```
In [75]: print ("kernel is working")
```

```
kernel is working
```

```
In [76]: !pip install pandas
```

```
Requirement already satisfied: pandas in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (2.3.3)
Requirement already satisfied: tzdata>=2022.7 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from pandas) (2025.3)
Requirement already satisfied: numpy>=1.22.4 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from pandas) (2.2.6)
Requirement already satisfied: pytz>=2020.1 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from pandas) (2025.2)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from pandas) (2.9.0.post0)
Requirement already satisfied: six>=1.5 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from python-dateutil>=2.8.2->pandas) (1.17.0)
[notice] A new release of pip is available: 23.0.1 -> 25.3
[notice] To update, run: python.exe -m pip install --upgrade pip
```

```
In [77]: import pandas as pd
```

```
In [78]: import zipfile
```

```
In [79]: zip_path = r"C:\Users\VENKATESAN\Downloads\Healthcare Providers.csv (1).zip"
```

```
In [80]: df = pd.read_csv(zip_path)
```

```
In [81]: df.head()
```

Out[81]:

	index	National Provider Identifier	Last Name/Organization Name of the Provider	First Name of the Provider	Middle Initial of the Provider	Credentials of the Provider	Gender of the Provider	Entity Type of the Provider	Street Address 1 of the Provider	Street Address 2 of the Provider	...	HCPCS Code	HCPCS Description	HCPCS Drug Indicator	Number of Services	Number of Medicare Beneficiaries	Number of Distinct Medicare Beneficiary/Per Day Services	Average Medicare Allowed Amount	Aver Submi Cha Amo
0	8774979	1891106191	UPADHYAYULA	SATYASREE	NaN	M.D.	F	I	1402 S GRAND BLVD	FDT 14TH FLOOR	...	99223	Initial hospital inpatient care, typically 70 ...	N	27	24	27	200.58777778	305.2111
1	3354385	1346202256	JONES	WENDY	P	M.D.	F	I	2950 VILLAGE DR	NaN	...	G0202	Screening mammography, bilateral (2-view study...	N	175	175	175	123.73	5
2	3001884	1306820956	DUROCHER	RICHARD	W	DPM	M	I	20 WASHINGTON AVE	STE 212	...	99348	Established patient home visit, typically 25 m...	N	32	13	32	90.65	
3	7594822	1770523540	FULLARD	JASPER	NaN	MD	M	I	5746 N BROADWAY ST	NaN	...	81002	Urinalysis, manual test	N	20	18	20	3.5	
4	746159	1073627758	PERROTTI	ANTHONY	E	DO	M	I	875 MILITARY TRL	SUITE 200	...	96372	Injection beneath the skin or into muscle for ...	N	33	24	31	26.52	

5 rows × 27 columns

```
In [82]: df.info()
         df.shape
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 27 columns):
 #   Column                                                    Non-Null Count   Dtype
---  ------                                                    --------------   -----
 0   index                                                     100000 non-null  int64
 1   National Provider Identifier                              100000 non-null  int64
 2   Last Name/Organization Name of the Provider               100000 non-null  object
 3   First Name of the Provider                                95745 non-null   object
 4   Middle Initial of the Provider                            70669 non-null   object
 5   Credentials of the Provider                               92791 non-null   object
 6   Gender of the Provider                                    95746 non-null   object
 7   Entity Type of the Provider                               100000 non-null  object
 8   Street Address 1 of the Provider                          100000 non-null  object
 9   Street Address 2 of the Provider                          40637 non-null   object
 10  City of the Provider                                      100000 non-null  object
 11  Zip Code of the Provider                                  100000 non-null  float64
 12  State Code of the Provider                                100000 non-null  object
 13  Country Code of the Provider                              100000 non-null  object
 14  Provider Type                                             100000 non-null  object
 15  Medicare Participation Indicator                          100000 non-null  object
 16  Place of Service                                          100000 non-null  object
 17  HCPCS Code                                                100000 non-null  object
 18  HCPCS Description                                         100000 non-null  object
 19  HCPCS Drug Indicator                                      100000 non-null  object
 20  Number of Services                                        100000 non-null  object
 21  Number of Medicare Beneficiaries                          100000 non-null  object
 22  Number of Distinct Medicare Beneficiary/Per Day Services  100000 non-null  object
 23  Average Medicare Allowed Amount                           100000 non-null  object
 24  Average Submitted Charge Amount                           100000 non-null  object
 25  Average Medicare Payment Amount                           100000 non-null  object
 26  Average Medicare Standardized Amount                      100000 non-null  object
dtypes: float64(1), int64(2), object(24)
memory usage: 20.6+ MB
```

```
Out[82]: (100000, 27)
```

```
In [83]: df.columns
```

```
Out[83]: Index(['index', 'National Provider Identifier',
       'Last Name/Organization Name of the Provider',
       'First Name of the Provider', 'Middle Initial of the Provider',
       'Credentials of the Provider', 'Gender of the Provider',
       'Entity Type of the Provider', 'Street Address 1 of the Provider',
       'Street Address 2 of the Provider', 'City of the Provider',
       'Zip Code of the Provider', 'State Code of the Provider',
       'Country Code of the Provider', 'Provider Type',
       'Medicare Participation Indicator', 'Place of Service', 'HCPCS Code',
       'HCPCS Description', 'HCPCS Drug Indicator', 'Number of Services',
       'Number of Medicare Beneficiaries',
       'Number of Distinct Medicare Beneficiary/Per Day Services',
       'Average Medicare Allowed Amount', 'Average Submitted Charge Amount',
       'Average Medicare Payment Amount',
       'Average Medicare Standardized Amount'],
      dtype='object')
```

```
In [84]: df.duplicated().sum()
```

```
Out[84]: np.int64(0)
```

```
In [85]: df.isnull().sum()
```

```
Out[85]: index                                                         0
National Provider Identifier                                  0
Last Name/Organization Name of the Provider                   0
First Name of the Provider                                 4255
Middle Initial of the Provider                            29331
Credentials of the Provider                                7209
Gender of the Provider                                     4254
Entity Type of the Provider                                   0
Street Address 1 of the Provider                              0
Street Address 2 of the Provider                          59363
City of the Provider                                          0
Zip Code of the Provider                                      0
State Code of the Provider                                    0
Country Code of the Provider                                  0
Provider Type                                                 0
Medicare Participation Indicator                              0
Place of Service                                              0
HCPCS Code                                                    0
HCPCS Description                                             0
HCPCS Drug Indicator                                          0
Number of Services                                            0
Number of Medicare Beneficiaries                              0
Number of Distinct Medicare Beneficiary/Per Day Services      0
Average Medicare Allowed Amount                               0
Average Submitted Charge Amount                               0
Average Medicare Payment Amount                               0
Average Medicare Standardized Amount                          0
dtype: int64
```

```
In [86]: df['Gender of the Provider'] = df['Gender of the Provider'].fillna('Unknown')
```

```
In [87]: df.dtypes
```

```
Out[87]: index                                                      int64
National Provider Identifier                               int64
Last Name/Organization Name of the Provider               object
First Name of the Provider                                object
Middle Initial of the Provider                            object
Credentials of the Provider                               object
Gender of the Provider                                    object
Entity Type of the Provider                               object
Street Address 1 of the Provider                          object
Street Address 2 of the Provider                          object
City of the Provider                                      object
Zip Code of the Provider                                 float64
State Code of the Provider                                object
Country Code of the Provider                              object
Provider Type                                             object
Medicare Participation Indicator                          object
Place of Service                                          object
HCPCS Code                                                object
HCPCS Description                                         object
HCPCS Drug Indicator                                      object
Number of Services                                        object
Number of Medicare Beneficiaries                          object
Number of Distinct Medicare Beneficiary/Per Day Services  object
Average Medicare Allowed Amount                           object
Average Submitted Charge Amount                           object
Average Medicare Payment Amount                           object
Average Medicare Standardized Amount                      object
dtype: object
```

```
In [ ]:
```

```
In [88]: df['Average Submitted Charge Amount'] = pd.to_numeric(
    df['Average Submitted Charge Amount'].astype(str).str.replace(',', ''),
    errors='coerce'
)

df['Average Medicare Allowed Amount'] = pd.to_numeric(
    df['Average Medicare Allowed Amount'].astype(str).str.replace(',', ''),
    errors='coerce'
)
```

```
In [89]: df['Average Submitted Charge Amount'].head(10)
```

```
Out[89]: 0    305.211111
1    548.800000
2    155.000000
3      5.000000
4     40.000000
5     63.112500
6    164.571429
7     50.000000
8     76.000000
9     29.650000
Name: Average Submitted Charge Amount, dtype: float64
```

```
In [90]: df['Average Submitted Charge Amount'] = (
    df['Average Submitted Charge Amount']
    .astype(str)                         # convert everything to string
    .str.replace(',', '', regex=True)    # remove commas
    .str.strip()                         # remove spaces
)

df['Average Submitted Charge Amount'] = pd.to_numeric(
    df['Average Submitted Charge Amount'],
    errors='coerce'
)
```

```
In [91]: df['Average Medicare Standardized Amount'] = (
    df['Average Medicare Standardized Amount']
    .astype(str)
    .str.replace(',', '', regex=True)
    .str.strip()
)

df['Average Medicare Standardized Amount'] = pd.to_numeric(
    df['Average Medicare Standardized Amount'],
    errors='coerce'
)
```

```
In [92]: df['Average Medicare Standardized Amount'] = (
    df['Average Medicare Standardized Amount'].fillna(0)
)
```

```
In [93]: df['Average Medicare Standardized Amount'].dtype
```

```
Out[93]: dtype('float64')
```

```
In [94]: df[['Average Submitted Charge Amount',
    'Average Medicare Allowed Amount']] = df[[
    'Average Submitted Charge Amount',
    'Average Medicare Allowed Amount'
]].fillna(0)
```

```
In [95]: pip install numpy
```

```
Requirement already satisfied: numpy in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (2.2.6)
Note: you may need to restart the kernel to use updated packages.
[notice] A new release of pip is available: 23.0.1 -> 25.3
[notice] To update, run: python.exe -m pip install --upgrade pip
```

```
In [96]: import numpy as np
print(np.__version__)
```

```
2.2.6
```

```
In [97]: df['Std_Charge_Ratio'] = (
    df['Average Submitted Charge Amount'] /
    df['Average Medicare Standardized Amount']
)
```

```
In [98]: df['Std_Charge_Ratio'] = df['Std_Charge_Ratio'].replace([np.inf, -np.inf], 0)
```

```
In [99]: df['Fraud_Flag_Std'] = np.where(
    df['Std_Charge_Ratio'] >= 2,
    1,
    0
)
```

```
In [100… df['Fraud_Flag_Std'].value_counts()
```

```
Out[100… Fraud_Flag_Std
1    81045
0    18955
Name: count, dtype: int64
```

```
In [101… !pip install matplotlib
```

```
Requirement already satisfied: matplotlib in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (3.10.8)
Requirement already satisfied: numpy>=1.23 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from matplotlib) (2.2.6)
Requirement already satisfied: cycler>=0.10 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from matplotlib) (0.12.1)
Requirement already satisfied: pillow>=8 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from matplotlib) (12.1.0)
Requirement already satisfied: pyparsing>=3 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from matplotlib) (3.3.2)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from matplotlib) (2.9.0.post0)
Requirement already satisfied: packaging>=20.0 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from matplotlib) (26.0)
Requirement already satisfied: kiwisolver>=1.3.1 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from matplotlib) (1.4.9)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from matplotlib) (1.3.2)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from matplotlib) (4.61.1)
Requirement already satisfied: six>=1.5 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from python-dateutil>=2.7->matplotlib) (1.17.0)
[notice] A new release of pip is available: 23.0.1 -> 25.3
[notice] To update, run: python.exe -m pip install --upgrade pip
```

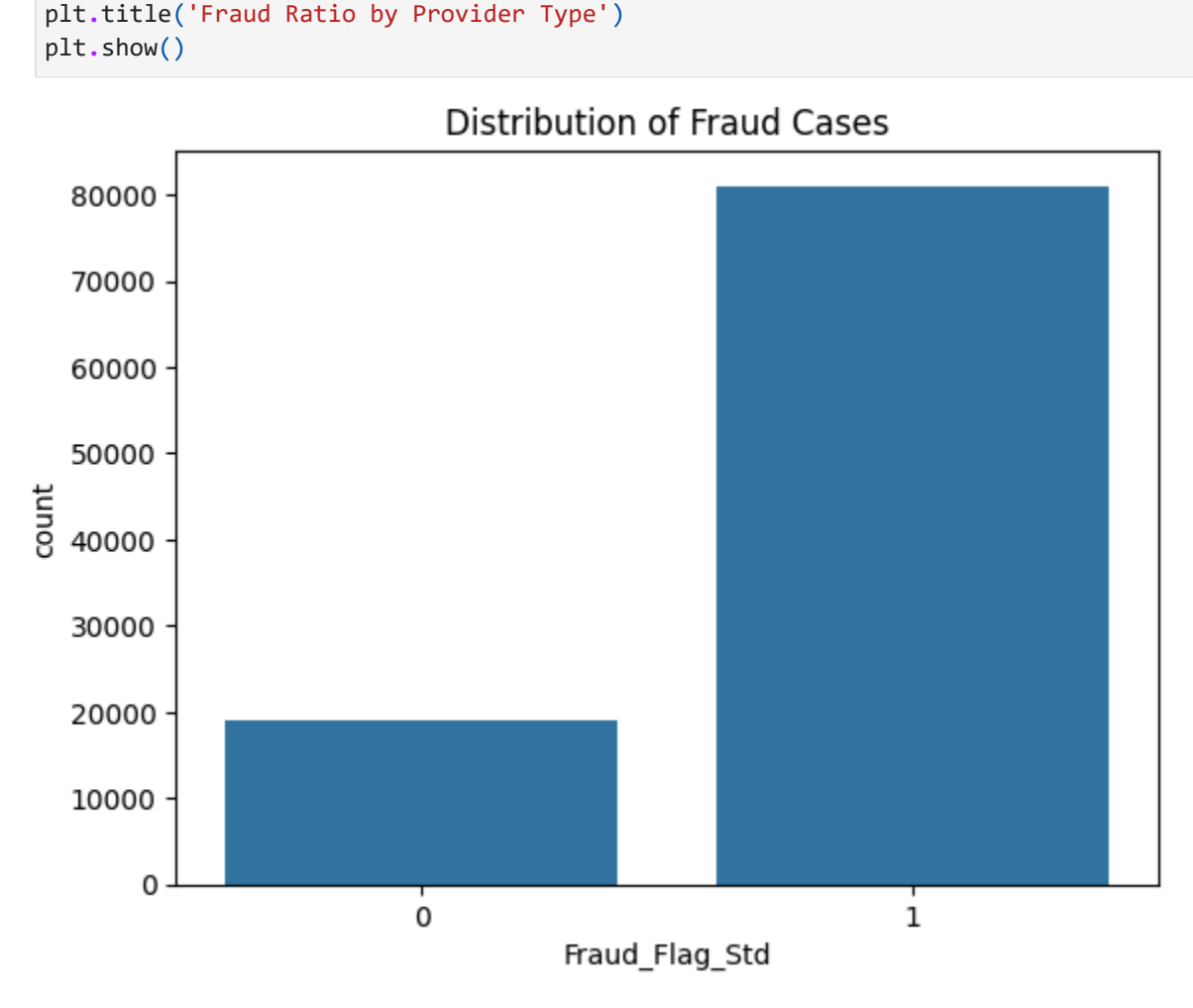
```
In [102… !pip install seaborn
```

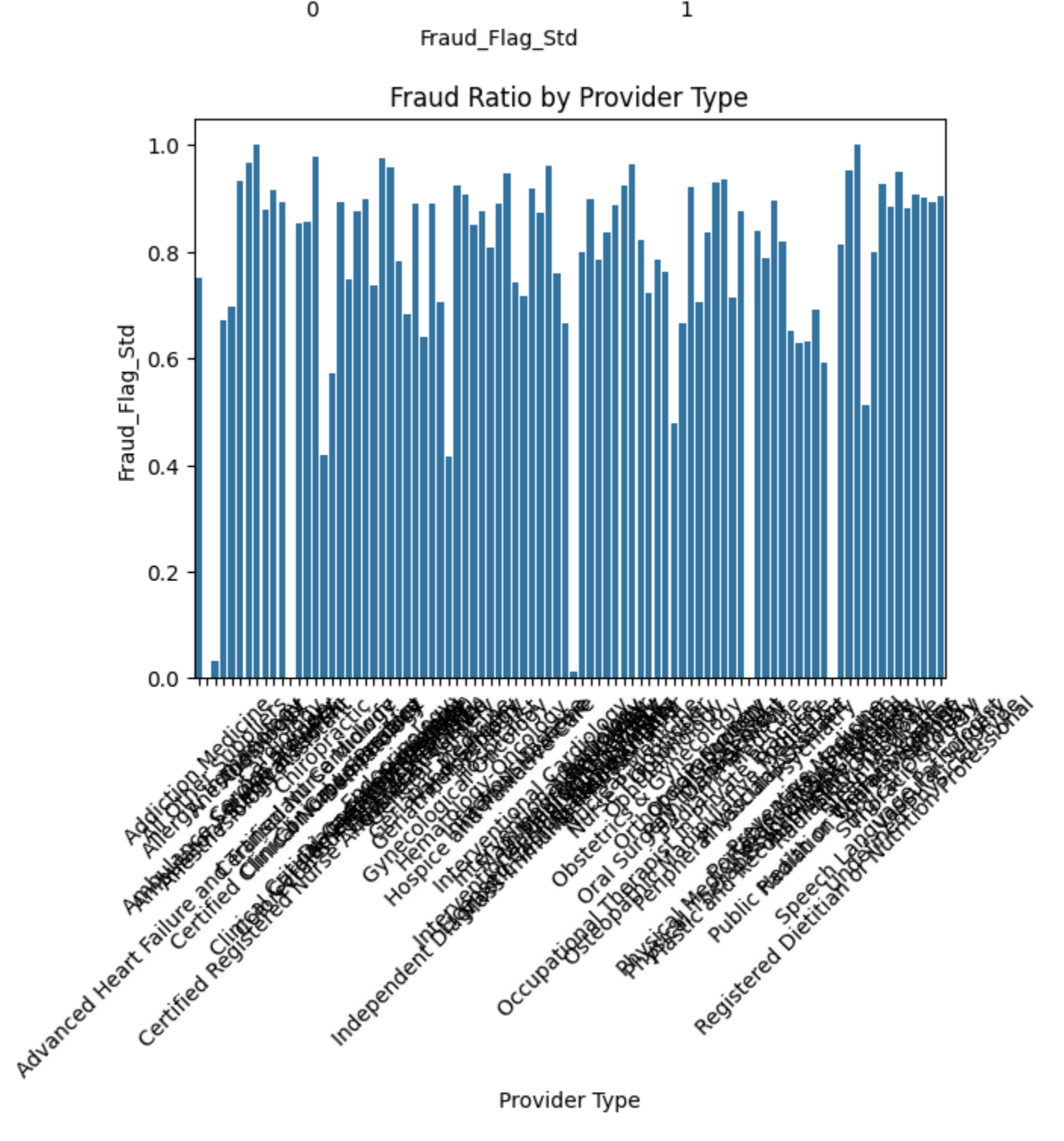
```
Requirement already satisfied: seaborn in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (0.13.2)
Requirement already satisfied: matplotlib!=3.6.1,>=3.4 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from seaborn) (3.10.8)
Requirement already satisfied: numpy!=1.24.0,>=1.20 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from seaborn) (2.2.6)
Requirement already satisfied: pandas>=1.2 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from seaborn) (2.3.3)
Requirement already satisfied: kiwisolver>=1.3.1 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (1.4.9)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (1.3.2)
Requirement already satisfied: packaging>=20.0 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (26.0)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (4.61.1)
Requirement already satisfied: cycler>=0.10 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (0.12.1)
Requirement already satisfied: pillow>=8 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (12.1.0)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (2.9.0.post0)
Requirement already satisfied: pyparsing>=3 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (3.3.2)
Requirement already satisfied: tzdata>=2022.7 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from pandas>=1.2->seaborn) (2025.3)
Requirement already satisfied: pytz>=2020.1 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from pandas>=1.2->seaborn) (2025.2)
Requirement already satisfied: six>=1.5 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from python-dateutil>=2.7->matplotlib!=3.6.1,>=3.4->seaborn) (1.17.0)
[notice] A new release of pip is available: 23.0.1 -> 25.3
[notice] To update, run: python.exe -m pip install --upgrade pip
```

```
In [103… import matplotlib.pyplot as plt
import seaborn as sns

# Fraud distribution
sns.countplot(x='Fraud_Flag_Std', data=df)
plt.title('Distribution of Fraud Cases')
plt.show()

# Check ratio by provider type
sns.barplot(
    x='Provider Type',
    y='Fraud_Flag_Std',
    data=df.groupby('Provider Type')['Fraud_Flag_Std'].mean().reset_index()
)
plt.xticks(rotation=45)
plt.title('Fraud Ratio by Provider Type')
plt.show()
```






```
In [104… fraud_gender = df.groupby('Gender of the Provider')['Fraud_Flag_Std'].mean()
print(fraud_gender)
```

```
Gender of the Provider
F          0.812369
M          0.831290
Unknown    0.470851
Name: Fraud_Flag_Std, dtype: float64
```

```
In [105… numeric_cols = [
    'Number of Services', 'Number of Medicare Beneficiaries',
    'Number of Distinct Medicare Beneficiary/Per Day Services',
    'Average Medicare Payment Amount'
]

for col in numeric_cols:
    df[col] = pd.to_numeric(df[col].astype(str).str.replace(',', ''), errors='coerce').fillna(0)
```

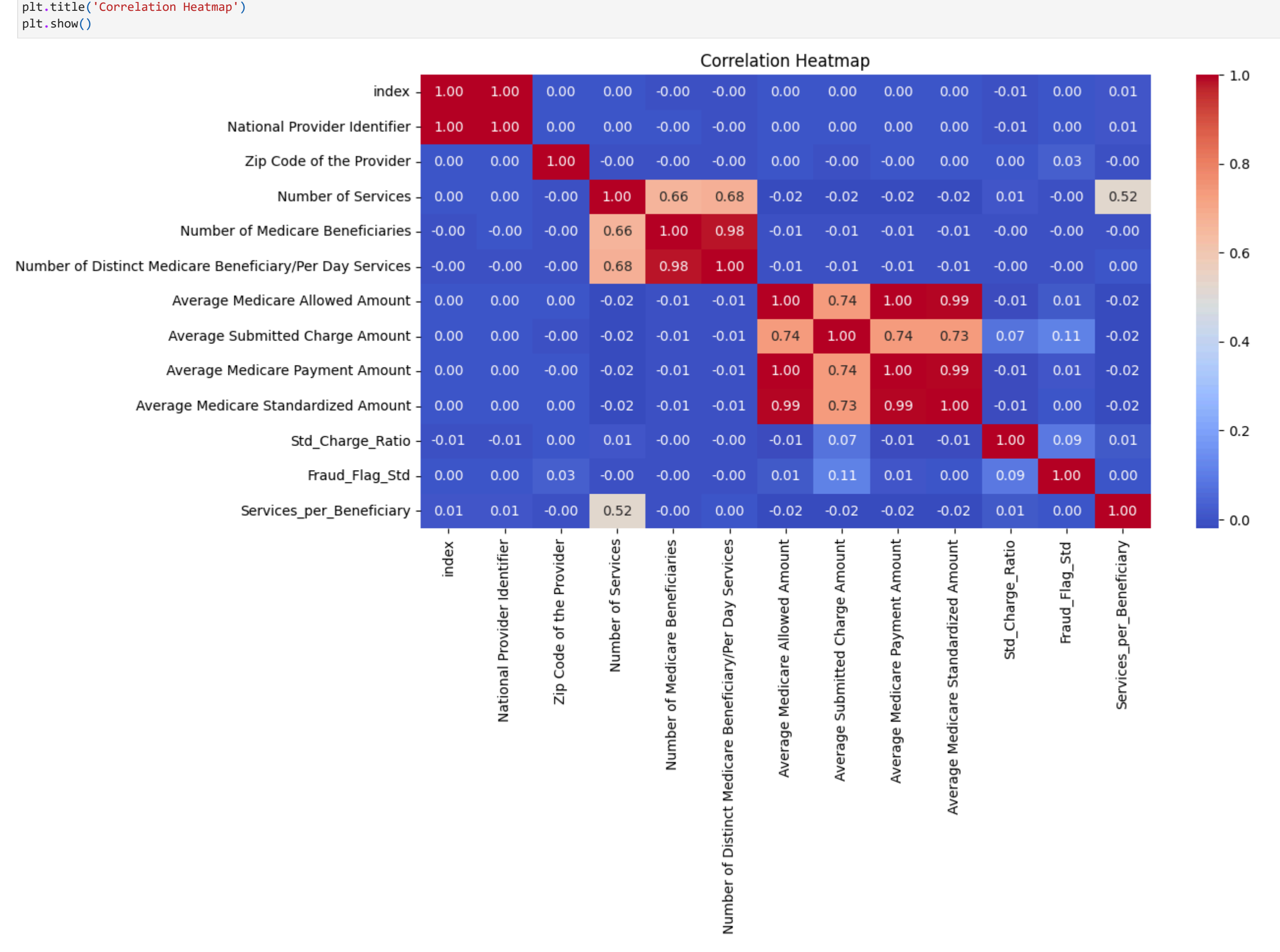
```
In [106… df['Services_per_Beneficiary'] = df['Number of Services'] / df['Number of Medicare Beneficiaries']
```

```
In [107… df['Services_per_Beneficiary'] = df['Number of Services'] / df['Number of Medicare Beneficiaries']
```

```
In [108… import matplotlib.pyplot as plt
import seaborn as sns

# Select only numeric columns
numeric_df = df.select_dtypes(include=['int64', 'float64'])

plt.figure(figsize=(12,6))
sns.heatmap(numeric_df.corr(), annot=True, fmt='.2f', cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```




```
In [109… !pip install scikit-learn
```

```
Requirement already satisfied: scikit-learn in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (1.7.2)
Requirement already satisfied: numpy>=1.22.0 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from scikit-learn) (2.2.6)
Requirement already satisfied: scipy>=1.8.0 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from scikit-learn) (1.15.3)
Requirement already satisfied: joblib>=1.2.0 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from scikit-learn) (1.5.3)
Requirement already satisfied: threadpoolctl>=3.1.0 in c:\users\venkatesan\appdata\local\programs\python\python310\lib\site-packages (from scikit-learn) (3.6.0)
[notice] A new release of pip is available: 23.0.1 -> 25.3
[notice] To update, run: python.exe -m pip install --upgrade pip
```

```
In [110… from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report
```

```
In [111… from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report

features = ['Average Submitted Charge Amount', 'Average Medicare Standardized Amount', 'Number of Services']
X = df[features]
y = df['Fraud_Flag_Std']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)
y_pred = model.predict(X_test)

print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.99      0.97      0.98      3762
           1       0.99      1.00      1.00     16238

    accuracy                           0.99     20000
   macro avg       0.99      0.98      0.99     20000
weighted avg       0.99      0.99      0.99     20000
```

```
In [ ]:
```