Week 1 - Data Visualization Project - Exploratory Data Analysis (EDA) Report

INTRODUCTION

Purpose of the exploratory data report:

An Exploratory Data Report is a preliminary report that uses visuals and quantitative measures to describe the major features, trends, and relationships in the data. It is intended to help data analysts and researchers better understand their data, identify possible irregularities or mistakes, and generate hypotheses for further examination or modeling. The report may also be used to communicate the findings and implications of data analysis with a wider audience, such as stakeholders or peers.

Specify the datasets you are working with:
User Data and Opportunity Sign Up and Completion Data.

DATA OVERVIEW

High-level summary of each data set: User Data:
This dataset contains de-identifying data on each of the users who have signed up for an account with Excelerate.

All user's data is holistic, irrespective of whether or not they interact with specific opportunities.

Each row stands for a specific user, and the dataset captures a comprehensive picture of how users are distributed across America.

Opportunity Sign Up and Completion Data:

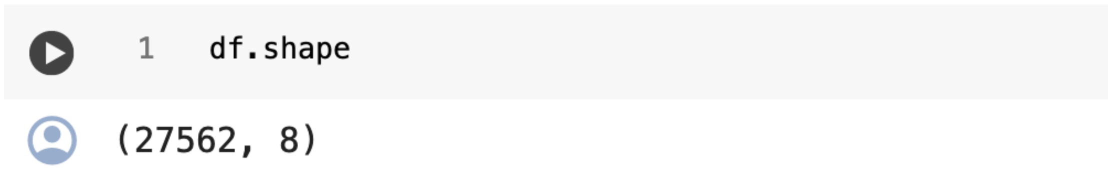
This dataset centers on the non-identifying information about users such as learners who have interacted with certain offers presented by Excelerate.

The rows represent learners enrolled in a specific opportunity.

Because learners can register for multiple opportunities, there may be more than one row with the same profile ID.

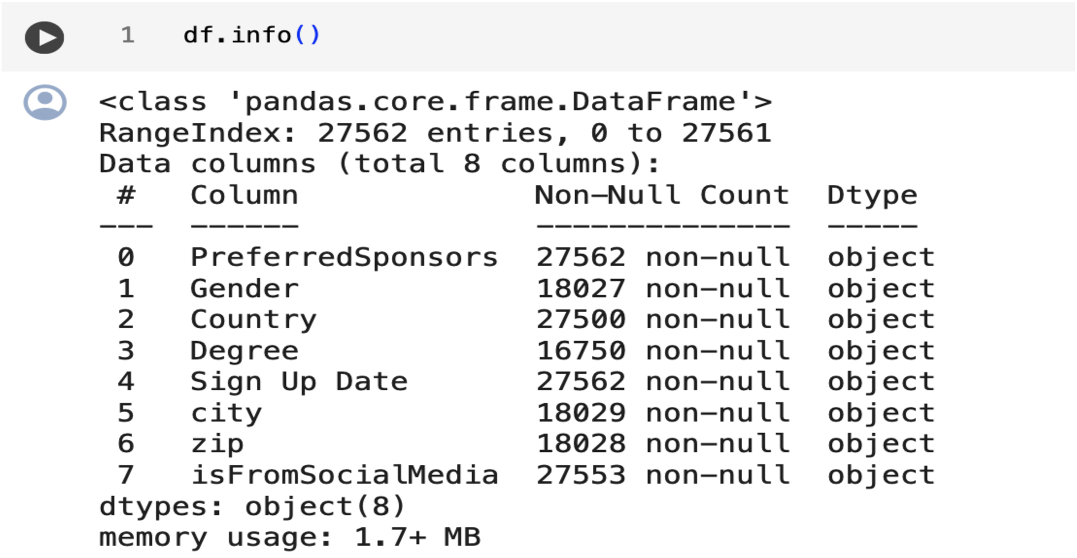
Key statistics such as the number of rows, columns, and unique identifiers:

User data



```
df.shape
```

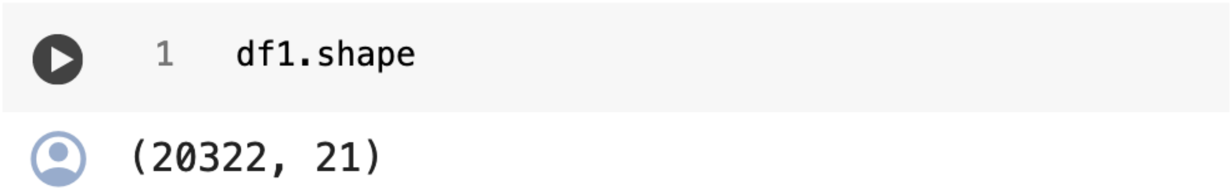
```
(27562, 8)
```



```
df.info()
```

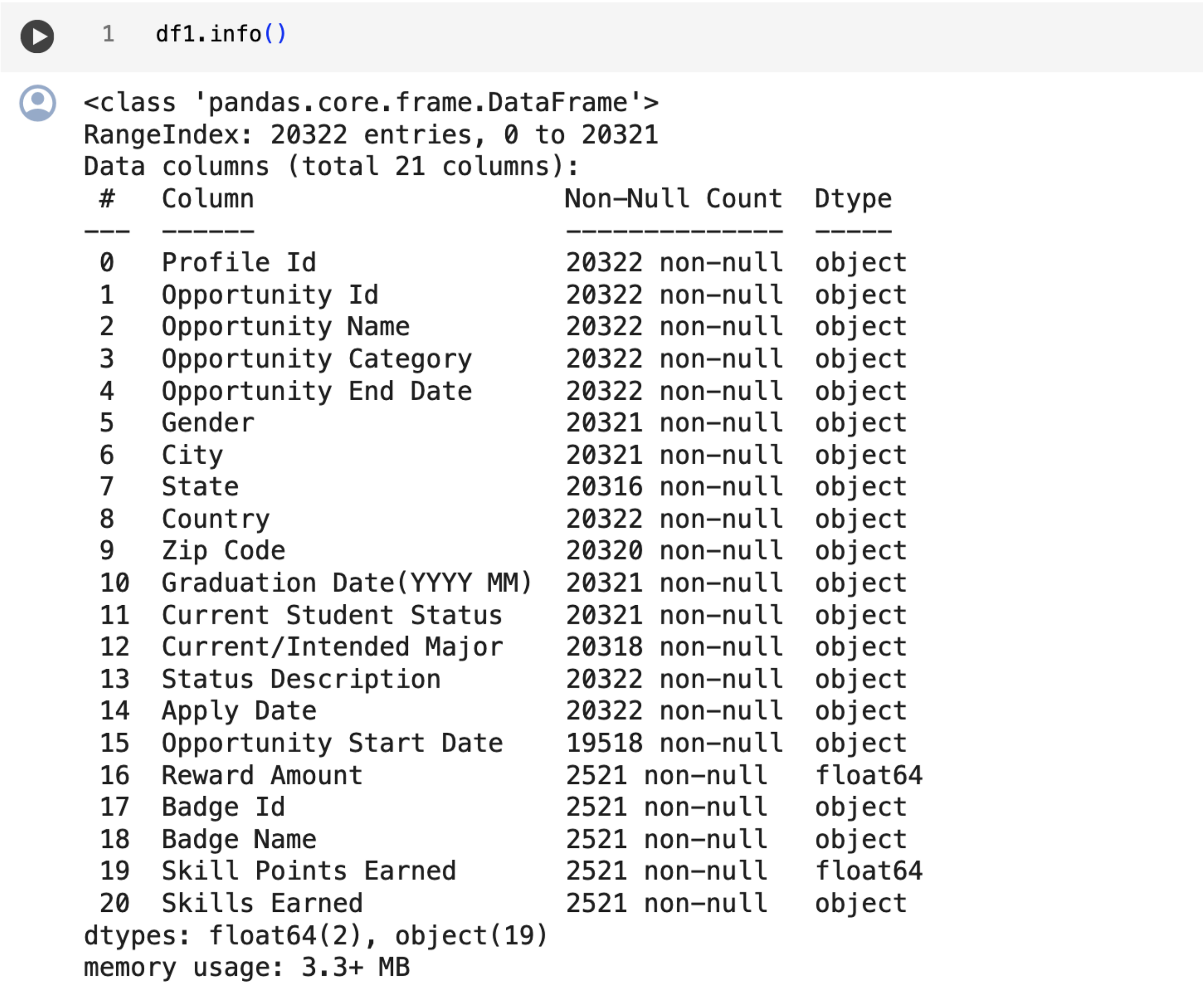
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27562 entries, 0 to 27561
Data columns (total 8 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   PreferredSponsors   27562 non-null  object
 1   Gender              18027 non-null  object
 2   Country             27500 non-null  object
 3   Degree              16750 non-null  object
 4   Sign Up Date        27562 non-null  object
 5   city                18029 non-null  object
 6   zip                 18028 non-null  object
 7   isFromSocialMedia   27553 non-null  object
dtypes: object(8)
memory usage: 1.7+ MB
```

Opportunity wise data



```
df1.shape
```

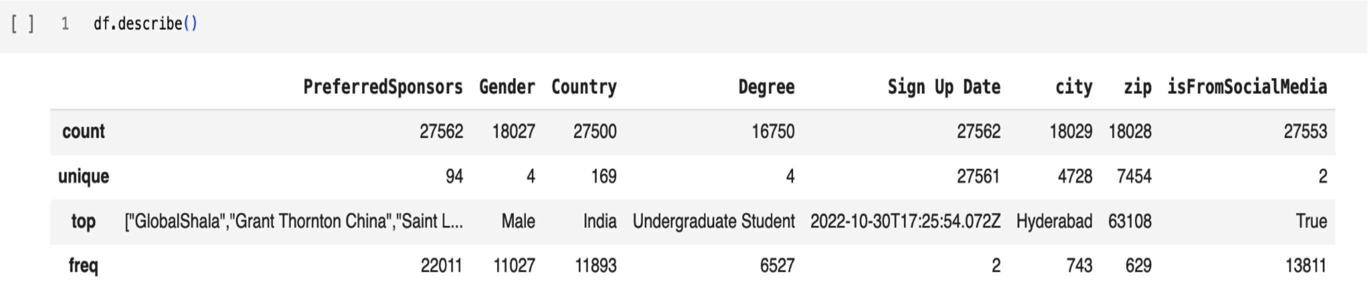
```
(20322, 21)
```



```
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                    20316 non-null  object 
 8   Country                  20322 non-null  object 
 9   Zip Code                 20320 non-null  object 
 10  Graduation Date(YYYY MM) 20321 non-null  object 
 11  Current Student Status   20321 non-null  object 
 12  Current/Intended Major   20318 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
```

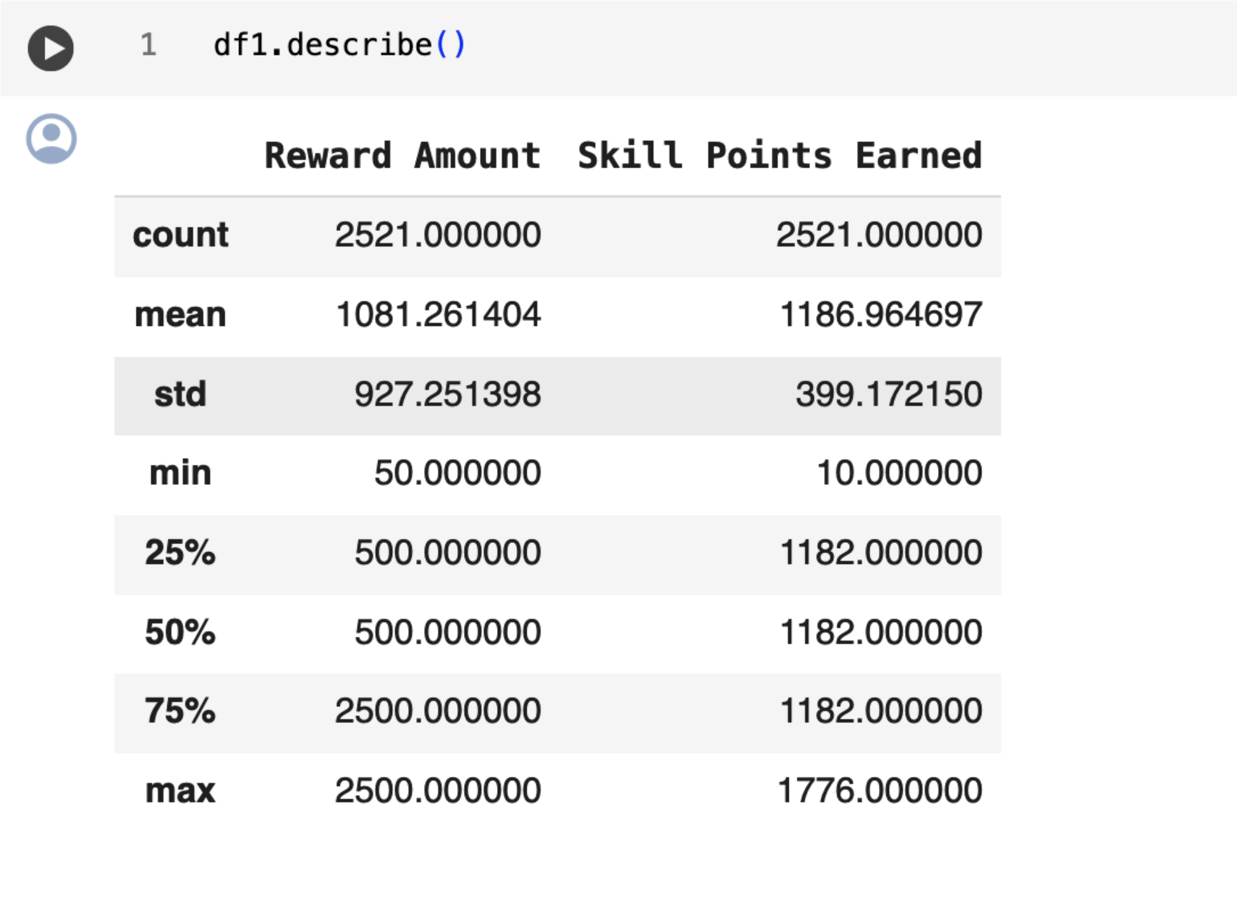
Explore summary statistics:
User data



```
[ ]  df.describe()
```

	PreferredSponsors	Gender	Country	Degree	Sign Up Date	city	zip	isFromSocialMedia
count	27562	18027	27500	16750	27562	18029	18028	27553
unique	94	4	169	4	27561	4728	7454	2
top	["GlobalShala","Grant Thornton China","Saint L...	Male	India	Undergraduate Student	2022-10-30T17:25:54.072Z	Hyderabad	63108	True
freq	22011	11027	11893	6527	2	743	629	13811

Opportunity wise data

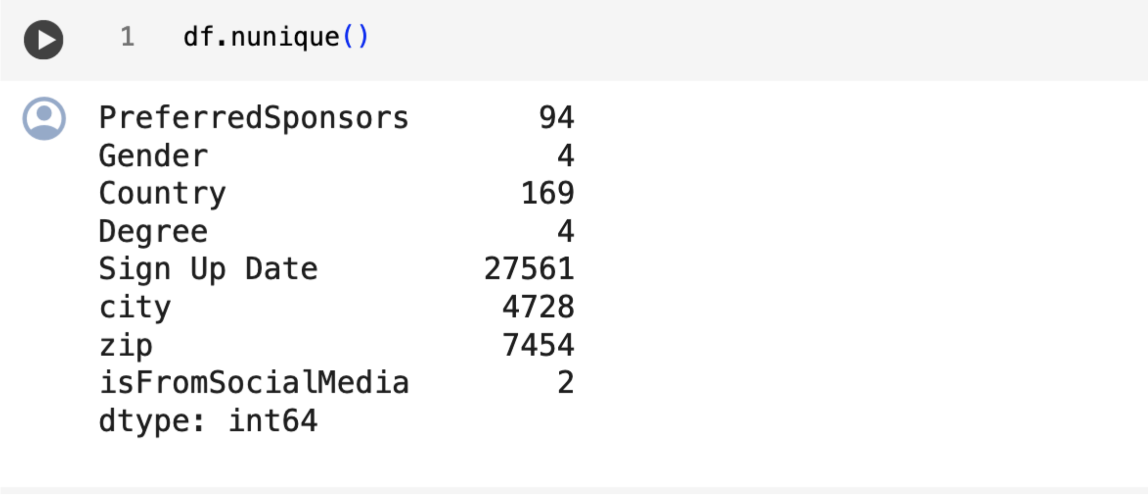


```
df1.describe()
```

	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

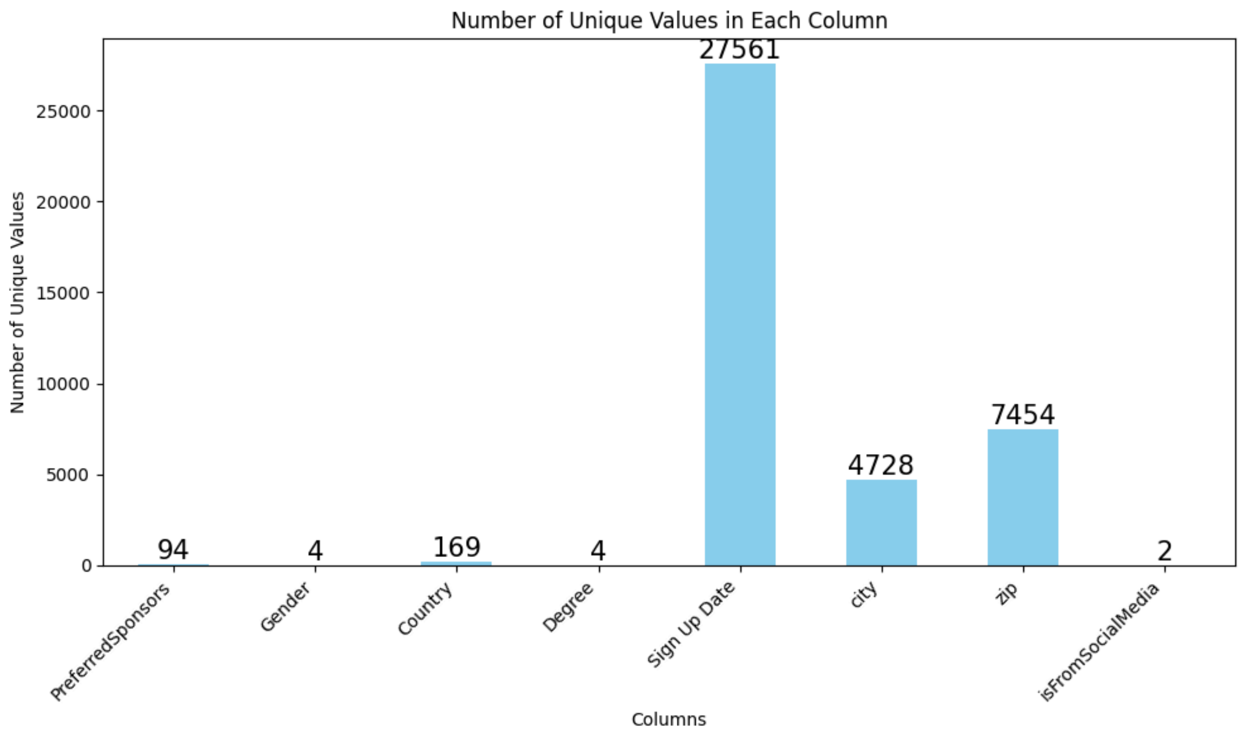
Identify unique values:

User data

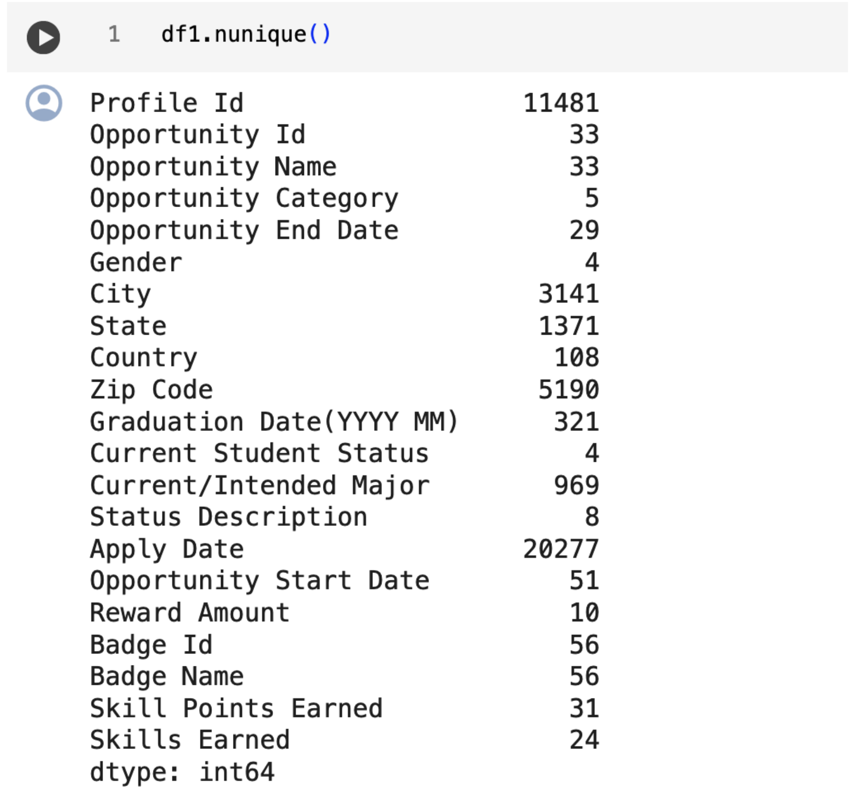


```
df.nunique()
```

```
PreferredSponsors        94
Gender                    4
Country                 169
Degree                    4
Sign Up Date          27561
city                   4728
zip                    7454
isFromSocialMedia         2
dtype: int64
```

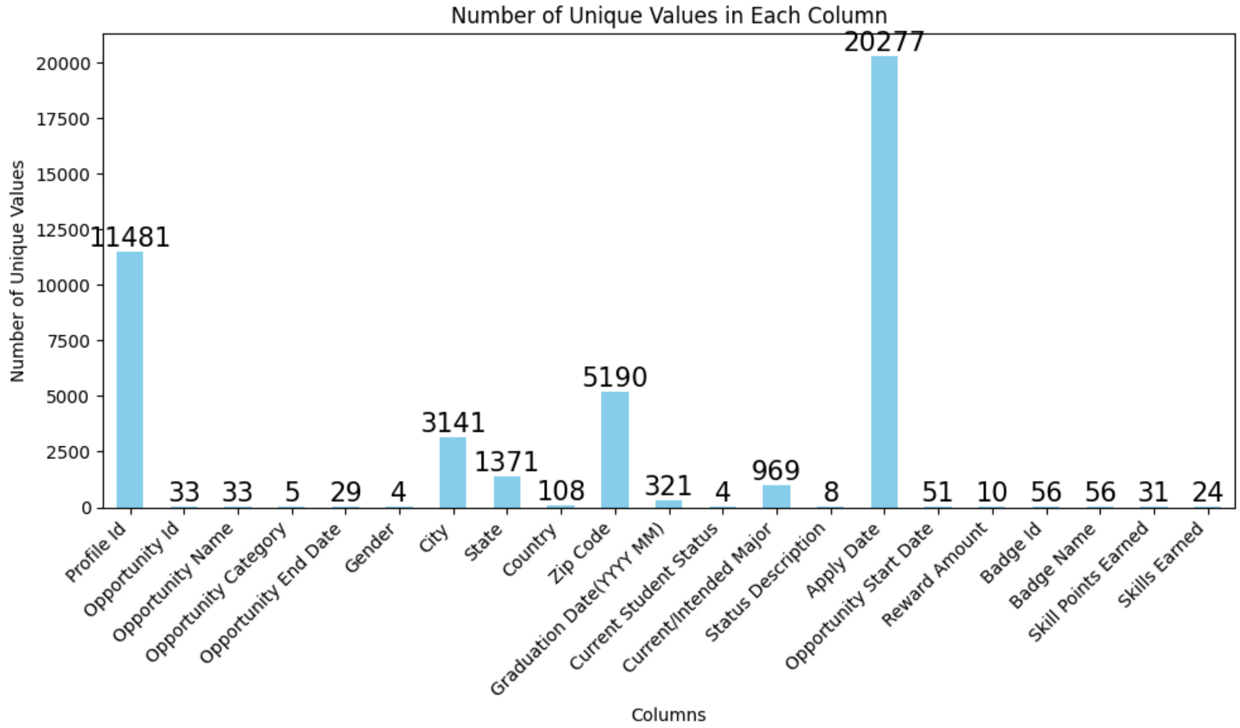


Opportunity wise data

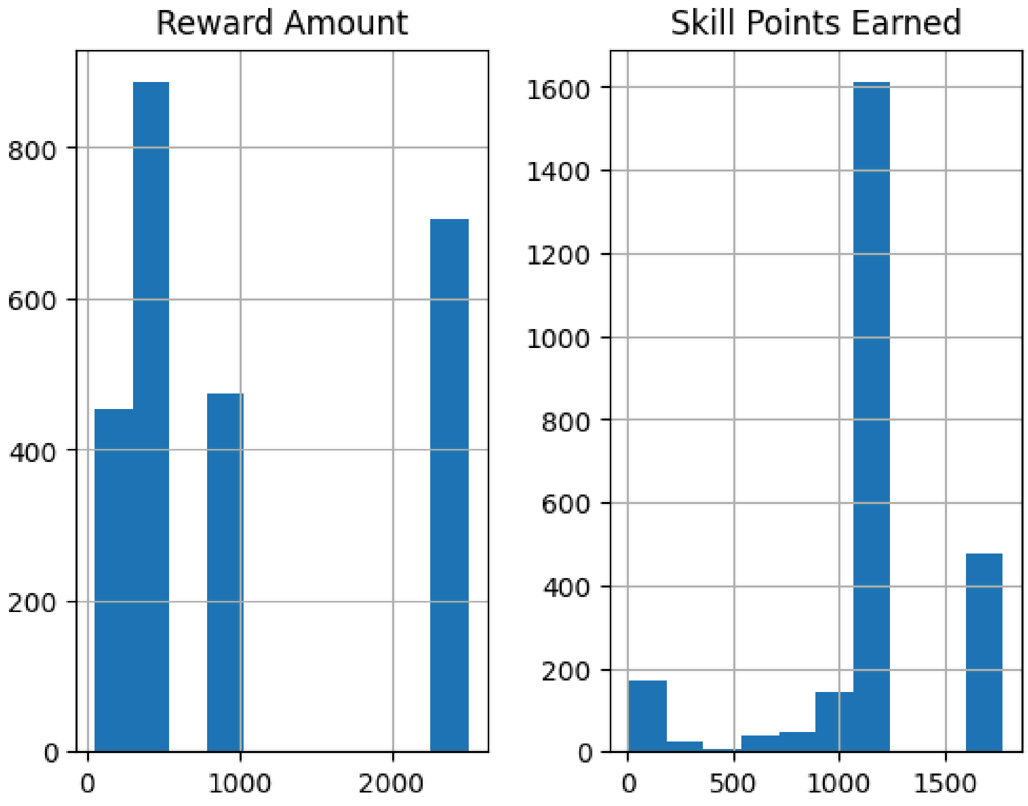


```
df1.nunique()
```

```
Profile Id                     11481
Opportunity Id                    33
Opportunity Name                  33
Opportunity Category               5
Opportunity End Date              29
Gender                             4
City                            3141
State                           1371
Country                          108
Zip Code                        5190
Graduation Date(YYYY MM)         321
Current Student Status             4
Current/Intended Major           969
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
```

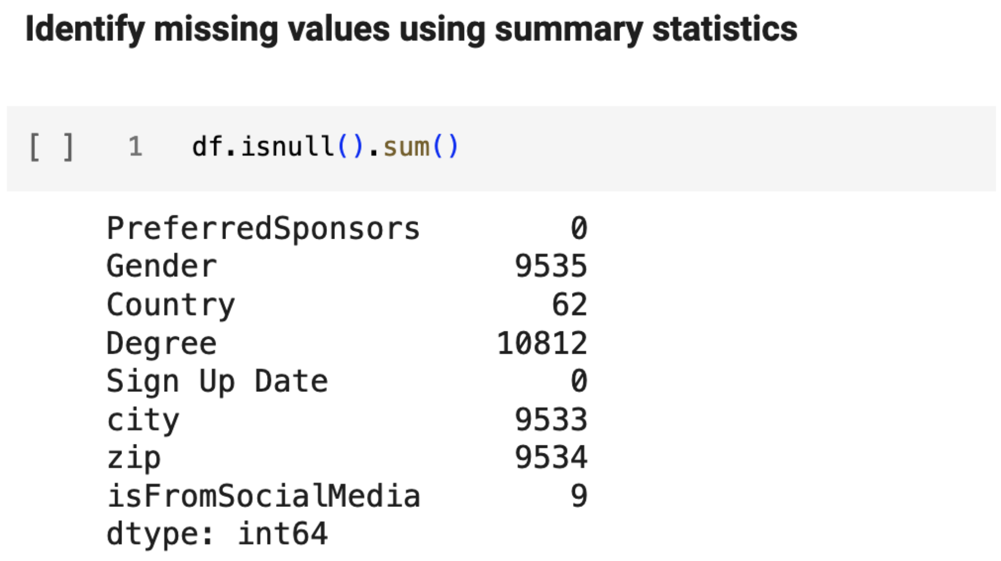


Assess data distributions



For both datasets, analyze each column's data type, and identify any potential issues (missing values, outliers).

User Data

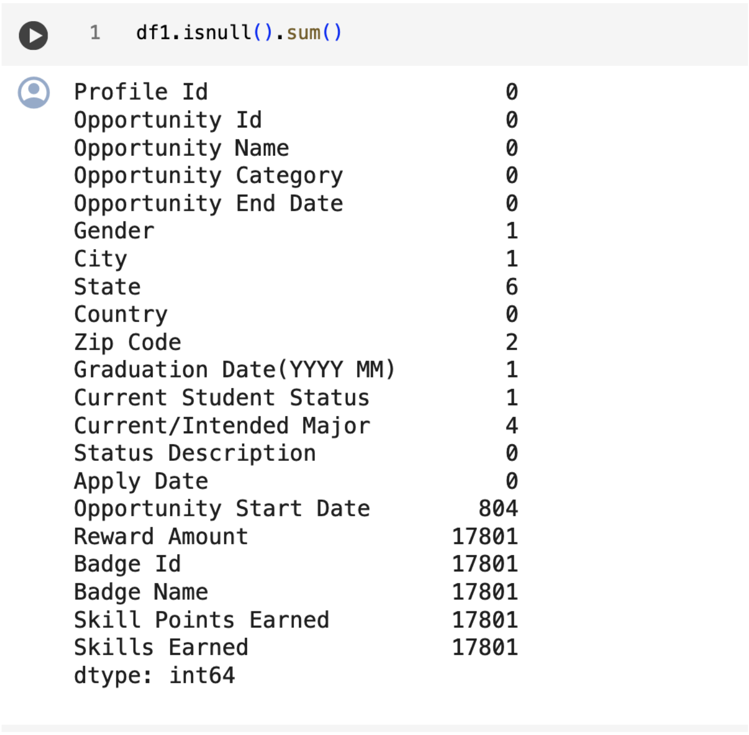


Identify missing values using summary statistics

```
[ ]   df.isnull().sum()
```

```
PreferredSponsors        0
Gender                9535
Country                 62
Degree               10812
Sign Up Date             0
city                  9533
zip                   9534
isFromSocialMedia        9
dtype: int64
```

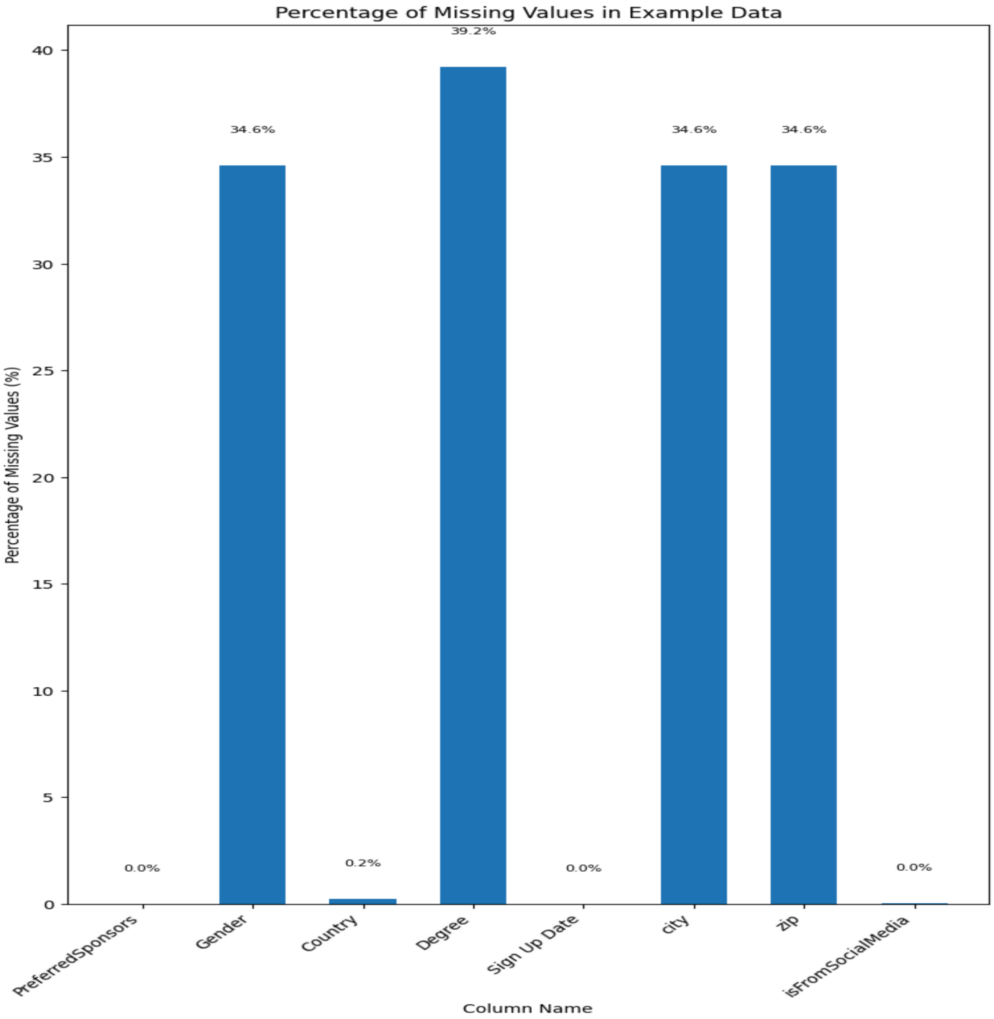
Opportunity wise Data

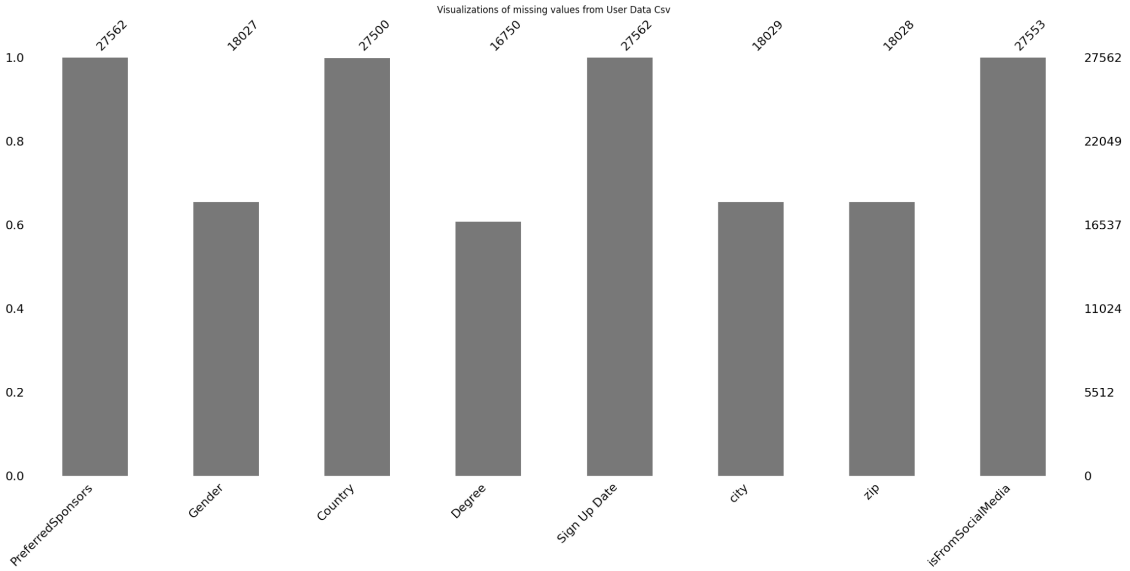


```
df1.isnull().sum()
```

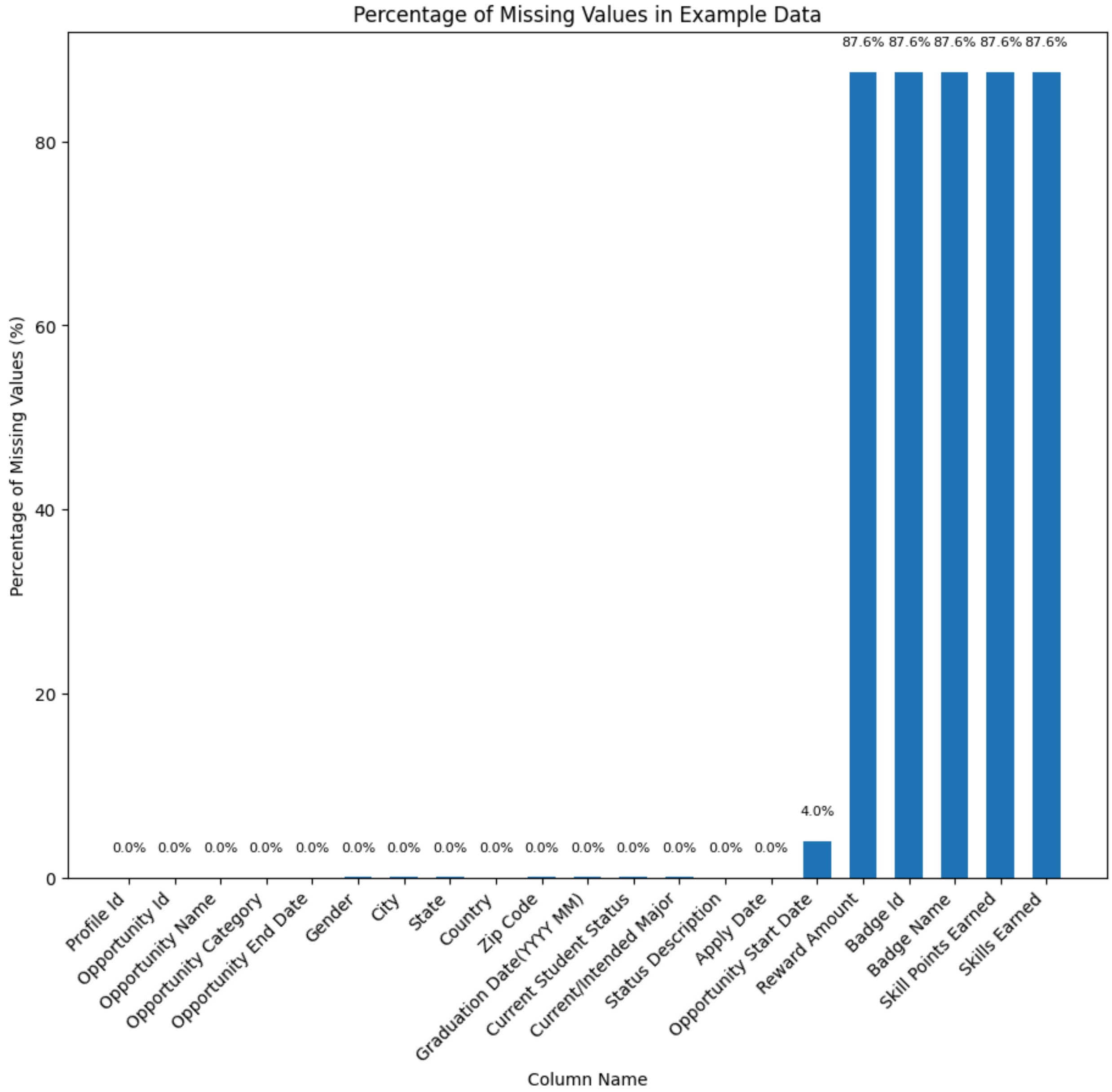
```
Profile Id                      0
Opportunity Id                  0
Opportunity Name                0
Opportunity Category            0
Opportunity End Date            0
Gender                          1
City                            1
State                           6
Country                         0
Zip Code                        2
Graduation Date(YYYY MM)        1
Current Student Status          1
Current/Intended Major          4
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
```

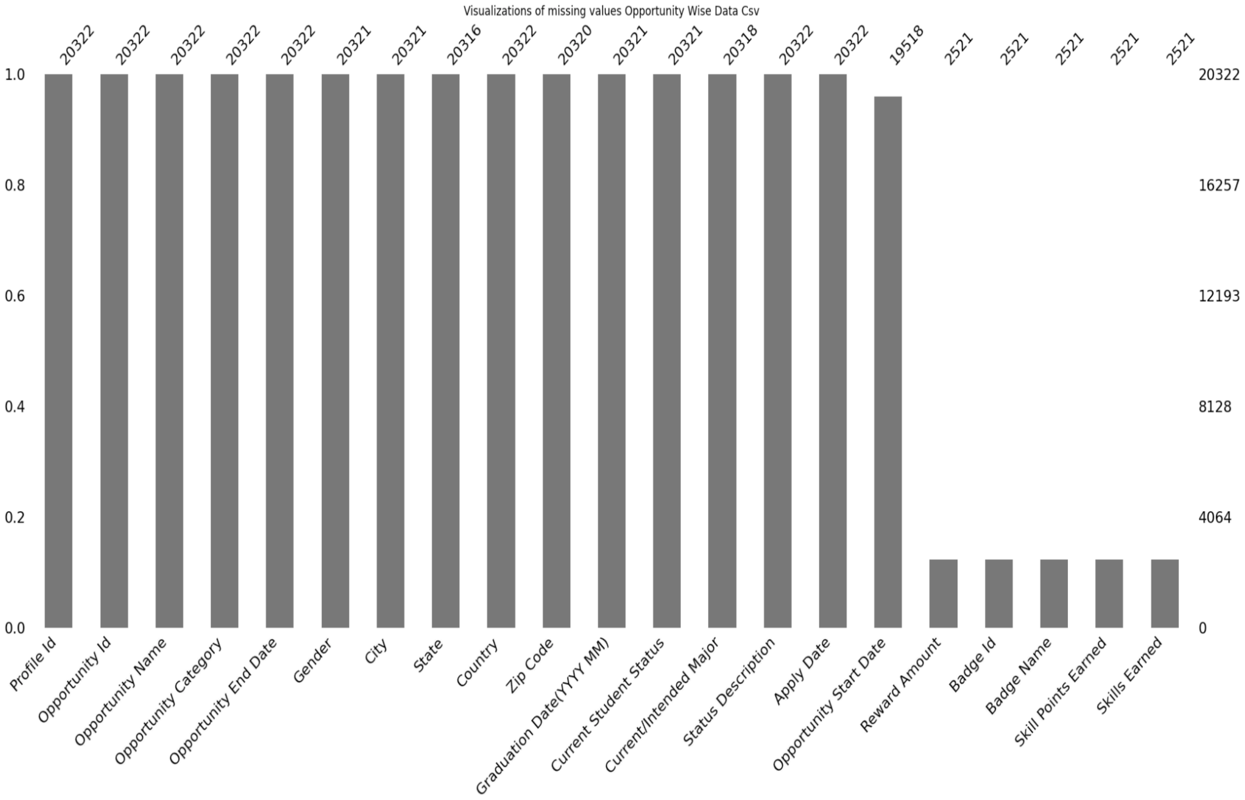
User Data


Percentage of Missing Values in Example Data
Percentage of Missing Values (%)
Column Name
0
5
10
15
20
25
30
35
40
0.0%
34.6%
0.2%
39.2%
0.0%
34.6%
34.6%
0.0%
PreferredSponsors
Gender
Country
Degree
Sign Up Date
city
zip
isFromSocialMedia

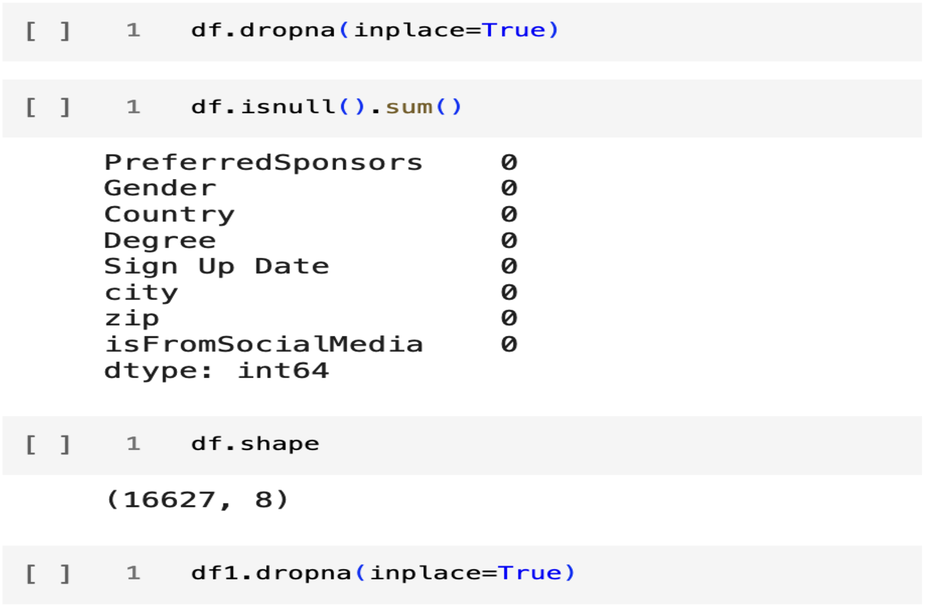

Visualizations of missing values from User Data Csv
27562
18027
27500
16750
27562
18029
18028
27553
1.0
0.8
0.6
0.4
0.2
0.0
27562
22049
16537
11024
5512
0
PreferredSponsors
Gender
Country
Degree
Sign Up Date
city
zip
isFromSocialMedia

Opportunity wise Data





Handling Missing Values



```
df.dropna(inplace=True)
```

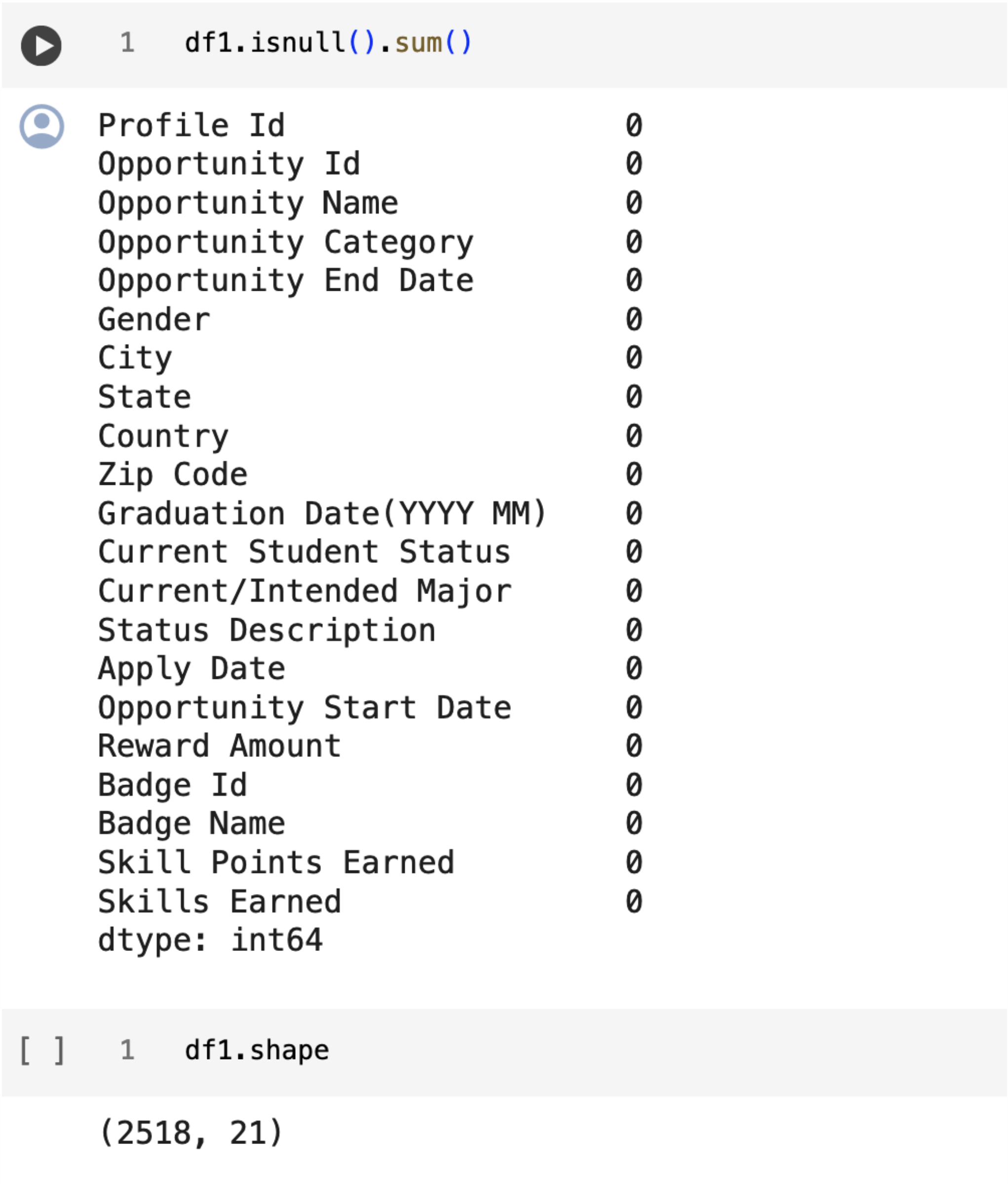
```
df.isnull().sum()
```

```
PreferredSponsors    0
Gender               0
Country              0
Degree               0
Sign Up Date         0
city                 0
zip                  0
isFromSocialMedia    0
dtype: int64
```

```
df.shape
```

```
(16627, 8)
```

```
df1.dropna(inplace=True)
```



```
df1.isnull().sum()
```

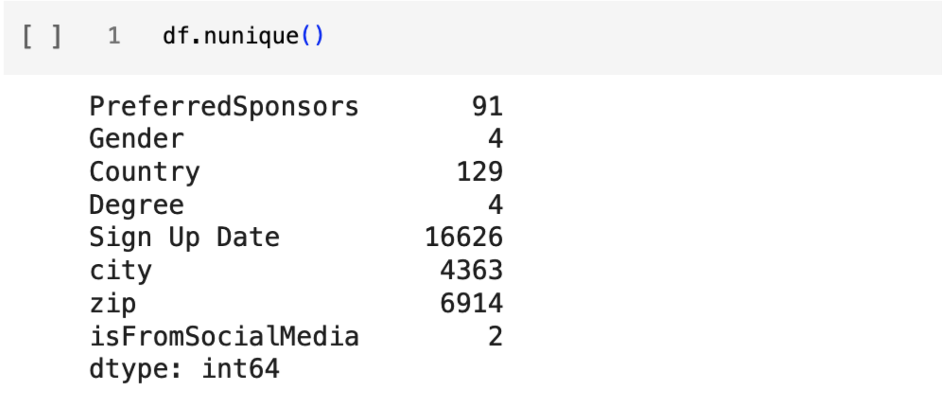
```
Profile Id                  0
Opportunity Id              0
Opportunity Name            0
Opportunity Category        0
Opportunity End Date        0
Gender                      0
City                        0
State                       0
Country                     0
Zip Code                    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
```

```
df1.shape
```

```
(2518, 21)
```

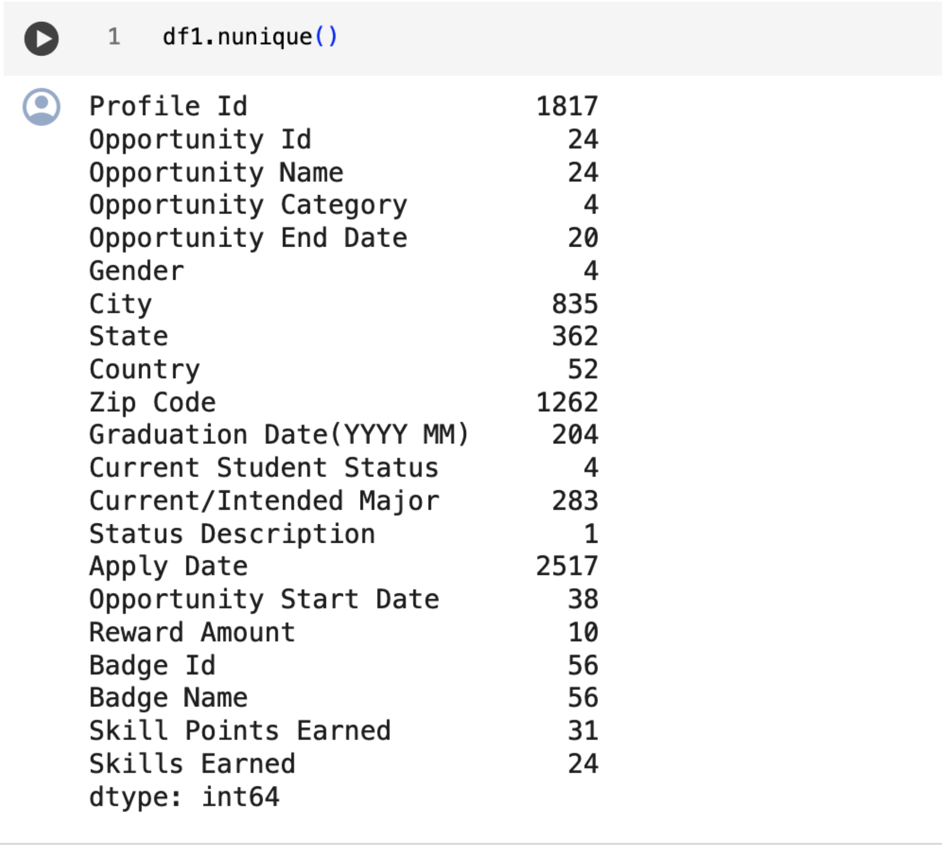
The unique values and their frequencies for categorical columns (e.g., Gender, City, Opportunity Category):

PROFILE ID ANALYSIS



```
df.nunique()
```

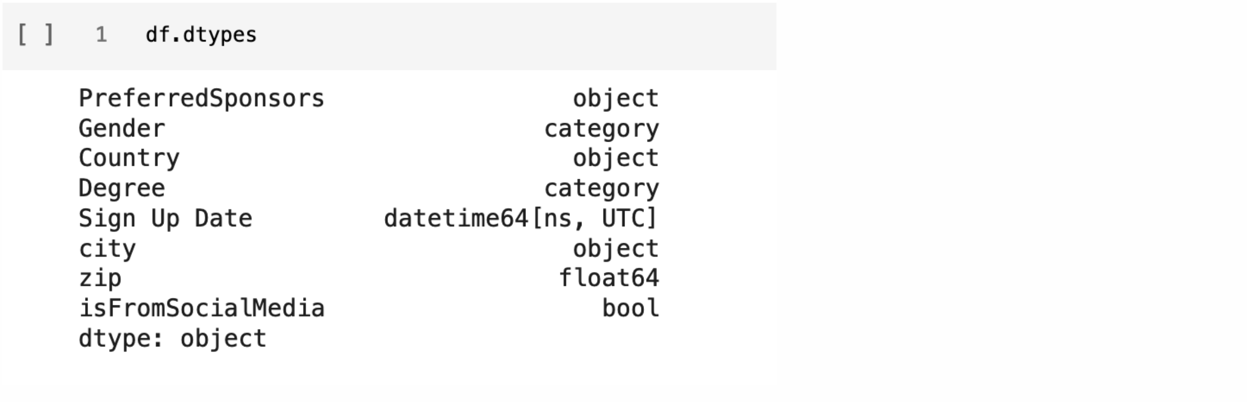
```
PreferredSponsors       91
Gender                   4
Country                129
Degree                   4
Sign Up Date         16626
city                  4363
zip                   6914
isFromSocialMedia        2
dtype: int64
```



```
df1.nunique()
```

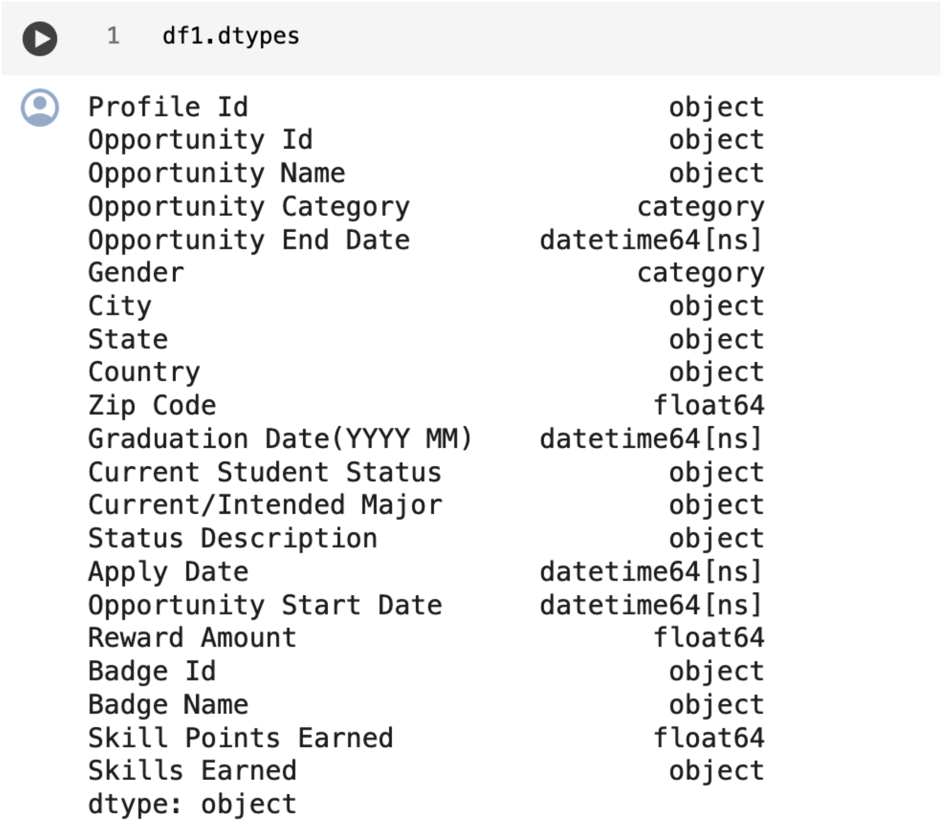
```
Profile Id                    1817
Opportunity Id                  24
Opportunity Name                24
Opportunity Category             4
Opportunity End Date            20
Gender                           4
City                           835
State                          362
Country                         52
Zip Code                      1262
Graduation Date(YYYY MM)       204
Current Student Status           4
Current/Intended Major         283
Status Description               1
Apply Date                    2517
Opportunity Start Date          38
Reward Amount                   10
Badge Id                        56
Badge Name                      56
Skill Points Earned             31
Skills Earned                   24
dtype: int64
```

We have uniqueness only for opportunity data because profile ID is provided only in opportunity-level data.



```
df.dtypes
```

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

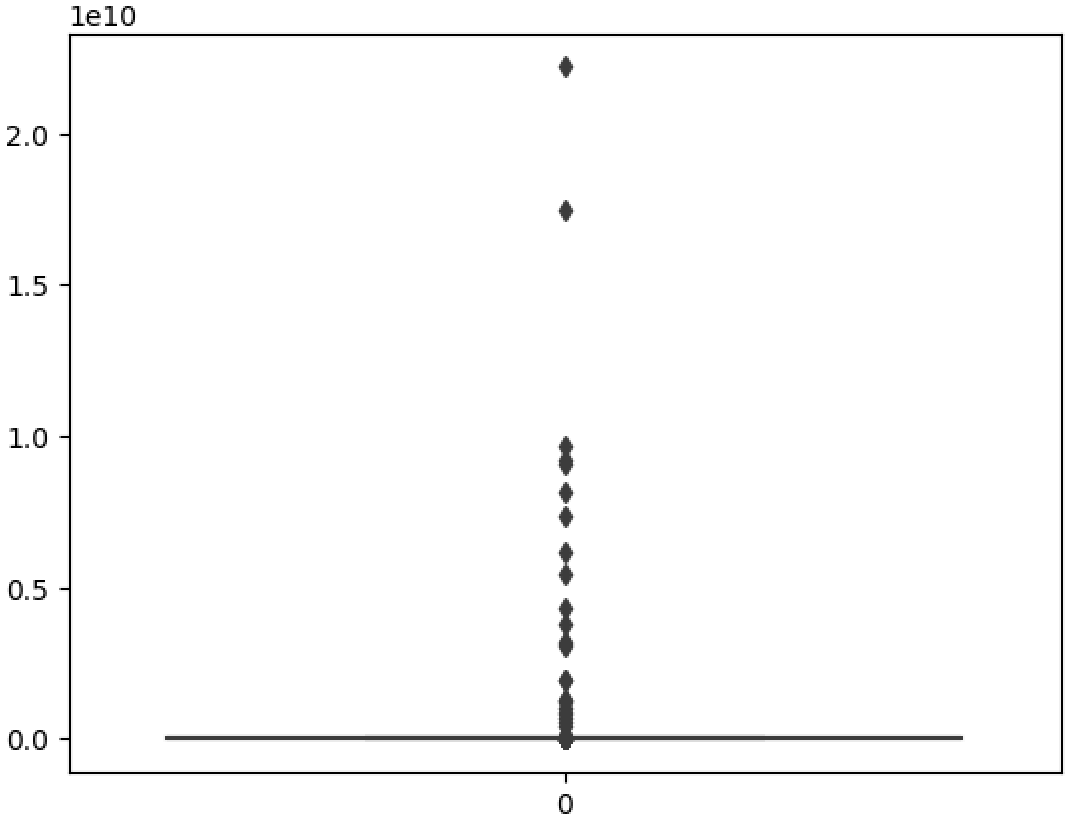


```
df1.dtypes
```

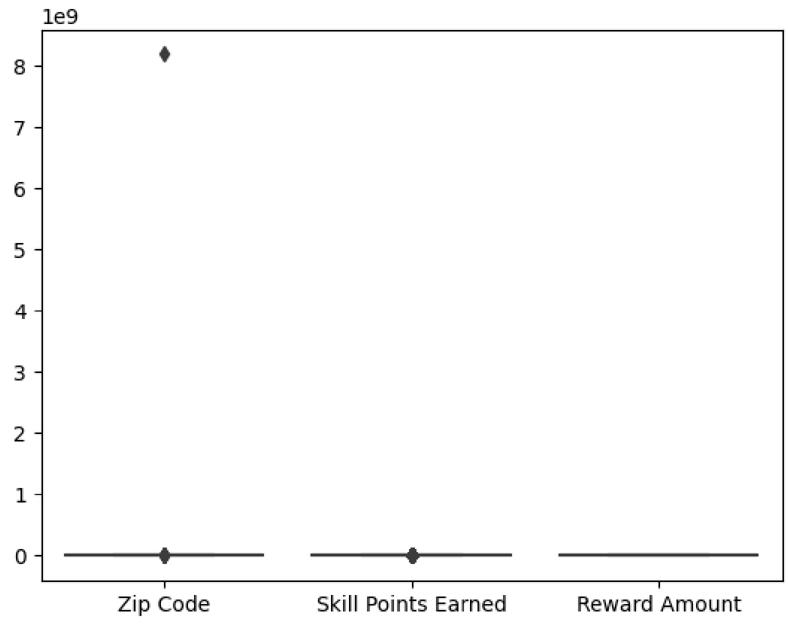
```
Profile Id                            object
Opportunity Id                        object
Opportunity Name                      object
Opportunity Category                category
Opportunity End Date          datetime64[ns]
Gender                              category
City                                  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
```

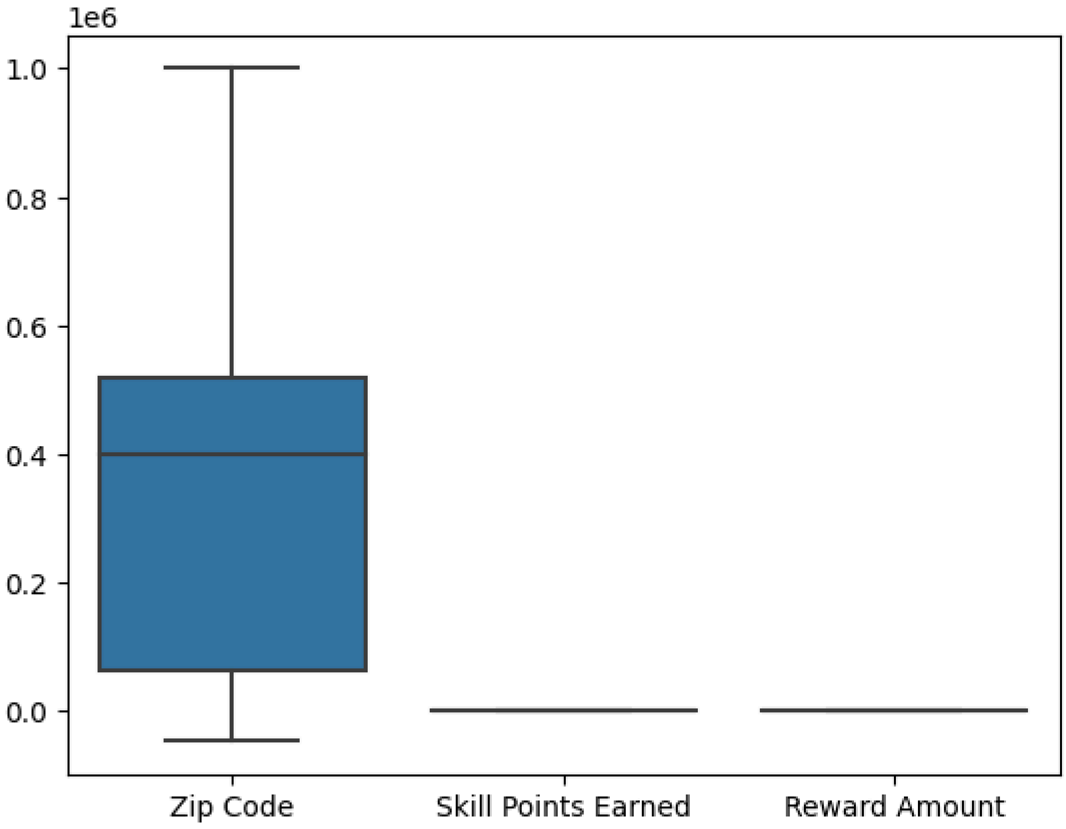
Validate Numeric Data:

userData

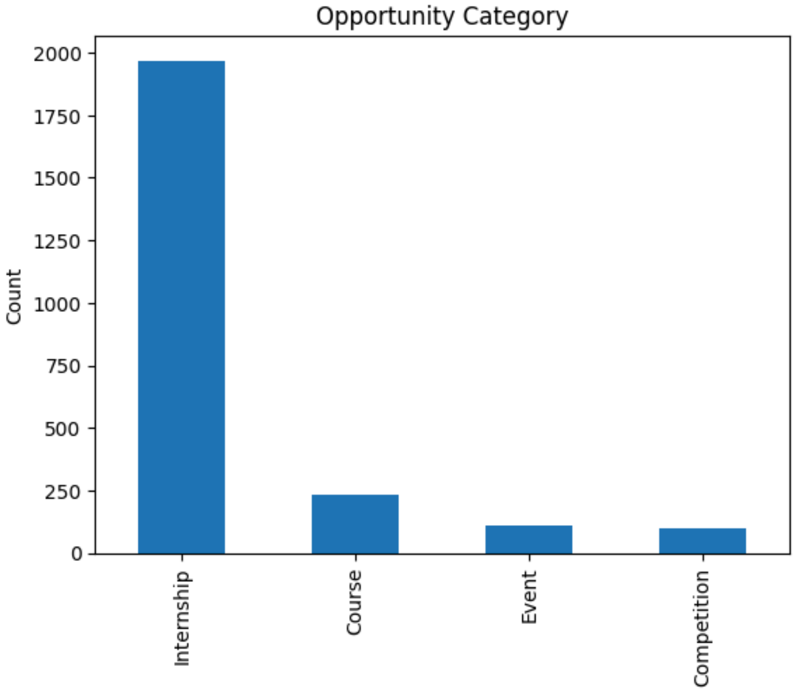


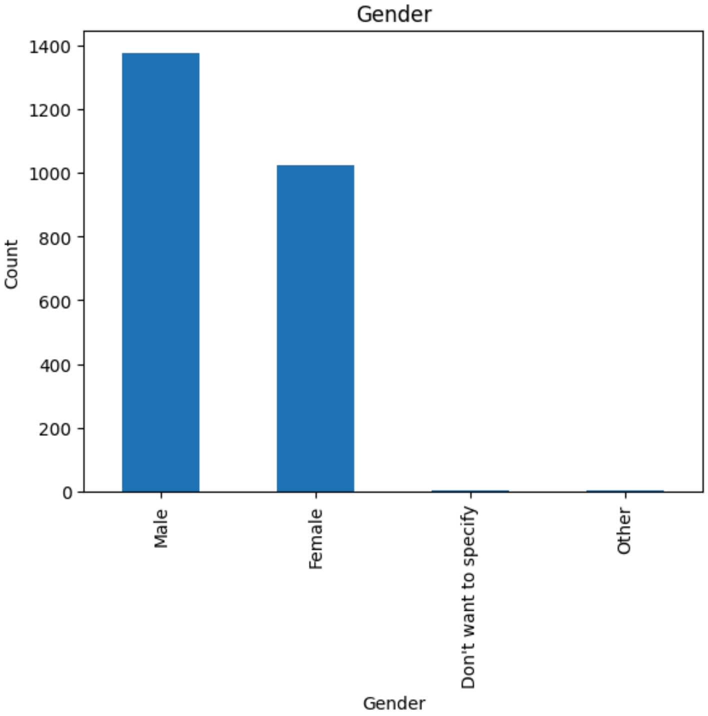
Opportunity Data

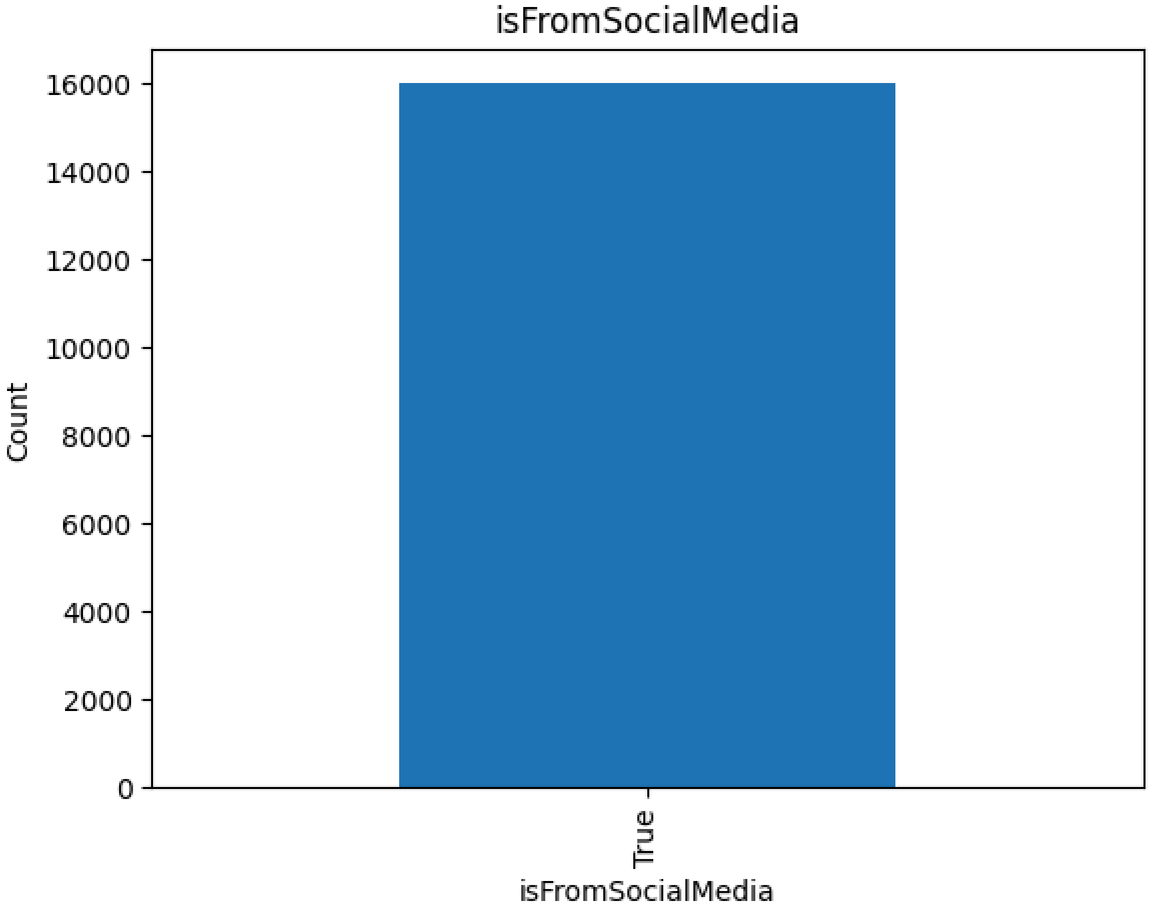




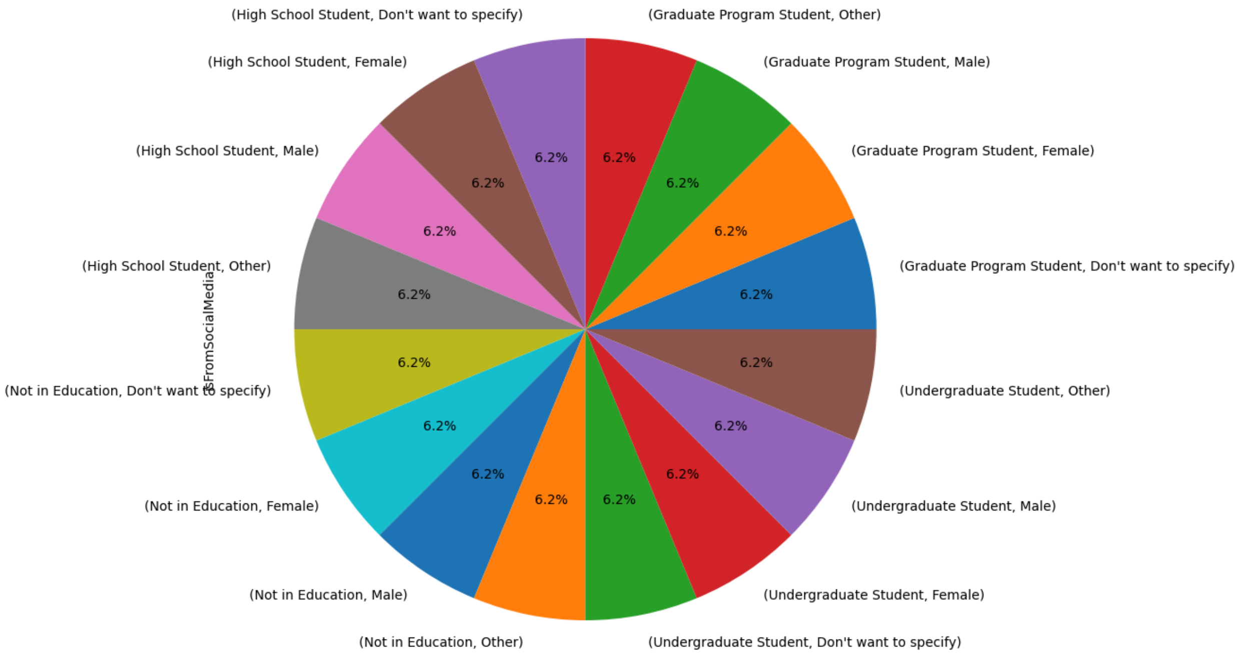
Validate Categorical Data:

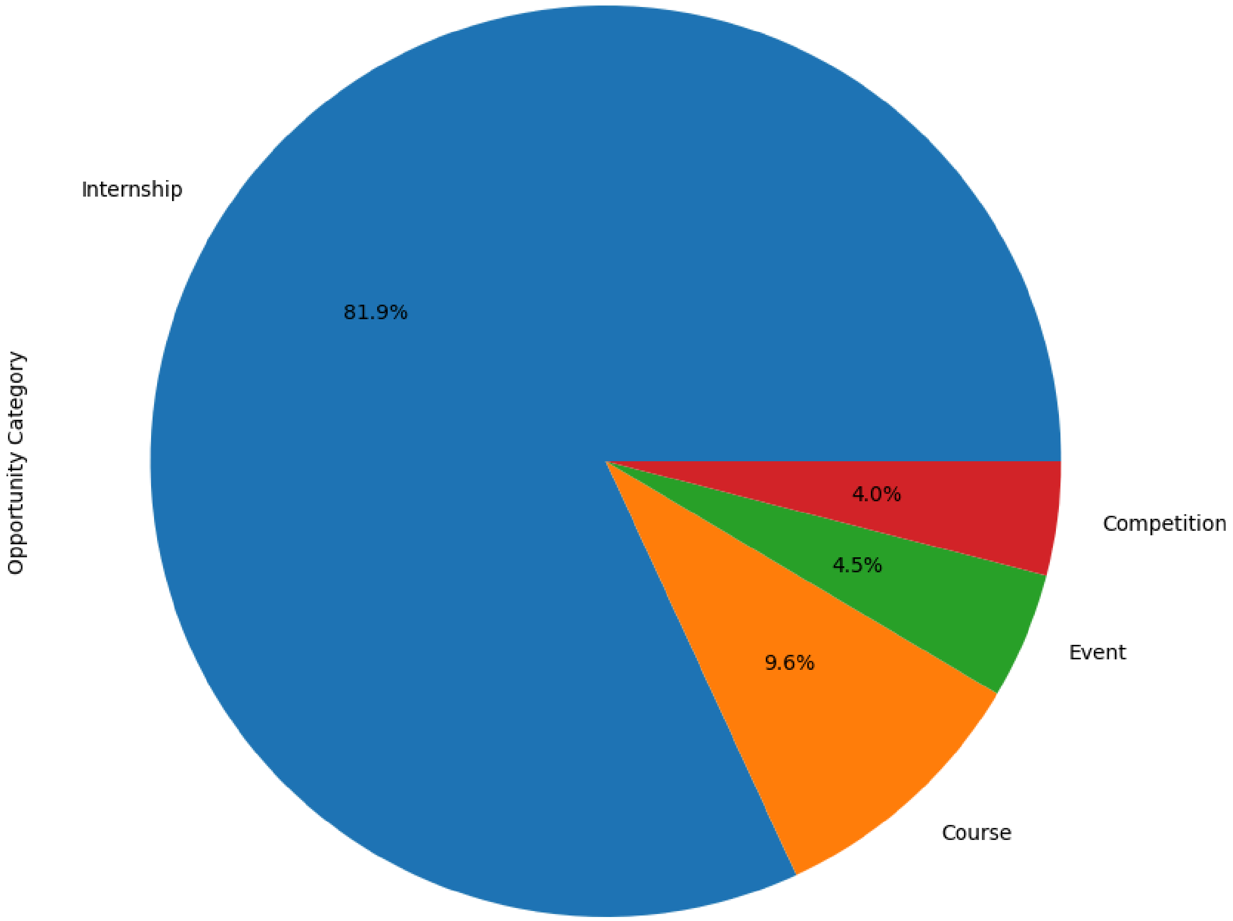

Opportunity Category
Count
2000
1750
1500
1250
1000
750
500
250
0
Internship
Course
Event
Competition
Opportunity Category


Gender
Count
1400
1200
1000
800
600
400
200
0
Male
Female
Don't want to specify
Other
Gender

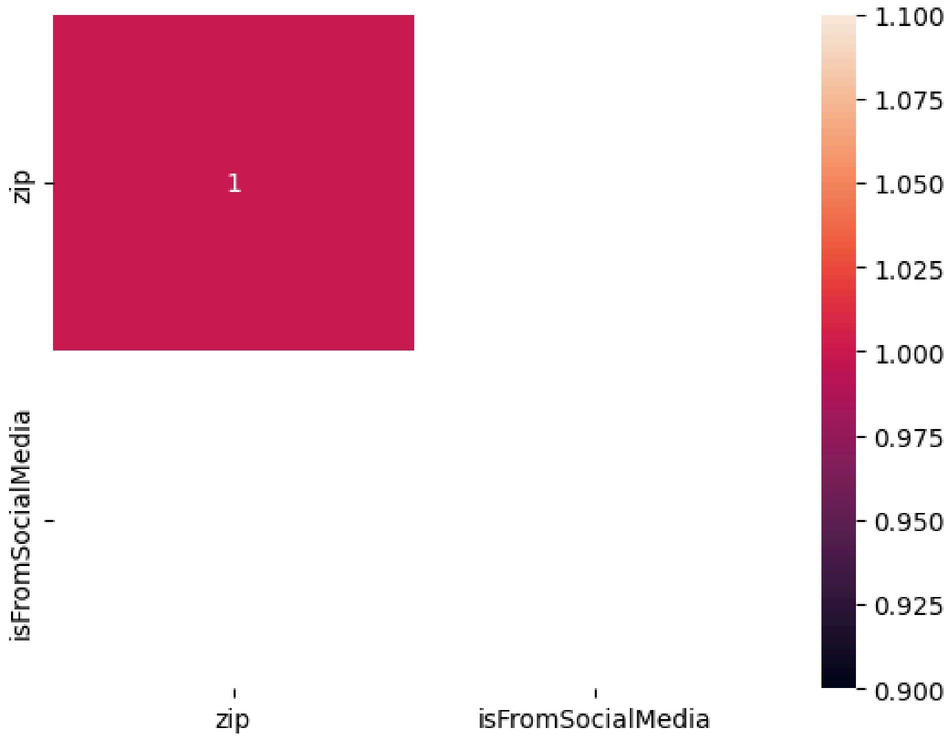


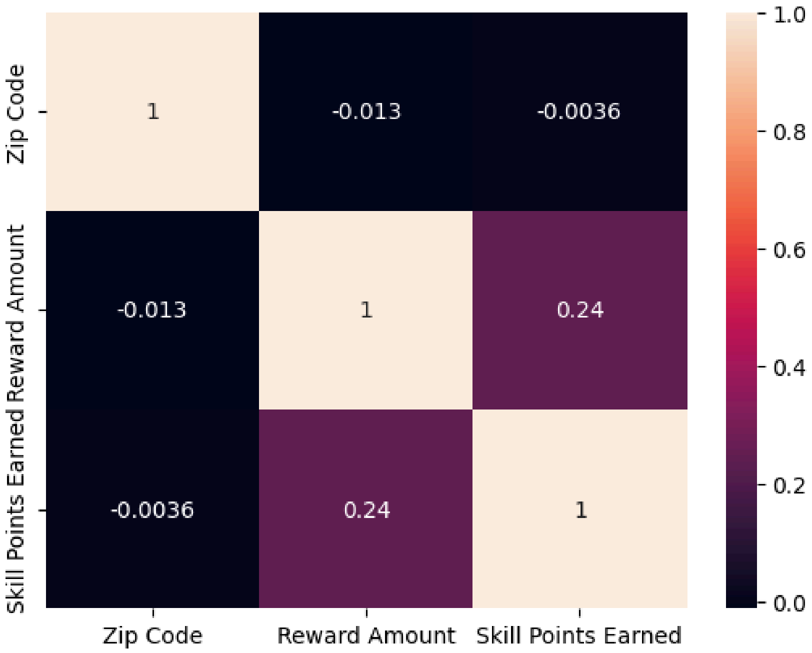
Cross-Check Relationships

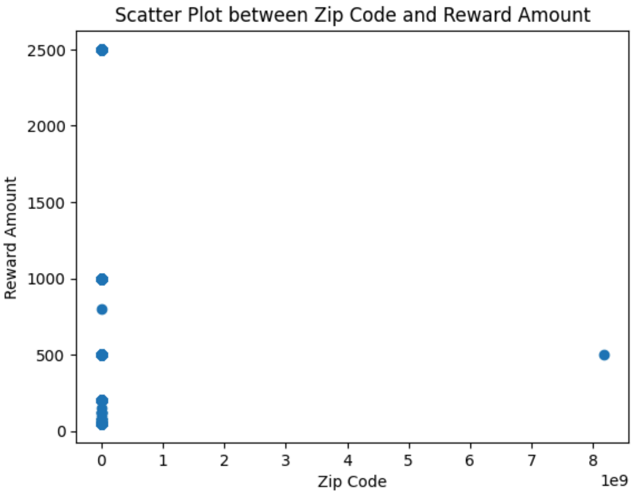

(High School Student, Don't want to specify)
(Graduate Program Student, Other)
(High School Student, Female)
(Graduate Program Student, Male)
(High School Student, Male)
(Graduate Program Student, Female)
(High School Student, Other)
(Graduate Program Student, Don't want to specify)
SFromSocialMedia
(Not in Education, Don't want to specify)
(Undergraduate Student, Other)
(Not in Education, Female)
(Undergraduate Student, Male)
(Not in Education, Male)
(Undergraduate Student, Female)
(Not in Education, Other)
(Undergraduate Student, Don't want to specify)
6.2%
6.2%
6.2%
6.2%
6.2%
6.2%
6.2%
6.2%
6.2%
6.2%
6.2%
6.2%
6.2%
6.2%
6.2%
6.2%


Opportunity Category
Internship
81.9%
Course
9.6%
Event
4.5%
Competition
4.0%

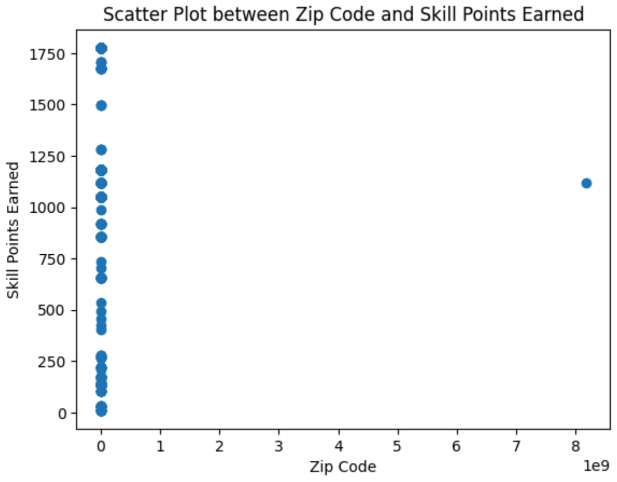
CORRELATION



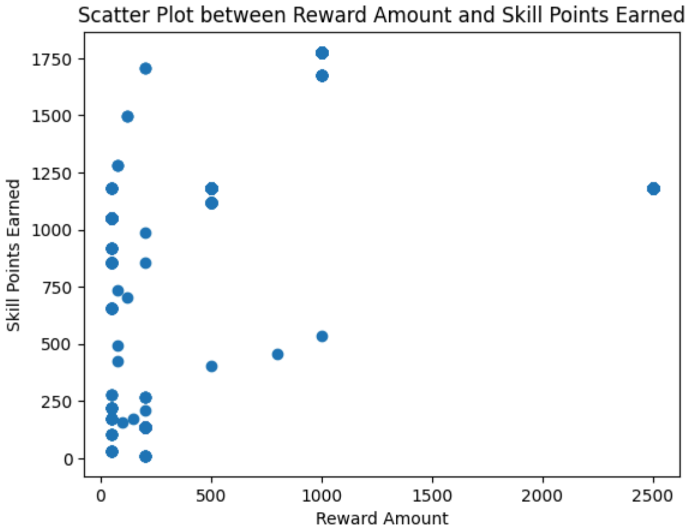




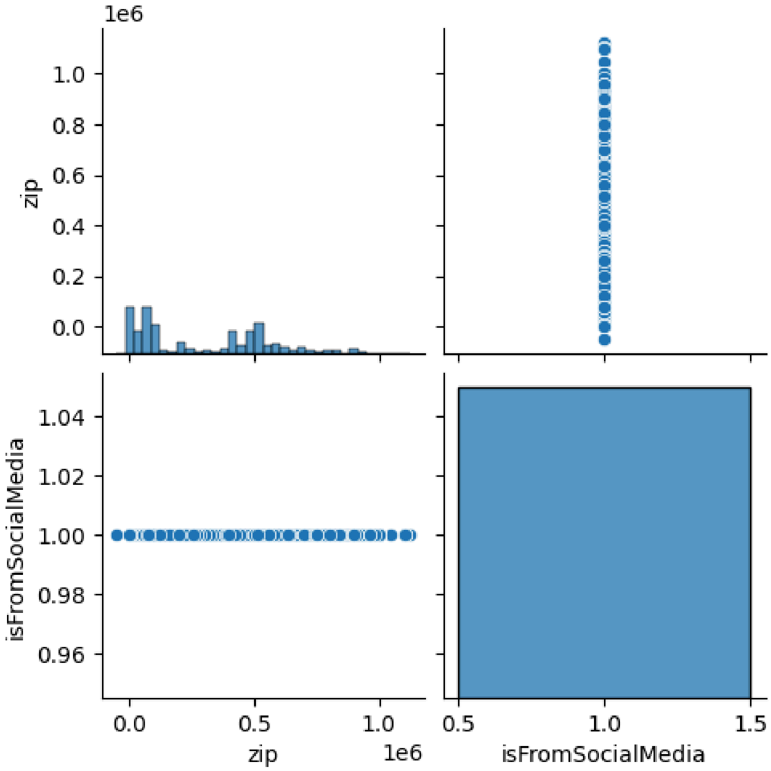
Scatter Plot between Zip Code and Reward Amount

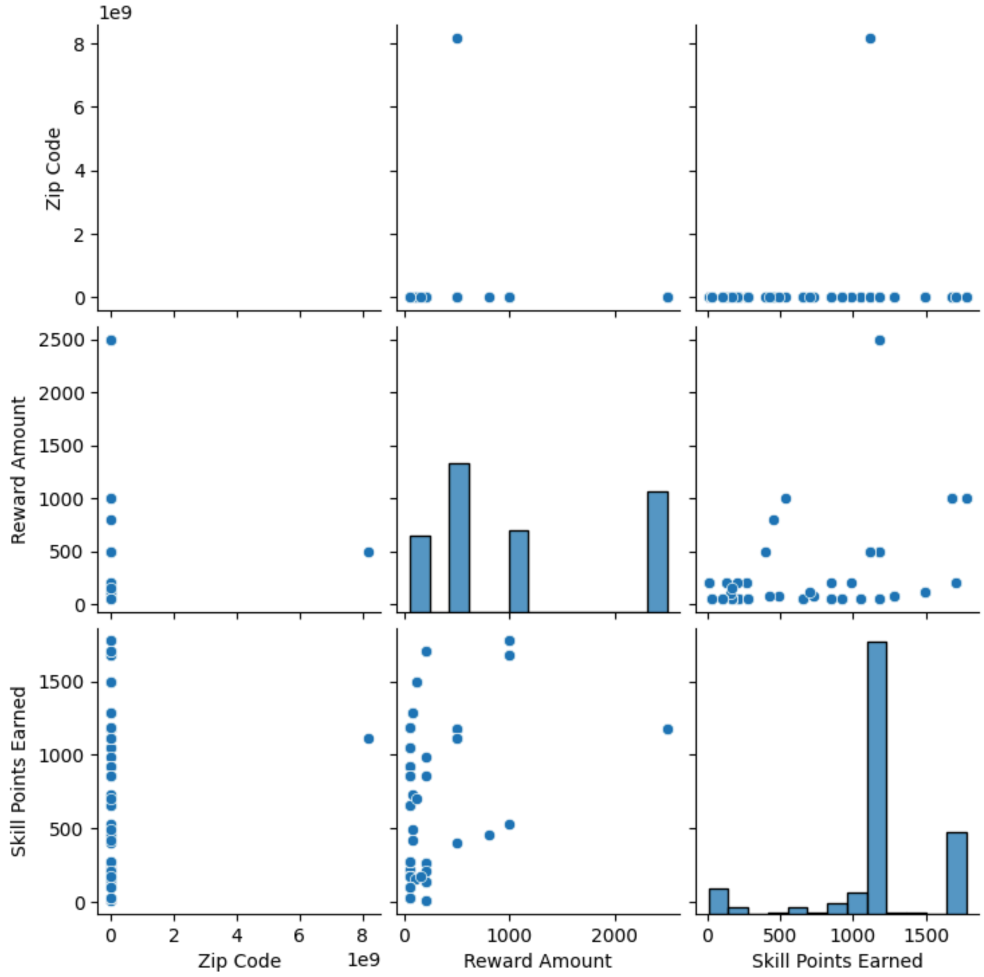


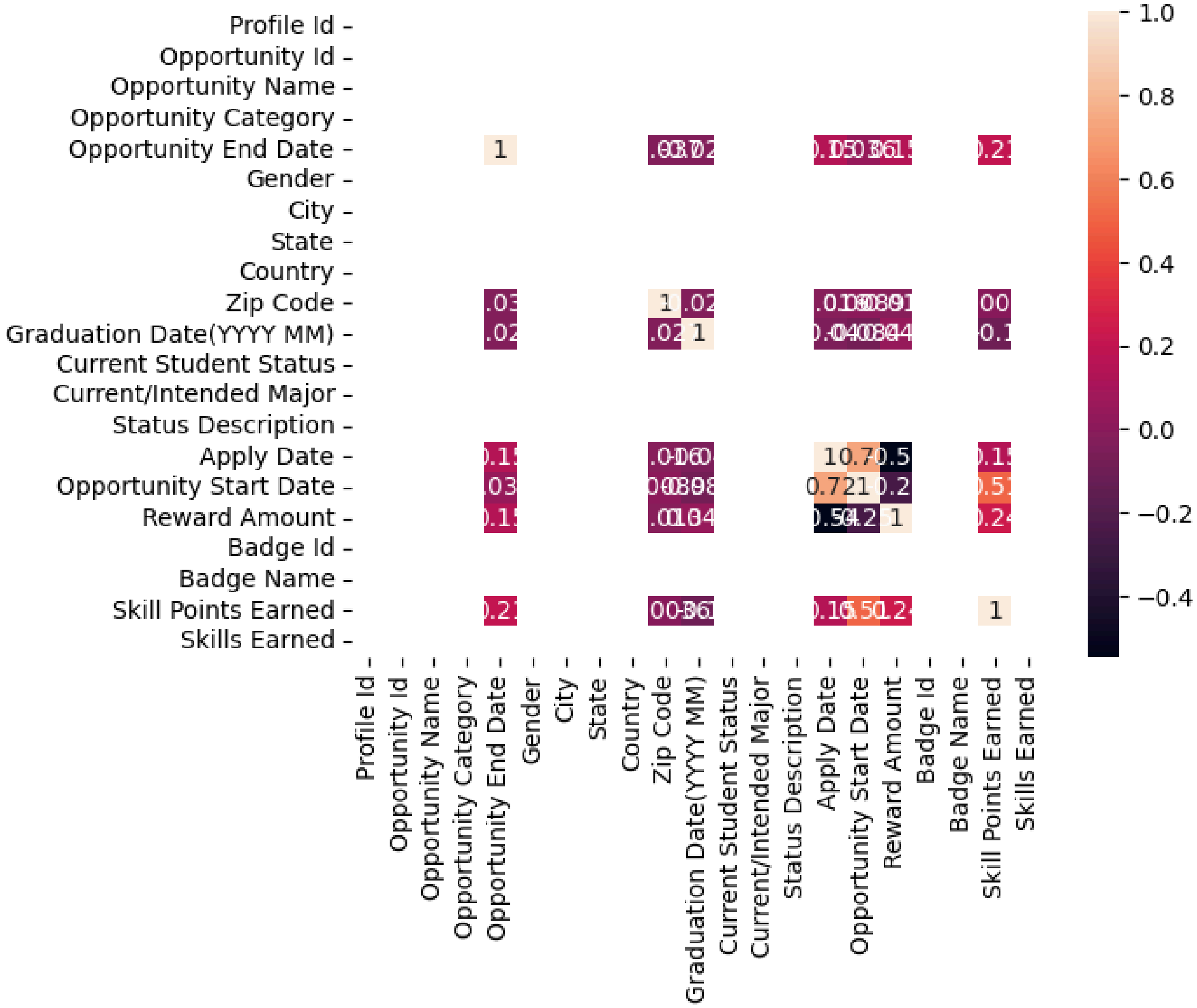
Scatter Plot between Zip Code and Skill Points Earned

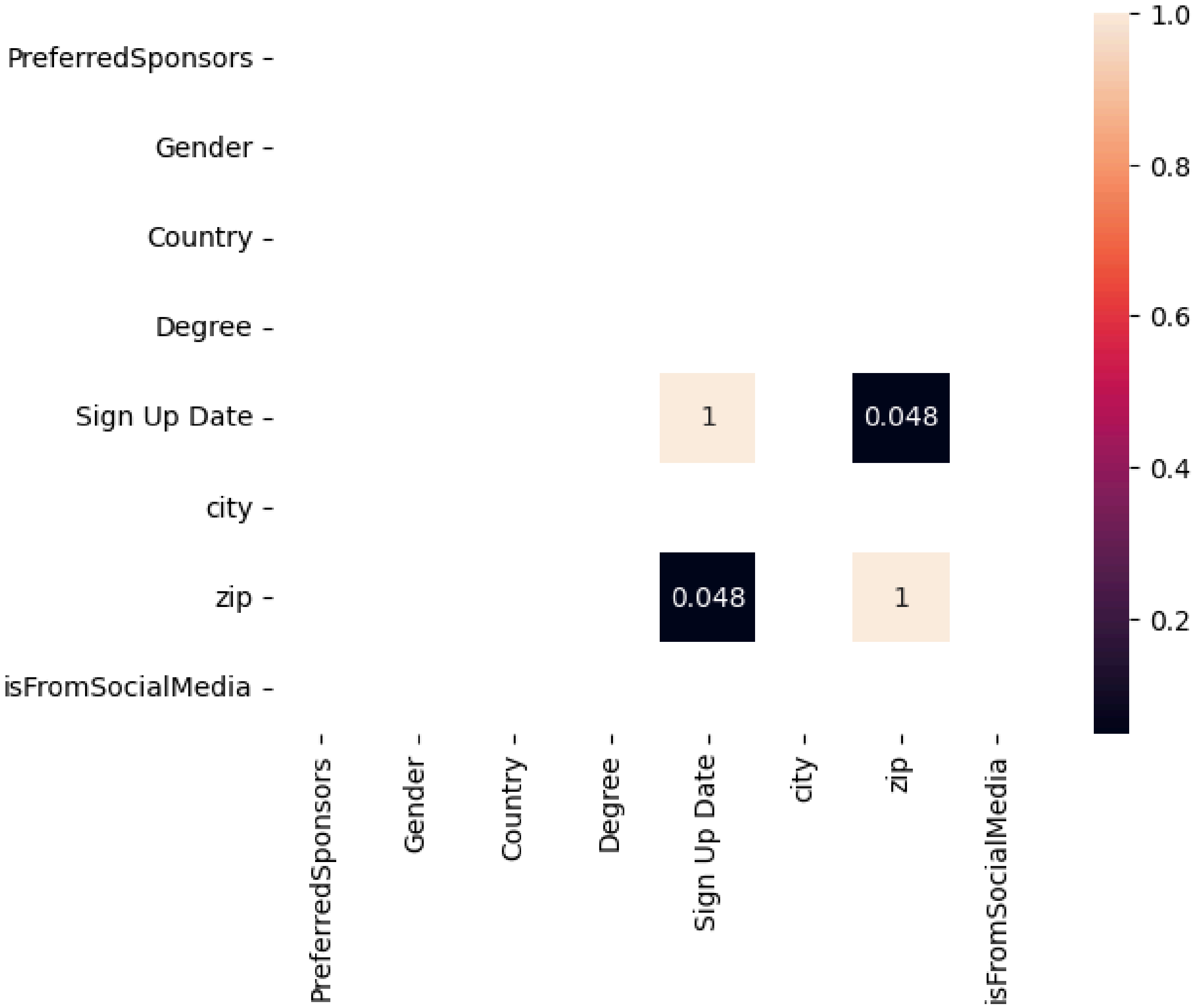


Scatter Plot between Reward Amount and Skill Points Earned


1e6
1.0
0.8
0.6
0.4
0.2
0.0
zip
1.04
1.02
1.00
0.98
0.96
isFromSocialMedia
0.0
0.5
1.0
zip
1e6
0.5
1.0
1.5
isFromSocialMedia


1e9
8
6
4
2
0
Zip Code
2500
2000
1500
1000
500
0
Reward Amount
1500
1000
500
0
Skill Points Earned
0
2
4
6
8
Zip Code
1e9
0
1000
2000
Reward Amount
0
500
1000
1500
Skill Points Earned


Profile Id
Opportunity Id
Opportunity Name
Opportunity Category
Opportunity End Date
Gender
City
State
Country
Zip Code
Graduation Date(YYYY MM)
Current Student Status
Current/Intended Major
Status Description
Apply Date
Opportunity Start Date
Reward Amount
Badge Id
Badge Name
Skill Points Earned
Skills Earned
1
0.21
.03
.02
1
1
0.15
1
0.72
1
0.2
0.24
0.51
0.15
1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


PreferredSponsors
Gender
Country
Degree
Sign Up Date
city
zip
isFromSocialMedia
1
0.048
0.048
1
PreferredSponsors
Gender
Country
Degree
Sign Up Date
city
zip
isFromSocialMedia
1.0
0.8
0.6
0.4
0.2

CHALLENGES FACED

During the exploration process, the following challenges and observations were encountered:

- There is less amount of numeric data
- Most of the attributes are of type object
- The count of the Male Applicants are more than female
- The Opportunity Category has the maximum Internship as the category
- There was the much need of validate the categorical and numeric data
- Too many missing values present in dataset
- The most of the application are from 2022
- Large amount of categorical data present