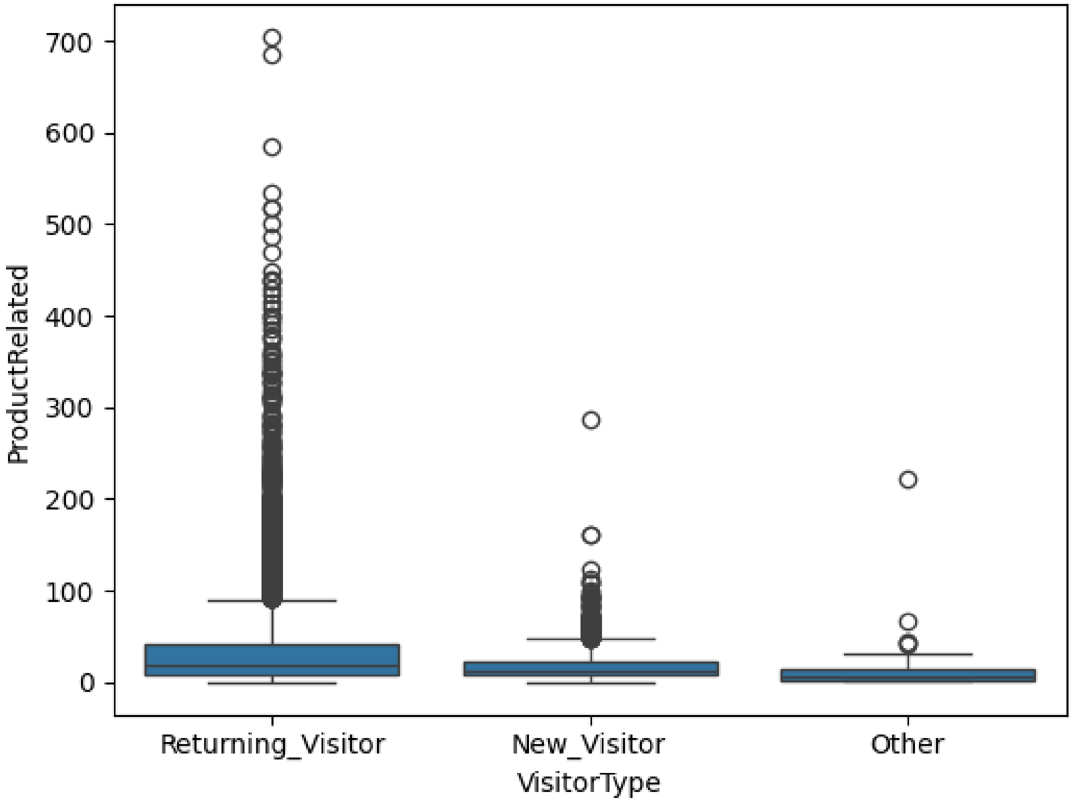
```
[53]: sns.boxplot(x='VisitorType', y='Informational', data=df1)
```

```
[53]: <Axes: xlabel='VisitorType', ylabel='Informational'>
```



```
[54]: sns.boxplot(x='VisitorType', y='ProductRelated', data=df1)
```

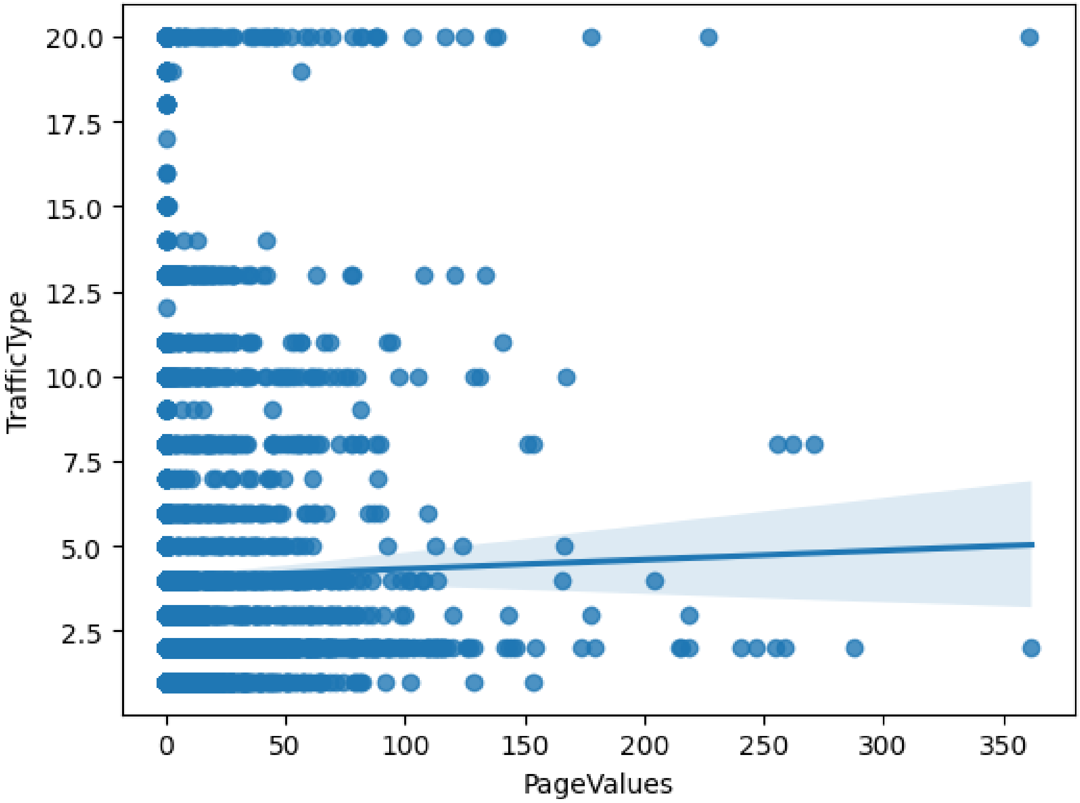
```
[54]: <Axes: xlabel='VisitorType', ylabel='ProductRelated'>
```



9 Q.8.Explore PageValues distribution and its relationship with TrafficType, VisitorType, and Region.

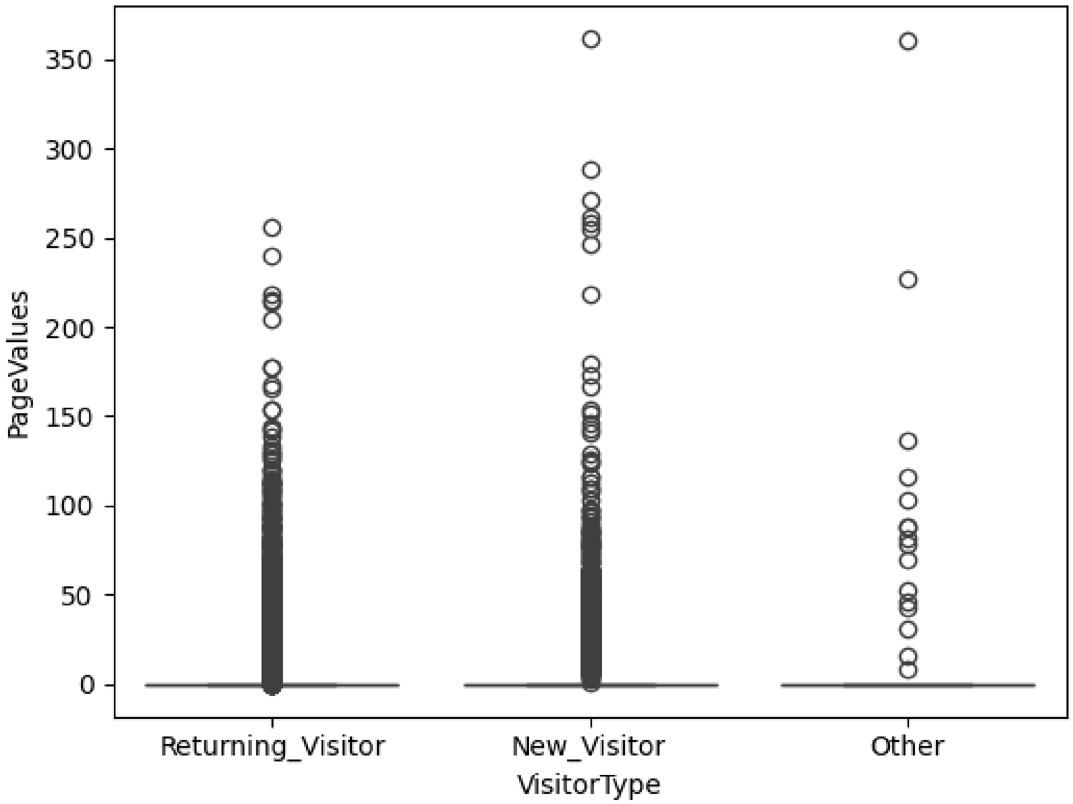
[55]:
```
sns.regplot(x='PageValues', y='TrafficType', data=df1)
```

[55]:
```
<Axes: xlabel='PageValues', ylabel='TrafficType'>
```



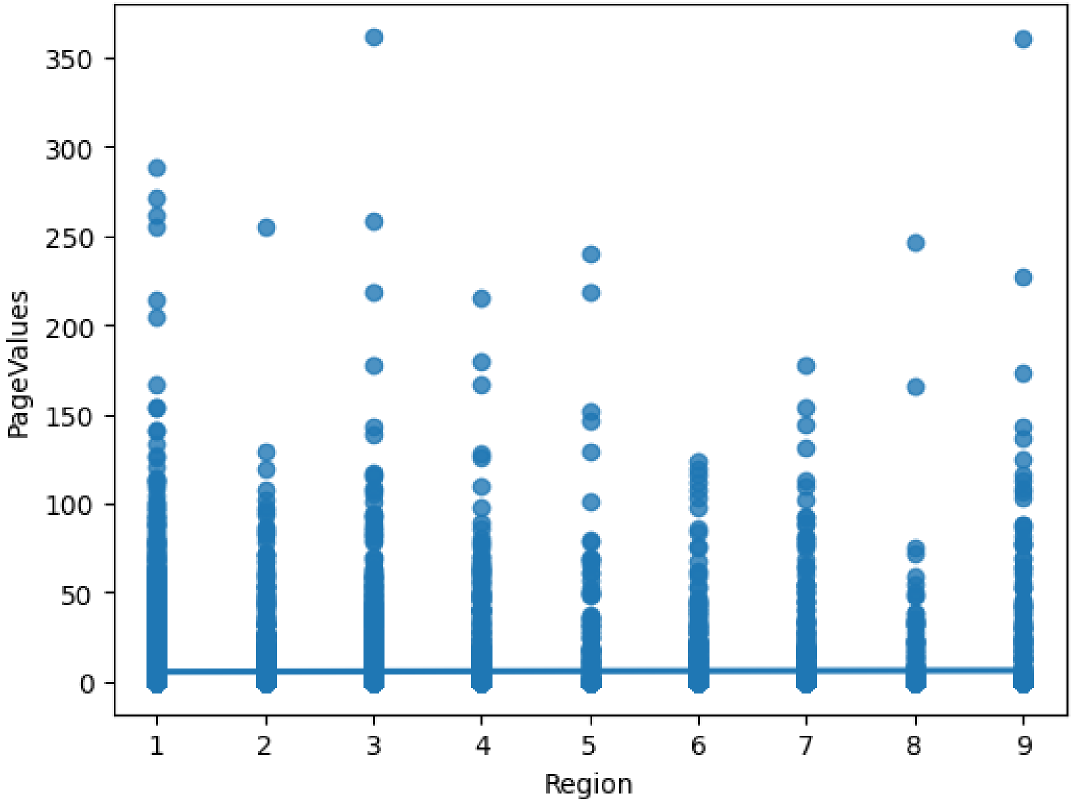
```
[56]: sns.boxplot(x='VisitorType', y='PageValues', data=df1)
```

```
[56]: <Axes: xlabel='VisitorType', ylabel='PageValues'>
```



```
[57]: sns.regplot(x='Region', y='PageValues', data=df1)
```

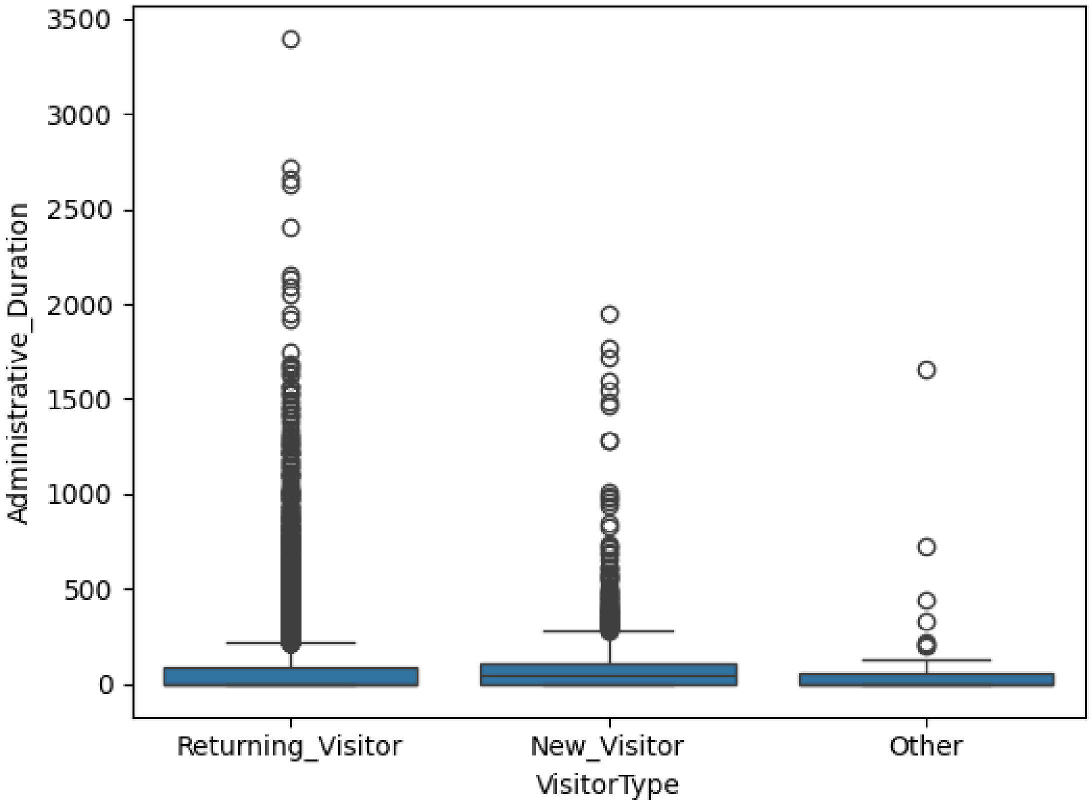
```
[57]: <Axes: xlabel='Region', ylabel='PageValues'>
```



9.1 Q.9. Investigate user session lengths and their impact on conversion rates.

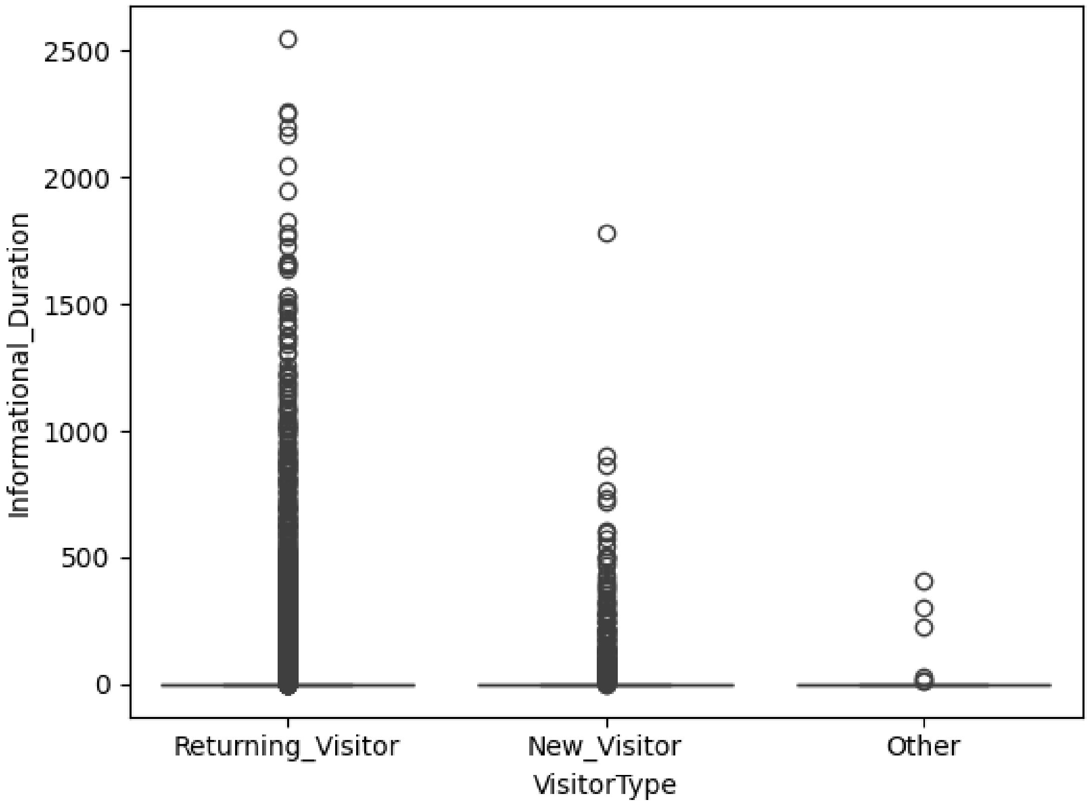
[58]:
```
sns.boxplot(x='VisitorType', y='Administrative_Duration', data=df1)
```

[58]:
```
<Axes: xlabel='VisitorType', ylabel='Administrative_Duration'>
```



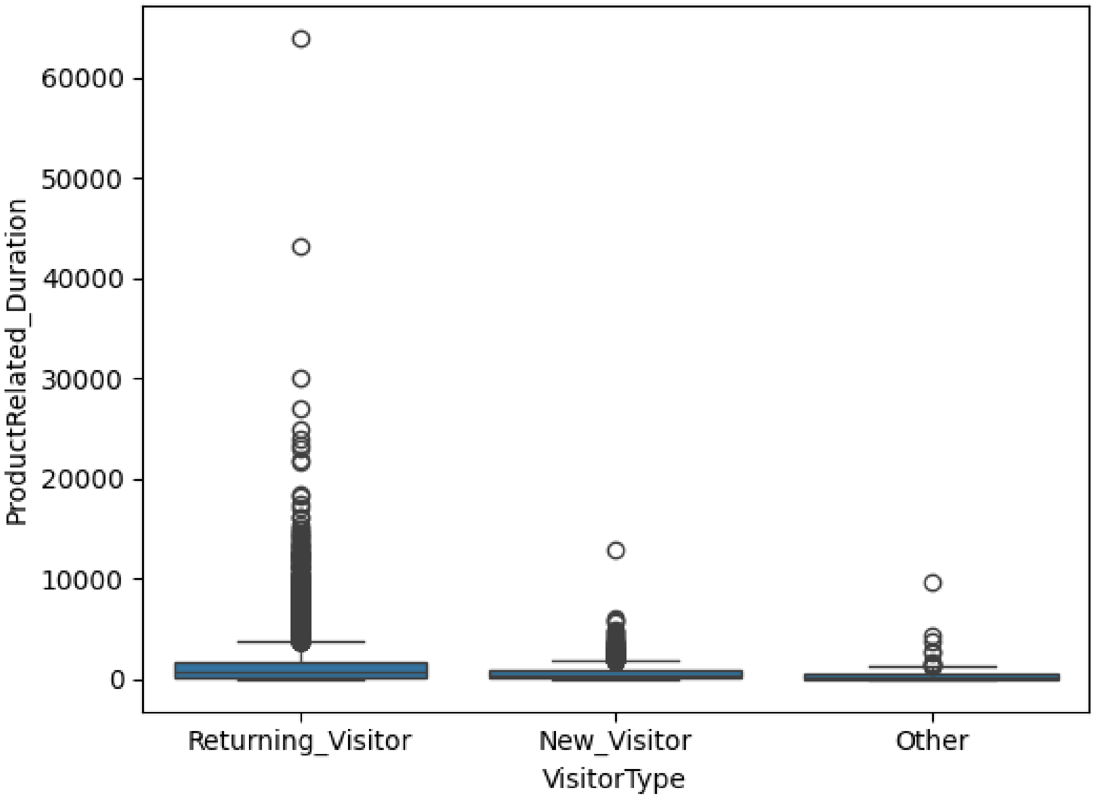
```
[59]: sns.boxplot(x='VisitorType', y='Informational_Duration', data=df1)
```

```
[59]: <Axes: xlabel='VisitorType', ylabel='Informational_Duration'>
```



```
[60]: sns.boxplot(x='VisitorType', y='ProductRelated_Duration', data=df1)
```

```
[60]: <Axes: xlabel='VisitorType', ylabel='ProductRelated_Duration'>
```



10 Q.10.Group users based on VisitorType, OperatingSystems, and Region to identify potential differences in behavior and conversion rates.

```
[61]: df1.groupby('VisitorType').size()
```

```
[61]: VisitorType
      New_Visitor           1694
      Other                   85
      Returning_Visitor    10551
      dtype: int64
```

```
[62]: df1.groupby('OperatingSystems').size()
```

```
[62]: OperatingSystems
      1    2585
      2    6601
      3    2555
      4     478
      5       6
```

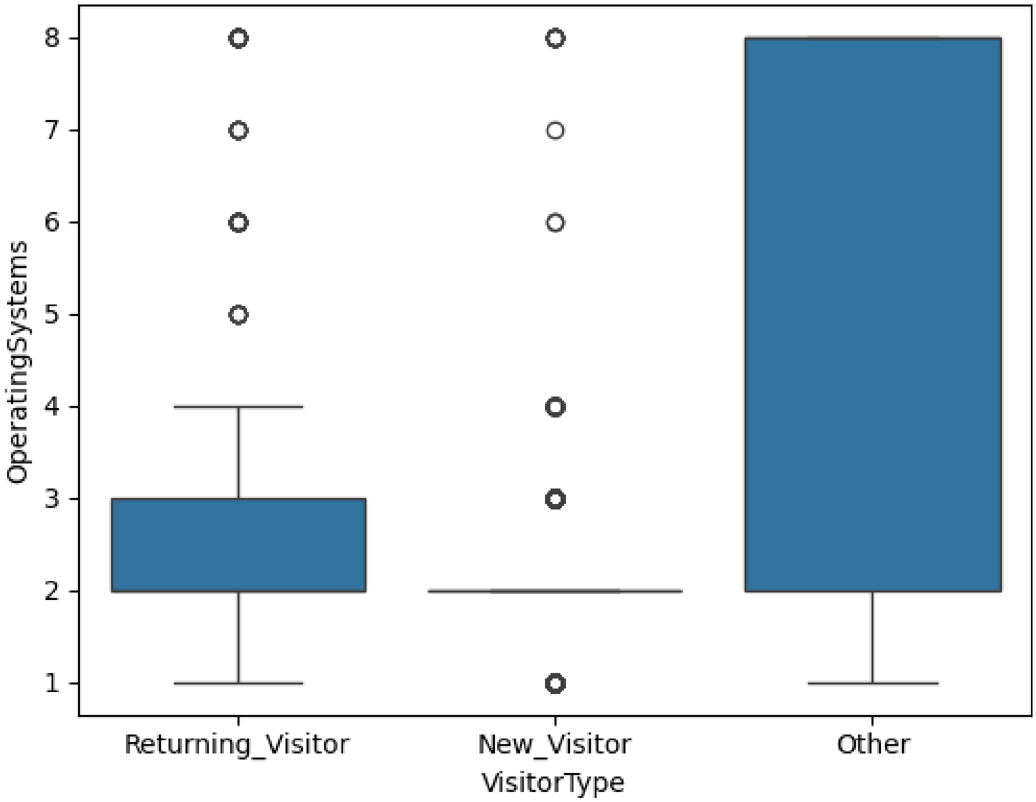
```
6      19
7       7
8      79
dtype: int64
```

```
[63]: df1.groupby('Region').size()
```

```
[63]: Region
      1    4780
      2    1136
      3    2403
      4    1182
      5     318
      6     805
      7     761
      8     434
      9     511
      dtype: int64
```

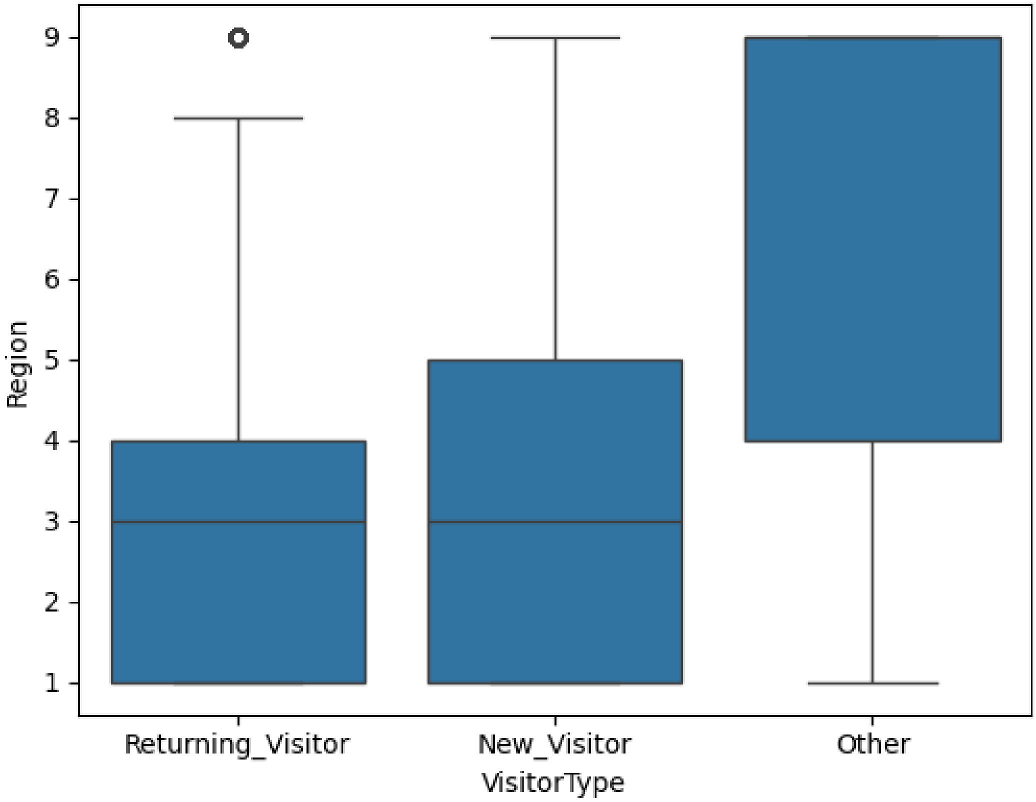
```
[64]: sns.boxplot(x='VisitorType', y='OperatingSystems', data=df1)
```

```
[64]: <Axes: xlabel='VisitorType', ylabel='OperatingSystems'>
```



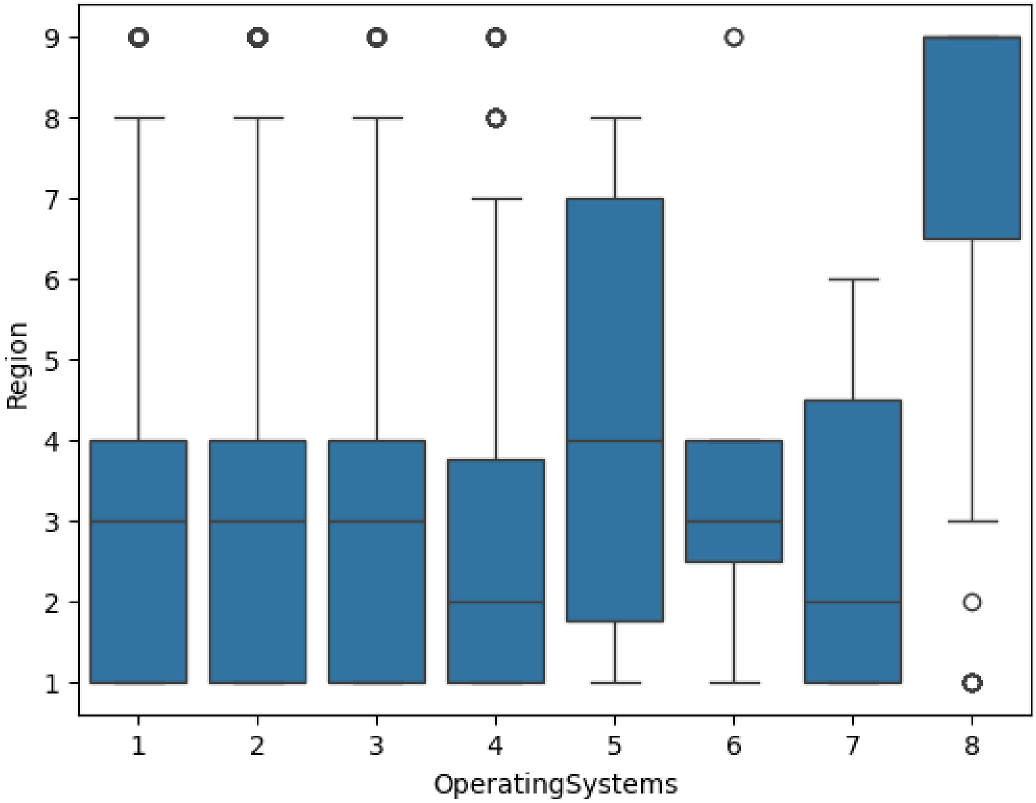
[65]:
```
sns.boxplot(x='VisitorType', y='Region', data=df1)
```

[65]:
```
<Axes: xlabel='VisitorType', ylabel='Region'>
```



[66]:
```
sns.boxplot(x='OperatingSystems', y='Region', data=df1)
```

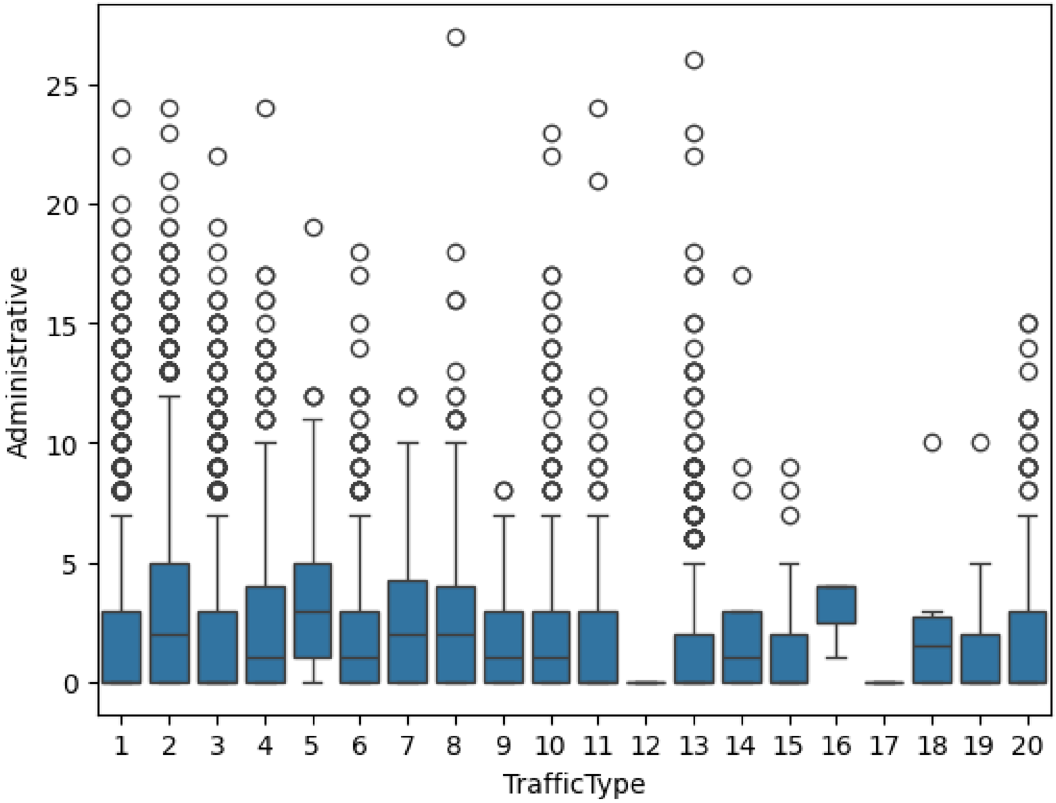
[66]:
```
<Axes: xlabel='OperatingSystems', ylabel='Region'>
```



11 Q.11. Segment users based on TrafficType and analyze their engagement patterns and purchase probability.

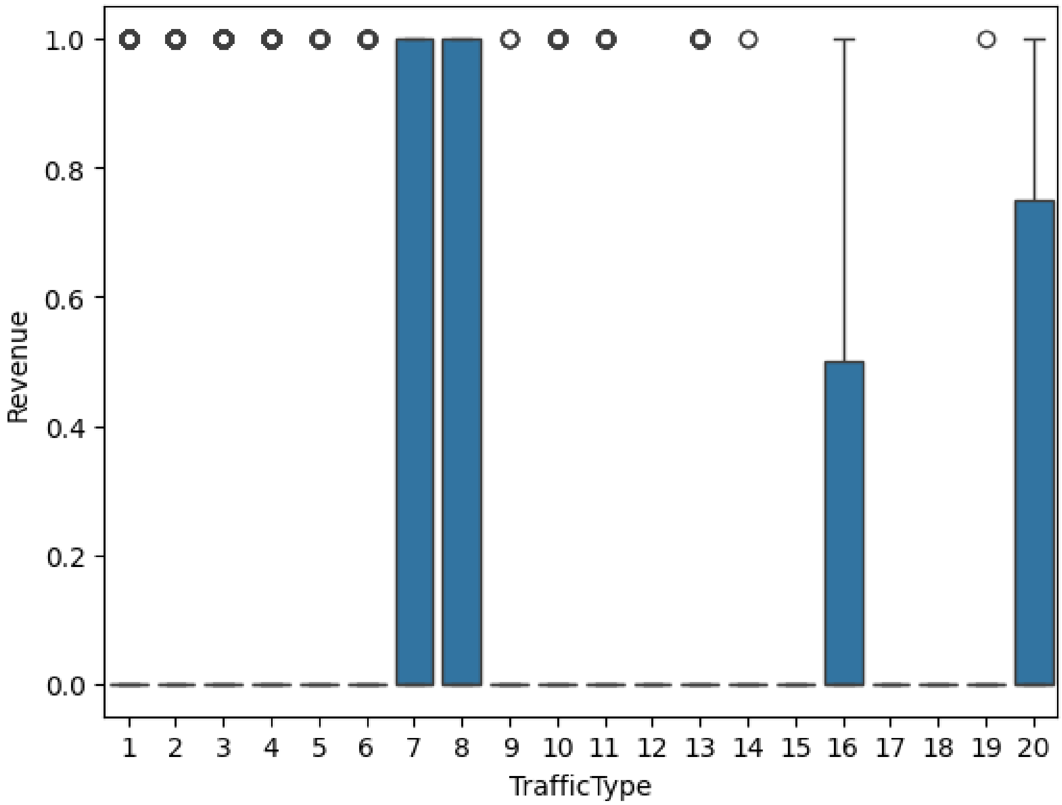
[67]:
```
sns.boxplot(x='TrafficType', y='Administrative', data=df1)
```

[67]:
```
<Axes: xlabel='TrafficType', ylabel='Administrative'>
```



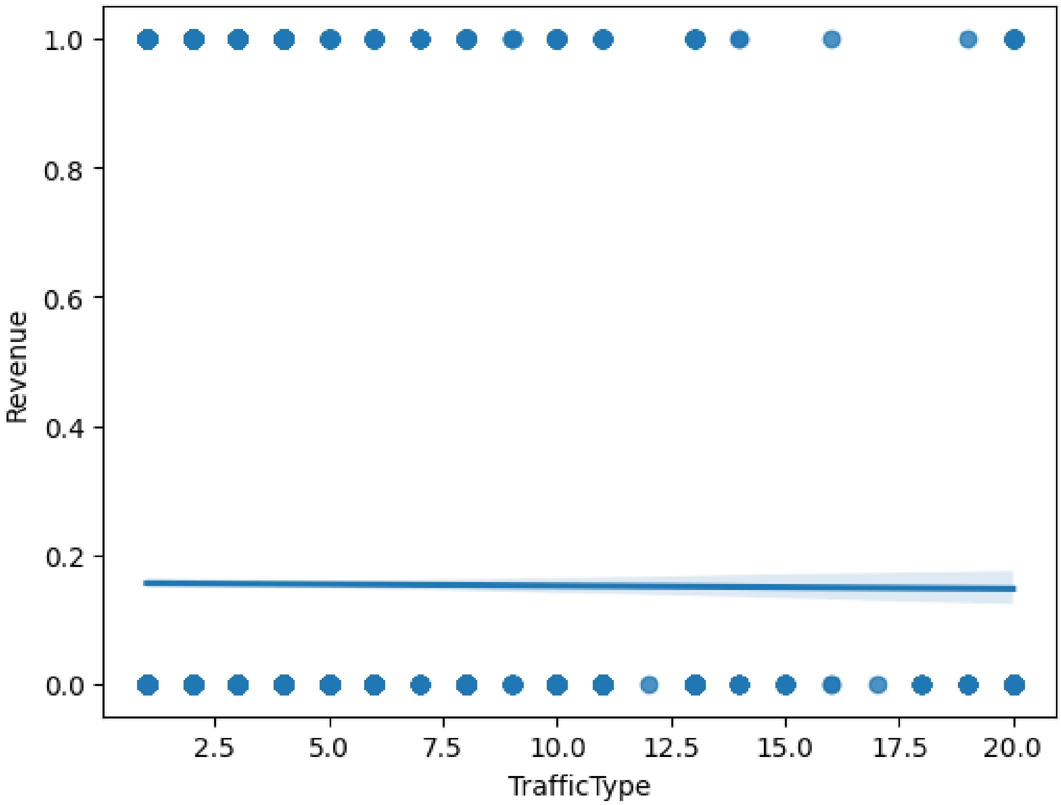
```
[68]: sns.boxplot(x='TrafficType', y='Revenue', data=df1)
```

```
[68]: <Axes: xlabel='TrafficType', ylabel='Revenue'>
```



```
[69]: sns.regplot(x='TrafficType', y='Revenue', data=df1)
```

```
[69]: <Axes: xlabel='TrafficType', ylabel='Revenue'>
```


1.0
0.8
0.6
0.4
0.2
0.0
Revenue
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
TrafficType

```
[69]:
```