1.Explore the Dataseta

Importing Libraries

In [110…
```
import numpy as np
import pandas as pd
import missingno as mp
from matplotlib import pyplot as pt
import seaborn as sns
```

Reading dataset & Loading into dataset

In [111…
```
df=pd.read_csv("UserData.csv")
df1=pd.read_csv("Opportunity Wise Data.csv")
```

Review data structure

In [112…
```
df
```

Out[112...

	PreferredSponsors	Gender	Country	Degree	Sign Up Date	
0	["GlobalShala","Grant Thornton China","Saint L...	Male	Nigeria	Undergraduate Student	2023-07-23T08:05:58.602Z	Ov
1	["GlobalShala","Grant Thornton China","Saint L...	Male	India	Undergraduate Student	2023-04-24T09:57:07.405Z	kotta
2	["GlobalShala","Illinois Institute of Technolo...	NaN	India	NaN	2022-10-14T17:13:36.303Z	
3	["GlobalShala","Grant Thornton China","Saint L...	NaN	Albania	NaN	2023-06-06T12:29:01.772Z	
4	["GlobalShala","Grant Thornton China","Saint L...	Female	Ghana	Not in Education	2023-06-15T16:31:42.719Z	Ku
...	...	...	...	...	...	
27557	["GlobalShala","Grant Thornton China","Saint L...	Female	Botswana	Undergraduate Student	2023-04-08T05:30:44.705Z	Gabo
27558	["GlobalShala","Saint Louis University","Illin...	Male	United States	Undergraduate Student	2023-02-01T20:46:32.637Z	Co
27559	["GlobalShala","Illinois Institute of Technolo...	Male	United States	High School Student	2022-09-22T14:06:56.114Z	A
27560	["GlobalShala","Grant Thornton China","Saint L...	Male	Pakistan	NaN	2023-06-16T04:18:38.811Z	Dara k
27561	["GlobalShala","Grant Thornton China","Saint L...	Male	Bangladesh	NaN	2023-05-05T04:03:14.765Z	D

27562 rows × 8 columns

In [113...

```
df1
```

Out[113...

	Profile Id	Opportunity Id	Opportunity Name	Opportunity Category	Opportunity End Date	Gender	
0	31ce84c2-2bd1-40ba-b2d8-f164fe125306	00000000-0G4F-19XB-EXPW-KS8F3N	Statement of Purpose (SOP) Writing Workshop	Event	Jan 05, 2023, 18:58:39	Male	
1	36814990-f854-4f76-8c63-91f27567d080	00000000-0G4F-19XB-EXPW-KS8F3N	Statement of Purpose (SOP) Writing Workshop	Event	Jan 05, 2023, 18:58:39	Female	
2	8154328c-f8fe-4bd1-af05-783e140f68b5	00000000-0G4F-19XB-EXPW-KS8F3N	Statement of Purpose (SOP) Writing Workshop	Event	Jan 05, 2023, 18:58:39	Female	
3	a83abad6-db1e-44c4-a8f4-9e397e282d73	00000000-0G4F-19XB-EXPW-KS8F3N	Statement of Purpose (SOP) Writing Workshop	Event	Jan 05, 2023, 18:58:39	Male	
4	c2b8a15f-2ba3-41e4-a553-7ca68b0d4a54	00000000-0G4F-19XB-EXPW-KS8F3N	Statement of Purpose (SOP) Writing Workshop	Event	Jan 05, 2023, 18:58:39	Male	
...	...	...	...	...	...	...	
20317	f386224b-4b64-4d70-a6c5-8d90e3653925	00000000-101Y-HSX2-0DFJ-QCKQBR	AI Ethics Challenge	Competition	Oct 31, 2023, 14:45:36	Male	
20318	f398b382-ac7a-4b14-8f76-cd41a51b1459	00000000-101Y-HSX2-0DFJ-QCKQBR	AI Ethics Challenge	Competition	Oct 31, 2023, 14:45:36	Male	
20319	f476e230-266d-491b-a693-f3f3bccac7d6	00000000-101Y-HSX2-0DFJ-QCKQBR	AI Ethics Challenge	Competition	Oct 31, 2023, 14:45:36	Female	N
20320	f92acfd4-3888-447a-a6dd-f996544eebbb	00000000-101Y-HSX2-0DFJ-QCKQBR	AI Ethics Challenge	Competition	Oct 31, 2023, 14:45:36	Female	
20321	fdccf84d-6011-4048-ad8d-73df5e7c431e	00000000-101Y-HSX2-0DFJ-QCKQBR	AI Ethics Challenge	Competition	Oct 31, 2023, 14:45:36	Male	

20322 rows × 21 columns

In [114…
```
df.shape
```

Out[114…
```
(27562, 8)
```

In [115…
```
df1.shape
```

Out[115…
```
(20322, 21)
```

column names

In [116…
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27562 entries, 0 to 27561
Data columns (total 8 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   PreferredSponsors  27562 non-null  object
 1   Gender             18027 non-null  object
 2   Country            27500 non-null  object
 3   Degree             16750 non-null  object
 4   Sign Up Date       27562 non-null  object
 5   city               18028 non-null  object
 6   zip                18018 non-null  object
 7   isFromSocialMedia  27553 non-null  object
dtypes: object(8)
memory usage: 1.7+ MB
```

In [117…
```
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20322 entries, 0 to 20321
Data columns (total 21 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Profile Id                20322 non-null  object
 1   Opportunity Id            20322 non-null  object
 2   Opportunity Name          20322 non-null  object
 3   Opportunity Category      20322 non-null  object
 4   Opportunity End Date      20322 non-null  object
 5   Gender                    20321 non-null  object
 6   City                      20321 non-null  object
 7   State                     20308 non-null  object
 8   Country                   20322 non-null  object
 9   Zip Code                  20309 non-null  object
 10  Graduation Date(YYYY MM)  20321 non-null  object
 11  Current Student Status    20321 non-null  object
 12  Current/Intended Major    20278 non-null  object
 13  Status Description        20322 non-null  object
 14  Apply Date                20322 non-null  object
 15  Opportunity Start Date    19518 non-null  object
 16  Reward Amount             2521 non-null   float64
 17  Badge Id                  2521 non-null   object
 18  Badge Name                2521 non-null   object
 19  Skill Points Earned       2521 non-null   float64
 20  Skills Earned             2521 non-null   object
dtypes: float64(2), object(19)
memory usage: 3.3+ MB
```

variable types

In [118…
```
df.dtypes
```

Out[118…
```
PreferredSponsors    object
Gender               object
Country              object
Degree               object
Sign Up Date         object
city                 object
zip                  object
isFromSocialMedia    object
dtype: object
```

In [119…
```
df1.dtypes
```

Out[119…

```
Profile Id                    object
Opportunity Id                object
Opportunity Name              object
Opportunity Category          object
Opportunity End Date          object
Gender                        object
City                          object
State                         object
Country                       object
Zip Code                      object
Graduation Date(YYYY MM)      object
Current Student Status        object
Current/Intended Major        object
Status Description            object
Apply Date                    object
Opportunity Start Date        object
Reward Amount                float64
Badge Id                      object
Badge Name                    object
Skill Points Earned          float64
Skills Earned                 object
dtype: object
```

Explore summary statistics

In [120…

```
df.describe()
```

Out[120…

	PreferredSponsors	Gender	Country	Degree	Sign Up Date	c
count	27562	18027	27500	16750	27562	180
unique	94	4	169	4	27561	47
top	["GlobalShala","Grant Thornton China","Saint L...	Male	India	Undergraduate Student	2022-10-30T17:25:54.072Z	Hyderab
freq	22011	11027	11893	6527	2	7

In [121…

```
df1.describe()
```

Out[121…

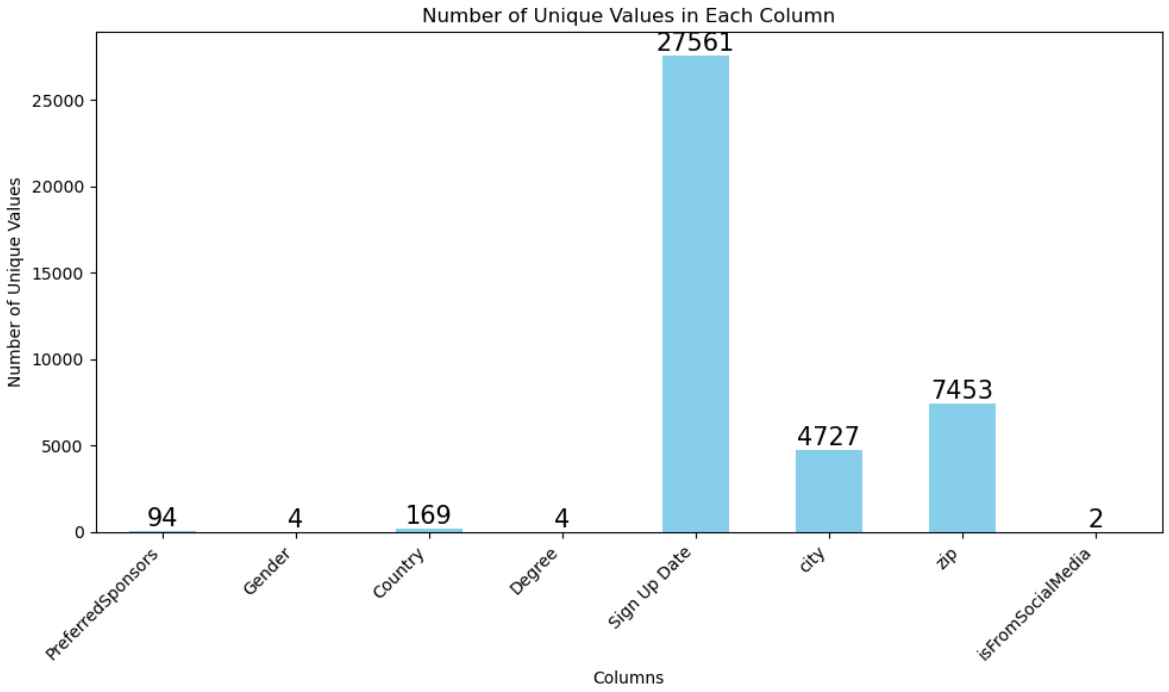
	Reward Amount	Skill Points Earned
count	2521.000000	2521.000000
mean	1081.261404	1186.964697
std	927.251398	399.172150
min	50.000000	10.000000
25%	500.000000	1182.000000
50%	500.000000	1182.000000
75%	2500.000000	1182.000000
max	2500.000000	1776.000000

Identify unique values

In [122…
```
df.nunique()
```

Out[122…
```
PreferredSponsors       94
Gender                   4
Country                169
Degree                   4
Sign Up Date         27561
city                  4727
zip                   7453
isFromSocialMedia        2
dtype: int64
```

In [123…
```
unique_value_counts1 = df.nunique()
unique_value_counts1

pt.figure(figsize=(10, 6))
ax = unique_value_counts1.plot(kind='bar', color='skyblue')
for p in ax.patches:
    ax.annotate(f'{p.get_height()}',
                (p.get_x() + p.get_width() / 2., p.get_height()),
                ha='center', va='bottom', fontsize=15, color='black')
pt.title('Number of Unique Values in Each Column')
pt.xlabel('Columns')
pt.ylabel('Number of Unique Values')
pt.xticks(rotation=45, ha='right')
pt.tight_layout()
pt.show()
```



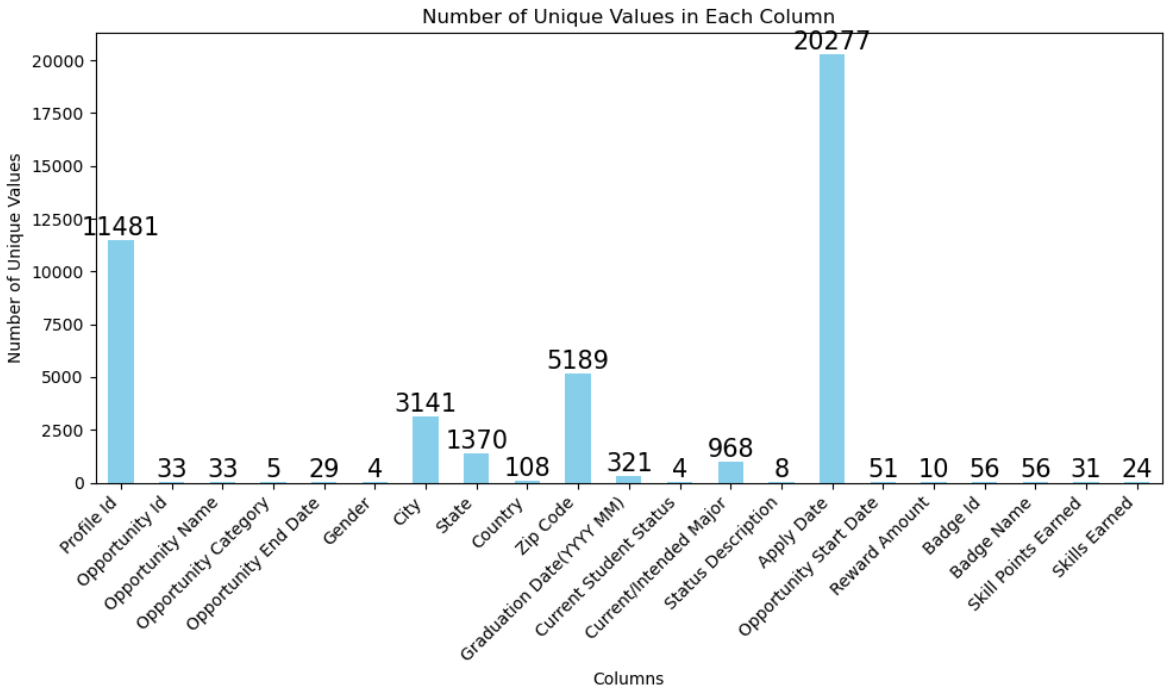
In [124…
```
df1.nunique()
```

Out[124…

```
Profile Id                    11481
Opportunity Id                   33
Opportunity Name                 33
Opportunity Category              5
Opportunity End Date             29
Gender                            4
City                           3141
State                          1370
Country                         108
Zip Code                       5189
Graduation Date(YYYY MM)        321
Current Student Status            4
Current/Intended Major          968
Status Description                8
Apply Date                    20277
Opportunity Start Date           51
Reward Amount                    10
Badge Id                         56
Badge Name                       56
Skill Points Earned              31
Skills Earned                    24
dtype: int64
```

In [125…

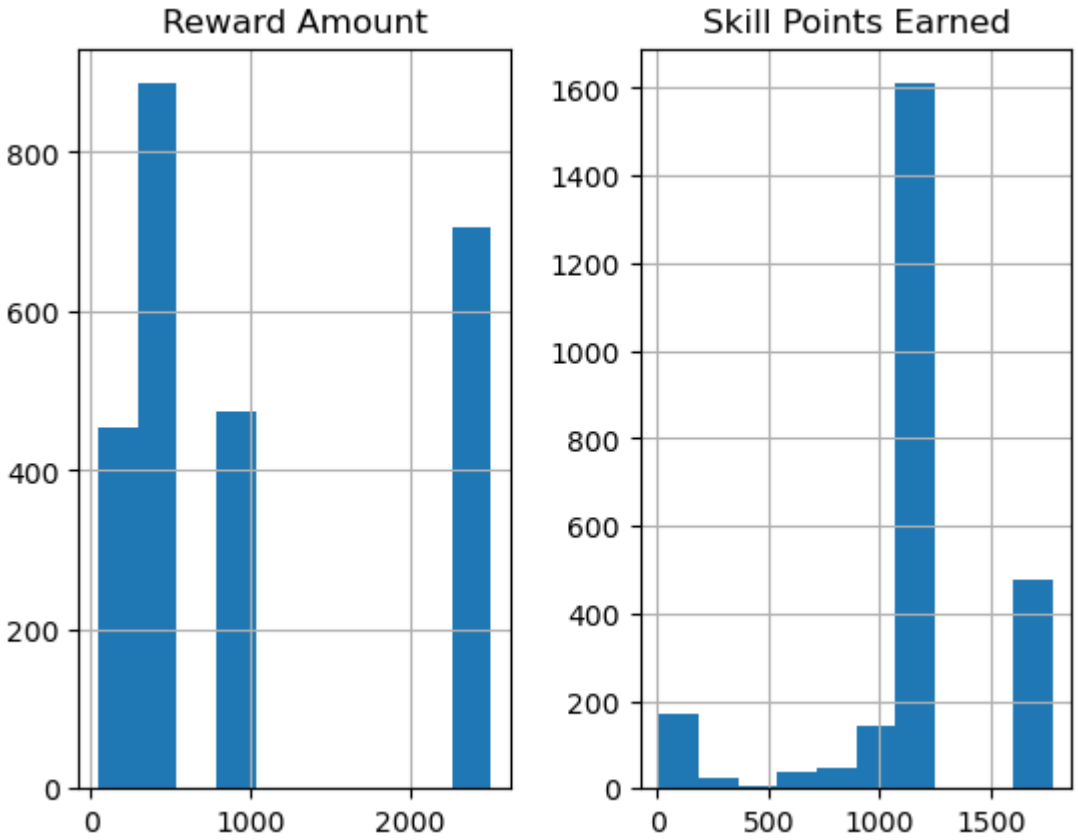
```
unique_value_counts2 = df1.nunique()
unique_value_counts2

pt.figure(figsize=(10, 6))
ax = unique_value_counts2.plot(kind='bar', color='skyblue')
for p in ax.patches:
    ax.annotate(f'{p.get_height()}',
                (p.get_x() + p.get_width() / 2., p.get_height()),
                ha='center', va='bottom', fontsize=15, color='black')
pt.title('Number of Unique Values in Each Column')
pt.xlabel('Columns')
pt.ylabel('Number of Unique Values')
pt.xticks(rotation=45, ha='right')
pt.tight_layout()
pt.show()
```



Assess data distributions

In [126…
```
#it will not work on UserData.csv as it has categorical data
#It will only work on numerical values which is present on the Opportunity Wise
df1.hist()
```

Out[126…
```
array([[<Axes: title={'center': 'Reward Amount'}>,
        <Axes: title={'center': 'Skill Points Earned'}>]], dtype=object)
```



2.Handling the Missing Values

Identify missing values using summary statistics

In [127…
```
df.isnull().sum()
```

Out[127…
```
PreferredSponsors        0
Gender                9535
Country                 62
Degree               10812
Sign Up Date             0
city                  9534
zip                   9544
isFromSocialMedia        9
dtype: int64
```

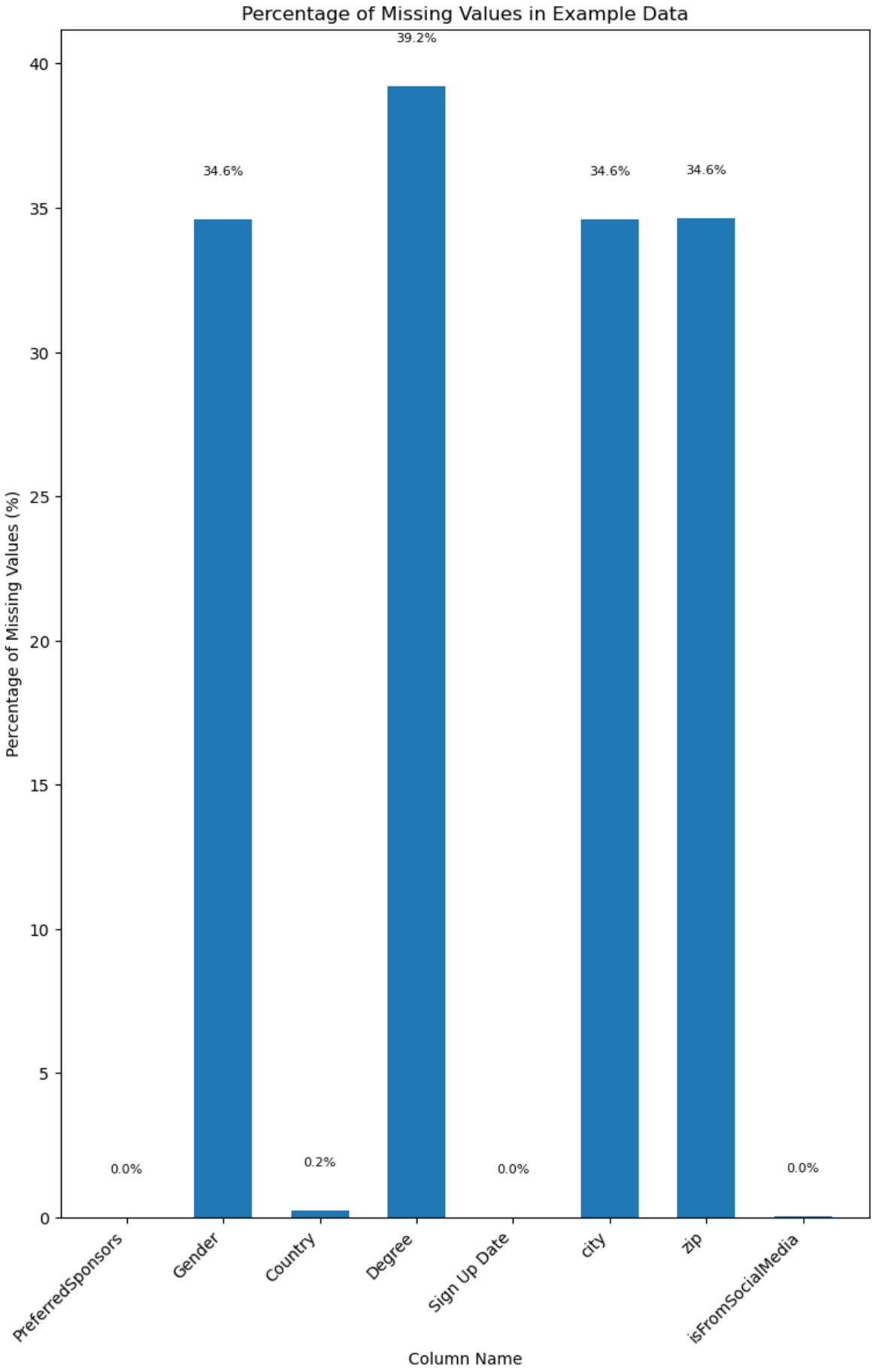
In [128…
```
df1.isnull().sum()
```

Out[128…

```
Profile Id                        0
Opportunity Id                    0
Opportunity Name                  0
Opportunity Category              0
Opportunity End Date              0
Gender                            1
City                              1
State                            14
Country                           0
Zip Code                         13
Graduation Date(YYYY MM)          1
Current Student Status            1
Current/Intended Major           44
Status Description                0
Apply Date                        0
Opportunity Start Date          804
Reward Amount                 17801
Badge Id                      17801
Badge Name                    17801
Skill Points Earned           17801
Skills Earned                 17801
dtype: int64
```

In [129…

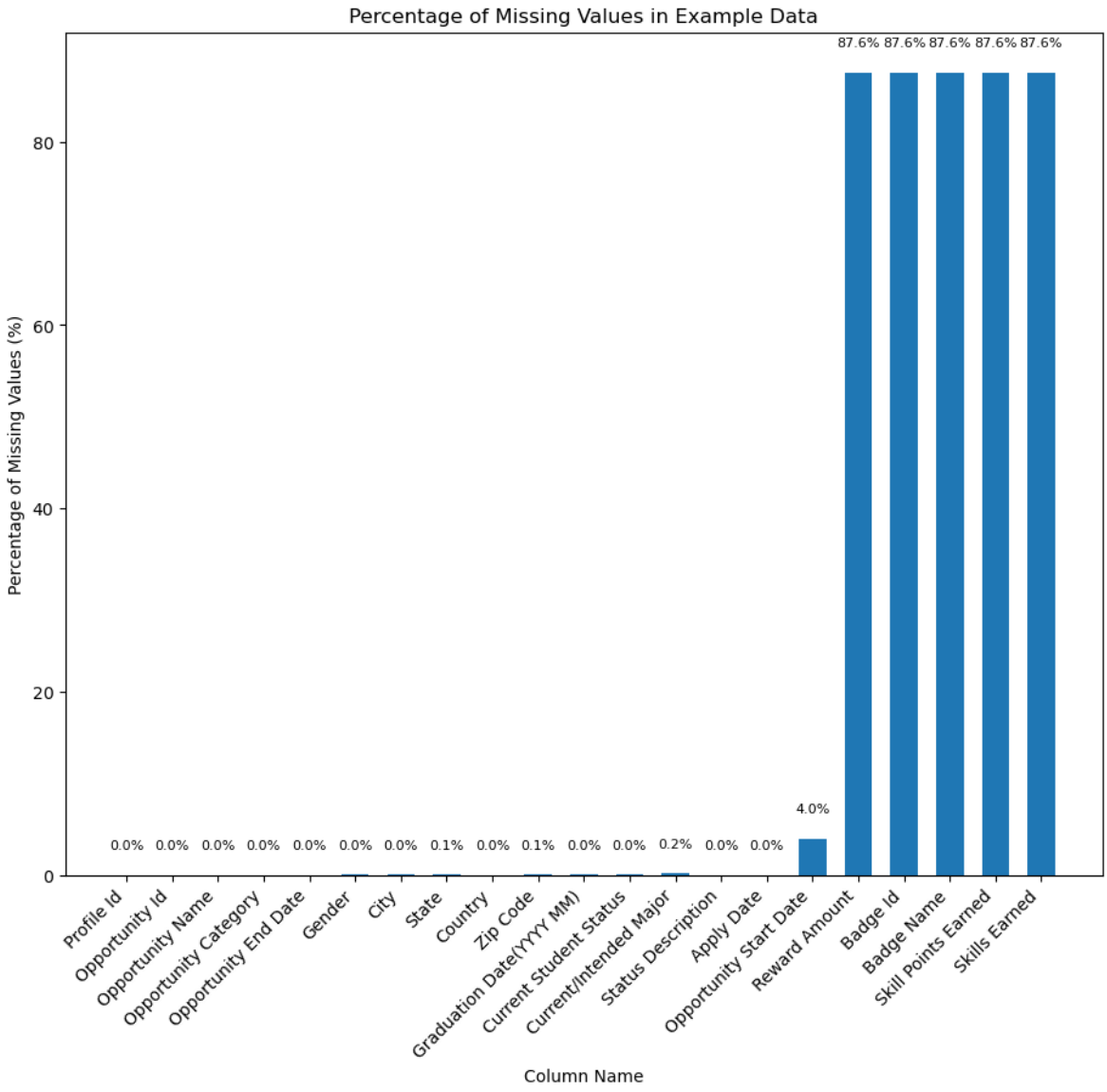
```
missing_values_percentages = (df.isnull().sum() / df.shape[0]) * 100

pt.figure(figsize=(9, 9))
pt.bar(missing_values_percentages.index, missing_values_percentages.values, widt

pt.xlabel("Column Name")
pt.ylabel("Percentage of Missing Values (%)")
pt.title("Percentage of Missing Values in Example Data")
pt.xticks(rotation=45, ha="right")
pt.subplots_adjust(top=1.25)
# pt.subplots_adjust(bottom=0.25)


for i, v in enumerate(missing_values_percentages):
    pt.text(i, v + 1, f"{v:.1f}%\n", ha="center", va="bottom", fontsize=8)

pt.show()
```



In [130…

```
missing_values_percentages = (df1.isnull().sum() / df1.shape[0]) * 100

pt.figure(figsize=(11, 9))
pt.bar(missing_values_percentages.index, missing_values_percentages.values, widt
```

```
pt.xlabel("Column Name")
pt.ylabel("Percentage of Missing Values (%)")
pt.title("Percentage of Missing Values in Example Data")
pt.xticks(rotation=45, ha="right")
# pt.subplots_adjust(top=1.0)
# pt.subplots_adjust(bottom=0.25)


for i, v in enumerate(missing_values_percentages):
    pt.text(i, v + 1, f"{v:.1f}%\n", ha="center", va="bottom", fontsize=8)

pt.show()
```



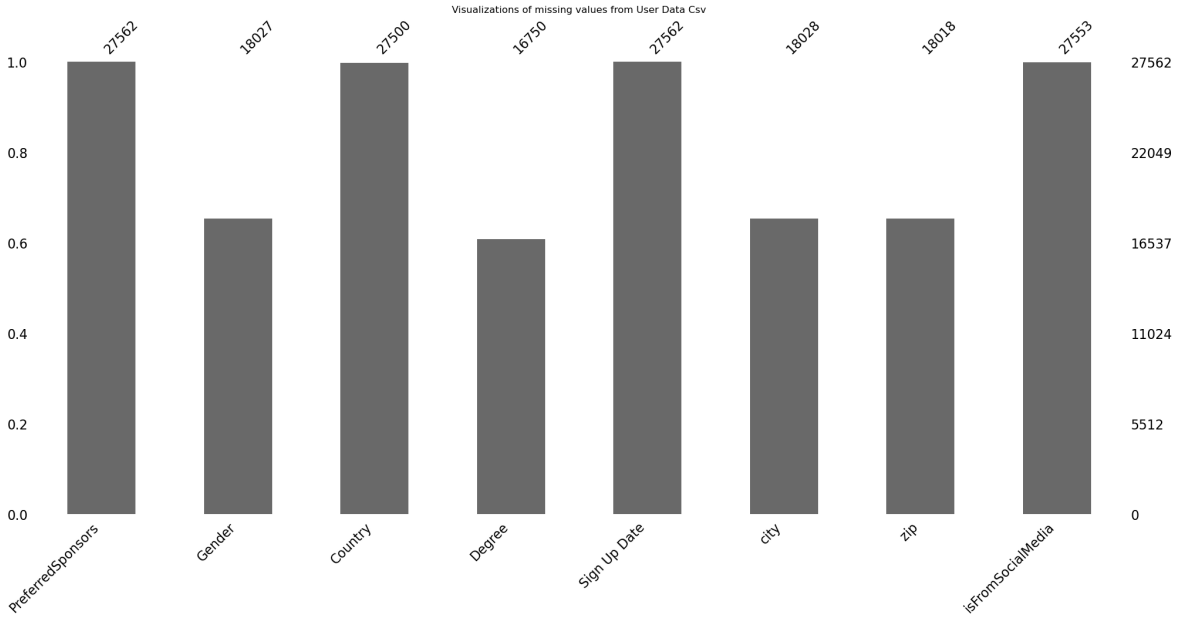
visualizations of missing values

In [131…

```
mp.bar(df)
pt.title("Visualizations of missing values from User Data Csv")
```

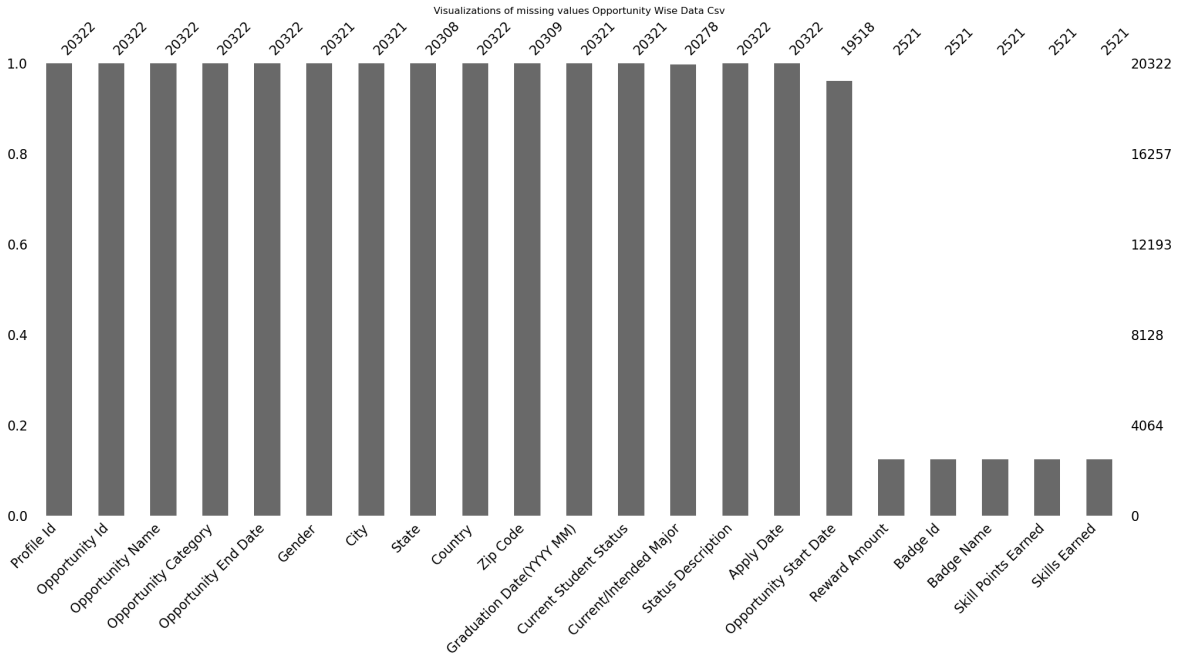
Out[131…

```
Text(0.5, 1.0, 'Visualizations of missing values from User Data Csv')
```



In [132…
```
pt.title("Visualizations of missing values Opportunity Wise Data Csv")
mp.bar(df1)
```

Out[132…
```
<Axes: title={'center': 'Visualizations of missing values Opportunity Wise Data Csv'}>
```



strategies (imputation, deletion) based on the nature and impact of missing data.

In [133…
```
# Imputation
#df.fillna(df.mean())
# Deletion
df.dropna()
```

Out[133...

	PreferredSponsors	Gender	Country	Degree	Sign Up Date	c
0	["GlobalShala","Grant Thornton China","Saint L...	Male	Nigeria	Undergraduate Student	2023-07-23T08:05:58.602Z	Owe
1	["GlobalShala","Grant Thornton China","Saint L...	Male	India	Undergraduate Student	2023-04-24T09:57:07.405Z	kottaya
4	["GlobalShala","Grant Thornton China","Saint L...	Female	Ghana	Not in Education	2023-06-15T16:31:42.719Z	Kum
8	["GlobalShala","Grant Thornton China","Saint L...	Male	Nigeria	Undergraduate Student	2023-07-27T18:02:17.535Z	Lag
9	["GlobalShala","Grant Thornton China","Saint L...	Male	India	High School Student	2023-05-05T04:47:25.446Z	R
...	...	...	...	...	...	
27555	["GlobalShala","Grant Thornton China","Saint L...	Male	India	Undergraduate Student	2023-03-31T18:01:16.166Z	Kada dis
27556	["Saint Louis University"]	Female	United States	High School Student	2023-05-16T00:34:56.486Z	N Len
27557	["GlobalShala","Grant Thornton China","Saint L...	Female	Botswana	Undergraduate Student	2023-04-08T05:30:44.705Z	Gaboro
27558	["GlobalShala","Saint Louis University","Illin...	Male	United States	Undergraduate Student	2023-02-01T20:46:32.637Z	Copp
27559	["GlobalShala","Illinois Institute of Technolo...	Male	United States	High School Student	2022-09-22T14:06:56.114Z	Aus

16618 rows × 8 columns

Handling Missing Values

In [134...
```
df.dropna(inplace=True)
```

In [135...
```
df.isnull().sum()
```

Out[135...
```
PreferredSponsors    0
Gender               0
Country              0
Degree               0
Sign Up Date         0
city                 0
zip                  0
isFromSocialMedia    0
dtype: int64
```

In [136…
```
df.shape
```

Out[136…
```
(16618, 8)
```

In [137…
```
df1.dropna(inplace=True)
```

In [138…
```
df1.isnull().sum()
```

Out[138…
```
Profile Id                  0
Opportunity Id              0
Opportunity Name            0
Opportunity Category        0
Opportunity End Date        0
Gender                      0
City                        0
State                       0
Country                     0
Zip Code                    0
Graduation Date(YYYY MM)    0
Current Student Status      0
Current/Intended Major      0
Status Description          0
Apply Date                  0
Opportunity Start Date      0
Reward Amount               0
Badge Id                    0
Badge Name                  0
Skill Points Earned         0
Skills Earned               0
dtype: int64
```

In [139…
```
df1.shape
```

Out[139…
```
(2514, 21)
```

3.Address Duplicate Data

In [140…
```
df.duplicated().sum()
```

Out[140…
```
0
```

In [141…
```
df1.duplicated().sum()
```

Out[141…
```
0
```

If there is duplicate then command to drop and check

In [142…
```
# df1.drop_duplicates(inplace=True)
# df1.shape

# df1.drop_duplicates(inplace=True)
# df1.shape
```

In [143…
```
df.nunique()
```

Out[143…
```
PreferredSponsors       91
Gender                   4
Country                129
Degree                   4
Sign Up Date         16617
city                  4359
zip                   6913
isFromSocialMedia        2
dtype: int64
```

In [144…
```
df1.nunique()
```

Out[144…
```
Profile Id                  1813
Opportunity Id                24
Opportunity Name              24
Opportunity Category           4
Opportunity End Date          20
Gender                         4
City                         833
State                        362
Country                       52
Zip Code                    1259
Graduation Date(YYYY MM)     204
Current Student Status         4
Current/Intended Major       282
Status Description             1
Apply Date                  2513
Opportunity Start Date        38
Reward Amount                 10
Badge Id                      56
Badge Name                    56
Skill Points Earned           31
Skills Earned                 24
dtype: int64
```

4. Standardize Formats:

Standardize date formats and categorical variables

In [145…
```
df['Sign Up Date'] = pd.to_datetime(df['Sign Up Date'])
```

In [146…
```
df['Sign Up Date'] = pd.to_datetime(df['Sign Up Date'], errors='coerce')
```

In [147…
```
df['Gender'] = df['Gender'].astype('category')
```

In [148…
```
duplicate_rows1 = df[df.duplicated()]
print("Duplicate Rows except first occurrence:")
print(duplicate_rows1)
```

```
Duplicate Rows except first occurrence:
Empty DataFrame
Columns: [PreferredSponsors, Gender, Country, Degree, Sign Up Date, city, zip, is
FromSocialMedia]
Index: []
```

In [149…
```
df['Degree'] = df['Degree'].astype('category')
```

In [150…
```
df['isFromSocialMedia'] = df['isFromSocialMedia'].astype(str)
```

In [151…
```
df['zip'] = pd.to_numeric(df['zip'], errors='coerce')
```

In [152…
```
df['isFromSocialMedia'] = df['isFromSocialMedia'].astype(bool)
```

In [153…
```
df.dtypes
```

Out[153…
```
PreferredSponsors                 object
Gender                          category
Country                           object
Degree                          category
Sign Up Date         datetime64[ns, UTC]
city                              object
zip                              float64
isFromSocialMedia                   bool
dtype: object
```

In [154…
```
df1['Opportunity Start Date'] = pd.to_datetime(df1['Opportunity Start Date'])
```

In [155…
```
df1['Apply Date'] = pd.to_datetime(df1['Apply Date'], errors='coerce')
```

In [156…
```
df1['Opportunity End Date'] = pd.to_datetime(df1['Opportunity End Date'], errors
```

In [157…
```
df1['Graduation Date(YYYY MM)'] = pd.to_datetime(df1['Graduation Date(YYYY MM)']
```

In [158…
```
df1['Opportunity Category'] = df1['Opportunity Category'].astype('category')
```

In [159…
```
df1['Gender'] = df1['Gender'].astype('category')
```

In [160…
```
df1['Zip Code'] = pd.to_numeric(df1['Zip Code'], errors='coerce')
```

In [161…
```
df1.dtypes
```

Out[161…
```
Profile Id                        object
Opportunity Id                    object
Opportunity Name                  object
Opportunity Category            category
Opportunity End Date      datetime64[ns]
Gender                          category
City                              object
State                             object
Country                           object
Zip Code                         float64
Graduation Date(YYYY MM)  datetime64[ns]
Current Student Status            object
Current/Intended Major            object
Status Description                object
Apply Date                datetime64[ns]
Opportunity Start Date    datetime64[ns]
Reward Amount                    float64
Badge Id                          object
Badge Name                        object
Skill Points Earned              float64
Skills Earned                     object
dtype: object
```

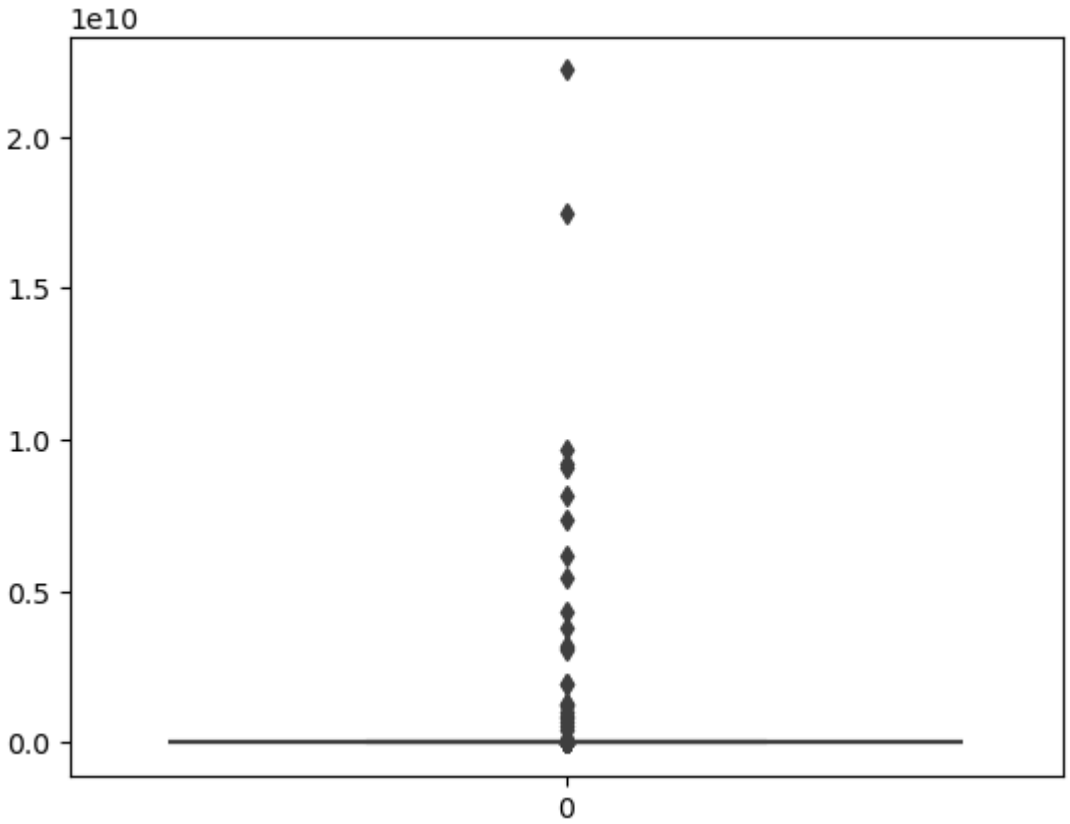
5. Validate Numeric Data:

Identify and handle outliers through statistical methods

On UserData.csv

In [162…
```
sns.boxplot(df['zip'])
```

Out[162…
```
<Axes: >
```



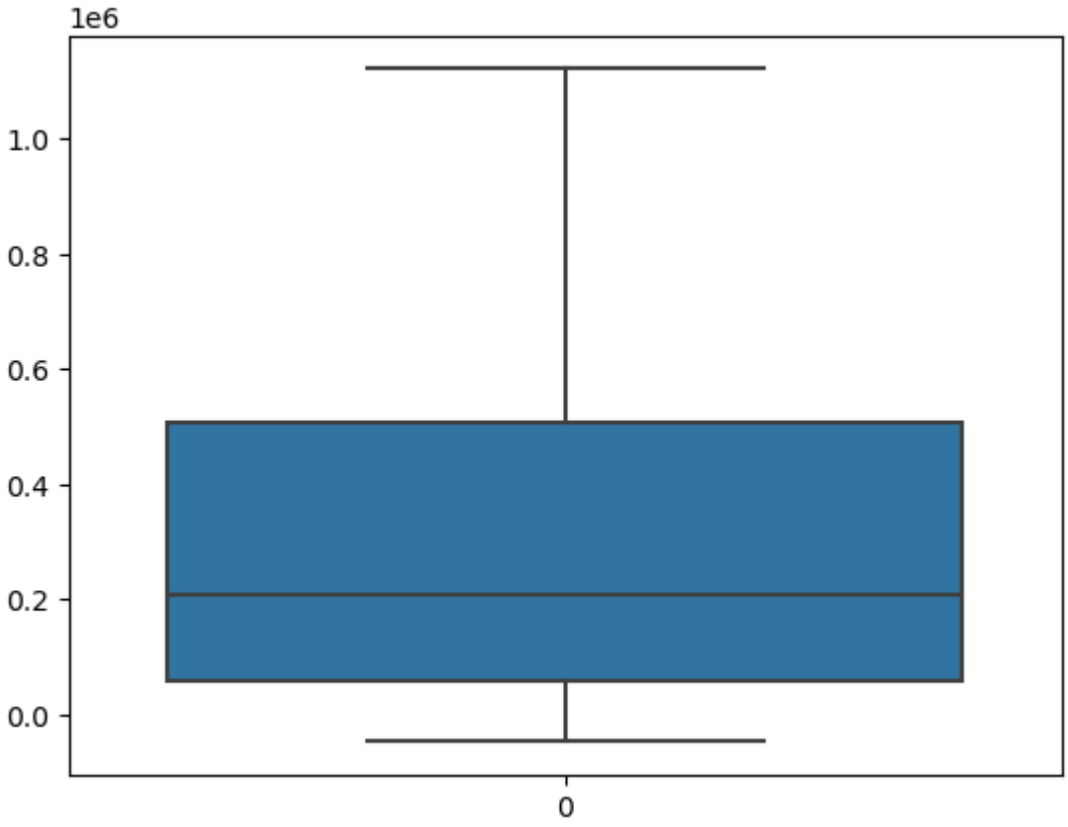
In [163…
```
Q1 = df['zip'].quantile(0.25)
Q3 = df['zip'].quantile(0.75)
IQR = Q3 - Q1

lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

outlier_indices = df[~df['zip'].between(lower_bound, upper_bound)].index

outlier_free_df = df.drop(outlier_indices)

df.drop(outlier_indices, inplace=True)
```

In [164…
```
df.shape
```

Out[164…
```
(16003, 8)
```

In [165…
```
sns.boxplot(df['zip'])
```

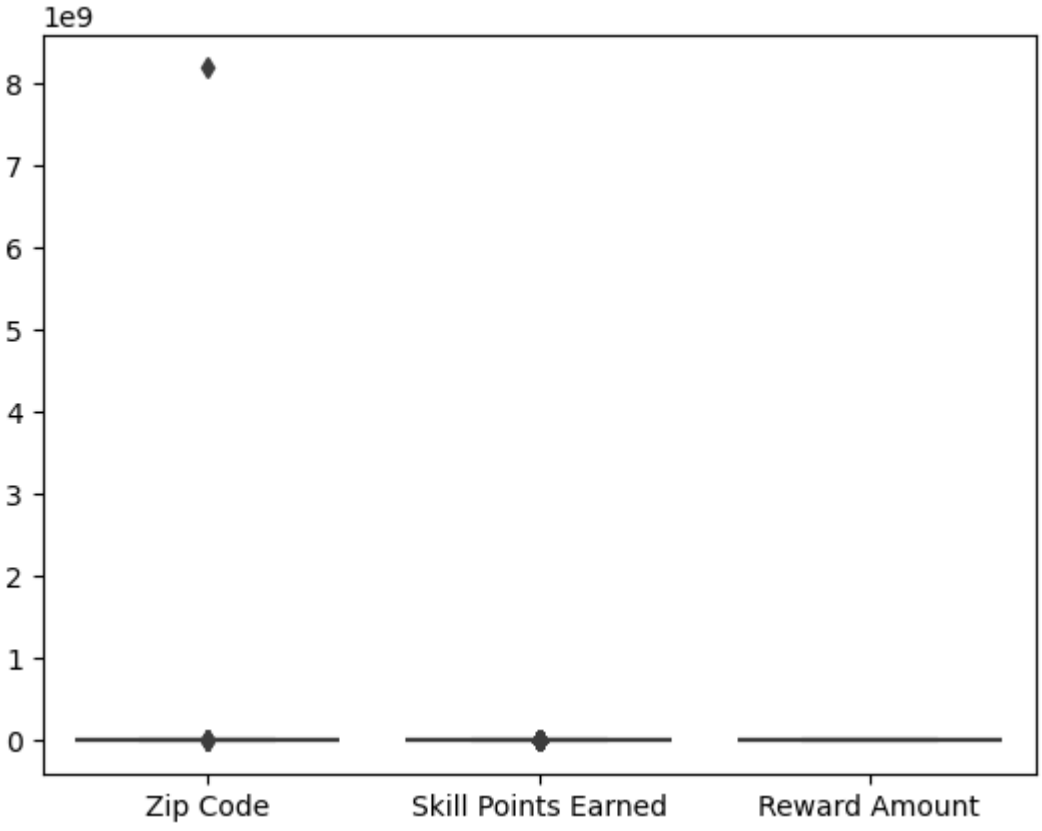
Out[165…
```
<Axes: >
```



On Opportunity Wise data.csv

In [166…
```
sns.boxplot(df1[['Zip Code','Skill Points Earned','Reward Amount']])
```

Out[166…
```
<Axes: >
```



In [167…
```
df2 = pd.DataFrame(df1)
```

```python
# Calculate IQR for each column
for column in ['Zip Code', 'Skill Points Earned', 'Reward Amount']:
    Q1 = df2[column].quantile(0.25)
    Q3 = df2[column].quantile(0.75)
    IQR = Q3 - Q1

    # Define the lower and upper bounds for outliers
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

    # Identify and filter the outliers
    outlier_indices = df2[~df2[column].between(lower_bound, upper_bound)].index

    # Drop outliers from the DataFrame
    df2.drop(outlier_indices, inplace=True)

# Display the DataFrame without outliers
print(df2)
```

	Profile Id	Opportunity Id
1378	01593fbd-baa7-4dee-8fd2-92c2c8268d67	00000000-0GNT-FT74-MZT8-93VC0G
1399	0371b96a-1f1a-488c-8a1e-a23511356837	00000000-0GNT-FT74-MZT8-93VC0G
1438	077c2a27-71f4-4ce2-a7f4-b04734631406	00000000-0GNT-FT74-MZT8-93VC0G
1464	09a609a6-684d-487b-8064-b1004cfeb7df	00000000-0GNT-FT74-MZT8-93VC0G
1542	11eae116-d2fd-4825-8227-ed1c3da2b955	00000000-0GNT-FT74-MZT8-93VC0G
...	...	...
16279	feb0d35e-228b-4686-9cd9-ff0d0da6cc1e	00000000-0GWQ-AXC5-X45C-2MHJ28
16283	fecf5e0c-5403-4fb7-ba5c-55b041208514	00000000-0GWQ-AXC5-X45C-2MHJ28
16288	ff089f51-accf-40ac-94ec-7f279ba87f2e	00000000-0GWQ-AXC5-X45C-2MHJ28
16300	ffaf1ffa-f108-47bf-9221-bb1f2e06eb97	00000000-0GWQ-AXC5-X45C-2MHJ28
16302	ffd92de8-4cf3-435a-be54-d48ca96ce1f9	00000000-0GWQ-AXC5-X45C-2MHJ28

	Opportunity Name	Opportunity Category	Opportunity End Date	Gender
1378	Digital Marketing	Internship	2024-01-01 03:30:46	Female
1399	Digital Marketing	Internship	2024-01-01 03:30:46	Female
1438	Digital Marketing	Internship	2024-01-01 03:30:46	Male
1464	Digital Marketing	Internship	2024-01-01 03:30:46	Male
1542	Digital Marketing	Internship	2024-01-01 03:30:46	Male
...	...	...	...	...
16279	Data Visualization	Internship	2024-01-01 03:30:46	Male
16283	Data Visualization	Internship	2024-01-01 03:30:46	Male
16288	Data Visualization	Internship	2024-01-01 03:30:46	Male
16300	Data Visualization	Internship	2024-01-01 03:30:46	Male
16302	Data Visualization	Internship	2024-01-01 03:30:46	Male

	City	State	Country	Zip Code	...
1378	Kadiri	Andhra Pradesh	India	515591.0	...
1399	Katsina	Katsina	Nigeria	820212.0	...
1438	Delhi	Delhi	India	110085.0	...
1464	Lahore	Punjab	Pakistan	54660.0	...
1542	Tema	WA	Ghana	233.0	...
...	...	...	...	...	...
16279	Bibiani	Western North	Ghana	233.0	...
16283	Hanamakonda	Telangana	India	506003.0	...
16288	Bantama	Ashanti	Ghana	233.0	...
16300	Erode	Tamilnadu	India	638009.0	...
16302	Hyderabad	Telangana	India	500079.0	...

	Current Student Status	Current/Intended Major	Status Description
1378	Graduate Program Student	Computer Science	Rewards Award
1399	Graduate Program Student	Mathematics	Rewards Award
1438	Undergraduate Student	Computer Science	Rewards Award
1464	Undergraduate Student	Business Administration	Rewards Award
1542	Not in Education	Digital Marketing	Rewards Award
...	...	...	...
16279	Not in Education	Public Health	Rewards Award
16283	Graduate Program Student	Artificial Intelligence	Rewards Award
16288	Undergraduate Student	Computer Science	Rewards Award
16300	Undergraduate Student	Computer Science	Rewards Award
16302	Graduate Program Student	Computer Science	Rewards Award

	Apply Date	Opportunity Start Date	Reward Amount
1378	2023-05-10 13:58:23	2023-05-25 02:30:00	2500.0
1399	2023-07-09 13:28:46	2023-07-24 02:30:00	500.0
1438	2023-03-12 03:52:13	2023-03-27 02:30:48	2500.0
1464	2023-06-14 20:08:19	2023-06-26 02:30:00	500.0
1542	2023-07-03 16:28:29	2023-07-24 02:30:00	500.0
...	...	...	...
16279	2023-06-16 10:14:00	2023-07-24 02:30:00	500.0

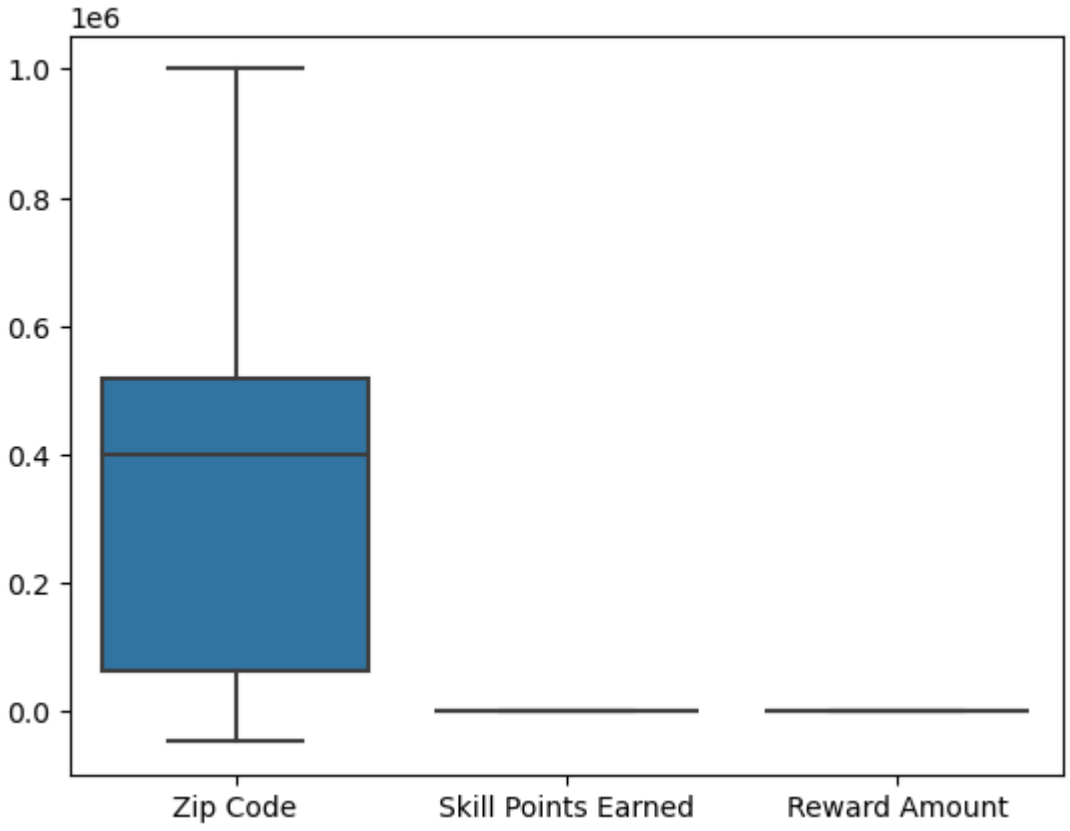
```
16283 2023-06-10 04:29:28    2023-06-26 02:30:00         500.0
16288 2023-05-23 04:41:29    2023-06-12 02:30:00        2500.0
16300 2023-07-24 02:12:27    2023-08-07 04:30:00         500.0
16302 2023-05-30 15:59:00    2023-06-12 02:30:00        2500.0

                             Badge Id  \
1378   00000000-0GFK-A0AE-P6B7-BQBNTK
1399   00000000-107V-NP8K-V0ZN-Z9ZEF5
1438   00000000-0GFK-A0AE-P6B7-BQBNTK
1464   00000000-107V-NP8K-V0ZN-Z9ZEF5
1542   00000000-107V-NP8K-V0ZN-Z9ZEF5
...                               ...
16279  00000000-10GX-D9CF-HZB9-KBFXGW
16283  00000000-10GX-D9CF-HZB9-KBFXGW
16288  00000000-0GGF-GHE1-MRMC-1B98GC
16300  00000000-10GX-D9CF-HZB9-KBFXGW
16302  00000000-0GGF-GHE1-MRMC-1B98GC

                                          Badge Name Skill Points Earned  \
1378                                 Digital Marketing              1182.0
1399    Digital Marketing Virtual Internship Completed              1182.0
1438                                 Digital Marketing              1182.0
1464    Digital Marketing Virtual Internship Completed              1182.0
1542    Digital Marketing Virtual Internship Completed              1182.0
...                                                ...                 ...
16279  Data Visualization Virtual Internship Completed              1182.0
16283  Data Visualization Virtual Internship Completed              1182.0
16288                               Data Visualization              1182.0
16300  Data Visualization Virtual Internship Completed              1182.0
16302                               Data Visualization              1182.0

                                        Skills Earned
1378   ["Critical Thinking","Creative Thinking","Coll...
1399   ["Critical Thinking","Creative Thinking","Coll...
1438   ["Critical Thinking","Creative Thinking","Coll...
1464   ["Critical Thinking","Creative Thinking","Coll...
1542   ["Critical Thinking","Creative Thinking","Coll...
...                                                  ...
16279  ["Critical Thinking","Creative Thinking","Coll...
16283  ["Critical Thinking","Creative Thinking","Coll...
16288  ["Critical Thinking","Creative Thinking","Coll...
16300  ["Critical Thinking","Creative Thinking","Coll...
16302  ["Critical Thinking","Creative Thinking","Coll...

[1467 rows x 21 columns]
```

In [168…
```
df2.shape
```

Out[168…
```
(1467, 21)
```

In [169…
```
sns.boxplot(df2[['Zip Code','Skill Points Earned','Reward Amount']])
```

Out[169…
```
<Axes: >
```



6. Validate Categorical Data:

In [170…
```
# **On User Data.csv**
df.nunique()
```

Out[170…
```
PreferredSponsors       89
Gender                   4
Country                117
Degree                   4
Sign Up Date         16002
city                  4155
zip                   6355
isFromSocialMedia        1
dtype: int64
```

In [171…
```
# **On Opportunity Wise Data.csv**
df1.nunique()
```

Out[171…

```
Profile Id                    1813
Opportunity Id                  24
Opportunity Name                24
Opportunity Category             4
Opportunity End Date            20
Gender                           4
City                           833
State                          362
Country                         52
Zip Code                      1218
Graduation Date(YYYY MM)       204
Current Student Status           4
Current/Intended Major         282
Status Description               1
Apply Date                    2459
Opportunity Start Date          38
Reward Amount                   10
Badge Id                        56
Badge Name                      56
Skill Points Earned             31
Skills Earned                   24
dtype: int64
```

In [172…

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 16003 entries, 0 to 27559
Data columns (total 8 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   PreferredSponsors  16003 non-null  object
 1   Gender             16003 non-null  category
 2   Country            16003 non-null  object
 3   Degree             16003 non-null  category
 4   Sign Up Date       16003 non-null  datetime64[ns, UTC]
 5   city               16003 non-null  object
 6   zip                16003 non-null  float64
 7   isFromSocialMedia  16003 non-null  bool
dtypes: bool(1), category(2), datetime64[ns, UTC](1), float64(1), object(3)
memory usage: 1.3+ MB
```

In [173…

```
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 2514 entries, 1 to 20061
Data columns (total 21 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Profile Id                2514 non-null   object
 1   Opportunity Id            2514 non-null   object
 2   Opportunity Name          2514 non-null   object
 3   Opportunity Category      2514 non-null   category
 4   Opportunity End Date      2514 non-null   datetime64[ns]
 5   Gender                    2514 non-null   category
 6   City                      2514 non-null   object
 7   State                     2514 non-null   object
 8   Country                   2514 non-null   object
 9   Zip Code                  2453 non-null   float64
 10  Graduation Date(YYYY MM)  2514 non-null   datetime64[ns]
 11  Current Student Status    2514 non-null   object
 12  Current/Intended Major    2514 non-null   object
 13  Status Description        2514 non-null   object
 14  Apply Date                2460 non-null   datetime64[ns]
 15  Opportunity Start Date    2514 non-null   datetime64[ns]
 16  Reward Amount             2514 non-null   float64
 17  Badge Id                  2514 non-null   object
 18  Badge Name                2514 non-null   object
 19  Skill Points Earned       2514 non-null   float64
 20  Skills Earned             2514 non-null   object
dtypes: category(2), datetime64[ns](4), float64(3), object(12)
memory usage: 462.7+ KB
```

In [174…

```
df.dtypes
```

Out[174…

```
PreferredSponsors                 object
Gender                          category
Country                           object
Degree                          category
Sign Up Date         datetime64[ns, UTC]
city                              object
zip                              float64
isFromSocialMedia                   bool
dtype: object
```

In [175…

```
df1.dtypes
```

Out[175…
```
Profile Id                          object
Opportunity Id                      object
Opportunity Name                    object
Opportunity Category              category
Opportunity End Date        datetime64[ns]
Gender                            category
City                                object
State                               object
Country                             object
Zip Code                           float64
Graduation Date(YYYY MM)    datetime64[ns]
Current Student Status              object
Current/Intended Major              object
Status Description                  object
Apply Date                  datetime64[ns]
Opportunity Start Date      datetime64[ns]
Reward Amount                      float64
Badge Id                            object
Badge Name                          object
Skill Points Earned                float64
Skills Earned                       object
dtype: object
```

In [176…
```
df.isnull().sum()
```

Out[176…
```
PreferredSponsors    0
Gender               0
Country              0
Degree               0
Sign Up Date         0
city                 0
zip                  0
isFromSocialMedia    0
dtype: int64
```

In [177…
```
df1.isnull().sum()
```

Out[177…
```
Profile Id                   0
Opportunity Id               0
Opportunity Name             0
Opportunity Category         0
Opportunity End Date         0
Gender                       0
City                         0
State                        0
Country                      0
Zip Code                    61
Graduation Date(YYYY MM)     0
Current Student Status       0
Current/Intended Major       0
Status Description           0
Apply Date                  54
Opportunity Start Date       0
Reward Amount                0
Badge Id                     0
Badge Name                   0
Skill Points Earned          0
Skills Earned                0
dtype: int64
```

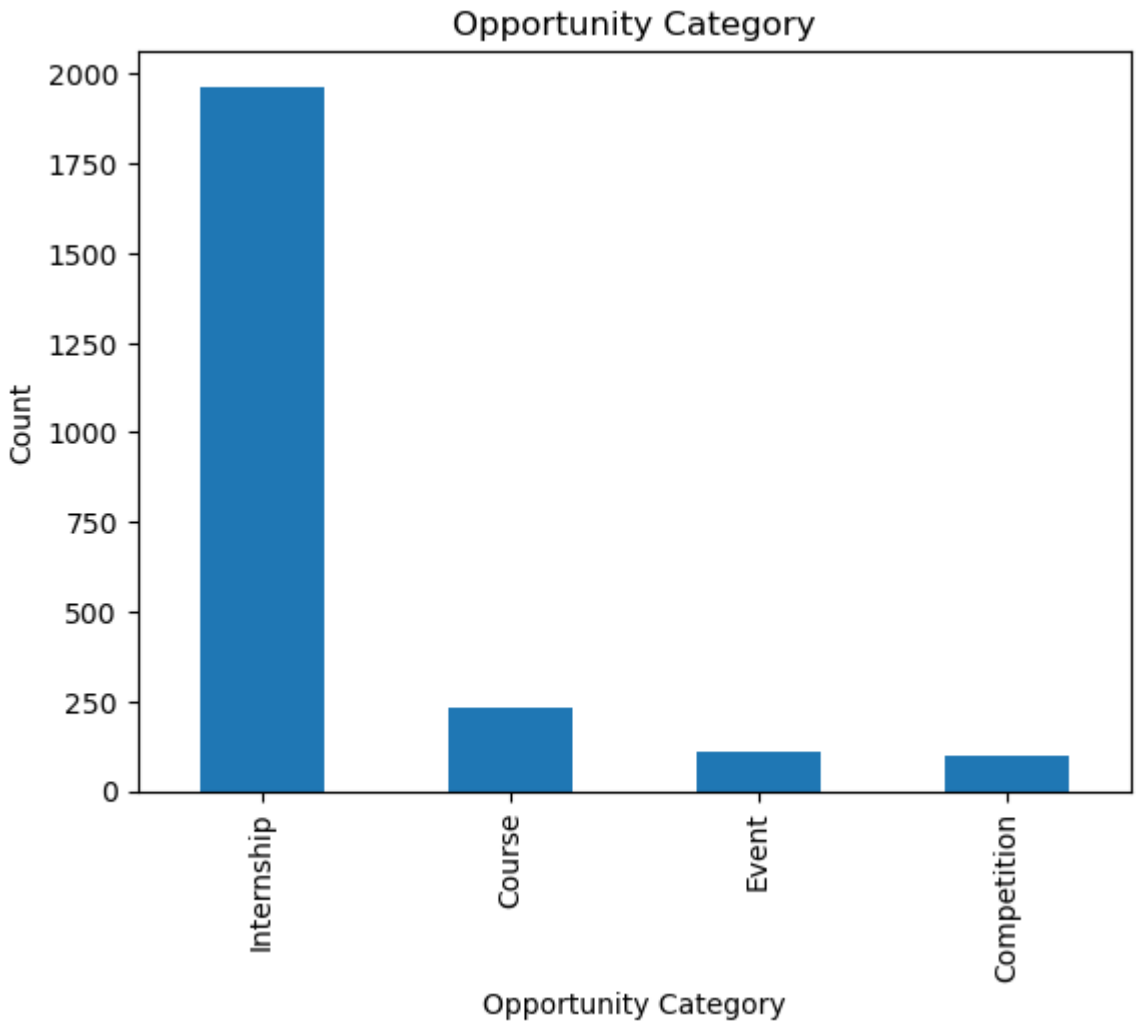
In [178…
```
df1.dropna(inplace=True)
df1.shape
```

Out[178…
```
(2401, 21)
```

In [179…
```
df1.isnull().sum()
```

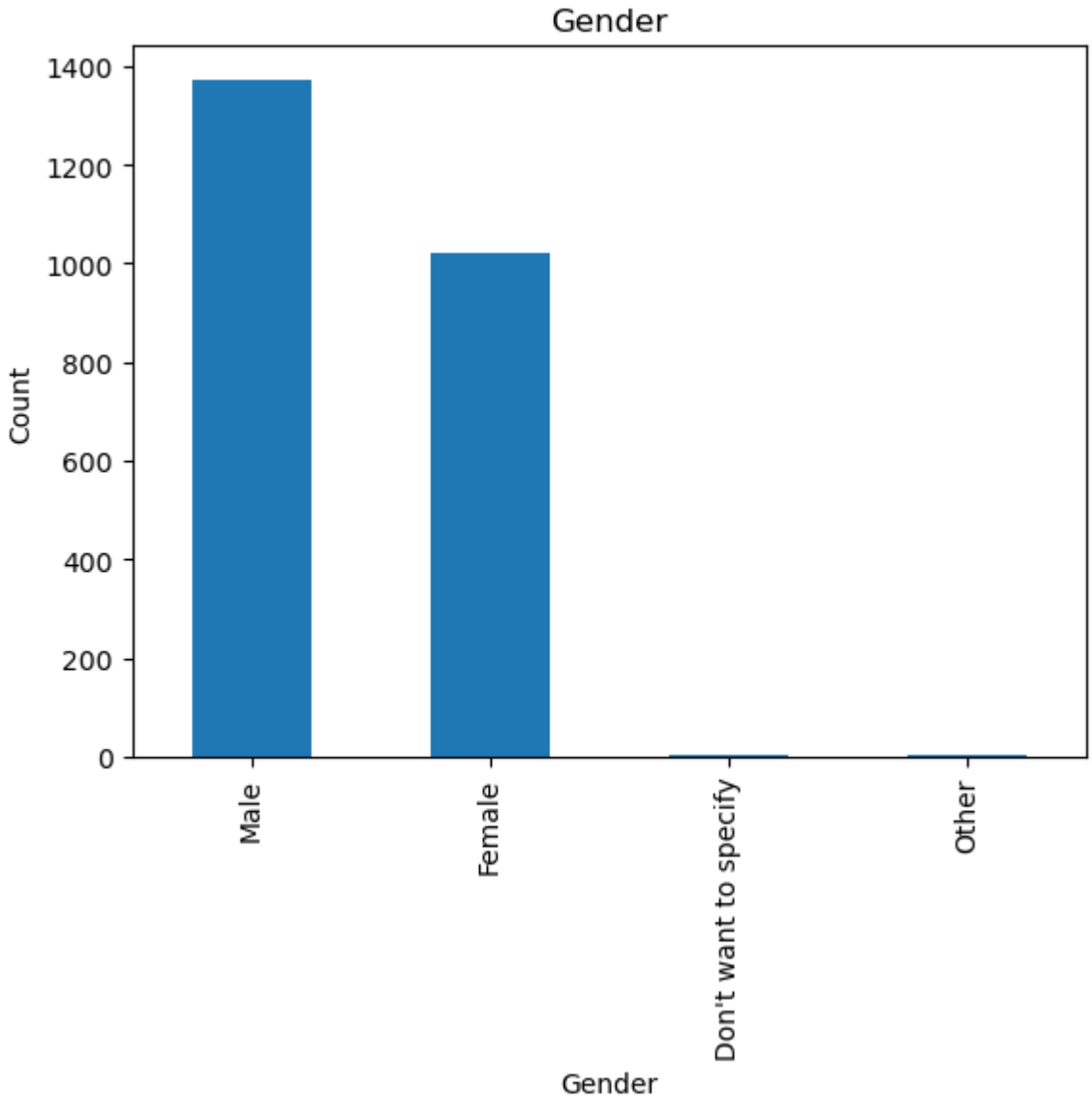
Out[179…
```
Profile Id                  0
Opportunity Id              0
Opportunity Name            0
Opportunity Category        0
Opportunity End Date        0
Gender                      0
City                        0
State                       0
Country                     0
Zip Code                    0
Graduation Date(YYYY MM)    0
Current Student Status      0
Current/Intended Major      0
Status Description          0
Apply Date                  0
Opportunity Start Date      0
Reward Amount               0
Badge Id                    0
Badge Name                  0
Skill Points Earned         0
Skills Earned               0
dtype: int64
```

In [180…
```
df1['Opportunity Category'].value_counts().plot(kind='bar')
pt.title('Opportunity Category')
pt.xlabel('Opportunity Category')
pt.ylabel('Count')
pt.show()
```



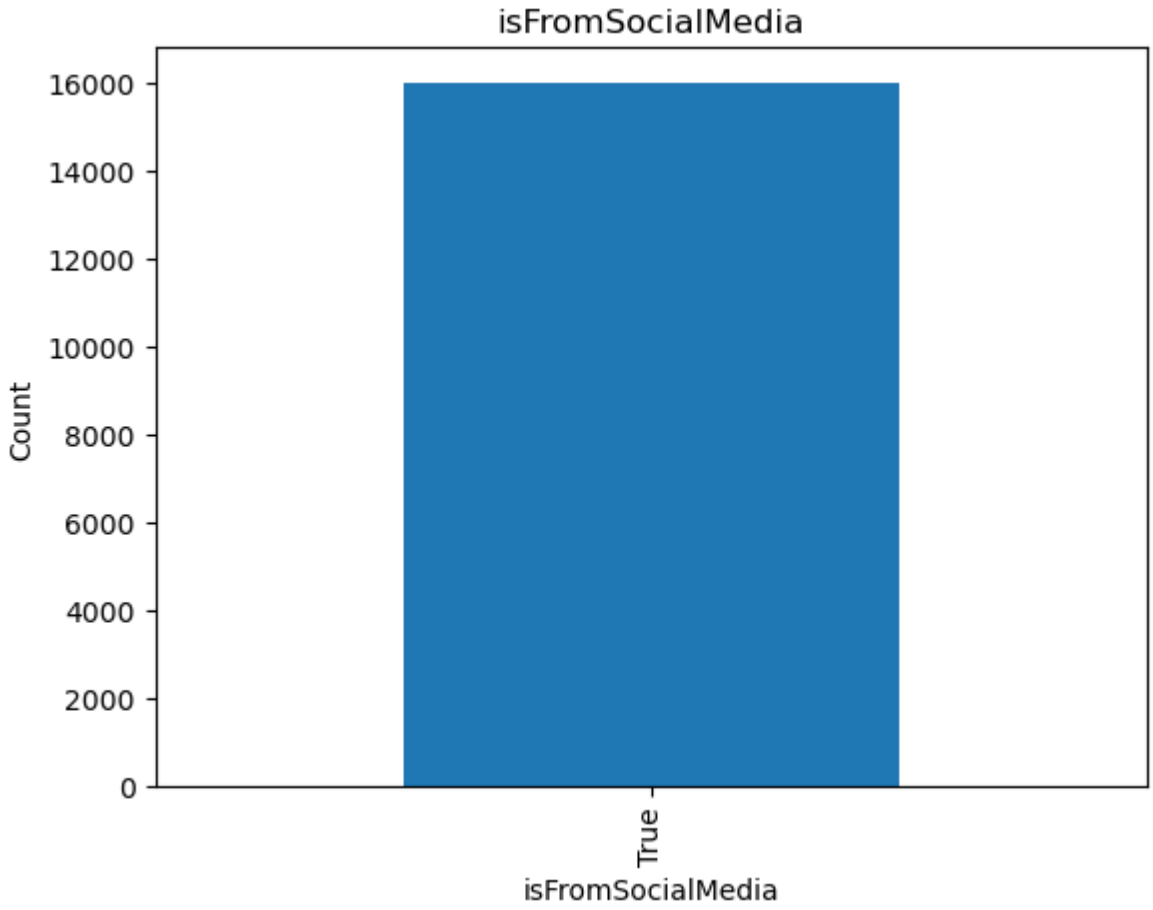
In [181…

```
df1['Gender'].value_counts().plot(kind='bar')
pt.title('Gender')
pt.xlabel('Gender')
pt.ylabel('Count')
pt.show()
```



In [182…

```
df['isFromSocialMedia'].value_counts().plot(kind='bar')
pt.title('isFromSocialMedia')
pt.xlabel('isFromSocialMedia')
pt.ylabel('Count')
pt.show()
```



7. Cross-Check Relationships

In [183…
```
date_check_failed = df1['Opportunity End Date'] < df1['Opportunity Start Date']
if any(date_check_failed):
    print("Inconsistencies found: Opportunity End Date should be after Opportuni
    df1[date_check_failed]
else:
    print("No inconsistencies found in date relationships.")
```

```
Inconsistencies found: Opportunity End Date should be after Opportunity Start Dat
e.
```

In [184…
```
reward_check_failed = df1['Reward Amount'] < 0
if any(reward_check_failed):
    print("Inconsistencies found: Reward Amount should be greater than or equal
    df1[reward_check_failed]
else:
    print("No inconsistencies found in reward relationships.")
```

```
No inconsistencies found in reward relationships.
```

In [185…
```
df.groupby(['Degree', 'Gender'])['isFromSocialMedia'].mean()
```

```
C:\Users\ve\AppData\Local\Temp\ipykernel_6428\378990288.py:1: FutureWarning: The
default of observed=False is deprecated and will be changed to True in a future v
ersion of pandas. Pass observed=False to retain current behavior or observed=True
to adopt the future default and silence this warning.
  df.groupby(['Degree', 'Gender'])['isFromSocialMedia'].mean()
```

Out[185…

Degree	Gender	
Graduate Program Student	Don't want to specify	1.0
	Female	1.0
	Male	1.0
	Other	1.0
High School Student	Don't want to specify	1.0
	Female	1.0
	Male	1.0
	Other	1.0
Not in Education	Don't want to specify	1.0
	Female	1.0
	Male	1.0
	Other	1.0
Undergraduate Student	Don't want to specify	1.0
	Female	1.0
	Male	1.0
	Other	1.0

Name: isFromSocialMedia, dtype: float64

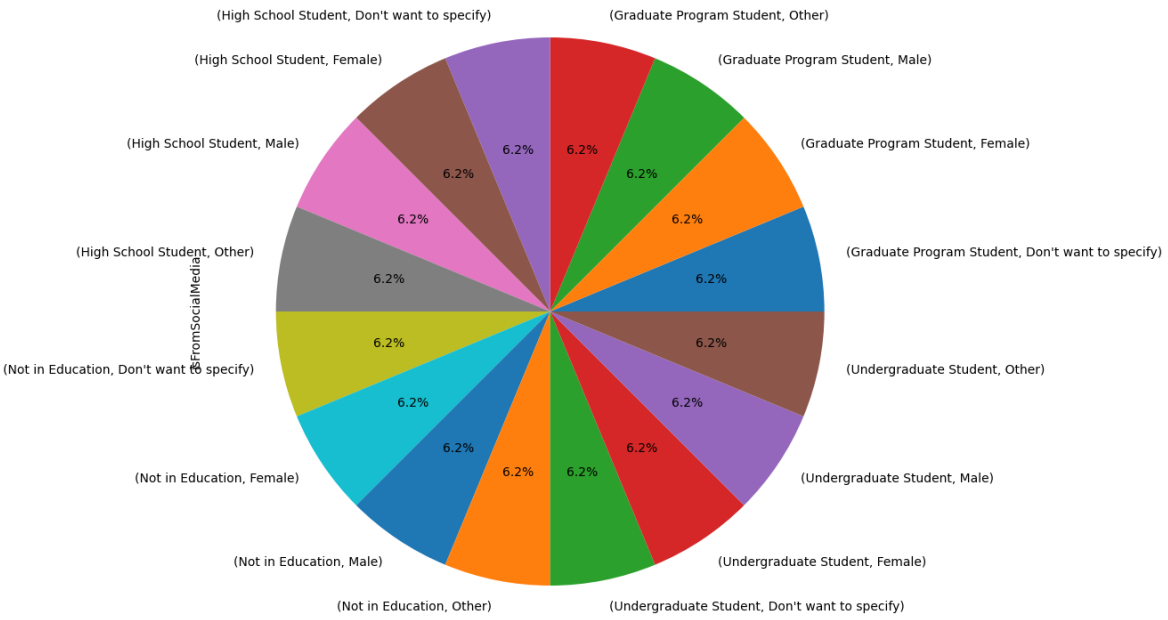
In [186…

```
# prompt: pie chart with big size on df data set

df.groupby(['Degree', 'Gender'])['isFromSocialMedia'].mean().plot(kind='pie', su
```

```
C:\Users\ve\AppData\Local\Temp\ipykernel_6428\2714612767.py:3: FutureWarning: The
default of observed=False is deprecated and will be changed to True in a future v
ersion of pandas. Pass observed=False to retain current behavior or observed=True
to adopt the future default and silence this warning.
  df.groupby(['Degree', 'Gender'])['isFromSocialMedia'].mean().plot(kind='pie', s
ubplots=True, figsize=(10, 10),autopct='%1.1f%%')
```

Out[186…

```
array([<Axes: ylabel='isFromSocialMedia'>], dtype=object)
```

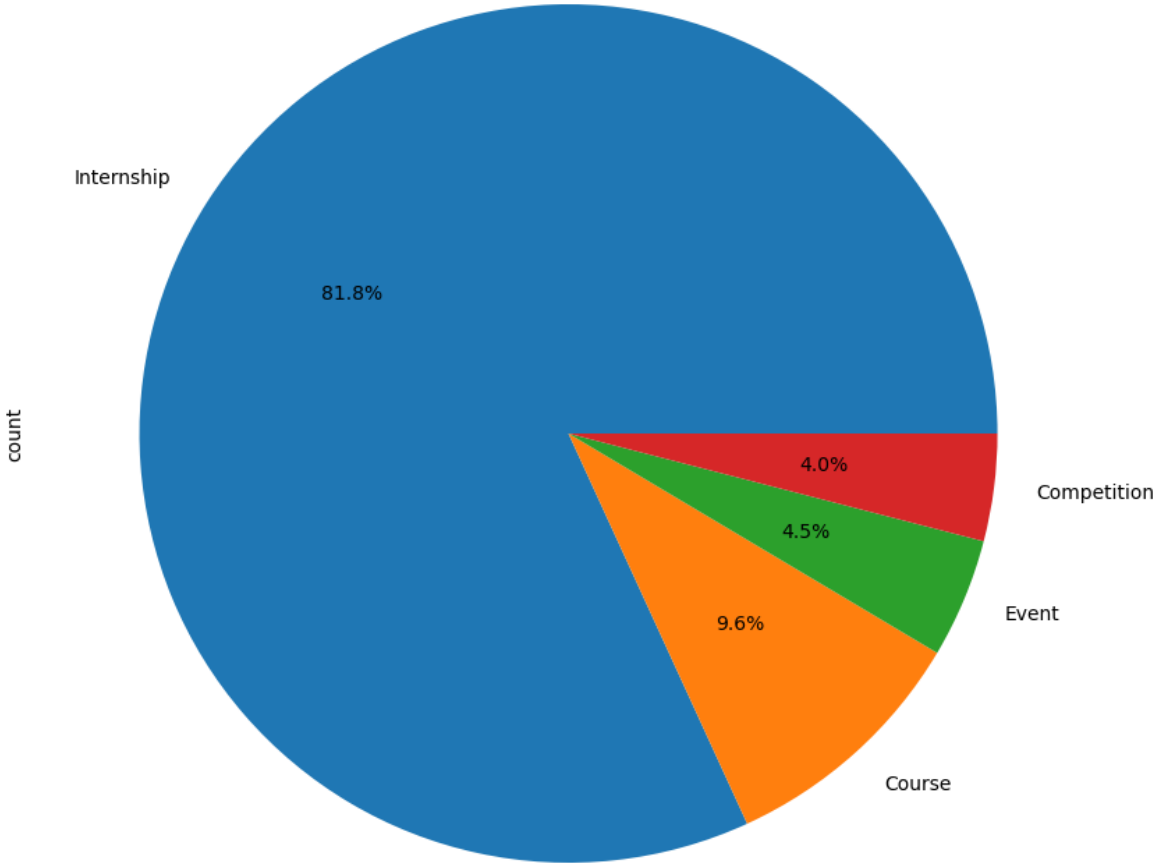


In [187…

```
df1['Opportunity Category'].value_counts().plot(kind='pie', subplots=True, figsi
```

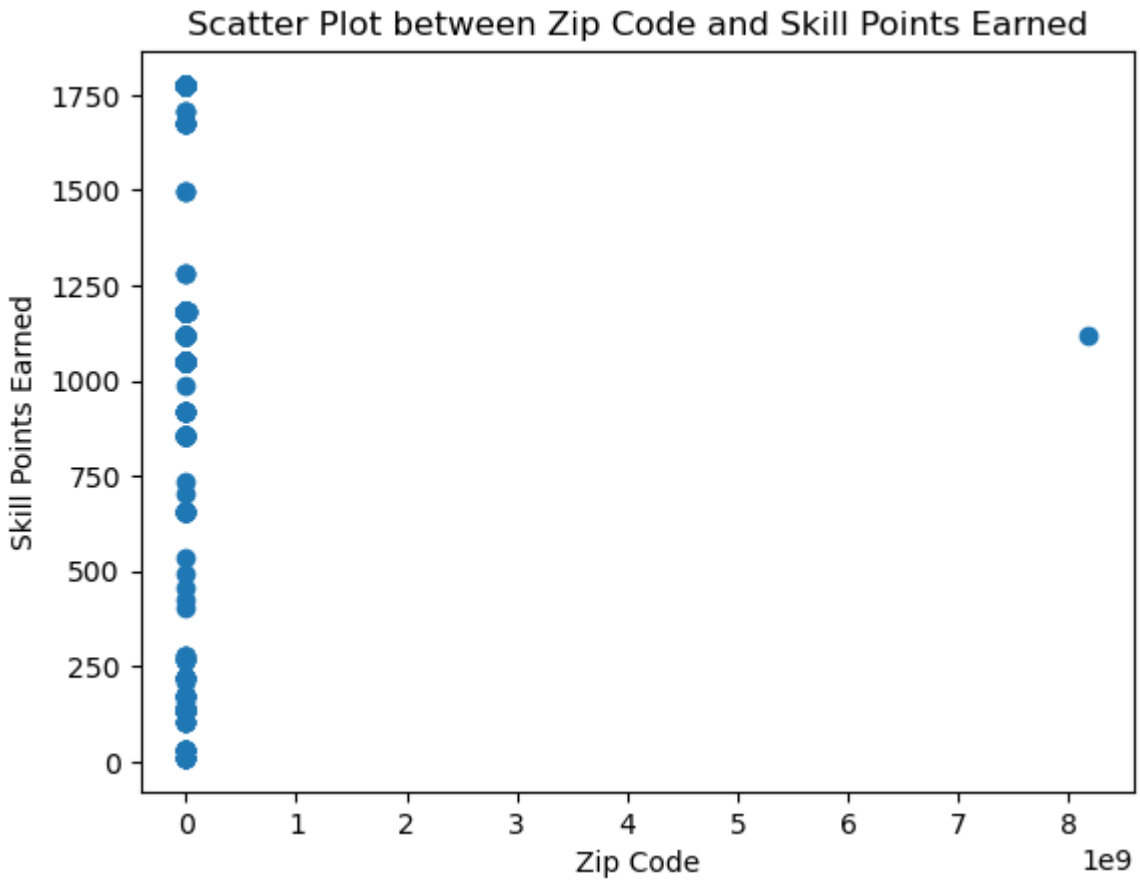
Out[187…

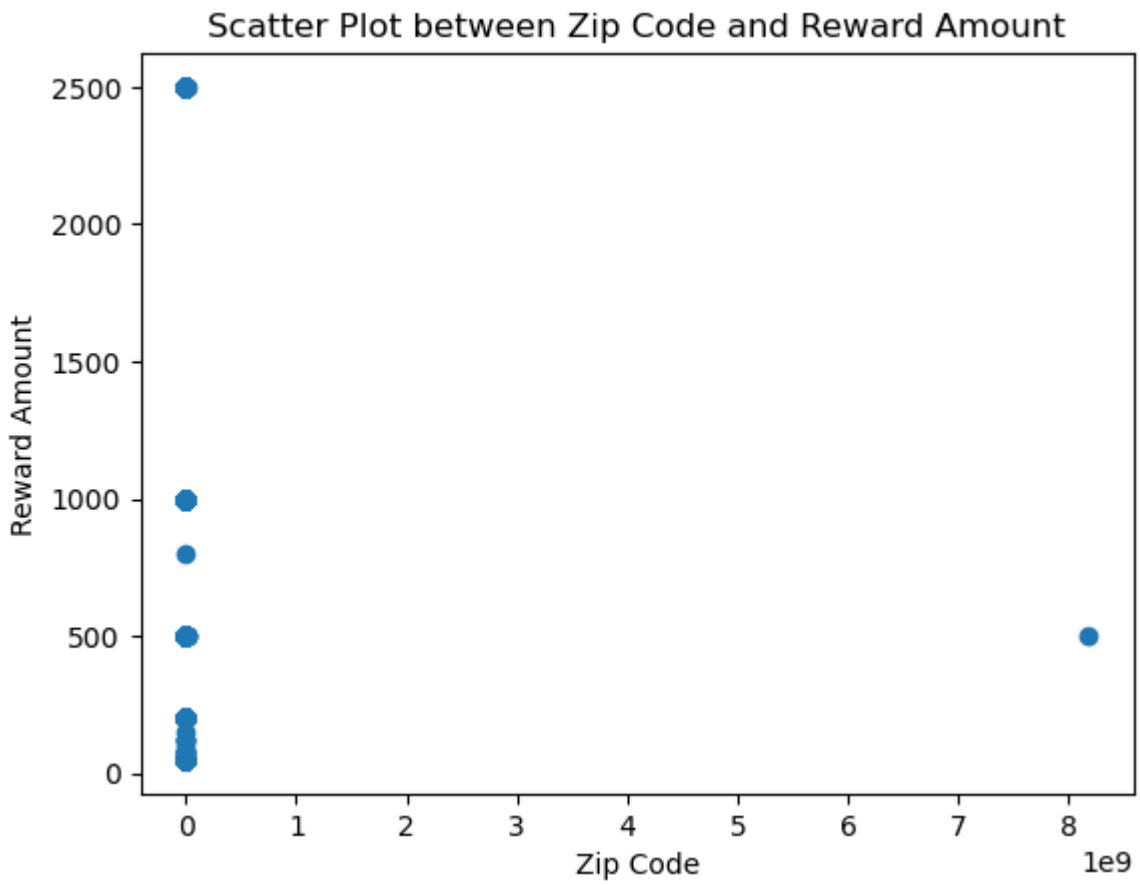
```
array([<Axes: ylabel='count'>], dtype=object)
```

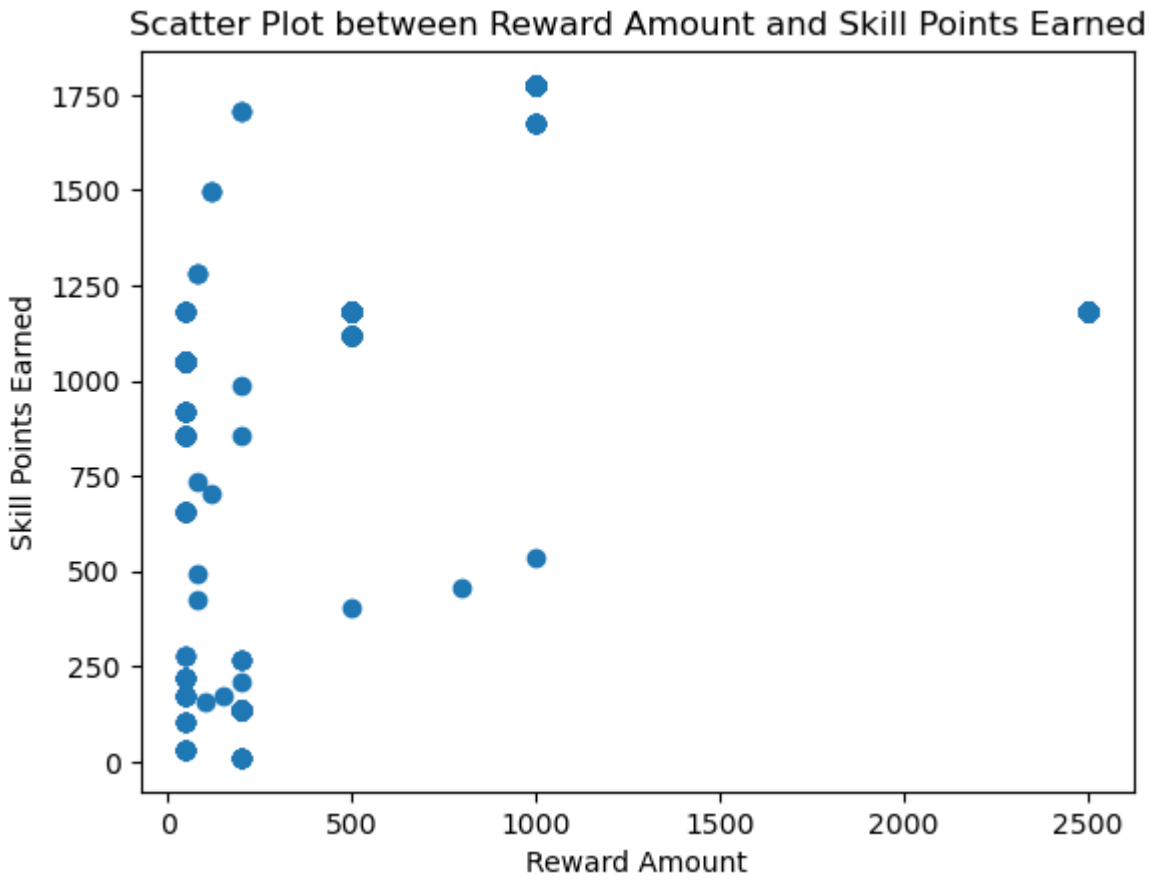


In [188…

```
pt.scatter(df1['Zip Code'], df1['Skill Points Earned'])
pt.title('Scatter Plot between Zip Code and Skill Points Earned')
pt.xlabel('Zip Code')
pt.ylabel('Skill Points Earned')
pt.show()

pt.scatter(df1['Zip Code'], df1['Reward Amount'])
pt.title('Scatter Plot between Zip Code and Reward Amount')
pt.xlabel('Zip Code')
pt.ylabel('Reward Amount')
pt.show()

pt.scatter(df1['Reward Amount'], df1['Skill Points Earned'])
pt.title('Scatter Plot between Reward Amount and Skill Points Earned')
pt.xlabel('Reward Amount')
pt.ylabel('Skill Points Earned')
pt.show()
```


Scatter Plot between Zip Code and Skill Points Earned
Skill Points Earned
1750
1500
1250
1000
750
500
250
0
0
1
2
3
4
5
6
7
8
Zip Code
1e9

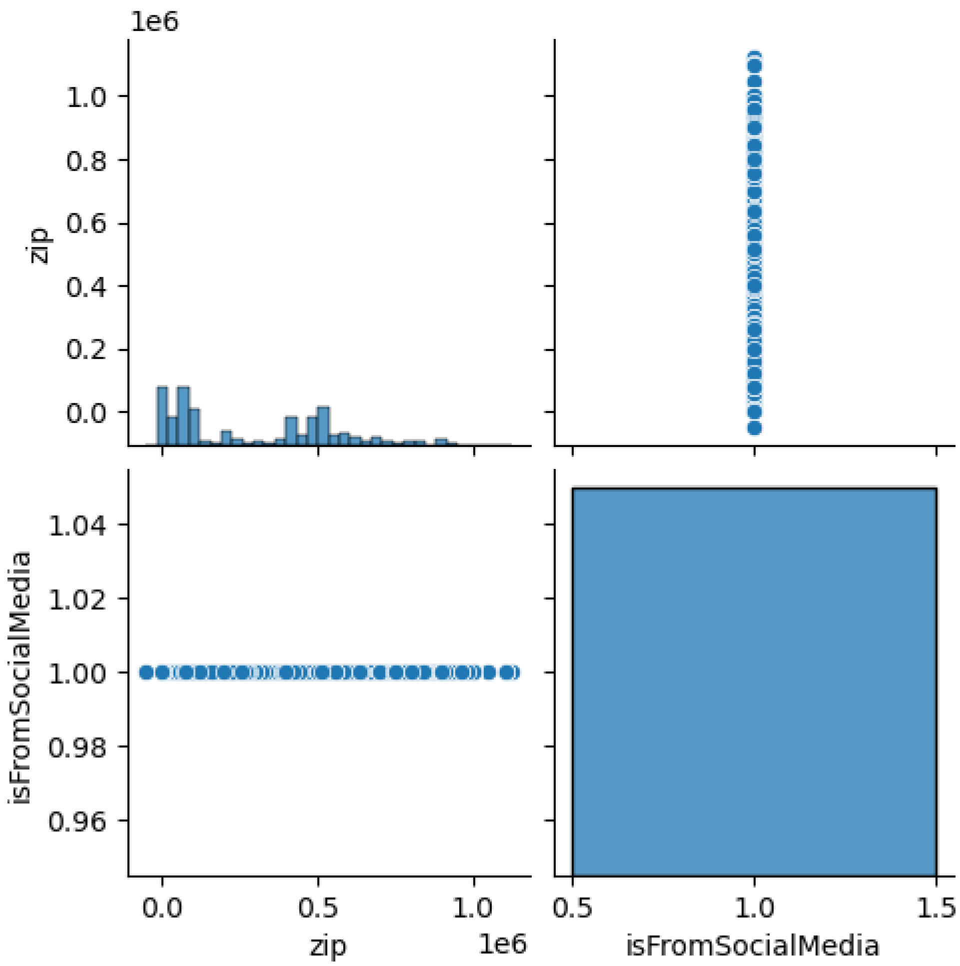

Scatter Plot between Zip Code and Reward Amount
Reward Amount
2500
2000
1500
1000
500
0
0
1
2
3
4
5
6
7
8
Zip Code
1e9



In [189…

```
sns.pairplot(df)
pt.show()
```

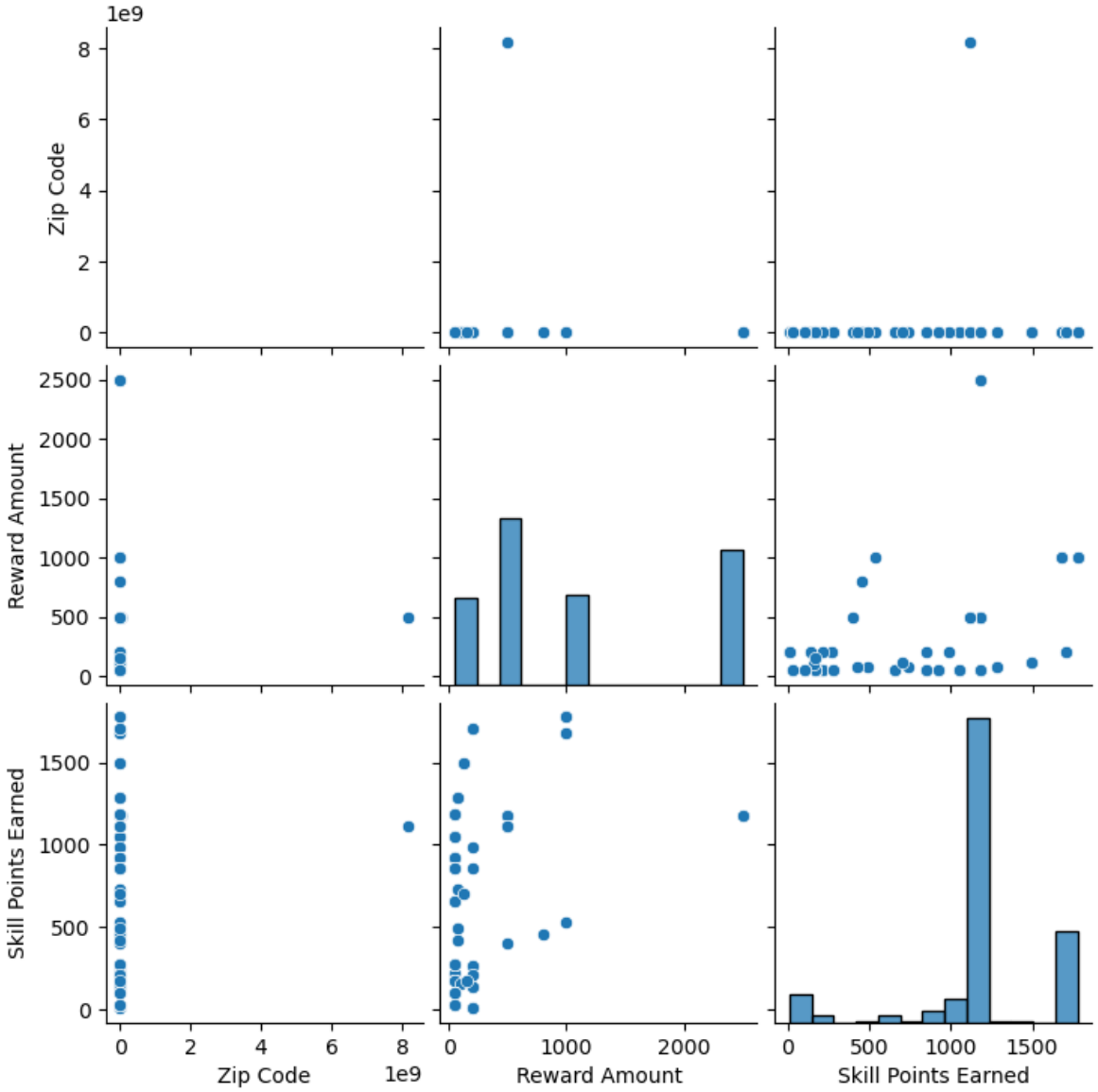
```
E:\anaconda\envs\ds_env\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarnin
g: use_inf_as_na option is deprecated and will be removed in a future version. Co
nvert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
E:\anaconda\envs\ds_env\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarnin
g: use_inf_as_na option is deprecated and will be removed in a future version. Co
nvert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
E:\anaconda\envs\ds_env\Lib\site-packages\seaborn\_stats\counting.py:137: Runtime
Warning: Converting input from bool to <class 'numpy.uint8'> for compatibility.
  bin_edges = np.histogram_bin_edges(vals, bins, binrange, weight)
E:\anaconda\envs\ds_env\Lib\site-packages\seaborn\_stats\counting.py:176: Runtime
Warning: Converting input from bool to <class 'numpy.uint8'> for compatibility.
  hist, edges = np.histogram(vals, **bin_kws, weights=weights, density=density)
```



In [190…

```
sns.pairplot(df1)
pt.show()
```

```
E:\anaconda\envs\ds_env\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarnin
g: use_inf_as_na option is deprecated and will be removed in a future version. Co
nvert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
E:\anaconda\envs\ds_env\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarnin
g: use_inf_as_na option is deprecated and will be removed in a future version. Co
nvert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
E:\anaconda\envs\ds_env\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarnin
g: use_inf_as_na option is deprecated and will be removed in a future version. Co
nvert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



In [206…

```
# #visualization on the correlation
df1_numeric = df1.apply(pd.to_numeric, errors='coerce')

corr_matrix = df1_numeric.corr()

print(corr_matrix)
sns.heatmap(corr_matrix, annot=True)
pt.show()
```

	Profile Id	Opportunity Id	Opportunity Name	\
Profile Id	NaN	NaN	NaN	
Opportunity Id	NaN	NaN	NaN	
Opportunity Name	NaN	NaN	NaN	
Opportunity Category	NaN	NaN	NaN	
Opportunity End Date	NaN	NaN	NaN	
Gender	NaN	NaN	NaN	
City	NaN	NaN	NaN	
State	NaN	NaN	NaN	
Country	NaN	NaN	NaN	
Zip Code	NaN	NaN	NaN	
Graduation Date(YYYY MM)	NaN	NaN	NaN	
Current Student Status	NaN	NaN	NaN	
Current/Intended Major	NaN	NaN	NaN	
Status Description	NaN	NaN	NaN	
Apply Date	NaN	NaN	NaN	
Opportunity Start Date	NaN	NaN	NaN	
Reward Amount	NaN	NaN	NaN	
Badge Id	NaN	NaN	NaN	
Badge Name	NaN	NaN	NaN	
Skill Points Earned	NaN	NaN	NaN	
Skills Earned	NaN	NaN	NaN	

	Opportunity Category	Opportunity End Date	Gender	\
Profile Id	NaN	NaN	NaN	
Opportunity Id	NaN	NaN	NaN	
Opportunity Name	NaN	NaN	NaN	
Opportunity Category	NaN	NaN	NaN	
Opportunity End Date	NaN	1.000000	NaN	
Gender	NaN	NaN	NaN	
City	NaN	NaN	NaN	
State	NaN	NaN	NaN	
Country	NaN	NaN	NaN	
Zip Code	NaN	-0.037218	NaN	
Graduation Date(YYYY MM)	NaN	-0.020788	NaN	
Current Student Status	NaN	NaN	NaN	
Current/Intended Major	NaN	NaN	NaN	
Status Description	NaN	NaN	NaN	
Apply Date	NaN	0.151497	NaN	
Opportunity Start Date	NaN	0.036470	NaN	
Reward Amount	NaN	0.152282	NaN	
Badge Id	NaN	NaN	NaN	
Badge Name	NaN	NaN	NaN	
Skill Points Earned	NaN	0.206973	NaN	
Skills Earned	NaN	NaN	NaN	

	City	State	Country	Zip Code	...	\
Profile Id	NaN	NaN	NaN	NaN	...	
Opportunity Id	NaN	NaN	NaN	NaN	...	
Opportunity Name	NaN	NaN	NaN	NaN	...	
Opportunity Category	NaN	NaN	NaN	NaN	...	
Opportunity End Date	NaN	NaN	NaN	-0.037218	...	
Gender	NaN	NaN	NaN	NaN	...	
City	NaN	NaN	NaN	NaN	...	
State	NaN	NaN	NaN	NaN	...	
Country	NaN	NaN	NaN	NaN	...	
Zip Code	NaN	NaN	NaN	1.000000	...	
Graduation Date(YYYY MM)	NaN	NaN	NaN	-0.027231	...	
Current Student Status	NaN	NaN	NaN	NaN	...	
Current/Intended Major	NaN	NaN	NaN	NaN	...	

Status Description	NaN	NaN	NaN	NaN	...
Apply Date	NaN	NaN	NaN	-0.016076	...
Opportunity Start Date	NaN	NaN	NaN	0.008948	...
Reward Amount	NaN	NaN	NaN	-0.012839	...
Badge Id	NaN	NaN	NaN	NaN	...
Badge Name	NaN	NaN	NaN	NaN	...
Skill Points Earned	NaN	NaN	NaN	-0.003593	...
Skills Earned	NaN	NaN	NaN	NaN	...

	Current Student Status	Current/Intended Major	\
Profile Id	NaN	NaN	
Opportunity Id	NaN	NaN	
Opportunity Name	NaN	NaN	
Opportunity Category	NaN	NaN	
Opportunity End Date	NaN	NaN	
Gender	NaN	NaN	
City	NaN	NaN	
State	NaN	NaN	
Country	NaN	NaN	
Zip Code	NaN	NaN	
Graduation Date(YYYY MM)	NaN	NaN	
Current Student Status	NaN	NaN	
Current/Intended Major	NaN	NaN	
Status Description	NaN	NaN	
Apply Date	NaN	NaN	
Opportunity Start Date	NaN	NaN	
Reward Amount	NaN	NaN	
Badge Id	NaN	NaN	
Badge Name	NaN	NaN	
Skill Points Earned	NaN	NaN	
Skills Earned	NaN	NaN	

	Status Description	Apply Date	\
Profile Id	NaN	NaN	
Opportunity Id	NaN	NaN	
Opportunity Name	NaN	NaN	
Opportunity Category	NaN	NaN	
Opportunity End Date	NaN	0.151497	
Gender	NaN	NaN	
City	NaN	NaN	
State	NaN	NaN	
Country	NaN	NaN	
Zip Code	NaN	-0.016076	
Graduation Date(YYYY MM)	NaN	-0.040344	
Current Student Status	NaN	NaN	
Current/Intended Major	NaN	NaN	
Status Description	NaN	NaN	
Apply Date	NaN	1.000000	
Opportunity Start Date	NaN	0.723474	
Reward Amount	NaN	-0.544052	
Badge Id	NaN	NaN	
Badge Name	NaN	NaN	
Skill Points Earned	NaN	0.146616	
Skills Earned	NaN	NaN	

	Opportunity Start Date	Reward Amount	Badge Id	\
Profile Id	NaN	NaN	NaN	
Opportunity Id	NaN	NaN	NaN	
Opportunity Name	NaN	NaN	NaN	
Opportunity Category	NaN	NaN	NaN	

Opportunity End Date	0.036470	0.152282	NaN
Gender	NaN	NaN	NaN
City	NaN	NaN	NaN
State	NaN	NaN	NaN
Country	NaN	NaN	NaN
Zip Code	0.008948	-0.012839	NaN
Graduation Date(YYYY MM)	-0.084337	0.049346	NaN
Current Student Status	NaN	NaN	NaN
Current/Intended Major	NaN	NaN	NaN
Status Description	NaN	NaN	NaN
Apply Date	0.723474	-0.544052	NaN
Opportunity Start Date	1.000000	-0.248550	NaN
Reward Amount	-0.248550	1.000000	NaN
Badge Id	NaN	NaN	NaN
Badge Name	NaN	NaN	NaN
Skill Points Earned	0.508630	0.240293	NaN
Skills Earned	NaN	NaN	NaN

	Badge Name	Skill Points Earned	Skills Earned
Profile Id	NaN	NaN	NaN
Opportunity Id	NaN	NaN	NaN
Opportunity Name	NaN	NaN	NaN
Opportunity Category	NaN	NaN	NaN
Opportunity End Date	NaN	0.206973	NaN
Gender	NaN	NaN	NaN
City	NaN	NaN	NaN
State	NaN	NaN	NaN
Country	NaN	NaN	NaN
Zip Code	NaN	-0.003593	NaN
Graduation Date(YYYY MM)	NaN	-0.102770	NaN
Current Student Status	NaN	NaN	NaN
Current/Intended Major	NaN	NaN	NaN
Status Description	NaN	NaN	NaN
Apply Date	NaN	0.146616	NaN
Opportunity Start Date	NaN	0.508630	NaN
Reward Amount	NaN	0.240293	NaN
Badge Id	NaN	NaN	NaN
Badge Name	NaN	NaN	NaN
Skill Points Earned	NaN	1.000000	NaN
Skills Earned	NaN	NaN	NaN

[21 rows x 21 columns]

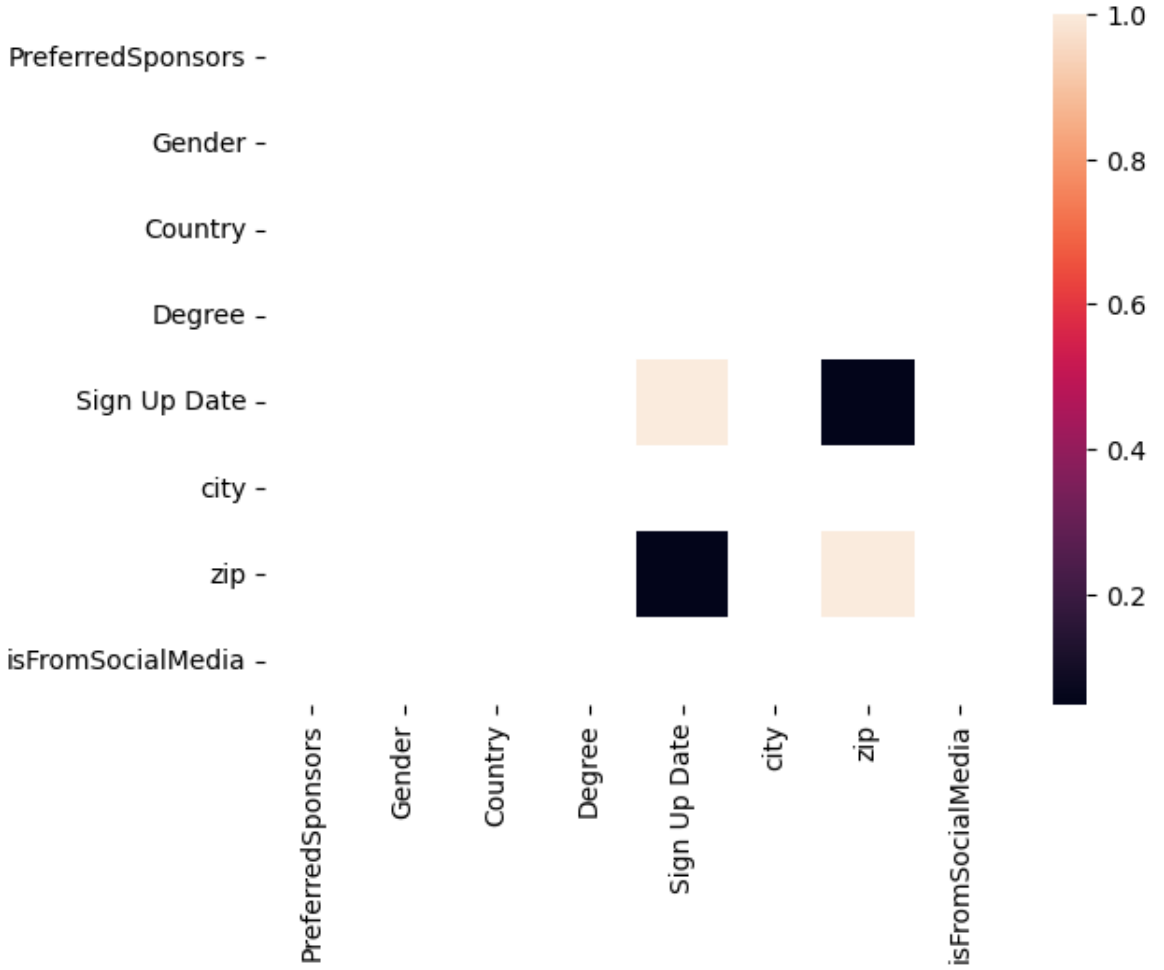
```
E:\anaconda\envs\ds_env\Lib\site-packages\seaborn\matrix.py:260: FutureWarning: F
ormat strings passed to MaskedConstant are ignored, but in future may error or pr
oduce different behavior
  annotation = ("{:" + self.fmt + "}").format(val)
```

In [208…

```
#visualization on the correlation
df_numeric = df.apply(pd.to_numeric, errors='coerce')

corr_matrix = df_numeric.corr()

print(corr_matrix)
sns.heatmap(corr_matrix, annot=True)
pt.show()
```

```
                  PreferredSponsors  Gender  Country  Degree  Sign Up Date  \
PreferredSponsors               NaN     NaN      NaN     NaN           NaN
Gender                          NaN     NaN      NaN     NaN           NaN
Country                         NaN     NaN      NaN     NaN           NaN
Degree                          NaN     NaN      NaN     NaN           NaN
Sign Up Date                    NaN     NaN      NaN     NaN       1.00000
city                            NaN     NaN      NaN     NaN           NaN
zip                             NaN     NaN      NaN     NaN       0.04773
isFromSocialMedia               NaN     NaN      NaN     NaN           NaN

                  city      zip  isFromSocialMedia
PreferredSponsors  NaN      NaN                NaN
Gender             NaN      NaN                NaN
Country            NaN      NaN                NaN
Degree             NaN      NaN                NaN
Sign Up Date       NaN  0.04773                NaN
city               NaN      NaN                NaN
zip                NaN  1.00000                NaN
isFromSocialMedia  NaN      NaN                NaN
```

```
E:\anaconda\envs\ds_env\Lib\site-packages\seaborn\matrix.py:260: FutureWarning: F
ormat strings passed to MaskedConstant are ignored, but in future may error or pr
oduce different behavior
  annotation = ("{:" + self.fmt + "}").format(val)
```



8. Document the Process

Understand the basic characteristics of the data.

Identify potential relationships between variables.

Detect outliers or anomalies.

Data cleaning and analysis steps.

Insights and Findings

1 - There is less amount of numeric data

2 - Most of the attributes are of type object

3 - The count of the Male Applicants are more than female

4 - The Opportunity Category has the maximum Internship as the category

5 - There was the much need of validate the categorical and numeric data

6 - Too many missing values present in dataset

7 - The most of the applicatoin are from 2022

8 - Large amount of categorical data present

In [ ]: