Week 1 Task

Data Cleaning and Validation

User Data

```
import pandas as pd
```

```
df=pd.read_csv('UserData.csv')
df
```

	PreferredSponsors	Gender	Country	Degree	Sign Up Date	city	zip	isFromSocialMedia
0	["GlobalShala","Grant Thornton China","Saint L...	Male	Nigeria	Undergraduate Student	2023-07-23T08:05:58.602Z	Owerri	460103	False
1	["GlobalShala","Grant Thornton China","Saint L...	Male	India	Undergraduate Student	2023-04-24T09:57:07.405Z	kottayam	686501	False
2	["GlobalShala","Illinois Institute of Technolo...	NaN	India	NaN	2022-10-14T17:13:36.303Z	NaN	NaN	False
3	["GlobalShala","Grant Thornton China","Saint L...	NaN	Albania	NaN	2023-06-06T12:29:01.772Z	NaN	NaN	True
4	["GlobalShala","Grant Thornton China","Saint L...	Female	Ghana	Not in Education	2023-06-15T16:31:42.719Z	Kumasi	AT-1214-9090	False
...	...	...	...	...	...	...	...	...
27557	["GlobalShala","Grant Thornton China","Saint	Female	Botswana	Undergraduate Student	2023-04-08T05:30:44.795Z	Gaborone	123456	True

Extracting Categorical variables of the dataset

```
[i for i in df.columns]
```

```
['PreferredSponsors',
 'Gender',
 'Country',
 'Degree',
 'Sign Up Date',
 'city',
 'zip',
 'isFromSocialMedia']
```

Checking descriptive statistics of the data

```
df.describe()
```

	PreferredSponsors	Gender	Country	Degree	Sign Up Date	city	zip	isFromSocialMedia
count	27562	18027	27500	16750	27562	18028	18018	27553
unique	94	4	169	4	27561	4727	7453	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

Checking count of Null Values in each column

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

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

Extracting rows containing even one null value

```
df[df.isnull().any(axis=1)]
```

	PreferredSponsors	Gender	Country	Degree	Sign Up Date	city	zip	isFromSocialMedia
2	["GlobalShala","Illinois Institute of Technolo...	NaN	India	NaN	2022-10-14T17:13:36.303Z	NaN	NaN	False
3	["GlobalShala","Grant Thornton China","Saint L...	NaN	Albania	NaN	2023-06-06T12:29:01.772Z	NaN	NaN	True
5	["GlobalShala","Grant Thornton China","Saint L...	Female	India	NaN	2023-07-06T18:49:16.691Z	Chennai	600033	False
6	["GlobalShala","Grant Thornton China","Saint L...	NaN	Nigeria	NaN	2023-05-15T21:30:04.370Z	NaN	NaN	True
7	["GlobalShala","Grant Thornton China","Saint L...	NaN	United States	NaN	2023-07-26T17:01:59.361Z	NaN	NaN	True
...	...	...	...	...	...	...	...	...
27548	["GlobalShala","Illinois Institute of Technolo...	NaN	Cameroon	NaN	2022-09-13T11:12:32.657Z	NaN	NaN	False
27549	["GlobalShala","Grant Thornton China","Saint L...	Female	India	NaN	2023-06-16T06:52:34.169Z	Karur	639117	True

Extracting rows with no null values

```
df.dropna()
```

	PreferredSponsors	Gender	Country	Degree	Sign Up Date	city	zip	isFromSocialMedia
0	["GlobalShala","Grant Thornton China","Saint L...	Male	Nigeria	Undergraduate Student	2023-07-23T08:05:58.602Z	Owerri	460103	False
1	["GlobalShala","Grant Thornton China","Saint L...	Male	India	Undergraduate Student	2023-04-24T09:57:07.405Z	kottayam	686501	False
4	["GlobalShala","Grant Thornton China","Saint L...	Female	Ghana	Not in Education	2023-06-15T16:31:42.719Z	Kumasi	AT-1214-9090	False
8	["GlobalShala","Grant Thornton China","Saint L...	Male	Nigeria	Undergraduate Student	2023-07-27T18:02:17.535Z	Lagos	100278	True
9	["GlobalShala","Grant Thornton China","Saint L...	Male	India	High School Student	2023-05-05T04:47:25.446Z	RAS	388570	True
...	...	...	...	...	...	...	...	...
	["GlobalShala","Grant							

Replacing null values with mean median and mod

```
newData=df
newData.describe()
```

	PreferredSponsors	Gender	Country	Degree	Sign Up Date	city	zip	isFromSocialMedia
count	27562	27562	27500	16750	27562	18028	18018	27553
unique	94	4	169	4	27561	4727	7453	2
top	["GlobalShala","Grant Thornton China","Saint L...	Male	India	Undergraduate Student	2022-10-30T17:25:54.072Z	Hyderabad	63108	True
freq	22011	20562	11893	6527	2	743	629	13811

```
newData["Gender"].fillna(df.Gender.mode()[0],inplace=True)
newData.Country.fillna(df.Country.mode()[0],inplace=True)
```

```
C:\Users\mahmo\AppData\Local\Temp\ipykernel_1572\1136611627.py:2: FutureWarning: A value is trying to be set on a copy
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[


  newData.Country.fillna(df.Country.mode()[0],inplace=True)
newData.Degree.fillna(df.Degree.mode()[0],inplace=True)
newData.city.fillna(df.city.mode()[0],inplace=True)
newData.zip.fillna(df.zip.mode()[0],inplace=True)
newData.isFromSocialMedia.fillna(df.isFromSocialMedia.mode()[0],inplace=True)
```

```
C:\Users\mahmo\AppData\Local\Temp\ipykernel_1572\2443661360.py:4: FutureWarning: A value is trying to be set on a copy
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[


  newData.isFromSocialMedia.fillna(df.isFromSocialMedia.mode()[0],inplace=True)
C:\Users\mahmo\AppData\Local\Temp\ipykernel_1572\2443661360.py:4: FutureWarning: Downcasting object dtype arrays on .fi
  newData.isFromSocialMedia.fillna(df.isFromSocialMedia.mode()[0],inplace=True)
```

```
newData
```

	PreferredSponsors	Gender	Country	Degree	Sign Up Date	city	zip	isFromSocialMedia
0	["GlobalShala","Grant Thornton China","Saint L...	Male	Nigeria	Undergraduate Student	2023-07-23T08:05:58.602Z	Owerri	460103	False
1	["GlobalShala","Grant Thornton China","Saint L...	Male	India	Undergraduate Student	2023-04-24T09:57:07.405Z	kottayam	686501	False
2	["GlobalShala","Illinois Institute of Technolo...	Male	India	Undergraduate Student	2022-10-14T17:13:36.303Z	Hyderabad	63108	False
3	["GlobalShala","Grant Thornton China","Saint L...	Male	Albania	Undergraduate Student	2023-06-06T12:29:01.772Z	Hyderabad	63108	True
4	["GlobalShala","Grant Thornton China","Saint L...	Female	Ghana	Not in Education	2023-06-15T16:31:42.719Z	Kumasi	AT-1214-9090	False
...	...	...	...	...	...	...	...	...
27557	["GlobalShala","Grant Thornton China","Saint	Female	Botswana	Undergraduate	2023-04-	Gaborone	123456	True

```
newData.isnull().sum()
```

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

Opportunity Wise Data

```
df=pd.read_csv("Opportunity Wise Data.csv")
df.head()
```

	Profile Id	Opportunity Id	Opportunity Name	Opportunity Category	Opportunity End Date	Gender	City	State	Country	Zip Code	..
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	Dhaka	Savar	Bangladesh	1342	
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	Amritsar	Punjab	Afghanistan	123242	
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	Satna	Madhya pradesh	India	485001	
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	Hyderabad	Telangana	India	500039	
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	Bangalore	Karnataka	India	560105	

5 rows × 21 columns

```
[i for i in df.columns]
```

```
['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']
```

```
df.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

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

```
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
```

```
df[df.isnull().any(axis=1)]
```

	Profile Id	Opportunity Id	Opportunity Name	Opportunity Category	Opportunity End Date	Gender	City	State	Country	C
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	Dhaka	Savar	Bangladesh	1
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	Satna	Madhya pradesh	India	485
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	Hyderabad	Telangana	India	500
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	Bangalore	Karnataka	India	560
6	2b39f489-0bb7-4ea2-9a5f-de98868c4ec3	00000000-0GT8-HCVB-01AE-6QEP8Y	Life Beyond Saint Louis University's Campus	Event	Oct 27, 2022, 18:29:00	Male	Agra	Uttar Pradesh	India	282
...	...	...	...	...	...	...	...	...	...	
20317	f386224b-4b64-4d70-a6c5-8d90e3653925	00000000-101Y-HSX2-0DFJ-QCKQBR	AI Ethics Challenge	Competition	Oct 31, 2023, 14:45:36	Male	Bijnor	Uttar Pradesh	India	246
20318	f398b382-ac7a-4b14-8f76-cd41a51b1459	00000000-101Y-HSX2-0DFJ-QCKQBR	AI Ethics Challenge	Competition	Oct 31, 2023, 14:45:36	Male	College Station	Texas	United States	77
20319	f476e230-266d-491b-a693-f3f3bccac7d6	00000000-101Y-HSX2-0DFJ-QCKQBR	AI Ethics Challenge	Competition	Oct 31, 2023, 14:45:36	Female	Narasaraopet	Andhra Pradesh	India	522
20320	f92acfd4-3888-447a-a6dd-f996544eebbb	00000000-101Y-HSX2-0DFJ-QCKQBR	AI Ethics Challenge	Competition	Oct 31, 2023, 14:45:36	Female	Saint Louis	Missouri	United States	63
20321	fdccf84d-6011-4048-ad8d-73df5e7c431e	00000000-101Y-HSX2-0DFJ-QCKQBR	AI Ethics Challenge	Competition	Oct 31, 2023, 14:45:36	Male	Rangpur	Rangpur	Bangladesh	5

17808 rows × 21 columns

```
newData2=df
```

```
newData2.isnull().sum()
```

```
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
```

```
newData2.Gender.fillna(df.Gender.mode()[0],inplace=True)
newData2.Gender.isnull().sum()
```

```
np.int64(0)
```

```
newData2.City.fillna(df.City.mode()[0],inplace=True)
newData2.State.fillna(df.State.mode()[0],inplace=True)
newData2["Zip Code"].fillna(df["Zip Code"].mode()[0],inplace=True)
newData2["Graduation Date(YYYY MM)"].fillna(df["Graduation Date(YYYY MM)"].mode()[0],inplace=True)
newData2["Current Student Status"].fillna(df["Current Student Status"].mode()[0],inplace=True)
newData2["Current/Intended Major"].fillna(df["Current/Intended Major"].mode()[0],inplace=True)
newData2["Opportunity Start Date"].fillna(df["Opportunity Start Date"].mode()[0],inplace=True)
newData2["Reward Amount"].fillna("0",inplace=True)
newData2["Badge Id"].fillna("0",inplace=True)
```

```
C:\Users\mahmo\AppData\Local\Temp\ipykernel_1572\3385531472.py:8: FutureWarning: A value is trying to be set on a copy
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[


  newData2["Reward Amount"].fillna("0",inplace=True)
C:\Users\mahmo\AppData\Local\Temp\ipykernel_1572\3385531472.py:8: FutureWarning: Setting an item of incompatible dtype
  newData2["Reward Amount"].fillna("0",inplace=True)
C:\Users\mahmo\AppData\Local\Temp\ipykernel_1572\3385531472.py:9: FutureWarning: A value is trying to be set on a copy
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[


  newData2["Badge Id"].fillna("0",inplace=True)
```

```
newData2["Badge Name"].fillna("No Badge",inplace=True)
newData2["Skill Points Earned"].fillna(0,inplace=True)
newData2["Skills Earned"].fillna(0,inplace=True)
```

```
C:\Users\mahmo\AppData\Local\Temp\ipykernel_1572\1584956032.py:1: FutureWarning: A value is trying to be set on a copy
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[


  newData2["Badge Name"].fillna("No Badge",inplace=True)
C:\Users\mahmo\AppData\Local\Temp\ipykernel_1572\1584956032.py:2: FutureWarning: A value is trying to be set on a copy
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[


  newData2["Skill Points Earned"].fillna(0,inplace=True)
C:\Users\mahmo\AppData\Local\Temp\ipykernel_1572\1584956032.py:3: FutureWarning: A value is trying to be set on a copy
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[


  newData2["Skills Earned"].fillna(0,inplace=True)
```

```
newData2.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
```

```
newData.to_csv('NewUserData.csv', index=False)
```

```
newData2.to_csv('NewOpportunityData.csv', index=False)
```