eda

June 7, 2024

1 IMPORTING DEPENDENCIES

[ ]:
```
from sklearn.preprocessing import StandardScaler
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import seaborn as sns
```

[ ]:
```
df = pd.read_csv("/content/HealthcareProviders.csv")
```

2 BASIC EXPLORATION OF THE DATASET

[ ]:
```
df.describe(include='all')
```

[ ]:

	index	National Provider Identifier	\
count	1.000000e+05	1.000000e+05	
unique	NaN	NaN	
top	NaN	NaN	
freq	NaN	NaN	
mean	4.907646e+06	1.498227e+09	
std	2.839633e+06	2.874125e+08	
min	2.090000e+02	1.003001e+09	
25%	2.458791e+06	1.245669e+09	
50%	4.901266e+06	1.497847e+09	
75%	7.349450e+06	1.740374e+09	
max	9.847440e+06	1.993000e+09	

	Last Name/Organization Name of the Provider	First Name of the Provider	\
count	100000	95745	
unique	42820	13022	
top	PATEL	MICHAEL	
freq	557	2350	
mean	NaN	NaN	
std	NaN	NaN	
min	NaN	NaN	
25%	NaN	NaN	
50%	NaN	NaN	

75%		NaN	NaN
max		NaN	NaN

	Middle Initial of the Provider	Credentials of the Provider	\
count	70669	92791	
unique	29	1854	
top	A	MD	
freq	8152	32874	
mean	NaN	NaN	
std	NaN	NaN	
min	NaN	NaN	
25%	NaN	NaN	
50%	NaN	NaN	
75%	NaN	NaN	
max	NaN	NaN	

	Gender of the Provider	Entity Type of the Provider	\
count	95746	100000	
unique	2	2	
top	M	I	
freq	66641	95746	
mean	NaN	NaN	
std	NaN	NaN	
min	NaN	NaN	
25%	NaN	NaN	
50%	NaN	NaN	
75%	NaN	NaN	
max	NaN	NaN	

	Street Address 1 of the Provider	Street Address 2 of the Provider	…	\
count	100000	40637	…	
unique	51928	10024	…	
top	200 1ST ST SW	SUITE 200	…	
freq	244	1624	…	
mean	NaN	NaN	…	
std	NaN	NaN	…	
min	NaN	NaN	…	
25%	NaN	NaN	…	
50%	NaN	NaN	…	
75%	NaN	NaN	…	
max	NaN	NaN	…	

	HCPCS Code	HCPCS Description	\
count	100000	100000	
unique	2631	2455	
top	99213	Established patient office or other outpatient…	
freq	4578	4578	

```
mean          NaN                                                    NaN
std           NaN                                                    NaN
min           NaN                                                    NaN
25%           NaN                                                    NaN
50%           NaN                                                    NaN
75%           NaN                                                    NaN
max           NaN                                                    NaN

       HCPCS Drug Indicator Number of Services  \
count                100000             100000
unique                    2               2748
top                       N                 13
freq                  93802               3018
mean                    NaN                NaN
std                     NaN                NaN
min                     NaN                NaN
25%                     NaN                NaN
50%                     NaN                NaN
75%                     NaN                NaN
max                     NaN                NaN

       Number of Medicare Beneficiaries  \
count                            100000
unique                             1274
top                                  11
freq                               4791
mean                                NaN
std                                 NaN
min                                 NaN
25%                                 NaN
50%                                 NaN
75%                                 NaN
max                                 NaN

       Number of Distinct Medicare Beneficiary/Per Day Services  \
count                                                 100000
unique                                                 1979
top                                                      12
freq                                                   3210
mean                                                    NaN
std                                                     NaN
min                                                     NaN
25%                                                     NaN
50%                                                     NaN
75%                                                     NaN
max                                                     NaN
```

```
       Average Medicare Allowed Amount Average Submitted Charge Amount  \
count                           100000                           100000
unique                           49629                            38088
top                                  3                              150
freq                              1017                              970
mean                               NaN                              NaN
std                                NaN                              NaN
min                                NaN                              NaN
25%                                NaN                              NaN
50%                                NaN                              NaN
75%                                NaN                              NaN
max                                NaN                              NaN

       Average Medicare Payment Amount Average Medicare Standardized Amount
count                           100000                              100000
unique                           83367                               76237
top                               2.94                               25.32
freq                               623                                1630
mean                               NaN                                 NaN
std                                NaN                                 NaN
min                                NaN                                 NaN
25%                                NaN                                 NaN
50%                                NaN                                 NaN
75%                                NaN                                 NaN
max                                NaN                                 NaN

[11 rows x 27 columns]
```

** ADDING A NEW COLUMN "MONEY DIFFERENCE" IN THE DATASET WHICH CALCULATES THE DIFFERENCE BETWEEN "AVERAGE SUBMITTED CHARGE AMOUNT" COLUMN AND THE "AVERAGE MEDICARE PAYMENT AMOUNT" COLUMN**

[ ]:
```
df["Average Submitted Charge Amount"] = df["Average Submitted Charge Amount"].
↪replace(',', '', regex=True)
```

[ ]:
```
df["Average Medicare Payment Amount"] = df["Average Medicare Payment Amount"].
↪replace(',', '', regex=True)
```

[ ]:
```
df["Money difference"] = df["Average Submitted Charge Amount"].astype(float) -␣
↪df["Average Medicare Payment Amount"].astype(float)
```

3 Basic exploration of the dataset with the new column

[ ]:
```
df.nunique()
```

[ ]:
```
index                                                    100000
National Provider Identifier                              89508
Last Name/Organization Name of the Provider               42820
First Name of the Provider                                13022
Middle Initial of the Provider                               29
Credentials of the Provider                                1854
Gender of the Provider                                        2
Entity Type of the Provider                                   2
Street Address 1 of the Provider                          51928
Street Address 2 of the Provider                          10024
City of the Provider                                       5846
Zip Code of the Provider                                  47827
State Code of the Provider                                   58
Country Code of the Provider                                  4
Provider Type                                                90
Medicare Participation Indicator                              2
Place of Service                                              2
HCPCS Code                                                 2631
HCPCS Description                                          2455
HCPCS Drug Indicator                                          2
Number of Services                                         2748
Number of Medicare Beneficiaries                           1274
Number of Distinct Medicare Beneficiary/Per Day Services   1979
Average Medicare Allowed Amount                           49629
Average Submitted Charge Amount                           38088
Average Medicare Payment Amount                           83367
Average Medicare Standardized Amount                      76237
Money difference                                          92772
dtype: int64
```

[ ]:
```
(df.isnull().sum()/(len(df)))*100
```

[ ]:
```
index                                                     0.000
National Provider Identifier                              0.000
Last Name/Organization Name of the Provider               0.000
First Name of the Provider                                4.255
Middle Initial of the Provider                           29.331
Credentials of the Provider                               7.209
Gender of the Provider                                    4.254
Entity Type of the Provider                               0.000
Street Address 1 of the Provider                          0.000
Street Address 2 of the Provider                         59.363
City of the Provider                                      0.000
Zip Code of the Provider                                  0.000
State Code of the Provider                                0.000
Country Code of the Provider                              0.000
Provider Type                                             0.000
```

```
Medicare Participation Indicator                                0.000
Place of Service                                                0.000
HCPCS Code                                                      0.000
HCPCS Description                                               0.000
HCPCS Drug Indicator                                            0.000
Number of Services                                              0.000
Number of Medicare Beneficiaries                                0.000
Number of Distinct Medicare Beneficiary/Per Day Services        0.000
Average Medicare Allowed Amount                                 0.000
Average Submitted Charge Amount                                 0.000
Average Medicare Payment Amount                                 0.000
Average Medicare Standardized Amount                            0.000
Money difference                                                0.000
dtype: float64
```

[ ]:
```
df.describe(include='all').T
```

[ ]:
```
                                                   count unique  \
index                                           100000.0    NaN
National Provider Identifier                    100000.0    NaN
Last Name/Organization Name of the Provider       100000  42820
First Name of the Provider                         95745  13022
Middle Initial of the Provider                     70669     29
Credentials of the Provider                        92791   1854
Gender of the Provider                             95746      2
Entity Type of the Provider                       100000      2
Street Address 1 of the Provider                  100000  51928
Street Address 2 of the Provider                   40637  10024
City of the Provider                              100000   5846
Zip Code of the Provider                        100000.0    NaN
State Code of the Provider                        100000     58
Country Code of the Provider                      100000      4
Provider Type                                     100000     90
Medicare Participation Indicator                  100000      2
Place of Service                                  100000      2
HCPCS Code                                        100000   2631
HCPCS Description                                 100000   2455
HCPCS Drug Indicator                              100000      2
Number of Services                                100000   2748
Number of Medicare Beneficiaries                  100000   1274
Number of Distinct Medicare Beneficiary/Per Day…  100000   1979
Average Medicare Allowed Amount                   100000  49629
Average Submitted Charge Amount                   100000  38088
Average Medicare Payment Amount                   100000  83367
Average Medicare Standardized Amount              100000  76237
Money difference                                100000.0    NaN
```

```
                  top  \
index
NaN
National Provider Identifier
NaN
Last Name/Organization Name of the Provider
PATEL
First Name of the Provider
MICHAEL
Middle Initial of the Provider
A
Credentials of the Provider
MD
Gender of the Provider
M
Entity Type of the Provider
I
Street Address 1 of the Provider
200 1ST ST SW
Street Address 2 of the Provider
SUITE 200
City of the Provider
NEW YORK
Zip Code of the Provider
NaN
State Code of the Provider
CA
Country Code of the Provider
US
Provider Type
Diagnostic Radiology
Medicare Participation Indicator
Y
Place of Service
O
HCPCS Code
99213
HCPCS Description                                    Established patient office
or other outpatient…
HCPCS Drug Indicator
N
Number of Services
13
Number of Medicare Beneficiaries
11
Number of Distinct Medicare Beneficiary/Per Day…
12
```

```
Average Medicare Allowed Amount
3
Average Submitted Charge Amount
150
Average Medicare Payment Amount
2.94
Average Medicare Standardized Amount
25.32
Money difference
NaN
```

	freq	mean	\
index	NaN	4907645.74603	
National Provider Identifier	NaN	1498226858.04662	
Last Name/Organization Name of the Provider	557	NaN	
First Name of the Provider	2350	NaN	
Middle Initial of the Provider	8152	NaN	
Credentials of the Provider	32874	NaN	
Gender of the Provider	66641	NaN	
Entity Type of the Provider	95746	NaN	
Street Address 1 of the Provider	244	NaN	
Street Address 2 of the Provider	1624	NaN	
City of the Provider	1061	NaN	
Zip Code of the Provider	NaN	416381950.78367	
State Code of the Provider	7775	NaN	
Country Code of the Provider	99994	NaN	
Provider Type	12537	NaN	
Medicare Participation Indicator	99969	NaN	
Place of Service	61616	NaN	
HCPCS Code	4578	NaN	
HCPCS Description	4578	NaN	
HCPCS Drug Indicator	93802	NaN	
Number of Services	3018	NaN	
Number of Medicare Beneficiaries	4791	NaN	
Number of Distinct Medicare Beneficiary/Per Day…	3210	NaN	
Average Medicare Allowed Amount	1017	NaN	
Average Submitted Charge Amount	970	NaN	
Average Medicare Payment Amount	623	NaN	
Average Medicare Standardized Amount	1630	NaN	
Money difference	NaN	277.191655	

	std	\
index	2839632.695465	
National Provider Identifier	287412506.095332	
Last Name/Organization Name of the Provider	NaN	
First Name of the Provider	NaN	
Middle Initial of the Provider	NaN	

```
Credentials of the Provider                                NaN
Gender of the Provider                                     NaN
Entity Type of the Provider                                NaN
Street Address 1 of the Provider                           NaN
Street Address 2 of the Provider                           NaN
City of the Provider                                       NaN
Zip Code of the Provider                   308256603.986241
State Code of the Provider                                 NaN
Country Code of the Provider                               NaN
Provider Type                                              NaN
Medicare Participation Indicator                           NaN
Place of Service                                           NaN
HCPCS Code                                                 NaN
HCPCS Description                                          NaN
HCPCS Drug Indicator                                       NaN
Number of Services                                         NaN
Number of Medicare Beneficiaries                           NaN
Number of Distinct Medicare Beneficiary/Per Day…           NaN
Average Medicare Allowed Amount                            NaN
Average Submitted Charge Amount                            NaN
Average Medicare Payment Amount                            NaN
Average Medicare Standardized Amount                       NaN
Money difference                                   924.898491

                                                         min  \
index                                                  209.0
National Provider Identifier                     1003001298.0
Last Name/Organization Name of the Provider               NaN
First Name of the Provider                                NaN
Middle Initial of the Provider                            NaN
Credentials of the Provider                               NaN
Gender of the Provider                                    NaN
Entity Type of the Provider                               NaN
Street Address 1 of the Provider                          NaN
Street Address 2 of the Provider                          NaN
City of the Provider                                      NaN
Zip Code of the Provider                                601.0
State Code of the Provider                                NaN
Country Code of the Provider                              NaN
Provider Type                                             NaN
Medicare Participation Indicator                          NaN
Place of Service                                          NaN
HCPCS Code                                                NaN
HCPCS Description                                         NaN
HCPCS Drug Indicator                                      NaN
Number of Services                                        NaN
Number of Medicare Beneficiaries                          NaN
```

```
Number of Distinct Medicare Beneficiary/Per Day…           NaN
Average Medicare Allowed Amount                            NaN
Average Submitted Charge Amount                            NaN
Average Medicare Payment Amount                            NaN
Average Medicare Standardized Amount                       NaN
Money difference                                    -16.984065

                                                           25%  \
index                                                2458790.75
National Provider Identifier                       1245669407.25
Last Name/Organization Name of the Provider                 NaN
First Name of the Provider                                  NaN
Middle Initial of the Provider                              NaN
Credentials of the Provider                                 NaN
Gender of the Provider                                      NaN
Entity Type of the Provider                                 NaN
Street Address 1 of the Provider                            NaN
Street Address 2 of the Provider                            NaN
City of the Provider                                        NaN
Zip Code of the Provider                           142630001.0
State Code of the Provider                                  NaN
Country Code of the Provider                                NaN
Provider Type                                               NaN
Medicare Participation Indicator                            NaN
Place of Service                                            NaN
HCPCS Code                                                  NaN
HCPCS Description                                           NaN
HCPCS Drug Indicator                                        NaN
Number of Services                                          NaN
Number of Medicare Beneficiaries                            NaN
Number of Distinct Medicare Beneficiary/Per Day…           NaN
Average Medicare Allowed Amount                            NaN
Average Submitted Charge Amount                            NaN
Average Medicare Payment Amount                            NaN
Average Medicare Standardized Amount                       NaN
Money difference                                     33.911165

                                                           50%  \
index                                                4901266.0
National Provider Identifier                       1497846612.0
Last Name/Organization Name of the Provider                 NaN
First Name of the Provider                                  NaN
Middle Initial of the Provider                              NaN
Credentials of the Provider                                 NaN
Gender of the Provider                                      NaN
Entity Type of the Provider                                 NaN
Street Address 1 of the Provider                            NaN
```

```
Street Address 2 of the Provider                          NaN
City of the Provider                                      NaN
Zip Code of the Provider                            363302500.0
State Code of the Provider                                NaN
Country Code of the Provider                              NaN
Provider Type                                             NaN
Medicare Participation Indicator                          NaN
Place of Service                                          NaN
HCPCS Code                                                NaN
HCPCS Description                                         NaN
HCPCS Drug Indicator                                      NaN
Number of Services                                        NaN
Number of Medicare Beneficiaries                          NaN
Number of Distinct Medicare Beneficiary/Per Day…          NaN
Average Medicare Allowed Amount                           NaN
Average Submitted Charge Amount                           NaN
Average Medicare Payment Amount                           NaN
Average Medicare Standardized Amount                      NaN
Money difference                                    89.912401

                                                          75%  \
index                                               7349450.5
National Provider Identifier                     1740373949.25
Last Name/Organization Name of the Provider               NaN
First Name of the Provider                                NaN
Middle Initial of the Provider                            NaN
Credentials of the Provider                               NaN
Gender of the Provider                                    NaN
Entity Type of the Provider                               NaN
Street Address 1 of the Provider                          NaN
Street Address 2 of the Provider                          NaN
City of the Provider                                      NaN
Zip Code of the Provider                          681988102.0
State Code of the Provider                                NaN
Country Code of the Provider                              NaN
Provider Type                                             NaN
Medicare Participation Indicator                          NaN
Place of Service                                          NaN
HCPCS Code                                                NaN
HCPCS Description                                         NaN
HCPCS Drug Indicator                                      NaN
Number of Services                                        NaN
Number of Medicare Beneficiaries                          NaN
Number of Distinct Medicare Beneficiary/Per Day…          NaN
Average Medicare Allowed Amount                           NaN
Average Submitted Charge Amount                           NaN
Average Medicare Payment Amount                           NaN
```

```
Average Medicare Standardized Amount                          NaN
Money difference                                           211.301618

                                                              max
index                                                    9847440.0
National Provider Identifier                          1992999874.0
Last Name/Organization Name of the Provider                    NaN
First Name of the Provider                                     NaN
Middle Initial of the Provider                                 NaN
Credentials of the Provider                                    NaN
Gender of the Provider                                         NaN
Entity Type of the Provider                                    NaN
Street Address 1 of the Provider                               NaN
Street Address 2 of the Provider                               NaN
City of the Provider                                           NaN
Zip Code of the Provider                               999016573.0
State Code of the Provider                                     NaN
Country Code of the Provider                                   NaN
Provider Type                                                  NaN
Medicare Participation Indicator                               NaN
Place of Service                                               NaN
HCPCS Code                                                     NaN
HCPCS Description                                              NaN
HCPCS Drug Indicator                                           NaN
Number of Services                                             NaN
Number of Medicare Beneficiaries                               NaN
Number of Distinct Medicare Beneficiary/Per Day…               NaN
Average Medicare Allowed Amount                                NaN
Average Submitted Charge Amount                                NaN
Average Medicare Payment Amount                                NaN
Average Medicare Standardized Amount                           NaN
Money difference                                         57038.775556
```

[ ]:
```
cat_cols=df.select_dtypes(include=['object']).columns
num_cols = df.select_dtypes(include=np.number).columns.tolist()
print("Categorical Variables:")
print(cat_cols)
print("Numerical Variables:")
print(num_cols)
```

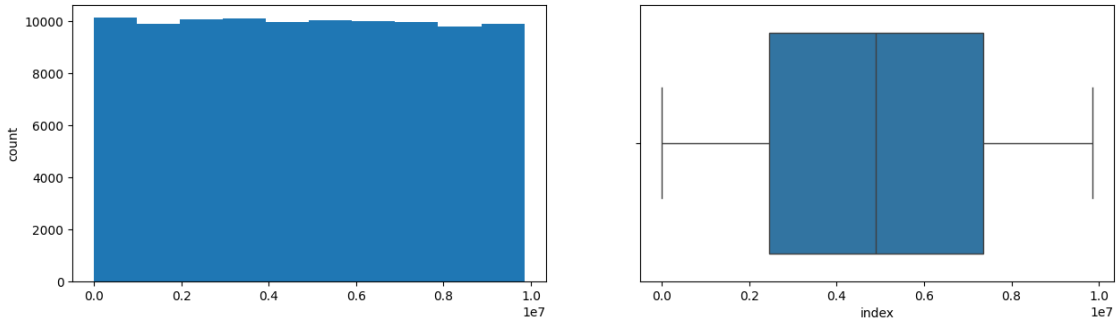
```
Categorical Variables:
Index(['Last Name/Organization Name of the Provider',
       'First Name of the Provider', 'Middle Initial of the Provider',
       'Credentials of the Provider', 'Gender of the Provider',
       'Entity Type of the Provider', 'Street Address 1 of the Provider',
       'Street Address 2 of the Provider', 'City of the Provider',
       'State Code of the Provider', 'Country Code of the Provider',
```

```
       'Provider Type', 'Medicare Participation Indicator', 'Place of Service',
       'HCPCS Code', 'HCPCS Description', 'HCPCS Drug Indicator',
       'Number of Services', 'Number of Medicare Beneficiaries',
       'Number of Distinct Medicare Beneficiary/Per Day Services',
       'Average Medicare Allowed Amount', 'Average Submitted Charge Amount',
       'Average Medicare Payment Amount',
       'Average Medicare Standardized Amount'],
      dtype='object')
Numerical Variables:
['index', 'National Provider Identifier', 'Zip Code of the Provider', 'Money
difference']
```

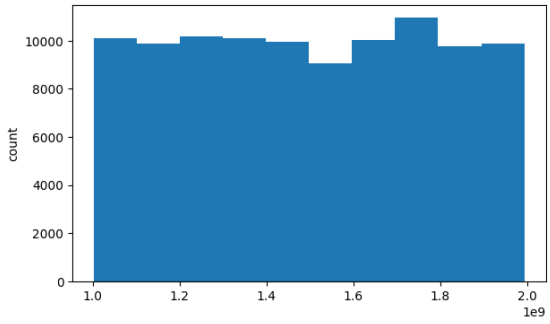
4 GRAPHS TO SEE THE DISTRIBUTION OF THE DATA IN THE DIFFERENT NUMERICAL FIELDS

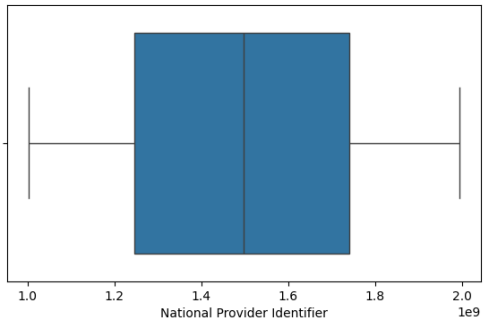
[ ]:
```
for col in num_cols:
    print(col)
    print('Skew :', round(df[col].skew(), 2))
    plt.figure(figsize = (15, 4))
    plt.subplot(1, 2, 1)
    df[col].hist(grid=False)
    plt.ylabel('count')
    plt.subplot(1, 2, 2)
    sns.boxplot(x=df[col])
    plt.show()
```

```
index
Skew : 0.01
```

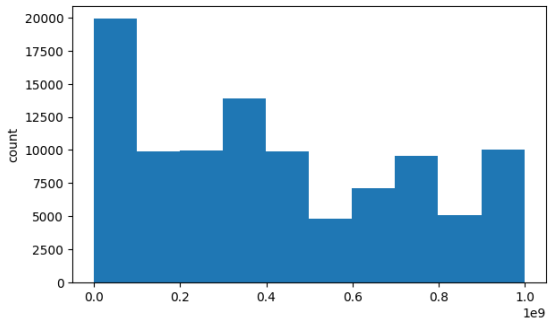


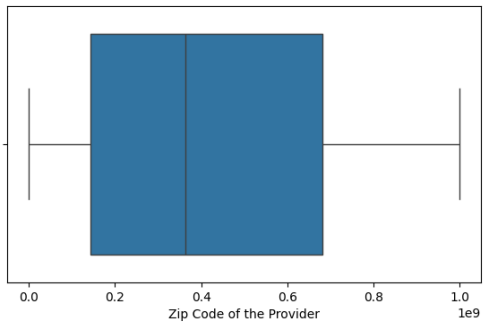

```
National Provider Identifier
Skew : -0.01
```



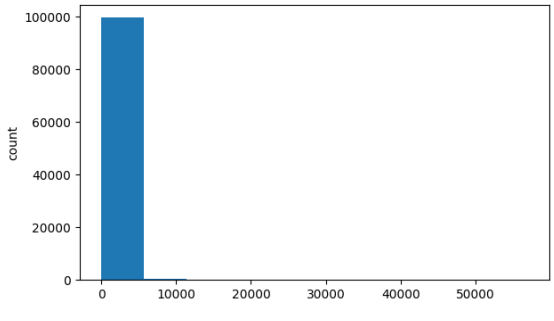


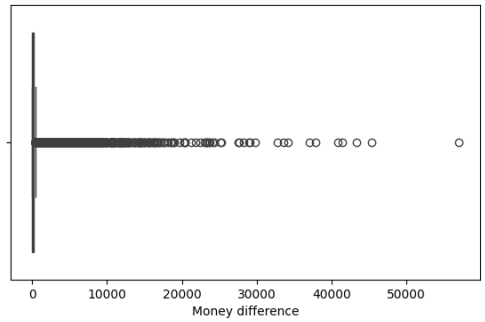
Zip Code of the Provider
Skew : 0.28





Money difference
Skew : 18.47





5 USING THE SWEETVIZ LIBRARY TO GENERATE A EDA REPORT ON THE WHOLE DATASET WITH SPECIAL FOCUS ON THE MONEY DIFFERENCE COLUMN AND HOW IT IS CORRELATED TO THE OTHER COLUMNS IN THE DATASET

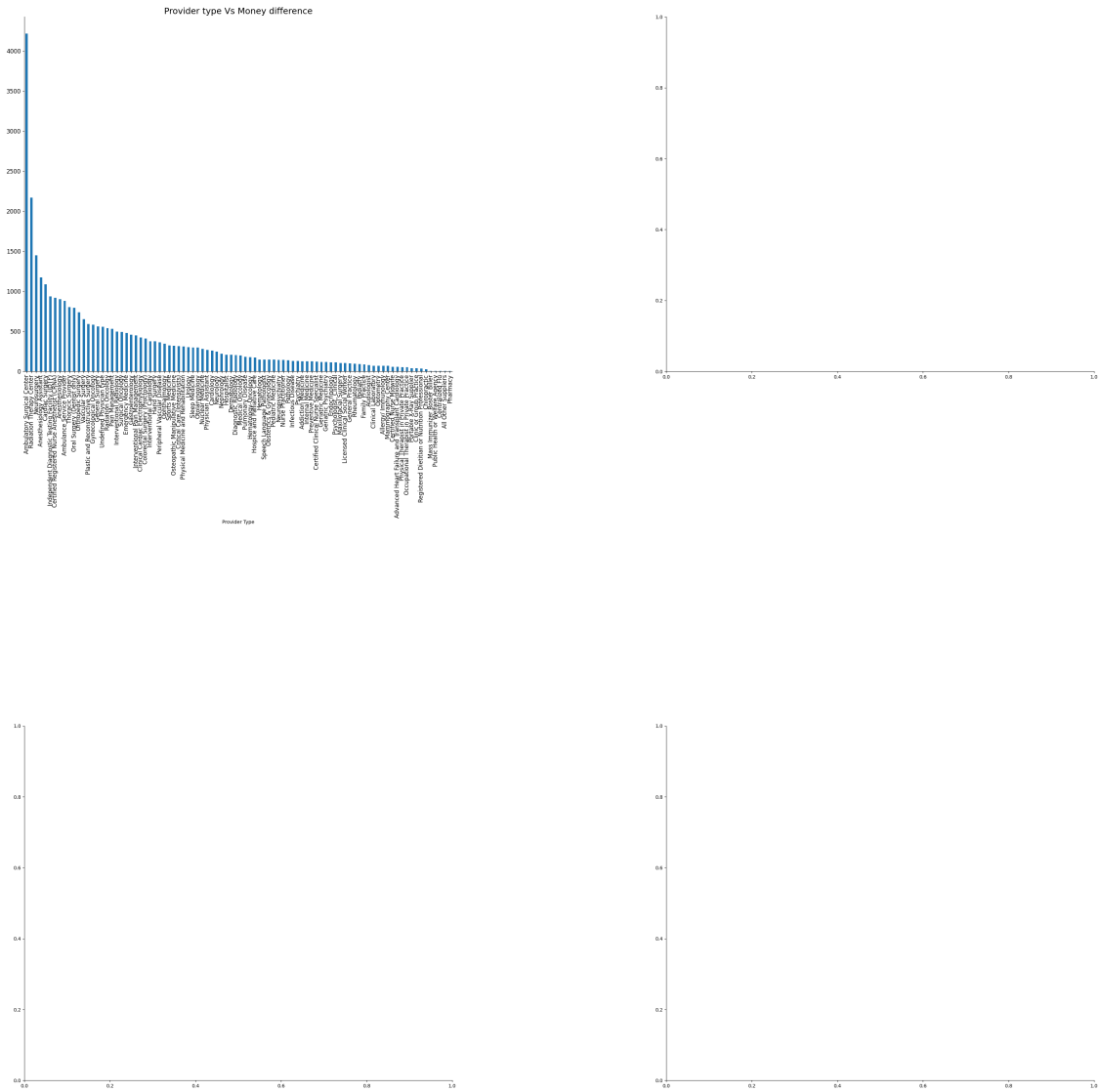
[ ]:
```
pip install sweetviz
import sweetviz as sv
report = sv.analyze(df, target_feat='Money difference')
report.show_html('Report.html')
```

The above line of code will generate a html file called "Report.html". The sweetviz library generates a eda report of the dataset provided. In the report we find out that the money difference column has the most correlation with the provider type column. This means the some specific providers are charging higher amounts of money than is actually necessary for the procedures.

We also target the sepcific Money difference feature using the sweetviz to find out what is its distribution in relation to the other features.

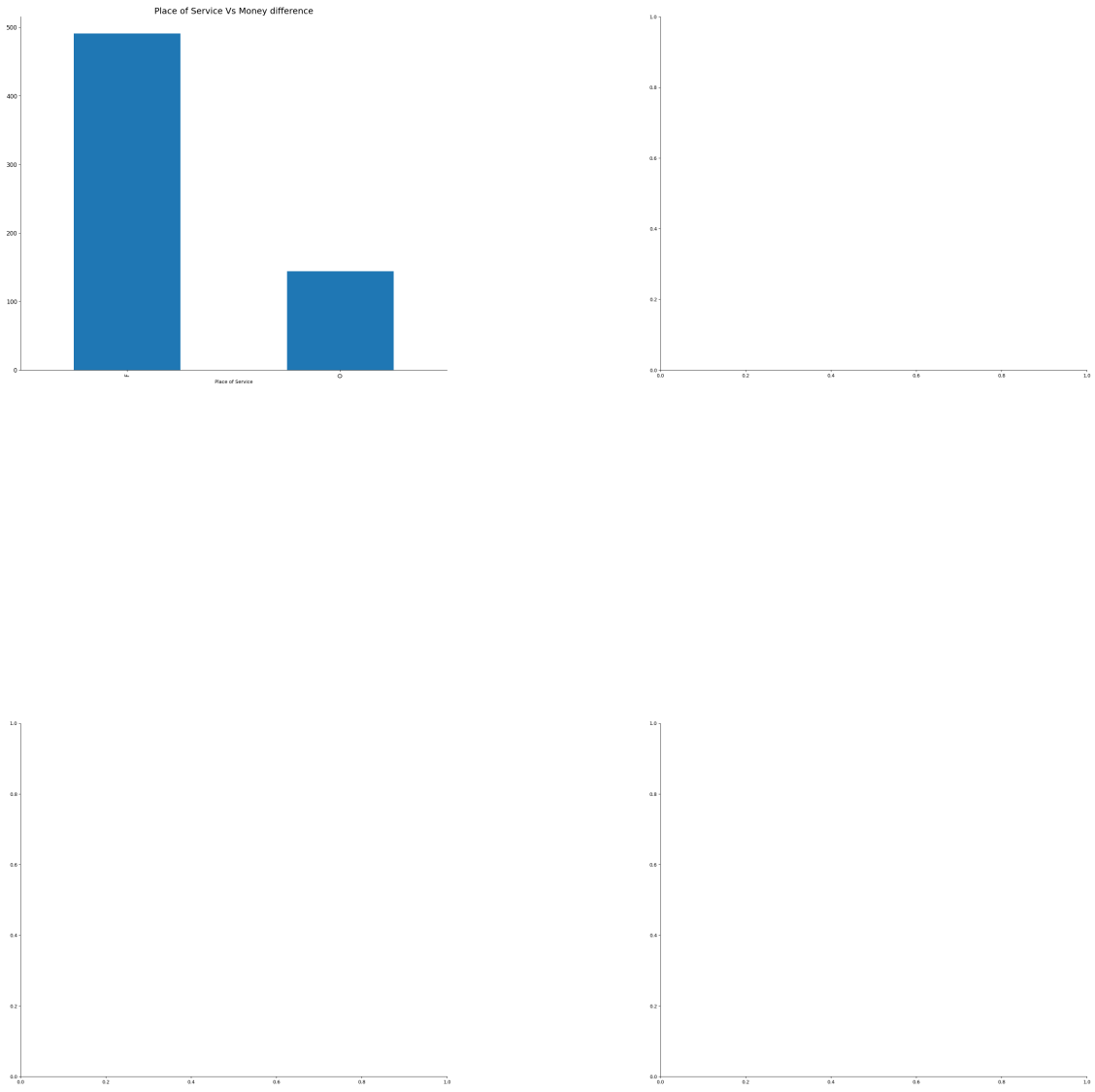
6 GRAPHS TO SEE THE CORRELATION BETWEEN THE MONEY DIFFERENCE COLUMN AND THE OTHER COLUMNS IN THE DATA SET THAT SHOW CORRELATION WITH IT

[ ]:
```
fig, axarr = plt.subplots(2, 2, figsize=(40, 40))
df.groupby('Provider Type')['Money difference'].mean().
  ↪sort_values(ascending=False).plot.bar(ax=axarr[0][0], fontsize=12)
axarr[0][0].set_title("Provider type Vs Money difference", fontsize=18)
plt.subplots_adjust(hspace=1.0)
plt.subplots_adjust(wspace=.5)
sns.despine()
```



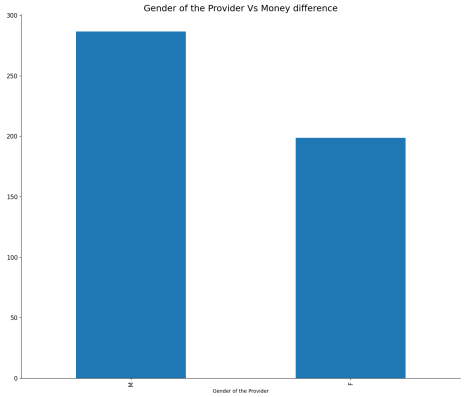
In the above graph we see that Ambulatory Surgical Center providers has the highest difference in the money difference column. This indicates that the anomaly lies here and we should be able to identify fraudulant behaviour if we study the Ambulatory Surgical Center rows.

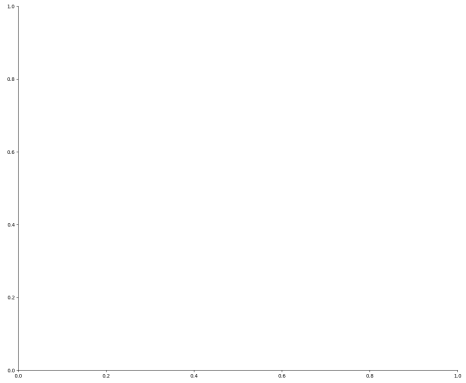
[ ]:
```
fig, axarr = plt.subplots(2, 2, figsize=(40, 40))
df.groupby('Place of Service')['Money difference'].mean().
  ↪sort_values(ascending=False).plot.bar(ax=axarr[0][0], fontsize=12)
axarr[0][0].set_title("Place of Service Vs Money difference", fontsize=18)
plt.subplots_adjust(hspace=1.0)
plt.subplots_adjust(wspace=.5)
sns.despine()
```

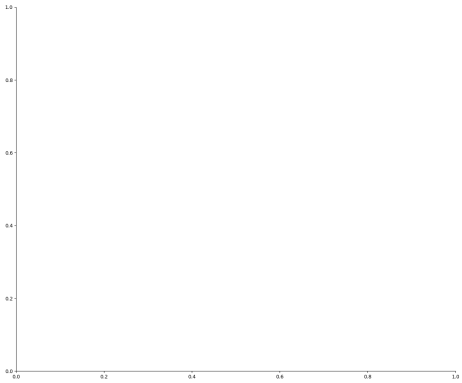


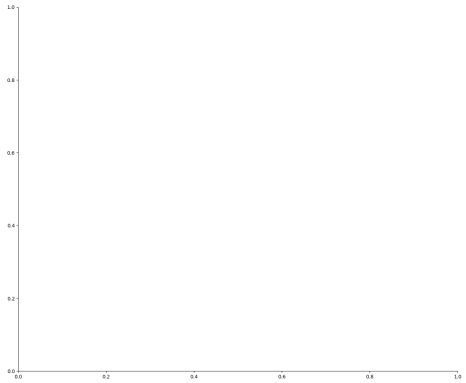
The above graph shows that procedures conducted in facilities have a higher money difference as compared to procedures conducted in offices.

[ ]:
```
fig, axarr = plt.subplots(2, 2, figsize=(40, 40))
df.groupby('Gender of the Provider')['Money difference'].mean().
  ↪sort_values(ascending=False).plot.bar(ax=axarr[0][0], fontsize=12)
axarr[0][0].set_title("Gender of the Provider Vs Money difference", fontsize=18)
plt.subplots_adjust(hspace=1.0)
plt.subplots_adjust(wspace=.5)
sns.despine()
```



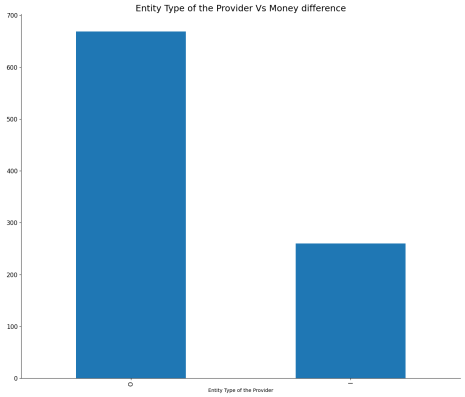


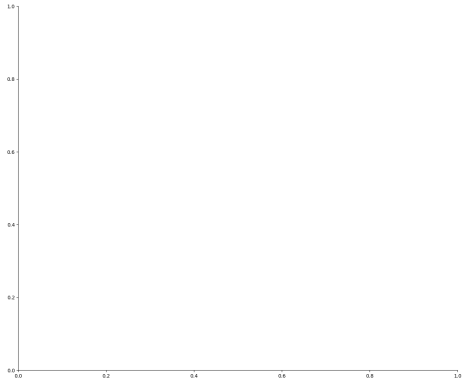


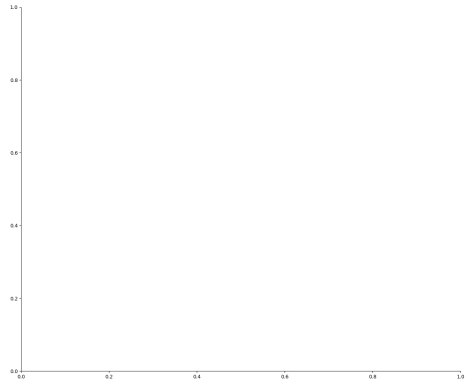


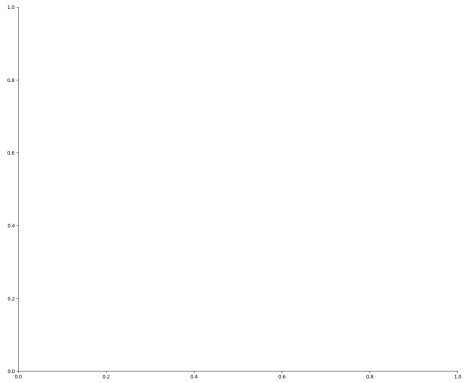
The above graph shows that procedures conducted by males have a higher money difference as compared to procedures conducted by females.

[ ]:
```
fig, axarr = plt.subplots(2, 2, figsize=(40, 40))
df.groupby('Entity Type of the Provider')['Money difference'].mean().
  ↪sort_values(ascending=False).plot.bar(ax=axarr[0][0], fontsize=12)
axarr[0][0].set_title("Entity Type of the Provider Vs Money difference",␣
  ↪fontsize=18)
plt.subplots_adjust(hspace=1.0)
plt.subplots_adjust(wspace=.5)
sns.despine()
```









The above graph shows that procedures conducted by organizations have a higher money difference as compared to procedures conducted by individuals.