

preprocess

June 20, 2024

1 IMPORTING DEPENDENCIES

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

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

2 BASIC EXPLORATION OF THE DATASET

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

```
[65]:
```

	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****

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

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

```
[68]: 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

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

```
[69]: 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
```

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

```
[70]: 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

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

```
[71]:
```

	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

0

HCPCS Code

99213

HCPCS Description

or other outpatient...

Established patient office

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

Dropping different irrelevant fields throughout the process

```
[72]: df = df.drop(columns=['index', 'National Provider Identifier', 'Street Address 1_
↳ of the Provider', 'Street Address 2 of the Provider', 'Country Code of the_
↳ Provider'])
```

```
[73]: df['Last Name/Organization Name of the Provider'] = df['First Name of the_
↳ Provider'].astype(str) + " " + df['Last Name/Organization Name of the_
↳ Provider']
```

```
[74]: df.head()
```

```
[74]: Last Name/Organization Name of the Provider First Name of the Provider \
0 SATYASREE UPADHYAYULA SATYASREE
```

1	WENDY JONES	WENDY
2	RICHARD DUROCHER	RICHARD
3	JASPER FULLARD	JASPER
4	ANTHONY PERROTTI	ANTHONY

	Middle Initial of the Provider	Credentials of the Provider	\
0	NaN	M.D.	
1	P	M.D.	
2	W	DPM	
3	NaN	MD	
4	E	DO	

	Gender of the Provider	Entity Type of the Provider	City of the Provider	\
0	F	I	SAINT LOUIS	
1	F	I	FAYETTEVILLE	
2	M	I	NORTH HAVEN	
3	M	I	KANSAS CITY	
4	M	I	JUPITER	

	Zip Code of the Provider	State Code of the Provider	\
0	631041004.0	MO	
1	283043815.0	NC	
2	64732343.0	CT	
3	641183998.0	MO	
4	334585700.0	FL	

	Provider Type	...	\
0	Internal Medicine	...	
1	Obstetrics & Gynecology	...	
2	Podiatry	...	
3	Internal Medicine	...	
4	Internal Medicine	...	

	HCPDS Description	HCPDS Drug Indicator	\
0	Initial hospital inpatient care, typically 70 ...	N	
1	Screening mammography, bilateral (2-view study...	N	
2	Established patient home visit, typically 25 m...	N	
3	Urinalysis, manual test	N	
4	Injection beneath the skin or into muscle for ...	N	

	Number of Services	Number of Medicare Beneficiaries	\
0	27	24	
1	175	175	
2	32	13	
3	20	18	
4	33	24	

	Number of Distinct Medicare Beneficiary/Per Day Services \
0	27
1	175
2	32
3	20
4	31

	Average Medicare Allowed Amount	Average Submitted Charge Amount \
0	200.58777778	305.21111111
1	123.73	548.8
2	90.65	155
3	3.5	5
4	26.52	40

	Average Medicare Payment Amount	Average Medicare Standardized Amount \
0	157.26222222	160.90888889
1	118.83	135.31525714
2	64.4396875	60.5959375
3	3.43	3.43
4	19.539393939	19.057575758

	Money difference
0	147.948889
1	429.970000
2	90.560312
3	1.570000
4	20.460606

[5 rows x 23 columns]

```
[75]: df=df.drop(columns=['First Name of the Provider'])
```

```
[76]: df.head()
```

```
[76]: Last Name/Organization Name of the Provider Middle Initial of the Provider \
0 SATYASREE UPADHYAYULA NaN
1 WENDY JONES P
2 RICHARD DUROCHER W
3 JASPER FULLARD NaN
4 ANTHONY PERROTTI E
```

	Credentials of the Provider	Gender of the Provider \
0	M.D.	F
1	M.D.	F
2	DPM	M
3	MD	M
4	DO	M

	Entity Type of the Provider	City of the Provider	Zip Code of the Provider	\
0	I	SAINT LOUIS	631041004.0	
1	I	FAYETTEVILLE	283043815.0	
2	I	NORTH HAVEN	64732343.0	
3	I	KANSAS CITY	641183998.0	
4	I	JUPITER	334585700.0	

	State Code of the Provider	Provider Type	\
0	MO	Internal Medicine	
1	NC	Obstetrics & Gynecology	
2	CT	Podiatry	
3	MO	Internal Medicine	
4	FL	Internal Medicine	

	Medicare Participation Indicator	...	\
0	Y	...	
1	Y	...	
2	Y	...	
3	Y	...	
4	Y	...	

	HCPDS Description	HCPDS Drug Indicator	\
0	Initial hospital inpatient care, typically 70 ...	N	
1	Screening mammography, bilateral (2-view study...	N	
2	Established patient home visit, typically 25 m...	N	
3	Urinalysis, manual test	N	
4	Injection beneath the skin or into muscle for ...	N	

	Number of Services	Number of Medicare Beneficiaries	\
0	27	24	
1	175	175	
2	32	13	
3	20	18	
4	33	24	

	Number of Distinct Medicare Beneficiary/Per Day Services	\
0	27	
1	175	
2	32	
3	20	
4	31	

	Average Medicare Allowed Amount	Average Submitted Charge Amount	\
0	200.58777778	305.21111111	
1	123.73	548.8	
2	90.65	155	

3	3.5	5
4	26.52	40

	Average Medicare Payment Amount	Average Medicare Standardized Amount \
0	157.26222222	160.90888889
1	118.83	135.31525714
2	64.4396875	60.5959375
3	3.43	3.43
4	19.539393939	19.057575758

	Money difference
0	147.948889
1	429.970000
2	90.560312
3	1.570000
4	20.460606

[5 rows x 22 columns]

```
[77]: df.rename(columns = {'Last Name/Organization Name of the Provider': 'Full_
↪name'}, inplace = True)
```

```
[78]: df.head()
```

```
[78]:
```

	Full name	Middle Initial of the Provider \
0	SATYASREE UPADHYAYULA	NaN
1	WENDY JONES	P
2	RICHARD DUROCHER	W
3	JASPER FULLARD	NaN
4	ANTHONY PERROTTI	E

	Credentials of the Provider	Gender of the Provider \
0	M.D.	F
1	M.D.	F
2	DPM	M
3	MD	M
4	DO	M

	Entity Type of the Provider	City of the Provider	Zip Code of the Provider \
0	I	SAINT LOUIS	631041004.0
1	I	FAYETTEVILLE	283043815.0
2	I	NORTH HAVEN	64732343.0
3	I	KANSAS CITY	641183998.0
4	I	JUPITER	334585700.0

	State Code of the Provider	Provider Type \
0	MO	Internal Medicine

1	NC	Obstetrics & Gynecology
2	CT	Podiatry
3	MO	Internal Medicine
4	FL	Internal Medicine

	Medicare Participation Indicator	...	\
0	Y	...	
1	Y	...	
2	Y	...	
3	Y	...	
4	Y	...	

	HCPCS Description	HCPCS Drug Indicator	\
0	Initial hospital inpatient care, typically 70 ...	N	
1	Screening mammography, bilateral (2-view study...	N	
2	Established patient home visit, typically 25 m...	N	
3	Urinalysis, manual test	N	
4	Injection beneath the skin or into muscle for ...	N	

	Number of Services	Number of Medicare Beneficiaries	\
0	27	24	
1	175	175	
2	32	13	
3	20	18	
4	33	24	

	Number of Distinct Medicare Beneficiary/Per Day Services	\
0	27	
1	175	
2	32	
3	20	
4	31	

	Average Medicare Allowed Amount	Average Submitted Charge Amount	\
0	200.58777778	305.21111111	
1	123.73	548.8	
2	90.65	155	
3	3.5	5	
4	26.52	40	

	Average Medicare Payment Amount	Average Medicare Standardized Amount	\
0	157.26222222	160.90888889	
1	118.83	135.31525714	
2	64.4396875	60.5959375	
3	3.43	3.43	
4	19.539393939	19.057575758	

	Money difference
0	147.948889
1	429.970000
2	90.560312
3	1.570000
4	20.460606

[5 rows x 22 columns]

```
[79]: df=df.drop(columns=['HCPCS Description'])
```

```
[80]: df.head()
```

```
[80]:
```

	Full name	Middle Initial of the Provider	\
0	SATYASREE UPADHYAYULA		NaN
1	WENDY JONES		P
2	RICHARD DUROCHER		W
3	JASPER FULLARD		NaN
4	ANTHONY PERROTTI		E

	Credentials of the Provider	Gender of the Provider	\
0	M.D.	F	
1	M.D.	F	
2	DPM	M	
3	MD	M	
4	DO	M	

	Entity Type of the Provider	City of the Provider	Zip Code of the Provider	\
0	I	SAINT LOUIS	631041004.0	
1	I	FAYETTEVILLE	283043815.0	
2	I	NORTH HAVEN	64732343.0	
3	I	KANSAS CITY	641183998.0	
4	I	JUPITER	334585700.0	

	State Code of the Provider	Provider Type	\
0	MO	Internal Medicine	
1	NC	Obstetrics & Gynecology	
2	CT	Podiatry	
3	MO	Internal Medicine	
4	FL	Internal Medicine	

	Medicare Participation Indicator	...	HCPCS Code	HCPCS Drug Indicator	\
0	Y	...	99223		N
1	Y	...	G0202		N
2	Y	...	99348		N
3	Y	...	81002		N
4	Y	...	96372		N

	Number of Services	Number of Medicare Beneficiaries	\
0	27	24	
1	175	175	
2	32	13	
3	20	18	
4	33	24	

	Number of Distinct Medicare Beneficiary/Per Day Services	\
0	27	
1	175	
2	32	
3	20	
4	31	

	Average Medicare Allowed Amount	Average Submitted Charge Amount	\
0	200.58777778	305.21111111	
1	123.73	548.8	
2	90.65	155	
3	3.5	5	
4	26.52	40	

	Average Medicare Payment Amount	Average Medicare Standardized Amount	\
0	157.26222222	160.90888889	
1	118.83	135.31525714	
2	64.4396875	60.5959375	
3	3.43	3.43	
4	19.539393939	19.057575758	

	Money difference
0	147.948889
1	429.970000
2	90.560312
3	1.570000
4	20.460606

[5 rows x 21 columns]

3.0.1 Finding the mean and standard deviation of the Money difference column and then finding the Z-score of each row

```
[81]: mean = np.mean(df["Money difference"])
std = np.std(df["Money difference"])
print('mean of the dataset is', mean)
print('std. deviation is', std)
```

mean of the dataset is 277.19165543803854

std. deviation is 924.8938668135768

```
[82]: threshold = 3
outlier = []
for i in df["Money difference"]:
    z = (i-mean)/std
    if z > threshold:
        outlier.append(i)
print('outlier in dataset is', outlier)
```

outlier in dataset is [4903.03071429, 7177.5219355, 5554.8092308, 5338.9915790000005, 5452.4853846, 4071.7436913, 3168.3871429, 7413.323333299999, 7159.16301205, 9933.1763636, 4822.53294118, 13832.885, 3533.14740331, 9377.5338095, 5171.43242153, 11642.71, 3343.0361290299998, 6398.92538462, 4982.326, 4006.542973, 6397.964507080001, 5873.700000000001, 5411.3152174, 4889.573276882, 29021.991875, 3521.7148571400003, 7810.418666699999, 3648.6304395800003, 5163.07307692, 3729.22, 3425.9775308, 7799.201111120001, 5765.6685714, 3412.1461538599997, 6550.96, 4314.96, 6214.62090909, 5258.2467164, 5112.11576475, 9790.5944182, 12284.515, 4128.2553333, 10987.428846499999, 5404.0697143, 3417.38, 6134.6764865000005, 3820.91071429, 5252.4815, 5809.91607143, 5941.11461537, 3081.77277773, 9390.56117647, 4038.14652174, 3716.751, 3122.0025, 3932.707727273, 3263.185, 6468.72307694, 4311.1666667, 4883.2008696, 7110.09470588, 4425.872, 5185.72, 4890.5146667, 4175.00899997, 21803.55333333, 15649.177878303, 4097.9817647, 6233.89352941, 3632.111875, 5095.5553125, 45411.3076, 3270.354, 3141.91153842, 5818.5363636, 4528.36, 3684.38117647, 5962.7474419, 5427.88631579, 57038.7755556, 5666.17, 23206.754078799997, 6549.7, 3316.4599999999999, 4305.948965500001, 5849.124210489999, 9437.6434997, 4106.754, 4212.48766667, 5346.549523799999, 10285.73731707, 5143.1084615400005, 5370.16923076, 3602.52, 6803.81555556, 3293.48206897, 14342.6, 3923.142549, 4786.29681823, 4537.59600003, 3081.4904348, 3520.19837838, 5168.47712121, 6404.6370833, 4299.6214286, 10752.28, 3631.40621723, 5261.916363636, 6458.2608333299995, 3948.1060000330003, 4169.1163158, 4805.52466667, 4618.76384615, 14475.052439, 3273.8118182000003, 3718.70760387, 10579.1322222, 5656.79454549, 4105.268, 5077.80666663, 3339.7624561000002, 4975.16, 8840.11, 4231.93999997, 7675.720192659999, 3214.4368000000004, 3335.66764708, 3279.49, 11696.330365189999, 3891.3809524, 4946.65258065, 5127.42125, 6147.7880000000005, 4422.844375, 4165.86, 3461.4, 3343.3281578700003, 3282.88733333, 18819.040344399997, 4404.279754601, 3629.340625, 4058.00842102, 17107.7616663, 5607.8410714, 6624.7009677, 3868.2699999999995, 5983.32736843, 10884.408, 12031.4385366, 5566.30866667, 9513.5966667, 43299.6, 3417.44380952, 7021.290000000001, 5912.67941175, 3223.4249999699996, 3307.68066667, 3922.6078261000002, 5017.823636409999, 3554.08771429, 6434.58, 4130.3936364, 4028.19, 3919.33612908, 5841.55, 11134.776923000001, 11387.358, 6301.77, 4123.56580645, 3619.22303031, 3623.62666667, 6914.974375, 4252.3916667, 3324.9804, 7101.0188, 3951.21315789, 6911.4817778, 7261.2811111, 7428.2784416, 6436.34333331, 3894.2030768900004, 5208.46714287, 3927.64, 12911.831875, 6771.295, 4063.7914286000005, 3987.51636368, 3254.1600000000003, 4719.46, 5685.42214286, 5428.0358823999995,

5614.24363636, 3465.60776978, 6487.24, 5396.14894735, 8018.346111079999,
5673.79, 3543.1168182, 3742.65878788, 11832.41516161, 9662.290769,
5379.79203297, 8843.9446667, 3489.69, 14357.1680459, 3135.84176471,
3465.31523814, 3132.4, 4088.7381250000003, 4944.542857099999, 22478.915263,
5167.7728, 4763.78, 4457.64, 4966.2229885, 3909.96846667, 6860.99,
8812.23133333, 4019.584, 3232.149375, 3352.1192499999997, 3320.3225714,
12672.41, 3486.74, 3790.0981818399996, 8068.4, 3260.73347826, 7096.22764706,
3233.42, 12570.2067861, 7463.77625, 3497.59416667, 14359.15625, 23728.11386364,
6372.1808088, 3854.4305, 3807.728, 8691.1167742, 3437.53538462, 6844.7454098,
3610.2255, 3497.52181818, 3144.27428567, 23211.352561599997, 3637.63666663,
7803.71, 4283.21217391, 6242.45875, 14975.1957895, 18125.806, 5841.9153571,
4337.2719443999995, 4091.6800000000003, 5143.703787799999, 5563.955333399999,
3067.890625, 3085.9745238200003, 6856.26, 4309.78890414, 3523.5732292000002,
4944.53214286, 8544.87761901, 3972.09826087, 3635.72487805, 3701.1342857099994,
3827.148, 8641.943571299998, 3819.67217391, 3850.94, 3344.30777778, 7756.85,
5043.84, 4754.5006897, 5139.1986667, 16616.23, 5290.423043520001, 8199.3105882,
4086.4145833000002, 8248.01945946, 7707.751875, 6287.88291667, 4629.7925,
3845.55090909, 3483.87416663, 3114.998125, 8184.58, 3553.06, 5687.429375,
3578.47815789, 41422.6775, 4288.1910714000005, 9016.65333333, 4567.946875,
6233.111875, 3108.3290908999998, 6137.172857109999, 6333.091,
3065.0709091100002, 4664.01478261, 3081.67058481, 7893.34135593, 6010.27142857,
4970.7209375, 3310.64727273, 3725.575384615, 4429.4038462, 3784.78684209,
3487.82, 3456.67, 3537.1251429, 3085.60555556, 34193.39000014,
16315.272727300002, 4522.3099999999995, 4270.3665116, 12877.501875,
3595.90416667, 3559.74775281, 5999.65, 4533.9617551, 3424.30615385,
4318.8415789499995, 6072.76571429, 12649.93909136, 12086.0986667, 4750.46,
3971.89454545, 4522.24020408, 7219.706744200001, 5222.4679058, 11303.25391304,
6627.5156522, 4497.7797222, 3773.2342857400004, 6091.09172419, 9494.4511538,
4386.5633333000005, 3119.4670968, 5544.6686538, 6702.86538459, 12476.8194118,
3454.06444442, 3986.91235294, 3800.09000003, 3301.73695657, 3113.1985713900003,
3517.94333337, 6999.38, 5424.38, 3458.95769231, 9801.92466667,
3367.1054544999997, 3088.35058824, 12858.07555556, 4185.78, 5670.025,
4550.52037037, 4216.18181816, 6918.325, 5942.492, 4025.09888892, 9189.92769231,
3084.78125, 3731.2, 4171.13, 5147.5382222, 3371.18241667, 10397.99461538,
3053.0606666699996, 3734.26882353, 3544.06692307, 3278.9905556000003,
3402.42859294, 12954.728627299999, 7410.7353846, 8957.2740625, 5723.5345,
3980.70714286, 3351.53611111, 7829.09875, 4410.6305, 15338.973333300002,
3377.0127273, 9254.72428571, 4141.2091561199995, 6909.271, 3731.3330769199997,
3240.02, 9888.77230775, 5364.7007692, 5036.56294118, 12210.35354832,
3278.84210525, 5189.9572499999995, 4691.86295455, 6548.75526311, 3630.92,
15016.51, 18928.69338, 7388.87630736, 25159.84, 4760.0240909, 4660.93,
19742.2674, 3242.976087, 11493.91689641, 4753.0517242, 7545.8099999999995,
4060.2512820499996, 16172.150385, 5828.5345, 3487.99145837, 5719.51,
8528.5795455, 3655.742, 5976.7898000000005, 5610.5599999999995, 3697.7609375,
7299.64, 3563.2136, 3181.47, 9038.00384615, 24288.9009091, 4085.0937037000003,
3686.42, 7002.89090909, 3294.43, 3329.7647169800002, 3625.8345, 4129.408125,
13633.808, 7668.319375, 3891.34, 5017.33608696, 4202.0025, 14626.24714286,
14416.294, 4644.08473773, 3157.12, 7958.49239072, 7128.610000000001,

3345.9543477999996, 4003.38411765, 12254.3287139, 5503.57, 3351.44,
5740.12458333, 6672.8225, 5613.2082353, 4151.21605261, 3155.86009804,
4574.09037593, 4070.4224242099995, 6866.91, 4511.6, 4670.0195652,
16266.84466667, 3438.2660784, 4651.58461542, 7410.15230769, 3379.49,
3678.76311111, 8613.54583333, 13535.39875, 4990.60157895, 28290.8378216,
4542.1173171, 10009.0575, 4711.8449153, 7621.6230435, 3933.8417283999997,
5426.82833333, 27502.6077275, 3457.84764706, 6846.5543939, 4624.6264706,
3375.11923076, 3390.5756, 8814.925454545, 3937.95941172, 5441.02035714,
7948.089090900001, 5082.171836735, 7676.025, 13295.43636364, 8250.148666699999,
3229.69, 3549.85485294, 11736.0680001, 3510.8615625, 3440.0378987100003,
6131.92, 6503.97, 3315.8086667, 28945.77, 3819.6, 3545.73529412, 6014.45791667,
3695.82128205, 3178.54069444, 4644.7513158, 5641.10999996, 3425.7235714,
10413.18, 33580.72555556, 5926.86, 6715.28, 3989.7508333, 3238.27846154,
9090.7201618, 7949.820930229999, 3778.13071431, 3661.08151515, 37919.7427272,
16869.1985938, 5101.06681817, 6261.75628529, 3614.62642857, 3565.8864286,
4289.38777783, 3911.0361539000005, 3144.6046667, 3113.50322577, 6523.03,
3700.0234, 4803.10732673, 7865.143999989999, 3359.45472222, 4429.828333333,
10449.25909064, 6839.7787234, 4312.625454589999, 3194.36, 7891.16368417,
3079.1673684469997, 3335.7057895000003, 3241.96486489, 15755.3433333,
14124.9212195, 6126.4147826, 4548.7, 6383.5317647, 7569.93, 5887.0994444,
3150.87, 10787.29964286, 3712.16166663, 3874.5203999999994, 4196.91957443,
7624.72958904, 3192.3247826300003, 3093.95, 12173.4830553, 3096.57,
3239.345263158, 20399.2328571, 3580.1415, 3562.0204546000004, 3143.515,
8845.718125, 16599.95652178, 12195.8615385, 3588.3868421, 5122.076875,
4258.88126579, 17616.57769231, 3918.94307692, 3918.21125, 3190.45428571,
7559.22818177, 3188.28405066, 4942.28, 3314.73482234, 3107.69607143, 6144.293,
6835.93, 3222.88, 7142.9933333, 4486.17, 3384.86087718, 10959.504444,
4417.1205555999995, 6530.0685714, 3586.99846154, 3091.0577273, 3177.3652174,
5696.71976195, 3207.89, 4980.3973171, 4361.4874, 5980.04214286, 10939.53928571,
3103.0991667, 5015.37133333, 6133.0356, 4105.44625, 4685.2861111, 23403.2824029,
12878.81030303, 3548.23, 14799.41818136, 4016.1683333, 4749.61114283, 6860.17,
4495.07666663, 8313.534, 7597.83083333, 3118.052, 4186.60761903, 11196.780625,
3569.31642853, 5122.86909091, 11963.8175754, 4599.15277779, 4679.69384618,
3889.6552381, 3467.63857143, 9650.063125, 4372.605, 15555.02799,
3304.5363636399998, 6739.03965514, 3204.7848387, 21206.4105, 3457.7809091400004,
3422.32642856, 4773.58878786, 3193.87115382, 5859.12769231, 4125.04384613,
3661.55823529, 3428.1542105, 3095.33, 5732.11807692, 6888.622608690001,
7107.754375, 6783.4417391, 6919.56192308, 3511.10545458, 3056.2295, 17512.73,
6957.25, 7909.06363635, 4528.36, 3202.3, 3751.3059375000003, 5168.198125,
4466.94969697, 40853.32400027, 4812.03823528, 4594.998, 3248.06, 7784.7273333,
3321.82947367, 4058.41, 4571.7853333699995, 3095.52826087, 8105.89833329,
6069.71970149, 3054.32121212, 9995.739, 4356.3169697, 3244.274,
4240.419999999999, 3714.84384615, 11652.3981818, 3807.4635, 3765.4555172100004,
4895.695, 4084.9199999800003, 3468.6593333, 9979.74, 14259.17227273,
6588.92363636, 3406.14666667, 16081.072885, 3863.19, 10722.89647059,
4523.28236363, 3281.13176471, 5486.875, 6191.4539394, 3218.15, 5096.16583333,
10569.7306452, 3522.1124, 5864.80175, 6488.6284772, 12619.577931299998,
3317.634058, 4048.3344927999997, 4203.35545454, 3122.43743587, 25225.5049254,

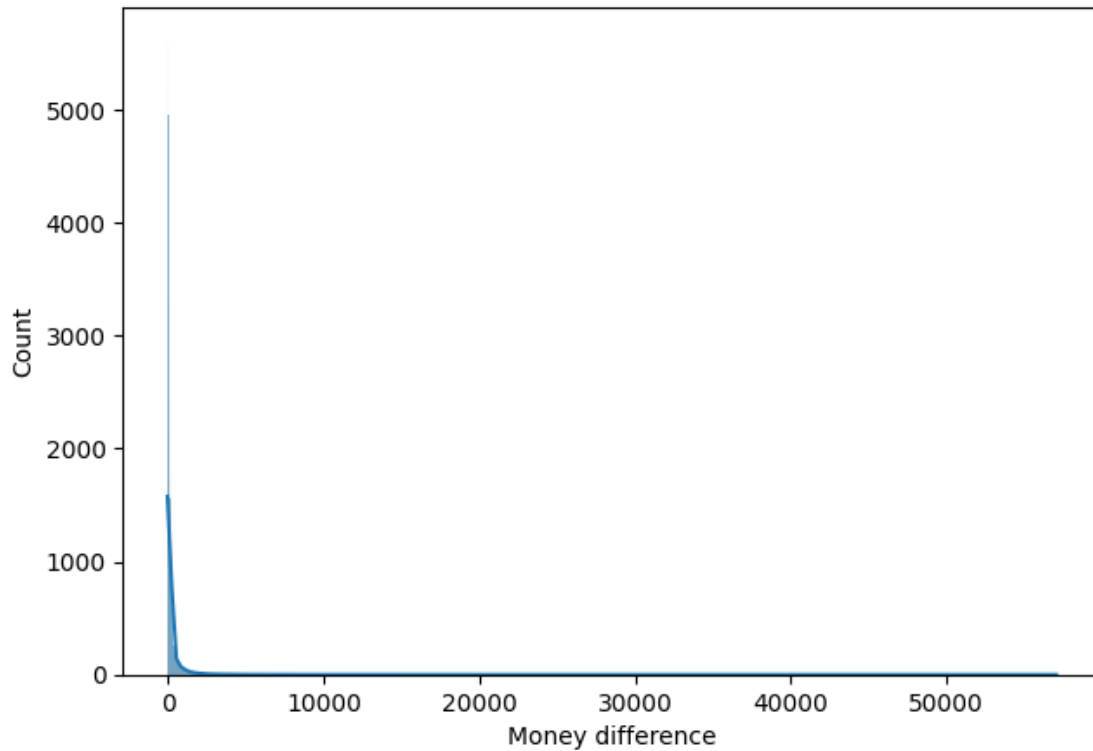
8785.0016129, 14755.619999999999, 5273.3992857, 10737.8584206, 3612.597142862,
3641.2068421, 7195.0705, 4406.61305556, 4641.8434783, 3783.4688000000006,
4975.4878571, 4156.2, 3887.8199999999997, 11621.0076471, 4493.6635713999995,
4315.56529408, 3395.84862069, 8794.14461538, 5490.1425, 3091.533125,
3277.32864865, 5229.0847414, 4507.04066667, 14316.630000000001,
3745.0299999999997, 5310.9342857, 4061.4837209300003, 3760.97565217,
6472.94619046, 3511.93999997, 3418.43238095, 3579.69, 3358.65, 4258.66192308,
6320.58466667, 3310.719048843, 3836.32058824, 4613.00871795, 7303.042381,
3356.66025642, 5572.1504, 8372.3236842, 11589.09260883, 5535.0342857, 9240.2725,
3127.78107143, 6597.81101449, 6258.09142857, 12054.604652330001, 7173.3076923,
37081.263, 4440.91875, 9274.48, 4195.14, 17207.8444444, 9858.56882388,
3501.4423809, 5326.0683333, 13588.6719355, 3627.7806557000004, 3311.70809524,
5073.6483721, 6222.06457627, 3227.1880769000004, 8008.562777800001, 8636.21,
5521.24, 4216.45829265, 9860.09708333, 5136.2744, 3248.0493377000003,
4126.758125, 4970.905625, 5025.3085715, 6159.30294116, 3564.1980000000003,
3378.13757576, 3123.2469230799998, 7030.49, 6222.15875, 13321.2725, 3269.06,
3140.9105814, 4468.61818182, 3584.26641791, 5568.2215789, 5413.18888889,
3230.8745455000003, 4185.345, 5145.691000000001, 3293.0950000000003,
5252.70769231, 4893.12428571, 14135.84947368, 3907.75318182, 7432.3975676,
5072.1593333, 23624.3241176, 3893.7643103, 5550.49, 3580.02375, 15824.6709802,
3645.25125, 4466.16176471, 3116.53230769, 4669.20470589, 4068.42153848,
8190.137, 5047.45785719, 4108.22615381, 17676.82, 5898.77288557, 4311.98939919,
10824.4882353, 6073.1518182, 3355.64076923, 13775.2432727, 5087.8466667,
3082.316363636, 4752.9782608, 3132.7672727000004, 8679.11452381, 7598.52484845,
3102.0437037399997, 3353.66, 3910.991176435, 16401.64538469, 5296.990833333,
3810.778, 8010.69090909, 3677.224375, 6150.97916667, 5515.51285716,
3286.17084507, 3109.29125, 5255.03133337, 4885.384545399999, 6975.64692311,
3104.9476923, 3609.8625, 3107.11375, 7873.7495, 3576.458125, 5092.18200003,
3287.5773333300003, 3604.813333333, 3457.86578947, 3658.61636364, 7284.786875,
7881.92, 4808.3263636, 5132.45272727, 29774.57777778, 5141.5966667,
4080.20771427, 9271.679166670001, 3268.60769231, 3874.4371429, 3133.408,
6789.10500005, 8184.643636400001, 6730.88315789, 4072.8500000300005, 5048.82,
12133.3866509, 3846.57529412, 3310.78615388, 6015.44, 11685.46, 3265.8165556,
3191.8916667000003, 3701.82984375, 6541.5366667, 3905.0003570999997,
3115.43444444, 5457.55, 3247.63448553, 3081.4704762, 7102.08578948,
4121.46818182, 3774.3055320000003, 5728.1805128099995, 15515.38,
3236.8509090899997, 5530.0178571, 7892.52971429, 3196.4771428999998,
4106.4163889, 4014.2411764200006, 4774.339999999999, 4915.966103896,
12039.591304829999, 4515.19636361, 3565.2392308, 6833.62238095, 3446.88076923,
3280.34, 3721.92166667, 6137.6727273, 10640.906363499998, 4087.43155844,
8932.860625, 3257.52727273, 3731.4018182, 4921.153076959999, 4835.1026415100005,
6100.2790909, 16308.786216199998, 5958.26, 6857.57, 4689.40560003,
15690.1702667, 4860.7241176, 16864.6575, 8429.74388889, 5182.35, 4638.67529415,
5101.19486486, 3141.348, 5346.7142105, 3133.43466019, 4542.7669231,
3587.31111111, 3277.94538462, 5998.0175, 4076.35083333, 3430.15246154,
8699.45363636, 3159.98733333, 3237.41342102, 6714.066000000001, 4807.692,
4875.37, 10519.400588530001, 7138.343125, 18666.40538462, 4413.84535717,
3056.16090909, 7666.86071429, 4791.68111111, 3137.496875, 4204.09228571,

4701.07117644, 3577.99837838, 27695.2433333, 4768.24823529, 4024.7336363599998, 3912.62521739, 3651.7473913000003, 3835.3364285299995, 4320.88177574, 4734.7375, 3555.2921429, 3448.61, 3895.01857146, 3896.75241135, 18468.639167, 3519.833, 6994.86588235, 7778.57, 5481.88, 5105.27, 3094.255, 4388.9084615, 4156.2975, 3142.84621622, 3205.4700000000003, 3485.81205128, 3884.3825, 14922.83, 3616.715, 5085.42825, 4103.532499999999, 5974.68499996, 4394.5770588, 5166.135625000001, 11598.5053845, 10690.96428571, 6531.67660377, 3317.95364706, 9388.037647, 4198.70583333, 5397.74, 4150.58, 4222.85230769, 5856.0, 3977.39, 7403.36, 3168.874, 3273.46611111, 5215.30807688, 9772.2701724, 5251.68111111, 20355.65166667, 3532.58357143, 4224.5, 7915.978888900001, 24099.501503500003, 4350.87567569, 3521.169, 4985.42615381, 10836.9896609, 5368.05909088, 6023.68125, 11054.722632, 5162.7986667000005, 4448.17400003, 3063.37384618, 5524.8496154, 3983.7000000000003, 4515.367894737, 5316.09153846, 4938.83, 3400.541875, 4737.91227275, 6964.17130433, 4262.47333333, 5465.03958333, 6272.43947368, 4971.53083337, 14378.89153846, 6698.13, 3586.87357146, 6723.93916663, 6452.41090905, 3979.5038461100003, 10793.82173913, 4848.98428571, 5159.85375, 3799.9185185300003, 3176.6475248, 20291.09, 6966.91466667, 4574.73057325, 5088.7390909000005, 3469.05839286, 3437.58868421, 3101.21, 6633.7307691999995, 3449.58571429, 15338.713809859999, 3247.4931221700003, 3718.00214286, 5841.54522293, 8373.274999969999, 4959.1385714, 6421.1383333, 4754.66076923, 22989.181558999997, 3258.2628125, 3521.31083337, 3328.85, 10191.58, 6182.95363635, 9044.23, 3350.1442857, 4610.639386, 20260.14692262, 7480.0221819, 3056.936686047, 4643.6348387, 3486.0194118, 3832.55, 3227.671875, 5554.95, 5401.09, 3809.7153333, 4409.2094594, 3395.94, 3934.93185185, 5673.23, 4061.5390909000002, 32799.61]

3.1 Plotting a distribution graph of the Money Difference column

```
[83]: plt.figure(figsize=(16,5))
plt.subplot(1,2,1)
sns.histplot(df['Money difference'], kde=True)

plt.show()
```

3.1.1 Scaling the numerical fields using the formula of min max scaler

```
[84]: column = 'Money difference'
df[column] = (df[column] - df[column].min()) / (df[column].max() - df[column].
↪min())

display(df)
```

	Full name	Middle Initial of the Provider	\
0	SATYASREE UPADHYAYULA		NaN
1	WENDY JONES		P
2	RICHARD DUROCHER		W
3	JASPER FULLARD		NaN
4	ANTHONY PERROTTI		E
...
99995	JOAN PAPES		NaN
99996	MARGARET HAYNER		S
99997	DANA VALENCIA		NaN
99998	RAFAELA GONZALEZ-LAMOS		NaN
99999	ELIIAN RAMEZANI		NaN

Credentials of the Provider Gender of the Provider \

0	M.D.	F
1	M.D.	F
2	DPM	M
3	MD	M
4	DO	M
...
99995	PT	F
99996	ARNP	F
99997	M.D.	M
99998	NaN	F
99999	NaN	F

	Entity Type of the Provider	City of the Provider	\
0	I	SAINT LOUIS	
1	I	FAYETTEVILLE	
2	I	NORTH HAVEN	
3	I	KANSAS CITY	
4	I	JUPITER	
...	
99995	I	WILMINGTON	
99996	I	REDMOND	
99997	I	SAINT LOUIS	
99998	I	LARCHMONT	
99999	I	GREAT NECK	

	Zip Code of the Provider	State Code of the Provider	\
0	631041004.0	MO	
1	283043815.0	NC	
2	64732343.0	CT	
3	641183998.0	MO	
4	334585700.0	FL	
...	
99995	604811236.0	IL	
99996	977561502.0	OR	
99997	631312322.0	MO	
99998	105383500.0	NY	
99999	11023.0	NY	

	Provider Type	\
0	Internal Medicine	
1	Obstetrics & Gynecology	
2	Podiatry	
3	Internal Medicine	
4	Internal Medicine	
...	...	
99995	Physical Therapist in Private Practice	
99996	Nurse Practitioner	
99997	Cardiology	

99998 Internal Medicine
 99999 Physical Therapist in Private Practice

	Medicare Participation Indicator	...	HCPSC Code	HCPSC Drug Indicator	\
0	Y	...	99223		N
1	Y	...	G0202		N
2	Y	...	99348		N
3	Y	...	81002		N
4	Y	...	96372		N
...	
99995	Y	...	97162		N
99996	Y	...	99213		N
99997	Y	...	93320		N
99998	Y	...	G0008		N
99999	Y	...	97112		N

	Number of Services	Number of Medicare Beneficiaries	\
0	27		24
1	175		175
2	32		13
3	20		18
4	33		24
...
99995	20		20
99996	136		107
99997	11		11
99998	12		12
99999	589		76

	Number of Distinct Medicare Beneficiary/Per Day Services	\
0	27	
1	175	
2	32	
3	20	
4	31	
...	...	
99995	20	
99996	136	
99997	11	
99998	12	
99999	587	

	Average Medicare Allowed Amount	Average Submitted Charge Amount	\
0	200.58777778	305.21111111	
1	123.73	548.8	
2	90.65	155	
3	3.5	5	
4	26.52	40	

```

...
99995      85.3725      214
99996      61.27      144.05147059
99997      17.98      109.54545455
99998      30.54      65
99999      38.601052632      46.867572156

Average Medicare Payment Amount Average Medicare Standardized Amount \
0      157.26222222      160.90888889
1      118.83      135.31525714
2      64.4396875      60.5959375
3      3.43      3.43
4      19.539393939      19.057575758
...
99995      60.7255      62.2485
99996      30.006176471      37.040220588
99997      14.09      14.62
99998      29.93      25.32
99999      29.736876061      25.912139219

```

```

Money difference
0      0.002891
1      0.007834
2      0.001885
3      0.000325
4      0.000656
...
99995      0.002984
99996      0.002297
99997      0.001971
99998      0.000912
99999      0.000598

```

[100000 rows x 21 columns]

```

[85]: column = 'Average Medicare Payment Amount'
df[column] = (df[column].astype(float) - df[column].astype(float).min()) / (
    df[column].astype(float).max() - df[column].astype(float).min())

display(df)

```

```

Full name Middle Initial of the Provider \
0      SATYASREE UPADHYAYULA      NaN
1      WENDY JONES      P
2      RICHARD DUROCHER      W
3      JASPER FULLARD      NaN
4      ANTHONY PERROTTI      E
...

```

99995	JOAN PAPES	NaN
99996	MARGARET HAYNER	S
99997	DANA VALENCIA	NaN
99998	RAFAELA GONZALEZ-LAMOS	NaN
99999	ELIIAN RAMEZANI	NaN

	Credentials of the Provider	Gender of the Provider	\
0	M.D.	F	
1	M.D.	F	
2	DPM	M	
3	MD	M	
4	DO	M	
...	
99995	PT	F	
99996	ARNP	F	
99997	M.D.	M	
99998	NaN	F	
99999	NaN	F	

	Entity Type of the Provider	City of the Provider	\
0	I	SAINT LOUIS	
1	I	FAYETTEVILLE	
2	I	NORTH HAVEN	
3	I	KANSAS CITY	
4	I	JUPITER	
...	
99995	I	WILMINGTON	
99996	I	REDMOND	
99997	I	SAINT LOUIS	
99998	I	LARCHMONT	
99999	I	GREAT NECK	

	Zip Code of the Provider	State Code of the Provider	\
0	631041004.0	MO	
1	283043815.0	NC	
2	64732343.0	CT	
3	641183998.0	MO	
4	334585700.0	FL	
...	
99995	604811236.0	IL	
99996	977561502.0	OR	
99997	631312322.0	MO	
99998	105383500.0	NY	
99999	11023.0	NY	

	Provider Type	\
0	Internal Medicine	
1	Obstetrics & Gynecology	

2	Podiatry
3	Internal Medicine
4	Internal Medicine
...	...
99995	Physical Therapist in Private Practice
99996	Nurse Practitioner
99997	Cardiology
99998	Internal Medicine
99999	Physical Therapist in Private Practice

	Medicare Participation Indicator	...	HCPCS Code	HCPCS Drug Indicator	\
0	Y	...	99223		N
1	Y	...	G0202		N
2	Y	...	99348		N
3	Y	...	81002		N
4	Y	...	96372		N
...	
99995	Y	...	97162		N
99996	Y	...	99213		N
99997	Y	...	93320		N
99998	Y	...	G0008		N
99999	Y	...	97112		N

	Number of Services	Number of Medicare Beneficiaries	\
0	27		24
1	175		175
2	32		13
3	20		18
4	33		24
...	
99995	20		20
99996	136		107
99997	11		11
99998	12		12
99999	589		76

	Number of Distinct Medicare Beneficiary/Per Day Services	\
0	27	
1	175	
2	32	
3	20	
4	31	
...	...	
99995	20	
99996	136	
99997	11	
99998	12	
99999	587	

	Average Medicare Allowed Amount	Average Submitted Charge Amount \
0	200.58777778	305.21111111
1	123.73	548.8
2	90.65	155
3	3.5	5
4	26.52	40
...
99995	85.3725	214
99996	61.27	144.05147059
99997	17.98	109.54545455
99998	30.54	65
99999	38.601052632	46.867572156

	Average Medicare Payment Amount	Average Medicare Standardized Amount \
0	0.009787	160.90888889
1	0.007395	135.31525714
2	0.004010	60.5959375
3	0.000213	3.43
4	0.001216	19.057575758
...
99995	0.003779	62.2485
99996	0.001867	37.040220588
99997	0.000876	14.62
99998	0.001862	25.32
99999	0.001850	25.912139219

	Money difference
0	0.002891
1	0.007834
2	0.001885
3	0.000325
4	0.000656
...	...
99995	0.002984
99996	0.002297
99997	0.001971
99998	0.000912
99999	0.000598

[100000 rows x 21 columns]

```
[86]: column = 'Average Submitted Charge Amount'
df[column] = (df[column].astype(float) - df[column].astype(float).min()) /
↳ (df[column].astype(float).max() - df[column].astype(float).min())

display(df)
```

	Full name	Middle Initial of the Provider	\
0	SATYASREE UPADHYAYULA		NaN
1	WENDY JONES		P
2	RICHARD DUROCHER		W
3	JASPER FULLARD		NaN
4	ANTHONY PERROTTI		E
...	
99995	JOAN PAPES		NaN
99996	MARGARET HAYNER		S
99997	DANA VALENCIA		NaN
99998	RAFAELA GONZALEZ-LAMOS		NaN
99999	ELIIAN RAMEZANI		NaN

	Credentials of the Provider	Gender of the Provider	\
0	M.D.		F
1	M.D.		F
2	DPM		M
3	MD		M
4	DO		M
...	
99995	PT		F
99996	ARNP		F
99997	M.D.		M
99998	NaN		F
99999	NaN		F

	Entity Type of the Provider	City of the Provider	\
0	I	SAINT LOUIS	
1	I	FAYETTEVILLE	
2	I	NORTH HAVEN	
3	I	KANSAS CITY	
4	I	JUPITER	
...	
99995	I	WILMINGTON	
99996	I	REDMOND	
99997	I	SAINT LOUIS	
99998	I	LARCHMONT	
99999	I	GREAT NECK	

	Zip Code of the Provider	State Code of the Provider	\
0	631041004.0		MO
1	283043815.0		NC
2	64732343.0		CT
3	641183998.0		MO
4	334585700.0		FL
...	
99995	604811236.0		IL
99996	977561502.0		OR

99997	631312322.0	MO
99998	105383500.0	NY
99999	11023.0	NY

	Provider Type \
0	Internal Medicine
1	Obstetrics & Gynecology
2	Podiatry
3	Internal Medicine
4	Internal Medicine
...	...
99995	Physical Therapist in Private Practice
99996	Nurse Practitioner
99997	Cardiology
99998	Internal Medicine
99999	Physical Therapist in Private Practice

	Medicare Participation Indicator	...	HCPCS Code	HCPCS Drug Indicator	\
0	Y	...	99223		N
1	Y	...	G0202		N
2	Y	...	99348		N
3	Y	...	81002		N
4	Y	...	96372		N
...	
99995	Y	...	97162		N
99996	Y	...	99213		N
99997	Y	...	93320		N
99998	Y	...	G0008		N
99999	Y	...	97112		N

	Number of Services	Number of Medicare Beneficiaries	\
0	27		24
1	175		175
2	32		13
3	20		18
4	33		24
...
99995	20		20
99996	136		107
99997	11		11
99998	12		12
99999	589		76

	Number of Distinct Medicare Beneficiary/Per Day Services	\
0	27	
1	175	
2	32	
3	20	

4	31
...	...
99995	20
99996	136
99997	11
99998	12
99999	587

	Average Medicare Allowed Amount	Average Submitted Charge Amount \
0	200.58777778	0.004868
1	123.73	0.008753
2	90.65	0.002472
3	3.5	0.000080
4	26.52	0.000638
...
99995	85.3725	0.003413
99996	61.27	0.002298
99997	17.98	0.001747
99998	30.54	0.001037
99999	38.601052632	0.000747

	Average Medicare Payment Amount	Average Medicare Standardized Amount \
0	0.009787	160.90888889
1	0.007395	135.31525714
2	0.004010	60.5959375
3	0.000213	3.43
4	0.001216	19.057575758
...
99995	0.003779	62.2485
99996	0.001867	37.040220588
99997	0.000876	14.62
99998	0.001862	25.32
99999	0.001850	25.912139219

	Money difference
0	0.002891
1	0.007834
2	0.001885
3	0.000325
4	0.000656
...	...
99995	0.002984
99996	0.002297
99997	0.001971
99998	0.000912
99999	0.000598

[100000 rows x 21 columns]

```
[87]: df.head()
```

```
[87]:
```

	Full name	Middle Initial of the Provider	\
0	SATYASREE UPADHYAYULA		NaN
1	WENDY JONES		P
2	RICHARD DUROCHER		W
3	JASPER FULLARD		NaN
4	ANTHONY PERROTTI		E

	Credentials of the Provider	Gender of the Provider	\
0	M.D.		F
1	M.D.		F
2	DPM		M
3	MD		M
4	DO		M

	Entity Type of the Provider	City of the Provider	Zip Code of the Provider	\
0	I	SAINT LOUIS	631041004.0	
1	I	FAYETTEVILLE	283043815.0	
2	I	NORTH HAVEN	64732343.0	
3	I	KANSAS CITY	641183998.0	
4	I	JUPITER	334585700.0	

	State Code of the Provider	Provider Type	\
0	MO	Internal Medicine	
1	NC	Obstetrics & Gynecology	
2	CT	Podiatry	
3	MO	Internal Medicine	
4	FL	Internal Medicine	

	Medicare Participation Indicator	...	HCPCS Code	HCPCS Drug Indicator	\
0	Y	...	99223		N
1	Y	...	G0202		N
2	Y	...	99348		N
3	Y	...	81002		N
4	Y	...	96372		N

	Number of Services	Number of Medicare Beneficiaries	\
0	27		24
1	175		175
2	32		13
3	20		18
4	33		24

	Number of Distinct Medicare Beneficiary/Per Day Services	\
0		27
1		175

2	32
3	20
4	31

	Average Medicare Allowed Amount	Average Submitted Charge Amount \
0	200.58777778	0.004868
1	123.73	0.008753
2	90.65	0.002472
3	3.5	0.000080
4	26.52	0.000638

	Average Medicare Payment Amount	Average Medicare Standardized Amount \
0	0.009787	160.90888889
1	0.007395	135.31525714
2	0.004010	60.5959375
3	0.000213	3.43
4	0.001216	19.057575758

	Money difference
0	0.002891
1	0.007834
2	0.001885
3	0.000325
4	0.000656

[5 rows x 21 columns]

```
[88]: df["Average Medicare Allowed Amount"] = df["Average Medicare Allowed Amount"].
      ↪replace('.', '', regex=True)
df["Average Medicare Standardized Amount"] = df["Average Medicare Standardized_
      ↪Amount"].replace('.', '', regex=True)
```

```
[89]: df.drop(columns=['Average Medicare Allowed Amount', 'Average Medicare_
      ↪Standardized Amount'], inplace=True)
```

```
[90]: df.head()
```

```
[90]:
```

	Full name	Middle Initial of the Provider \
0	SATYASREE UPADHYAYULA	NaN
1	WENDY JONES	P
2	RICHARD DUROCHER	W
3	JASPER FULLARD	NaN
4	ANTHONY PERROTTI	E

	Credentials of the Provider	Gender of the Provider \
0	M.D.	F
1	M.D.	F

2	DPM	M
3	MD	M
4	DO	M

	Entity Type of the Provider	City of the Provider	Zip Code of the Provider \
0	I	SAINT LOUIS	631041004.0
1	I	FAYETTEVILLE	283043815.0
2	I	NORTH HAVEN	64732343.0
3	I	KANSAS CITY	641183998.0
4	I	JUPITER	334585700.0

	State Code of the Provider	Provider Type \
0	MO	Internal Medicine
1	NC	Obstetrics & Gynecology
2	CT	Podiatry
3	MO	Internal Medicine
4	FL	Internal Medicine

	Medicare Participation Indicator	Place of Service	HCPSC Code \
0	Y	F	99223
1	Y	O	G0202
2	Y	O	99348
3	Y	O	81002
4	Y	O	96372

	HCPSC Drug Indicator	Number of Services	Number of Medicare Beneficiaries \
0	N	27	24
1	N	175	175
2	N	32	13
3	N	20	18
4	N	33	24

	Number of Distinct Medicare Beneficiary/Per Day Services \
0	27
1	175
2	32
3	20
4	31

	Average Submitted Charge Amount	Average Medicare Payment Amount \
0	0.004868	0.009787
1	0.008753	0.007395
2	0.002472	0.004010
3	0.000080	0.000213
4	0.000638	0.001216

Money difference

```

0          0.002891
1          0.007834
2          0.001885
3          0.000325
4          0.000656

```

```

[91]: mean = np.mean(df["Money difference"])
      std = np.std(df["Money difference"])
      print('mean of the dataset is', mean)
      print('std. deviation is', std)

```

```

mean of the dataset is 0.005155933816918636
std. deviation is 0.01621035059322907

```

```

[92]: threshold = 3
      outlier = []
      for i in df["Money difference"]:
          z = (i-mean)/std
          if z > threshold:
              outlier.append(i)
      print('outlier in dataset is', outlier)

```

```

outlier in dataset is [0.08623169355807118, 0.1260960514481938,
0.09765522942620496, 0.09387265509479728, 0.09586182860426613,
0.07166196337627757, 0.05582909121041109, 0.13022887518707357,
0.12577427984139958, 0.17439361941367468, 0.08482083208422132,
0.24274269867103462, 0.06222214009513836, 0.1646550310959583,
0.0909358936077171, 0.20435612710381107, 0.05889011409909453,
0.11244981211921393, 0.0876214793787706, 0.07051920901223939,
0.11243297109306007, 0.10324433684183285, 0.09514025084465452,
0.08599582889694884, 0.5089578358629917, 0.06202176512430257,
0.1371886516590768, 0.06424617828230543, 0.09078938176271528,
0.06565864848611717, 0.060343804354407435, 0.13699204476689653,
0.10135090085362947, 0.06010138576192941, 0.11511447939191116,
0.07592474614031629, 0.10921956022605839, 0.09245746295404116,
0.08989626750906567, 0.17189462638741376, 0.21560485999716522,
0.07265242678217632, 0.19287120151773665, 0.09501326098161221,
0.060193117888095116, 0.10781839716869931, 0.06726568544259004,
0.09235641765312268, 0.10212641414666418, 0.10442589354738255,
0.05431102597482588, 0.16488335803537016, 0.07107311538295691,
0.06544010788508224, 0.055016120824802194, 0.0692251197525761,
0.057490586171310816, 0.11367313633370306, 0.07585826147119222,
0.08588414153489514, 0.12491427365628781, 0.07786866908056714,
0.09118630791409033, 0.08601232836736619, 0.0734718649419846,
0.3824423256031174, 0.27457634509655615, 0.07212183059344295,
0.10955734593688131, 0.06395666211969833, 0.08960601719323236,
0.7962086907104512, 0.05761623518637374, 0.055365060853861024,
0.10227749954500637, 0.07966494697926545, 0.06487277123501878,

```

0.10480504591821829, 0.09543068775234237, 1.0, 0.09960701781602627,
0.4070358242227097, 0.11509239573184989, 0.05842432187747302,
0.07576681231277786, 0.10281360399955006, 0.16570855646481963,
0.0722755790556197, 0.07412874282686761, 0.09400512103426162,
0.18057285453058758, 0.09043946765215496, 0.09441909689081047,
0.0634380137799557, 0.11954620648206332, 0.05802159424431595,
0.2516763278679652, 0.06905747360542816, 0.08418573190869247,
0.07982682371338326, 0.05430607743094744, 0.06199518623428194,
0.09088409690323627, 0.11254991943034214, 0.07565591138109154,
0.18874981485908052, 0.06394429425754573, 0.09252177981285435,
0.11348976757865425, 0.06949500088048295, 0.0733685855498745,
0.08452273291556443, 0.08124942936546256, 0.25399778392920536,
0.05767683937818524, 0.06547440072217721, 0.1857151032199533,
0.09944269690307485, 0.07224953435811886, 0.08929494174724742,
0.05883273736882635, 0.08749588294385069, 0.1552357575103765,
0.07446967831575317, 0.13482782998330556, 0.05663619039559732,
0.058760968820879105, 0.057776359248532305, 0.20529591592699944,
0.06850079717501725, 0.08699624154849388, 0.09016452237679094,
0.10804819892030473, 0.0778156047620806, 0.07331151303306548,
0.06096464385309297, 0.058895232405150606, 0.057835903339305765,
0.3301336190189998, 0.07749022796362177, 0.06390809121259929, 0.071421229217775,
0.30014052648148604, 0.09858470334702567, 0.11640691626752721,
0.06809573811430782, 0.10516574441153141, 0.1910655845706488,
0.2111692611184072, 0.09785677675370458, 0.16703976592561742,
0.7591973948475864, 0.060194236259323956, 0.12335781894478355,
0.10392751785649589, 0.056793724008851465, 0.05827044901015213,
0.06904810166991182, 0.0882436363116731, 0.06258915494375716,
0.11307472037768011, 0.07268990420977027, 0.07089861027065623,
0.06899075957085375, 0.10268085297598364, 0.1954537291622674,
0.19988064554510748, 0.11074699744693581, 0.07257023478471314,
0.06373076302071697, 0.06380794429723102, 0.12149445535613121,
0.07482812883619111, 0.058573656494287384, 0.12475520284660364,
0.0695494591486343, 0.12143324160272187, 0.12756407459181524,
0.1304909890980869, 0.11310562581688073, 0.06855025974476946,
0.09158499058909535, 0.06913629914433786, 0.22659966366240408,
0.11897622799476902, 0.07152258633962301, 0.07018573506593656,
0.05733240757444336, 0.08301430208855874, 0.09994444462425749,
0.09543330916358979, 0.09869691927408668, 0.06103839223201172,
0.11399767715452672, 0.09487443596197064, 0.1408329365789954,
0.09974057137925398, 0.06239687118199595, 0.0658941862822898,
0.20768103527910028, 0.16964588497983135, 0.09458775299624106,
0.1553029666181948, 0.06146047460161158, 0.2519316578468641,
0.05525867766397174, 0.06103326511352368, 0.055198354837093545,
0.07195981996101143, 0.08695926502650857, 0.3942792012167353,
0.09087175246658896, 0.08379108606785689, 0.0784254577415406,
0.08733924649628744, 0.0688265752278131, 0.12054828663698806,
0.15474713608362897, 0.07074777536709485, 0.05694663363774797,
0.059049311365600016, 0.05849202006299685, 0.2224033848538859,

0.06140877079432526, 0.06672564298842115, 0.14171021679141926,
0.0574476190500894, 0.1246712296776921, 0.05696890351914905,
0.22061209831980866, 0.131113149045409, 0.06159900867288772, 0.2519665045321632,
0.4161735482370158, 0.11198106757882356, 0.06785317715127766,
0.06703463577507791, 0.15262439580402734, 0.060546375556628694,
0.12026357234507418, 0.063573065877259, 0.06159774064157097,
0.05540647204995604, 0.40711642051711666, 0.06405349356435412,
0.13707107077426156, 0.0753683110617004, 0.10970746611137289,
0.26276365355964093, 0.3179834986979443, 0.10268725648550536,
0.07631580121605885, 0.07201138136374378, 0.0904499017654497,
0.09781553055374775, 0.05406771744958142, 0.05438466877825076,
0.12046538527818662, 0.07583411382038255, 0.06205433628052511,
0.08695907724108479, 0.1500613039065163, 0.06991550638240814,
0.0640199861920448, 0.06516639819488032, 0.0673750045674509,
0.15176255112389425, 0.06724397789915822, 0.06779200015489363,
0.05891240192346941, 0.13624976894055418, 0.08869961768433349,
0.08362845024700906, 0.09037094179488839, 0.2915255913799614,
0.09302140824780185, 0.14400464927414172, 0.07191909590939159,
0.1448583557462641, 0.1353892401293272, 0.11050360250447364,
0.08144272543098387, 0.06769754709799224, 0.061358542151519445,
0.054893357145086565, 0.14374647046278344, 0.062571142488979,
0.09997962480857797, 0.06301663928122231, 0.7263011103623821,
0.07545557477570762, 0.15832998208128438, 0.08035877482877986,
0.10954364610332905, 0.05477647088952985, 0.10786215034325343,
0.1112959516666018, 0.05401829709467338, 0.08204252960215856,
0.05430923486871004, 0.13864201394504663, 0.1056379852566126,
0.0874180807635009, 0.058322443867108296, 0.06559477035340573,
0.07793057073998737, 0.06663255265389026, 0.061427699645806355,
0.06088174249429152, 0.0622918568006287, 0.05437820197695929,
0.5995954535114866, 0.2862507992345051, 0.07955891035754266,
0.07514316880795757, 0.22599797155930604, 0.06332205995909715,
0.06268835682201859, 0.10545182651224355, 0.07976312734067412,
0.060314510608059785, 0.07599277746573913, 0.10673330474995595,
0.22200954365731945, 0.212127273604148, 0.08355763023292342,
0.06991193592043629, 0.07955768706439112, 0.1268354125395277,
0.09183037796143234, 0.1984065772386087, 0.1164562484387, 0.07912897518599994,
0.06643007429891239, 0.10705449949036162, 0.16670420799023464,
0.07717971730845226, 0.05497168353719341, 0.09747749842994055,
0.11777688167346886, 0.218975324488018, 0.060836075665958604,
0.0701751487422417, 0.06690076673151799, 0.05816627530114328,
0.054861816883034764, 0.061955662704580576, 0.12297380863371804,
0.09536923355712505, 0.060921838223893764, 0.1720932084156846,
0.05931197029082759, 0.05442631338058009, 0.22565749200791343,
0.07366064518260562, 0.09967458329978515, 0.08005334546028371,
0.07419348916474064, 0.12155318080334772, 0.10445003457432107,
0.07084443324978135, 0.16136691227259356, 0.054363754608884964,
0.06569335138049918, 0.07340387881760747, 0.0905171068018123,
0.05938342604213311, 0.18254035613001227, 0.053807797006341265,

0.06574713777385201, 0.06241352339864011, 0.05776760562920763,
0.05993106884765798, 0.22735150278583732, 0.13018351694949676,
0.1572892585640653, 0.10061242902045878, 0.07006639179778495,
0.05903909085685775, 0.13751605214281765, 0.07760153566404021,
0.2691394786510732, 0.059485612231025616, 0.162502583654954,
0.07287946473428018, 0.12139449393176456, 0.06569568378166052,
0.05708457983375572, 0.17361536221150545, 0.0943232527272, 0.08857207475320113,
0.21430505341894956, 0.057765003782330335, 0.09126057298440352,
0.08253061655648465, 0.11507583759825774, 0.0639357724670511,
0.26348775592502344, 0.3320554764498548, 0.1298003991470989, 0.4412670032340127,
0.08372525732199544, 0.08198846349856954, 0.3463147558882393,
0.05713639032615219, 0.20174827288226863, 0.08360305464261023,
0.13255093114743405, 0.07146053920233943, 0.28374233482615063,
0.10245273402556498, 0.06143070474767758, 0.10054189275862022,
0.1497756522279064, 0.0643708205702582, 0.10505116231721758,
0.09863235723189334, 0.06510727448445339, 0.1282363798797484,
0.06274910173564421, 0.056058390709479836, 0.15870418642037232,
0.42600230258519445, 0.07189594522997234, 0.06490850511260722,
0.12303534333439232, 0.05803820836068742, 0.058657509851280806,
0.06384664036130377, 0.07267263143298923, 0.2392535329614963,
0.13469811796633815, 0.06850007941400116, 0.0882350911717391,
0.07394497230589453, 0.2566477303126255, 0.2529679415541204,
0.08169322139887612, 0.055631615215438544, 0.13978389752039794,
0.12523878592714544, 0.05894126088584144, 0.07046384465689642,
0.21507579358387574, 0.0967571740652618, 0.05903740634488842,
0.10090319867176667, 0.11725032861748026, 0.09867877209548657,
0.07305485275043377, 0.055609533273696346, 0.08046645021459276,
0.07163880590543176, 0.12065204478584739, 0.07937119924702184,
0.08214777371125334, 0.28540201445007674, 0.06055918221777664,
0.08182466961269129, 0.13017329752706025, 0.059529030681966785,
0.06477430500834237, 0.1512648320827525, 0.23752874214888917,
0.08776652308697856, 0.4961431076346516, 0.07990606754608415,
0.17572356641472295, 0.08288083467439, 0.1338796847036758, 0.06924499506638386,
0.0954121447959952, 0.48232802394562724, 0.06090238277789881,
0.12029527789263024, 0.08135218190944725, 0.05945242545807596,
0.0597233248263916, 0.15479435517654638, 0.0693171645256528, 0.0956608843431393,
0.13960156185631828, 0.08937144883669858, 0.13483317225448074,
0.23332298995216977, 0.1448956737533182, 0.05690352887468195,
0.06251496679205883, 0.20599238610238954, 0.06183154252956081,
0.06059023640620181, 0.10777008501724733, 0.1142908990853403,
0.05841290614478749, 0.5076219168338385, 0.06724271292765525,
0.06244276446143716, 0.10571136063760542, 0.06507327869747417,
0.056007049607731715, 0.08170490432229056, 0.09916779835410652,
0.060339353280551115, 0.1828065060283042, 0.5888574588085228,
0.10417605697584659, 0.11799446909133068, 0.07022489797665507,
0.05705405638666503, 0.159628130225529, 0.13963191530952815,
0.06651589260371729, 0.06446440472697615, 0.6649061732676737,
0.29595929965905493, 0.08970261577866258, 0.11004568849975765,

0.06365019969511124, 0.06279594763898107, 0.07547654910744829,
0.06884528827689175, 0.05541226254389033, 0.05486715647332372,
0.11462495826055291, 0.06514692801838305, 0.08448036488898704,
0.13814780694250242, 0.05917787809170748, 0.07793801060470607,
0.1834388539433944, 0.12017652265135309, 0.07588382922981464,
0.05628431005724954, 0.13860384651419516, 0.054265361710598675,
0.05876163733217847, 0.057118666925101653, 0.2764370766984643,
0.24786113406549312, 0.10767359664452907, 0.08002144034882602,
0.11218001254030462, 0.13297367549717845, 0.10347918507607057,
0.05552207325084889, 0.1893635941355785, 0.06535967195012117,
0.06820528708958319, 0.0738558853214464, 0.133934132239922, 0.05624863938379726,
0.05452445267093798, 0.21365883482042156, 0.05457037266249397,
0.05707275391387513, 0.3578292017823632, 0.06304579220322457,
0.06272818981705908, 0.05539316426691978, 0.15533404951520277,
0.29124037077597337, 0.2140510561027039, 0.06319030595880021,
0.09007085304287032, 0.07494187018558596, 0.3090583996180998,
0.06898387065792694, 0.06897104413655791, 0.05621585571861715,
0.13278610778620317, 0.056177818628857415, 0.08691960457653743,
0.05839408518145022, 0.05476537613811109, 0.10798694305370618,
0.1201090671757694, 0.056784171950065056, 0.12549087850107019,
0.07892549490149944, 0.0596231645120255, 0.19238177849216193,
0.07771528501455542, 0.1147483212908026, 0.06316597220933773,
0.05447376063354806, 0.05598644735744505, 0.10014245476670615,
0.05652144650219323, 0.08758767590454077, 0.07674021858883608,
0.10510816520144213, 0.1920318548661731, 0.05468480750208618,
0.08820065549612872, 0.10778963781975873, 0.07225265849494895,
0.0824153461001148, 0.4104803200178179, 0.2260209040036149,
0.062486488458744126, 0.2596828496353968, 0.07068790995258263,
0.08354275255579417, 0.12053391473123391, 0.0790815995031947,
0.14600661038305454, 0.13346268543264353, 0.05494688153982436,
0.07367515062492211, 0.19654045033489317, 0.06285606426792101,
0.09008473798481602, 0.20998408784844097, 0.08090571177252553,
0.08231733207037999, 0.06847055107345847, 0.06107398551240098,
0.16943157455646132, 0.07693507358811957, 0.27292620689965896,
0.058215339709164704, 0.11841089777965202, 0.05646702322712606,
0.3719763737465413, 0.06090121308143972, 0.06027981252844147,
0.0839630018906453, 0.056275742189899246, 0.10298893216775803,
0.07259613996395933, 0.06447276006468736, 0.0603819543977067,
0.05454863953671938, 0.1007628709211025, 0.12103259547581044,
0.12487325534550621, 0.11918912042106852, 0.12157486000082444,
0.06183581715637537, 0.05386333624288933, 0.30723829078070153,
0.12223540815881209, 0.13891757386265244, 0.07966494697926545,
0.056423472169064244, 0.06604574240348587, 0.09087920701636322,
0.0785886261408731, 0.716322214217355, 0.08463689437188905, 0.0808328921690775,
0.05722549461700385, 0.13673836699770658, 0.05851843110861291,
0.07142826757784013, 0.08042605039211512, 0.054552114398351545,
0.14236743936700716, 0.1066799181537516, 0.05382989022634888,
0.17549013686986095, 0.07664959793398234, 0.057159138476534026,

0.07461830485383421, 0.06540668174436164, 0.20452592909863887,
0.06702999995913649, 0.0662937380450141, 0.08610312258927198,
0.07189290077449308, 0.061091876114099865, 0.17520972696722578,
0.25021411392454024, 0.11577985720148583, 0.059996234463797725,
0.28214604550136435, 0.06800670240548935, 0.1882348181329766,
0.07957595269711479, 0.05780513398960869, 0.09646456556944992,
0.10881351936630933, 0.056701270591263606, 0.08961671761741494,
0.18555032446558192, 0.06202873274444544, 0.10308837975600724,
0.11402201259777087, 0.22147741227808848, 0.058444899256650994,
0.07125167704137209, 0.07396868515362447, 0.055023743821549446,
0.44241789362328454, 0.15426988855224827, 0.2589152114223362,
0.09272303770759363, 0.18849705195668662, 0.06361463298420436,
0.0641160670099403, 0.12640362012516312, 0.07753112306298623,
0.08165393948509737, 0.06660945170669819, 0.08750162920158487,
0.0731422049725957, 0.06843838537954425, 0.20397575616414965,
0.07905683258677125, 0.07593535495674471, 0.05981574355370432,
0.15443013534487346, 0.09652183410853739, 0.054482092793231164,
0.05773847786084471, 0.09194634934879085, 0.07929128911419187,
0.25122115909670223, 0.06593574583974317, 0.09338090293013325,
0.07148213980652375, 0.06621522073020158, 0.11374715363796961,
0.06185044399502495, 0.06021156266837667, 0.0630378788917026,
0.059163773955239043, 0.07493802582856647, 0.1110767567349568,
0.058323701866536876, 0.06753576990123994, 0.08114856087754362,
0.12829601243959202, 0.05912890028791379, 0.0979591631449969,
0.1470370003838343, 0.2034163904054217, 0.0973086396124599, 0.1622492913351937,
0.055117400195516406, 0.11593562373915164, 0.10998145560294992,
0.21157528701104328, 0.12602218961149195, 0.6502103786139, 0.07813238916947743,
0.1628488364157858, 0.07382469522877509, 0.30189464874303235,
0.17308599437781397, 0.06166645522439429, 0.09364615305914083,
0.23846244605282824, 0.06388075007630785, 0.058341036600200506,
0.08922206050690205, 0.10935002325432218, 0.05685967838322175,
0.14066146689135917, 0.15166206045760355, 0.09706687110754968,
0.0741983348541476, 0.17311277974462289, 0.09031968900780005,
0.057225307741917614, 0.07262618564000323, 0.08742131772855452,
0.08837482263080665, 0.10825001800459283, 0.06276635503323494,
0.05950532712865346, 0.055037931472636414, 0.12351906471665951,
0.10935167381038569, 0.23377581253365465, 0.057593555618025094,
0.05534751736611053, 0.07861786919821023, 0.06311808846108576,
0.09789030381990646, 0.09517309014961617, 0.05692429006527648,
0.07365302106187019, 0.09048473106600086, 0.05801481019705108,
0.09236038206112468, 0.08605806641427594, 0.2480526704546791,
0.06878774856307796, 0.13056318384279605, 0.08919596254922169,
0.4143544550073386, 0.06854256960808706, 0.09757952750182926,
0.06304372843261169, 0.2776521625611529, 0.06418695219185516,
0.07857481627723727, 0.05492024632683089, 0.08213349190490428,
0.07160373695282639, 0.14384386641433938, 0.08876302683380508,
0.07230138107490379, 0.31011424933582415, 0.10368378214475593,
0.07587268126851804, 0.19001538797177073, 0.10674007188288989,

0.05911103202716713, 0.24173242858290933, 0.08947090996039112,
0.05432055324971651, 0.08360176707058439, 0.05520479192078925,
0.15241403579007254, 0.13347484923739547, 0.05466630870429578,
0.05907631565071066, 0.0688444999697112, 0.2877646281268778,
0.09313652002366231, 0.06708809225379767, 0.14069876604054496,
0.06474733602010646, 0.1081041295869231, 0.09696649311104101,
0.057893452511586216, 0.054793334377219755, 0.09240110785419885,
0.08592241419684055, 0.12255784577479197, 0.05471720608221711,
0.06356670367995564, 0.054755169956756715, 0.1382986330828819,
0.0629812347418133, 0.0895468941092102, 0.057918103629533854,
0.06347820837815663, 0.06090270075473622, 0.06442119872066951,
0.12797605340090107, 0.13844183510235067, 0.08457183745730795,
0.09025270764158197, 0.5221481939930717, 0.09041297085586214,
0.07181030988899209, 0.1627997470101128, 0.05758562815035124,
0.0682038278661752, 0.05521602176514256, 0.11928837877794003,
0.14374758579978753, 0.1182679411823824, 0.07168135333335388,
0.08878690072171852, 0.2129560765946684, 0.06771550120178142,
0.05832487799735076, 0.10572857333158285, 0.2051053941416043,
0.057536708694566965, 0.0562410482846185, 0.06517858904195035,
0.11494931932108474, 0.06873950059059798, 0.054901004391269964,
0.09595059467159527, 0.05721803674644303, 0.05430572762226673,
0.12477390366641736, 0.07253347031704244, 0.06644884972609315,
0.10069385835977335, 0.272231307904998, 0.05702903608267658,
0.09722071810140907, 0.1386277885338805, 0.056321416616062825,
0.0722696618423139, 0.07065413322446941, 0.08397616817122756, 0.086458408436499,
0.2113121523582085, 0.07943423168465674, 0.06278460438802266,
0.12006862219588754, 0.06071016944303275, 0.057791256955716505,
0.06553073268272544, 0.10787091142527781, 0.18679780094776782,
0.07193692014215551, 0.1568613712190834, 0.057391424801667895,
0.06569688859043805, 0.08654931903165165, 0.08504113763117355,
0.10721552384217266, 0.28613711200742875, 0.10472639580594503,
0.12048834527396461, 0.0824875472057944, 0.27529480697787284,
0.08549019792342721, 0.2958797092052569, 0.1480433879084555,
0.09112724288678359, 0.08159841162654172, 0.08970486001641965,
0.05535518387695726, 0.0940080074510072, 0.055216489030676795,
0.07991745300487603, 0.06317145192903788, 0.057749287216010615,
0.10542321415109274, 0.07174271143853514, 0.0604169771728531,
0.15277051360555047, 0.055681870147614804, 0.05703889506858335,
0.11797319166012882, 0.08456071914770664, 0.08574689212042642,
0.1846682039398955, 0.12540937562860088, 0.3274584296815016,
0.07765788154728452, 0.053862134069603836, 0.13467255243676327,
0.08428010087189267, 0.0552876863092164, 0.07398159938305371,
0.08269200643107795, 0.06300823032280707, 0.4857042931791073,
0.08386939955136336, 0.07083803157250233, 0.06887313933874012,
0.06430080820469582, 0.0675185208151434, 0.07602853541206332,
0.0832820664668017, 0.06261026461994074, 0.0607404771767567,
0.06856455268513904, 0.06859494120159257, 0.323992237680292,
0.061988782351791734, 0.12289469098319487, 0.13663044917589612,

0.09637701963134987, 0.08977628374589229, 0.05452979831880996,
0.07722081970049108, 0.07314391382724329, 0.05538144268465628,
0.056479031853504114, 0.0613925069021929, 0.06837813729901995,
0.26184585332618204, 0.06368680548993172, 0.0894285230617488,
0.0722191167453916, 0.10501427208818556, 0.07732017158604461,
0.09084305816804865, 0.20358136543568306, 0.18767515185051337,
0.11477650481479197, 0.05845050060280999, 0.16483912885523516,
0.07388719257091389, 0.0949023218872581, 0.07304370483803668,
0.07431040092920319, 0.10293411399811493, 0.07000825318247152,
0.13005425069050663, 0.05583762421572443, 0.057670780268775444,
0.09170488968526871, 0.17157346256588482, 0.09234238946569126,
0.3570653667073459, 0.062212257974845, 0.07433927956163146, 0.13903877551864616,
0.42268275330814864, 0.07655422992825989, 0.06201219804214242,
0.08767581488899101, 0.190234496887034, 0.09438211307192822,
0.10587301536609077, 0.19405064047267187, 0.09078457225317893,
0.07825954986417748, 0.05398855315749186, 0.09713013580552234,
0.07011884674992123, 0.07943723806171911, 0.09347129262600778,
0.08685913741208394, 0.05989800088129412, 0.0833377098018674,
0.12235671588262506, 0.07500482732722921, 0.09608186245910344,
0.11023292969084918, 0.08743227557705495, 0.25231239929530747,
0.11769388594049668, 0.06316378329658238, 0.11814623583122819,
0.11338723762692114, 0.07004530195938825, 0.18947790505376388,
0.08528443724355796, 0.09073295753943958, 0.06689776122425388,
0.05597386856426498, 0.35593381281867165, 0.12240479801067385,
0.08047767077013801, 0.08948655122440957, 0.06109887031701037,
0.06054730972331676, 0.05465169661700533, 0.11656517901891354,
0.06075757824238946, 0.2691349300578772, 0.05721555911121498,
0.06546203630858091, 0.1026807692496424, 0.1470536738235768,
0.08721507994155038, 0.11283913212525175, 0.08363125591520215,
0.4032224928211715, 0.05740431639709824, 0.06201468391510149,
0.05864147786807557, 0.1789225861318134, 0.10866453698299662,
0.15881331044030292, 0.05901469673049339, 0.08110703427329273,
0.3553914823408068, 0.1313978868529155, 0.05387573089071634,
0.08168533614709803, 0.06139614125079116, 0.06746968387828503,
0.05686815777431594, 0.09765769664776043, 0.09496103638027815,
0.06706946719811416, 0.07757662949138651, 0.05981734513276676,
0.06926410134868118, 0.09973075641922673, 0.07148311026017064,
0.5751670696041098]

3.1.2 Appending a column called Z-score to the dataset to store the Z-score of each row

```
[93]: df["Z-score"] = (df["Money difference"] - mean)/std
```

3.1.3 Making a column called Fraud that represents if we consider the data to be a fraudulent data or not depending on the Z-score. If the Z-score is above 3 then we consider it to be anomalous/fraudulent or if the Z-score is less then we consider it normal.

```
[94]: df['Fraud'] = df['Z-score'].apply(lambda x: 0 if x <= 3 else 1)
```

```
[95]: df.head()
```

```
[95]:
```

	Full name	Middle Initial of the Provider	\
0	SATYASREE UPADHYAYULA		NaN
1	WENDY JONES		P
2	RICHARD DUROCHER		W
3	JASPER FULLARD		NaN
4	ANTHONY PERROTTI		E

	Credentials of the Provider	Gender of the Provider	\
0	M.D.		F
1	M.D.		F
2	DPM		M
3	MD		M
4	DO		M

	Entity Type of the Provider	City of the Provider	Zip Code of the Provider	\
0	I	SAINT LOUIS	631041004.0	
1	I	FAYETTEVILLE	283043815.0	
2	I	NORTH HAVEN	64732343.0	
3	I	KANSAS CITY	641183998.0	
4	I	JUPITER	334585700.0	

	State Code of the Provider	Provider Type	\
0	MO	Internal Medicine	
1	NC	Obstetrics & Gynecology	
2	CT	Podiatry	
3	MO	Internal Medicine	
4	FL	Internal Medicine	

	Medicare Participation Indicator	... HCPCS Code	HCPCS Drug Indicator	\
0	Y	99223		N
1	Y	G0202		N
2	Y	99348		N
3	Y	81002		N
4	Y	96372		N

	Number of Services	Number of Medicare Beneficiaries	\
0	27		24
1	175		175

2	32	13
3	20	18
4	33	24

	Number of Distinct Medicare Beneficiary/Per Day Services \
0	27
1	175
2	32
3	20
4	31

	Average Submitted Charge Amount	Average Medicare Payment Amount \
0	0.004868	0.009787
1	0.008753	0.007395
2	0.002472	0.004010
3	0.000080	0.000213
4	0.000638	0.001216

	Money difference	Z-score	Fraud
0	0.002891	-0.139738	0
1	0.007834	0.165185	0
2	0.001885	-0.201787	0
3	0.000325	-0.298004	0
4	0.000656	-0.277579	0

[5 rows x 21 columns]

```
[96]: df['Fraud'].value_counts()
```

```
[96]: Fraud
0    98933
1     1067
Name: count, dtype: int64
```

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

```
[97]: Full name                        84197
Middle Initial of the Provider        29
Credentials of the Provider          1854
Gender of the Provider                 2
Entity Type of the Provider           2
City of the Provider                 5846
Zip Code of the Provider             47827
State Code of the Provider            58
Provider Type                        90
Medicare Participation Indicator       2
Place of Service                      2
```

HCPCS Code	2631
HCPCS Drug Indicator	2
Number of Services	2748
Number of Medicare Beneficiaries	1274
Number of Distinct Medicare Beneficiary/Per Day Services	1979
Average Submitted Charge Amount	38088
Average Medicare Payment Amount	83367
Money difference	92382
Z-score	91956
Fraud	2
dtype: int64	

```
[98]: df.drop(columns=['Number of Services','Number of Medicare_
↳Beneficiaries','Number of Distinct Medicare Beneficiary/Per Day_
↳Services','Zip Code of the Provider'], inplace=True)
```

```
[99]: df.head()
```

```
[99]:
```

	Full name	Middle Initial of the Provider	\
0	SATYASREE UPADHYAYULA		NaN
1	WENDY JONES		P
2	RICHARD DUROCHER		W
3	JASPER FULLARD		NaN
4	ANTHONY PERROTTI		E

	Credentials of the Provider	Gender of the Provider	\
0	M.D.		F
1	M.D.		F
2	DPM		M
3	MD		M
4	DO		M

	Entity Type of the Provider	City of the Provider	State	Code of the Provider	\
0	I	SAINT LOUIS			MO
1	I	FAYETTEVILLE			NC
2	I	NORTH HAVEN			CT
3	I	KANSAS CITY			MO
4	I	JUPITER			FL

	Provider Type	Medicare Participation Indicator	Place of Service	\
0	Internal Medicine	Y		F
1	Obstetrics & Gynecology	Y		0
2	Podiatry	Y		0
3	Internal Medicine	Y		0
4	Internal Medicine	Y		0

HCPCS Code	HCPCS Drug Indicator	Average Submitted Charge Amount	\
------------	----------------------	---------------------------------	---

0	99223	N	0.004868
1	G0202	N	0.008753
2	99348	N	0.002472
3	81002	N	0.000080
4	96372	N	0.000638

	Average Medicare Payment Amount	Money difference	Z-score	Fraud
0	0.009787	0.002891	-0.139738	0
1	0.007395	0.007834	0.165185	0
2	0.004010	0.001885	-0.201787	0
3	0.000213	0.000325	-0.298004	0
4	0.001216	0.000656	-0.277579	0

3.1.4 Applying One hot encoding the the columns with few categories

```
[100]: df = pd.get_dummies(df, columns=['Gender of the Provider', 'Entity Type of the
↳ Provider', 'Place of Service', 'Medicare Participation Indicator', 'HCPCS Drug
↳ Indicator'], dtype='int')
```

```
[101]: df.head()
```

```
[101]:      Full name Middle Initial of the Provider \
0  SATYASREE UPADHYAYULA      NaN
1      WENDY JONES      P
2  RICHARD DUROCHER      W
3      JASPER FULLARD      NaN
4  ANTHONY PERROTTI      E
```

	Credentials of the Provider	City of the Provider	State Code of the Provider
0	M.D.	SAINT LOUIS	MO
1	M.D.	FAYETTEVILLE	NC
2	DPM	NORTH HAVEN	CT
3	MD	KANSAS CITY	MO
4	DO	JUPITER	FL

	Provider Type	HCPCS Code	Average Submitted Charge Amount
0	Internal Medicine	99223	0.004868
1	Obstetrics & Gynecology	G0202	0.008753
2	Podiatry	99348	0.002472
3	Internal Medicine	81002	0.000080
4	Internal Medicine	96372	0.000638

	Average Medicare Payment Amount	Money difference	...
0	0.009787	0.002891	...
1	0.007395	0.007834	...
2	0.004010	0.001885	...
3	0.000213	0.000325	...

4	0.001216	0.000656	...
---	----------	----------	-----

	Gender of the Provider_F	Gender of the Provider_M	\
0	1	0	
1	1	0	
2	0	1	
3	0	1	
4	0	1	

	Entity Type of the Provider_I	Entity Type of the Provider_O	\
0	1	0	
1	1	0	
2	1	0	
3	1	0	
4	1	0	

	Place of Service_F	Place of Service_O	Medicare Participation Indicator_N	\
0	1	0	0	
1	0	1	0	
2	0	1	0	
3	0	1	0	
4	0	1	0	

	Medicare Participation Indicator_Y	HCPCS Drug Indicator_N	\
0	1	1	
1	1	1	
2	1	1	
3	1	1	
4	1	1	

	HCPCS Drug Indicator_Y
0	0
1	0
2	0
3	0
4	0

[5 rows x 22 columns]

[102]: df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 22 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Full name                            100000 non-null object
```

```

1 Middle Initial of the Provider      70669 non-null object
2 Credentials of the Provider         92791 non-null object
3 City of the Provider                100000 non-null object
4 State Code of the Provider          100000 non-null object
5 Provider Type                      100000 non-null object
6 HCPCS Code                         100000 non-null object
7 Average Submitted Charge Amount     100000 non-null float64
8 Average Medicare Payment Amount     100000 non-null float64
9 Money difference                    100000 non-null float64
10 Z-score                           100000 non-null float64
11 Fraud                             100000 non-null int64
12 Gender of the Provider_F           100000 non-null int64
13 Gender of the Provider_M           100000 non-null int64
14 Entity Type of the Provider_I      100000 non-null int64
15 Entity Type of the Provider_O      100000 non-null int64
16 Place of Service_F                 100000 non-null int64
17 Place of Service_O                 100000 non-null int64
18 Medicare Participation Indicator_N  100000 non-null int64
19 Medicare Participation Indicator_Y  100000 non-null int64
20 HCPCS Drug Indicator_N              100000 non-null int64
21 HCPCS Drug Indicator_Y              100000 non-null int64
dtypes: float64(4), int64(11), object(7)
memory usage: 16.8+ MB

```

```
[103]: df.drop(columns=['State Code of the Provider'], inplace=True)
```

3.1.5 Applying label encoding the fields with large no. of categories

```
[104]: from sklearn import preprocessing

label_encoder = preprocessing.LabelEncoder()

df['Provider Type'] = label_encoder.fit_transform(df['Provider Type'])
```

```
[105]: df.head()
```

```
[105]:
```

	Full name	Middle Initial of the Provider	\
0	SATYASREE UPADHYAYULA		NaN
1	WENDY JONES		P
2	RICHARD DUROCHER		W
3	JASPER FULLARD		NaN
4	ANTHONY PERROTTI		E

	Credentials of the Provider	City of the Provider	Provider Type	HCPCS Code	\
0	M.D.	SAINT LOUIS	39	99223	
1	M.D.	FAYETTEVILLE	54	G0202	
2	DPM	NORTH HAVEN	71	99348	

3	MD	KANSAS CITY	39	81002
4	DO	JUPITER	39	96372

	Average Submitted Charge Amount	Average Medicare Payment Amount \
0	0.004868	0.009787
1	0.008753	0.007395
2	0.002472	0.004010
3	0.000080	0.000213
4	0.000638	0.001216

	Money difference	Z-score	...	Gender of the Provider_F \
0	0.002891	-0.139738	...	1
1	0.007834	0.165185	...	1
2	0.001885	-0.201787	...	0
3	0.000325	-0.298004	...	0
4	0.000656	-0.277579	...	0

	Gender of the Provider_M	Entity Type of the Provider_I \
0	0	1
1	0	1
2	1	1
3	1	1
4	1	1

	Entity Type of the Provider_0	Place of Service_F	Place of Service_0 \
0	0	1	0
1	0	0	1
2	0	0	1
3	0	0	1
4	0	0	1

	Medicare Participation Indicator_N	Medicare Participation Indicator_Y \
0	0	1
1	0	1
2	0	1
3	0	1
4	0	1

	HCPCS Drug Indicator_N	HCPCS Drug Indicator_Y
0	1	0
1	1	0
2	1	0
3	1	0
4	1	0

[5 rows x 21 columns]

```
[106]: df.drop(columns=['City of the Provider'], inplace=True)
```

```
[107]: df.head()
```

```
[107]:
```

	Full name	Middle Initial of the Provider	\
0	SATYASREE UPADHYAYULA		NaN
1	WENDY JONES		P
2	RICHARD DUROCHER		W
3	JASPER FULLARD		NaN
4	ANTHONY PERROTTI		E

	Credentials of the Provider	Provider Type	HCPCS Code	\
0	M.D.	39	99223	
1	M.D.	54	G0202	
2	DPM	71	99348	
3	MD	39	81002	
4	DO	39	96372	

	Average Submitted Charge Amount	Average Medicare Payment Amount	\
0	0.004868	0.009787	
1	0.008753	0.007395	
2	0.002472	0.004010	
3	0.000080	0.000213	
4	0.000638	0.001216	

	Money difference	Z-score	Fraud	Gender of the Provider_F	\
0	0.002891	-0.139738	0	1	
1	0.007834	0.165185	0	1	
2	0.001885	-0.201787	0	0	
3	0.000325	-0.298004	0	0	
4	0.000656	-0.277579	0	0	

	Gender of the Provider_M	Entity Type of the Provider_I	\
0	0	1	
1	0	1	
2	1	1	
3	1	1	
4	1	1	

	Entity Type of the Provider_0	Place of Service_F	Place of Service_0	\
0	0	1	0	
1	0	0	1	
2	0	0	1	
3	0	0	1	
4	0	0	1	

	Medicare Participation Indicator_N	Medicare Participation Indicator_Y	\
--	------------------------------------	------------------------------------	---

0	0	1
1	0	1
2	0	1
3	0	1
4	0	1

	HCPCS Drug Indicator_N	HCPCS Drug Indicator_Y
0	1	0
1	1	0
2	1	0
3	1	0
4	1	0

```
[108]: df.drop(columns=['Full name', 'Middle Initial of the Provider', 'Credentials of the Provider'], inplace=True)
```

```
[109]: df.head()
```

```
[109]:
```

	Provider Type	HCPCS Code	Average Submitted Charge Amount \
0	39	99223	0.004868
1	54	G0202	0.008753
2	71	99348	0.002472
3	39	81002	0.000080
4	39	96372	0.000638

	Average Medicare Payment Amount	Money difference	Z-score	Fraud \
0	0.009787	0.002891	-0.139738	0
1	0.007395	0.007834	0.165185	0
2	0.004010	0.001885	-0.201787	0
3	0.000213	0.000325	-0.298004	0
4	0.001216	0.000656	-0.277579	0

	Gender of the Provider_F	Gender of the Provider_M \
0	1	0
1	1	0
2	0	1
3	0	1
4	0	1

	Entity Type of the Provider_I	Entity Type of the Provider_O \
0	1	0
1	1	0
2	1	0
3	1	0
4	1	0

	Place of Service_F	Place of Service_O	Medicare Participation Indicator_N \
--	--------------------	--------------------	--------------------------------------

0	1	0	0
1	0	1	0
2	0	1	0
3	0	1	0
4	0	1	0

	Medicare Participation Indicator_Y	HCPCS Drug Indicator_N \
0	1	1
1	1	1
2	1	1
3	1	1
4	1	1

	HCPCS Drug Indicator_Y
0	0
1	0
2	0
3	0
4	0

```
[110]: df['Fraud'].value_counts()
```

```
[110]: Fraud
0      98933
1       1067
Name: count, dtype: int64
```

3.1.6 Dropping the Fraud column to avoid bias in the model while training

```
[111]: df.drop(columns=['Fraud'], inplace=True)
```

```
[112]: df.drop(columns=['HCPCS Code'], inplace=True)
```

3.1.7 Applying pca to the dataset to find the first two principal components

```
[113]: from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
import matplotlib.pyplot as plt

scaler = StandardScaler()
X_std = scaler.fit_transform(df)

pca = PCA()
X_pca = pca.fit_transform(X_std)

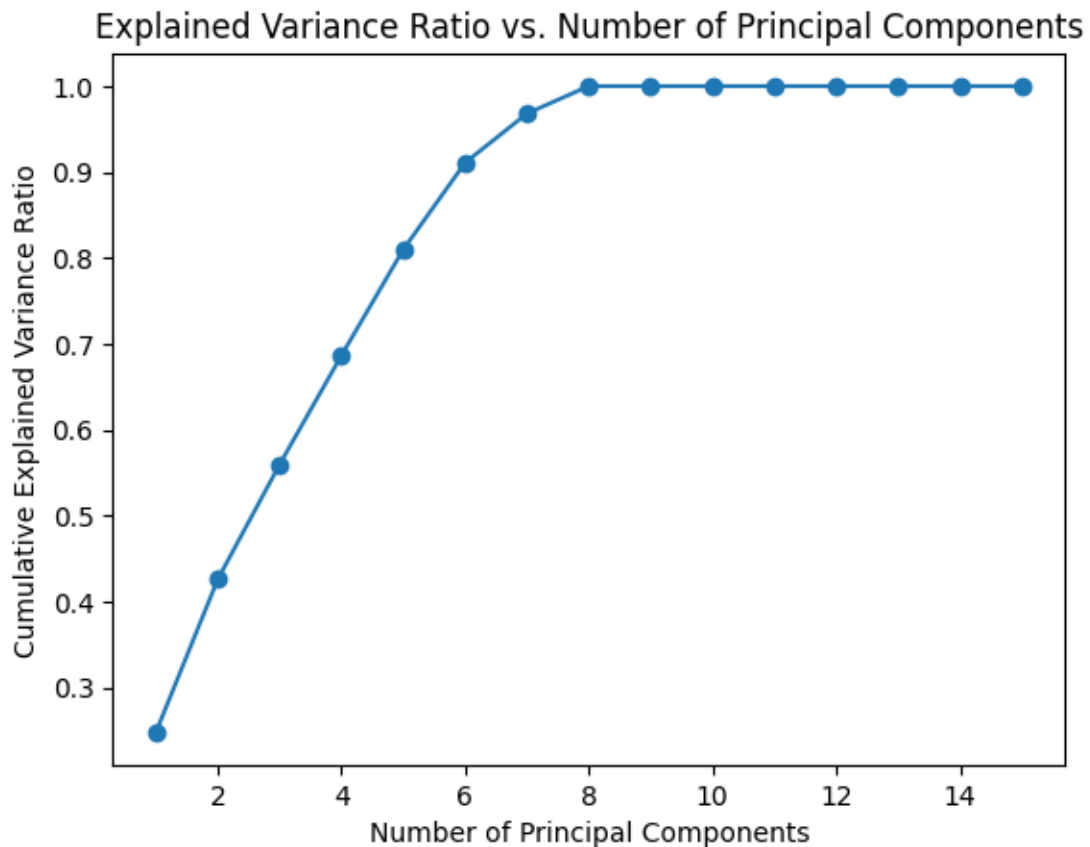
explained_var_ratio = pca.explained_variance_ratio_
```

```

cumulative_var_ratio = np.cumsum(explained_var_ratio)

plt.plot(range(1, len(cumulative_var_ratio) + 1), cumulative_var_ratio,
         marker='o')
plt.xlabel('Number of Principal Components')
plt.ylabel('Cumulative Explained Variance Ratio')
plt.title('Explained Variance Ratio vs. Number of Principal Components')
plt.show()

```



```
[114]: pca = PCA(n_components=8)
```

```
pca.fit(df)
```

```
X1=pca.fit_transform(df)
X1
```

```
[114]: array([[ -8.17503227e-01,  1.49308243e-03, -2.55628531e-01, ...,
                5.79848111e-02, -1.90651197e-04,  6.86814364e-03],
               [ 1.41905617e+01,  6.82606153e-02,  8.74523122e-01, ...,
                2.95552848e-02, -8.44388260e-04,  1.48919083e-03],
```



```
[ 3.11873179e+01, -1.65266953e-01, -2.53621286e-02, ...,
 -8.06726757e-02, -5.39170596e-04,  3.77080421e-04],
...,
 [-2.98201280e+01, -1.12108683e-01, -8.18641713e-01, ...,
  4.36599782e-02, -1.78205037e-04, -2.03053964e-03],
 [-8.07896826e-01, -4.08762517e-01,  8.80901136e-01, ...,
  4.18222114e-02, -9.66905755e-04, -6.08119637e-04],
 [ 2.81912661e+01, -3.12251292e-01,  7.03029656e-01, ...,
 -1.22970456e-02, -7.75550645e-04, -7.39638576e-04]])
```

```
[115]: scaler = StandardScaler()
scaled_data = scaler.fit_transform(df)

pca = PCA(n_components=2)
pca_fit = pca.fit(scaled_data)
u = pca.components_.T

pca_component = pd.DataFrame(u,
                              index=df.columns,
                              columns=['PC1', 'PC2']
                              )

print(pca_component)
```

	PC1	PC2
Provider Type	-0.074173	0.007580
Average Submitted Charge Amount	0.498565	0.092612
Average Medicare Payment Amount	0.385175	0.096741
Money difference	0.489624	0.085511
Z-score	0.489624	0.085511
Gender of the Provider_F	-0.070783	0.086619
Gender of the Provider_M	0.040798	-0.266392
Entity Type of the Provider_I	-0.064003	-0.427396
Entity Type of the Provider_O	0.064003	0.427396
Place of Service_F	0.207384	-0.311517
Place of Service_O	-0.207384	0.311517
Medicare Participation Indicator_N	-0.007491	0.001393
Medicare Participation Indicator_Y	0.007491	-0.001393
HCPCS Drug Indicator_N	0.092503	-0.405887
HCPCS Drug Indicator_Y	-0.092503	0.405887

3.1.8 Plotting the first two Pca components

```
[116]: scaler = StandardScaler()
scaled_data = scaler.fit_transform(df)

# Performing PCA
```

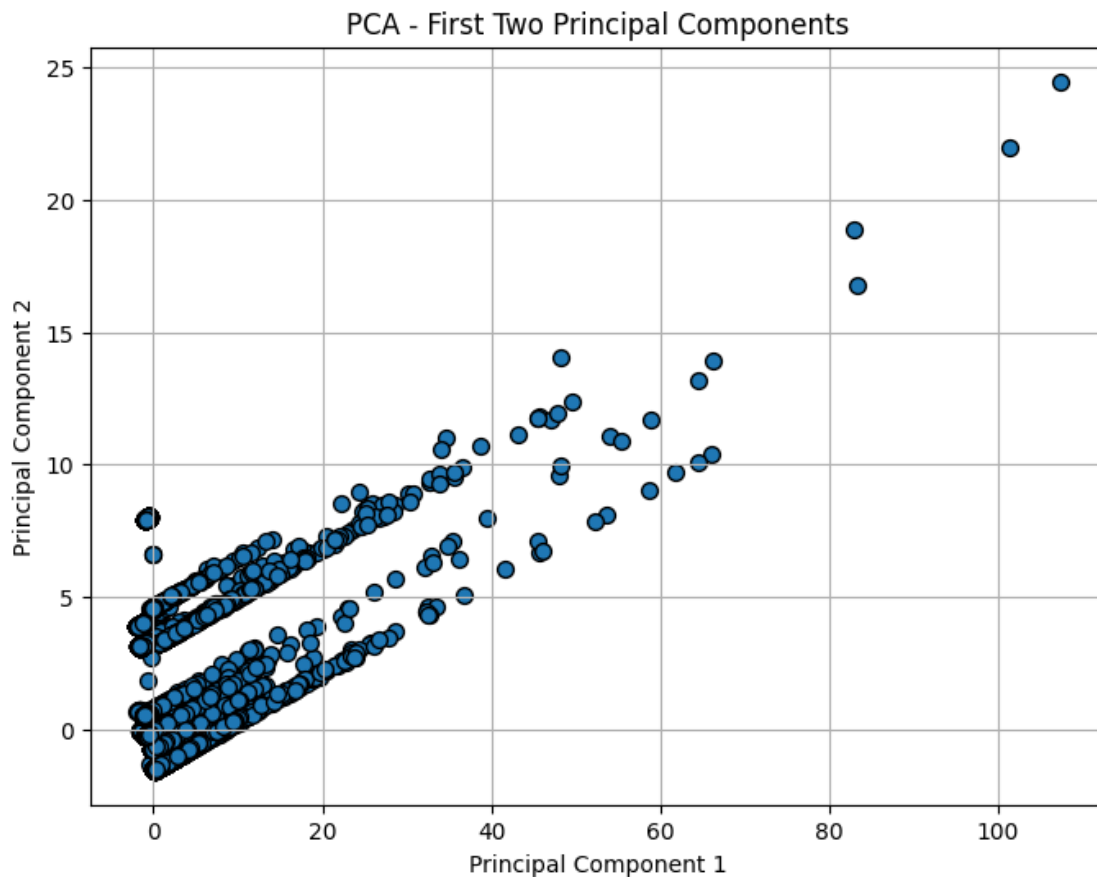
```

pca = PCA(n_components=2)
pca_transformed = pca.fit_transform(scaled_data)

# Creating a DataFrame for the transformed data
pca_df = pd.DataFrame(pca_transformed, columns=['PC1', 'PC2'])

# Scatter plot of the first two PCA components
plt.figure(figsize=(8, 6))
plt.scatter(pca_df['PC1'], pca_df['PC2'], edgecolor='k', s=50)
plt.title('PCA - First Two Principal Components')
plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.grid(True)
plt.show()

```



3.1.9 Using the elbow method to find the optimal no. of clusters in the k-means algorithm

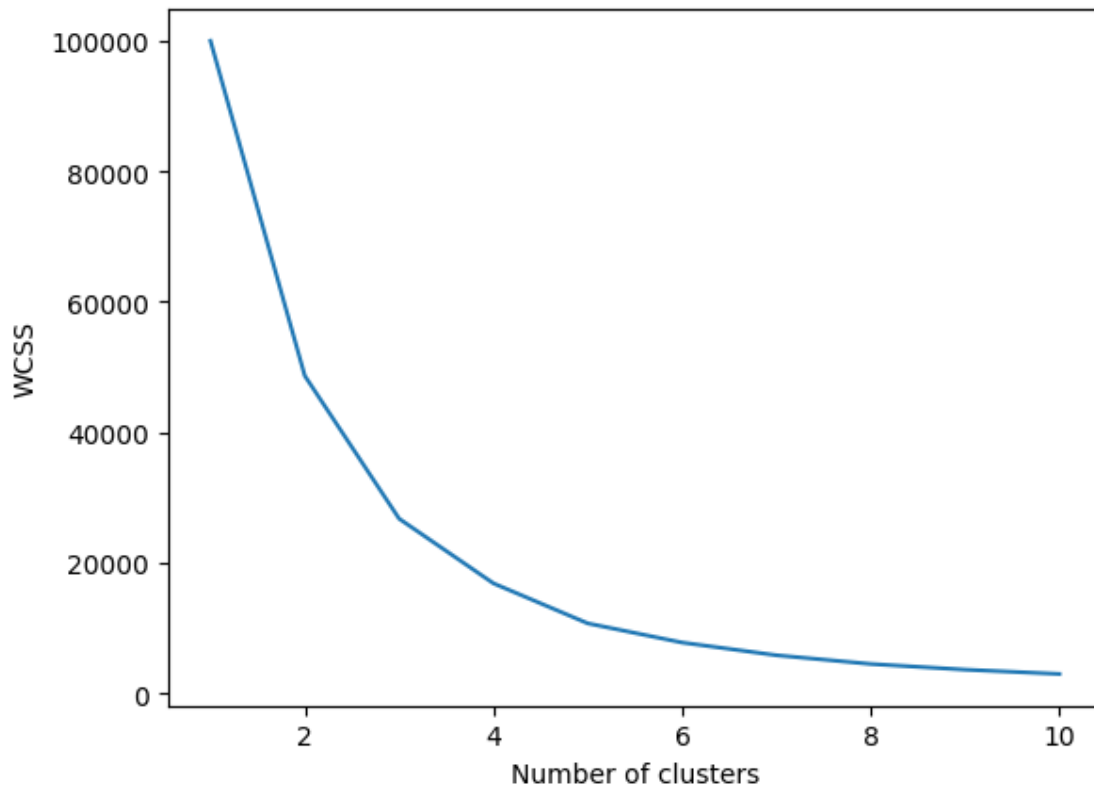
```
[131]: X = df.iloc[:, [2, 4]].values

from sklearn.cluster import KMeans
wcss = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters = i, init = 'k-means++', random_state = 42)
    kmeans.fit(X)
    wcss.append(kmeans.inertia_)

plt.plot(range(1, 11), wcss)
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870:  
FutureWarning: The default value of `n_init` will change from 10 to 'auto' in  
1.4. Set the value of `n_init` explicitly to suppress the warning  
    warnings.warn(  
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870:  
FutureWarning: The default value of `n_init` will change from 10 to 'auto' in  
1.4. Set the value of `n_init` explicitly to suppress the warning  
    warnings.warn(  
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870:  
FutureWarning: The default value of `n_init` will change from 10 to 'auto' in  
1.4. Set the value of `n_init` explicitly to suppress the warning  
    warnings.warn(  
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870:  
FutureWarning: The default value of `n_init` will change from 10 to 'auto' in  
1.4. Set the value of `n_init` explicitly to suppress the warning  
    warnings.warn(  
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870:  
FutureWarning: The default value of `n_init` will change from 10 to 'auto' in  
1.4. Set the value of `n_init` explicitly to suppress the warning  
    warnings.warn(  
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870:  
FutureWarning: The default value of `n_init` will change from 10 to 'auto' in  
1.4. Set the value of `n_init` explicitly to suppress the warning  
    warnings.warn(  
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870:  
FutureWarning: The default value of `n_init` will change from 10 to 'auto' in  
1.4. Set the value of `n init` explicitly to suppress the warning
```

```
warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870:
FutureWarning: The default value of `n_init` will change from 10 to 'auto' in
1.4. Set the value of `n_init` explicitly to suppress the warning
warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870:
FutureWarning: The default value of `n_init` will change from 10 to 'auto' in
1.4. Set the value of `n_init` explicitly to suppress the warning
warnings.warn(
```



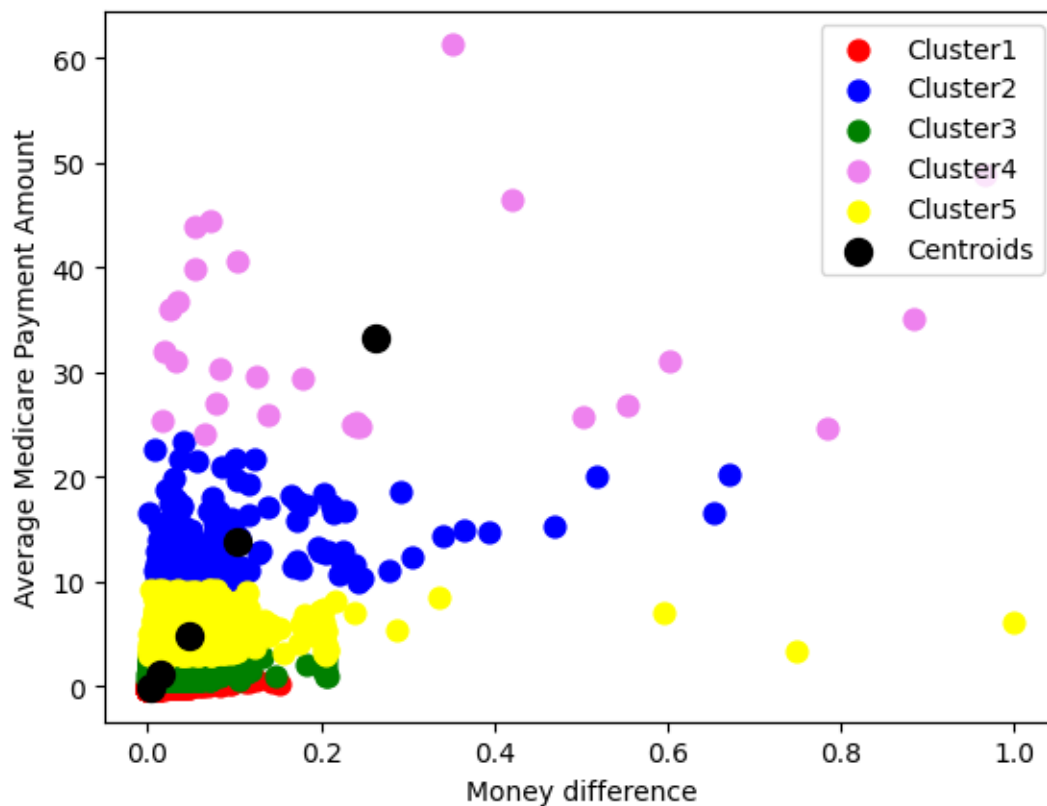
3.1.10 Applying K-means algorithm to the data and making a scatter plot to visualize the clusters

```
[132]: kmeans = KMeans(n_clusters = 5, init = "k-means++", random_state = 42)
y_kmeans = kmeans.fit_predict(X)
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870:
FutureWarning: The default value of `n_init` will change from 10 to 'auto' in
1.4. Set the value of `n_init` explicitly to suppress the warning
warnings.warn(
```

```
[133]: plt.scatter(X[y_kmeans == 0, 0], X[y_kmeans == 0, 1], s = 60, c = 'red', label_
        ↪= 'Cluster1')
plt.scatter(X[y_kmeans == 1, 0], X[y_kmeans == 1, 1], s = 60, c = 'blue', label_
        ↪= 'Cluster2')
plt.scatter(X[y_kmeans == 2, 0], X[y_kmeans == 2, 1], s = 60, c = 'green',
        ↪label = 'Cluster3')
plt.scatter(X[y_kmeans == 3, 0], X[y_kmeans == 3, 1], s = 60, c = 'violet',
        ↪label = 'Cluster4')
plt.scatter(X[y_kmeans == 4, 0], X[y_kmeans == 4, 1], s = 60, c = 'yellow',
        ↪label = 'Cluster5')
plt.scatter(kmeans.cluster_centers[:, 0], kmeans.cluster_centers[:, 1], s =
        ↪100, c = 'black', label = 'Centroids')
plt.xlabel('Money difference')
plt.ylabel('Average Medicare Payment Amount')
plt.legend()

plt.show()
```

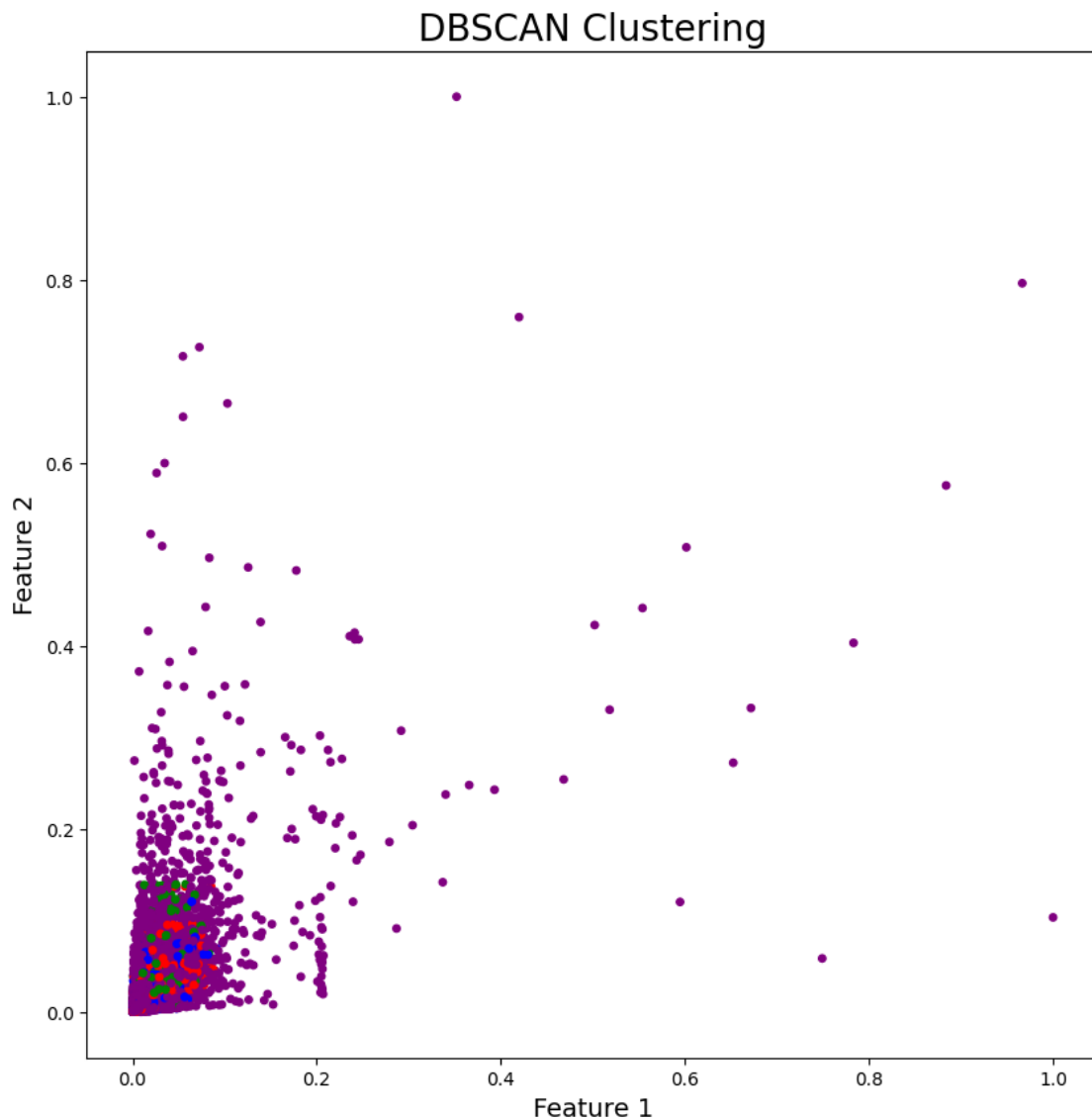


3.1.11 Applying Dbscan to the data and making a scatter plot to see the clusters

```
[122]: from sklearn.cluster import DBSCAN
```

```
[123]: db_default = DBSCAN(eps = 0.0375, min_samples = 3).fit(df)
labels = db_default.labels_
```

```
[124]: colors=['purple','red','blue','green']
plt.figure(figsize=(10,10))
plt.scatter(df["Average Medicare Payment Amount"],df["Money_␣
↵difference"],c=labels,cmap=matplotlib.colors.ListedColormap(colors),s=15)
plt.title('DBSCAN Clustering',fontsize=20)
plt.xlabel('Feature 1',fontsize=14)
plt.ylabel('Feature 2',fontsize=14)
plt.show()
```



[124] :