om-eda-2

June 7, 2024

EXPLORATORY DATA ANALYSIS

[4]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
data = pd.read_csv("/content/Healthcare Providers.csv")
data.head()
```

[4]:
```
     index  National Provider Identifier  \
0  8774979                    1891106191
1  3354385                    1346202256
2  3001884                    1306820956
3  7594822                    1770523540
4   746159                    1073627758

  Last Name/Organization Name of the Provider First Name of the Provider  \
0                                 UPADHYAYULA                  SATYASREE
1                                       JONES                      WENDY
2                                    DUROCHER                    RICHARD
3                                     FULLARD                     JASPER
4                                    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  \
0                      F                           I
1                      F                           I
2                      M                           I
3                      M                           I
4                      M                           I
```

```
  Street Address 1 of the Provider Street Address 2 of the Provider  …  \
0                1402 S GRAND BLVD                   FDT 14TH FLOOR  …
1                  2950 VILLAGE DR                              NaN  …
2                20 WASHINGTON AVE                          STE 212  …
3               5746 N BROADWAY ST                              NaN  …
4                 875 MILITARY TRL                        SUITE 200  …

  HCPCS Code                                  HCPCS Description  \
0      99223  Initial hospital inpatient care, typically 70 …
1      G0202  Screening mammography, bilateral (2-view study…
2      99348  Established patient home visit, typically 25 m…
3      81002                            Urinalysis, manual test
4      96372  Injection beneath the skin or into muscle for …

  HCPCS 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 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

[5 rows x 27 columns]
```

[5]:

```
# Descriptive statistics
data.describe()
```

```
[5]:                index  National Provider Identifier  Zip Code of the Provider
     count  2.244500e+04                  2.244500e+04              2.244500e+04
     mean   4.910642e+06                  1.498512e+09              4.188710e+08
     std    2.841408e+06                  2.875684e+08              3.071427e+08
     min    3.900000e+02                  1.003002e+09              6.030000e+02
     25%    2.447611e+06                  1.245521e+09              1.521325e+08
     50%    4.914401e+06                  1.497926e+09              3.706759e+08
     75%    7.349263e+06                  1.740373e+09              6.850629e+08
     max    9.847437e+06                  1.993000e+09              9.970939e+08
```

```
<google.colab._quickchart_helpers.SectionTitle at 0x7b414798e350>
```

```
from matplotlib import pyplot as plt
_df_0['index'].plot(kind='hist', bins=20, title='index')
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_1['National Provider Identifier'].plot(kind='hist', bins=20, title='National␣
 ↪Provider Identifier')
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_2['Zip Code of the Provider'].plot(kind='hist', bins=20, title='Zip Code of␣
 ↪the Provider')
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
<google.colab._quickchart_helpers.SectionTitle at 0x7b417dc1b310>
```

```
from matplotlib import pyplot as plt
_df_3.plot(kind='scatter', x='index', y='National Provider Identifier', s=32,␣
 ↪alpha=.8)
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_4.plot(kind='scatter', x='National Provider Identifier', y='Zip Code of the␣
 ↪Provider', s=32, alpha=.8)
plt.gca().spines[['top', 'right',]].set_visible(False)
```

```
<google.colab._quickchart_helpers.SectionTitle at 0x7b414798dde0>
```

```
from matplotlib import pyplot as plt
_df_5['index'].plot(kind='line', figsize=(8, 4), title='index')
plt.gca().spines[['top', 'right']].set_visible(False)
```

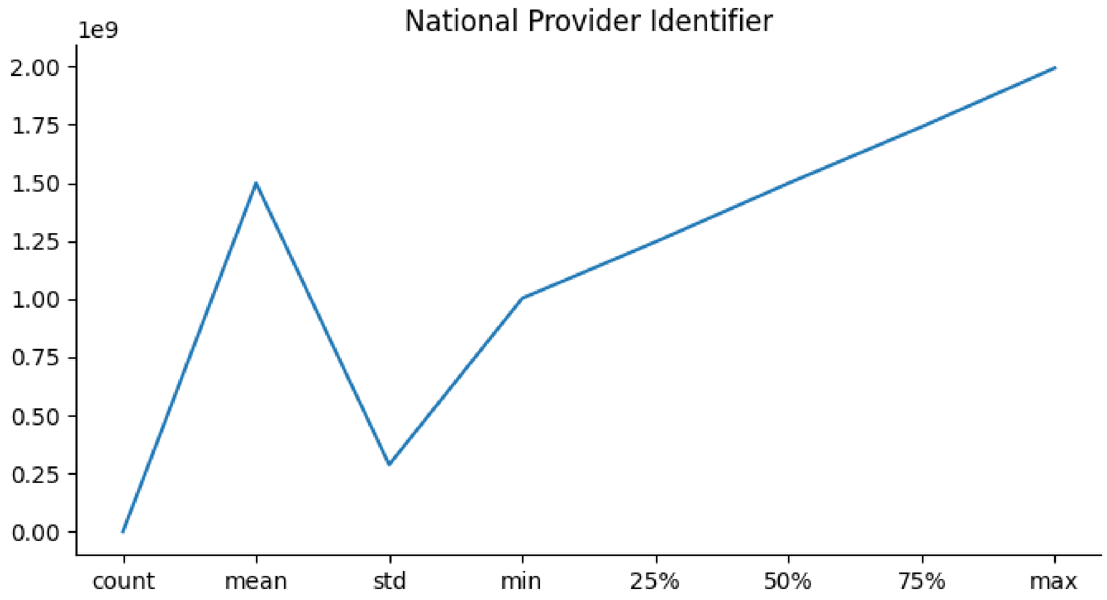
```
from matplotlib import pyplot as plt
_df_6['National Provider Identifier'].plot(kind='line', figsize=(8, 4),␣
 ↪title='National Provider Identifier')
plt.gca().spines[['top', 'right']].set_visible(False)
```

```
from matplotlib import pyplot as plt
_df_7['Zip Code of the Provider'].plot(kind='line', figsize=(8, 4), title='Zip␣
 ↪Code of the Provider')
```

```
plt.gca().spines[['top', 'right']].set_visible(False)
```

[27]:

```
from matplotlib import pyplot as plt
_df_6['National Provider Identifier'].plot(kind='line', figsize=(8, 4),␣
 ↪title='National Provider Identifier')
plt.gca().spines[['top', 'right']].set_visible(False)
```



[6]:

```
# information about the dataset
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 22445 entries, 0 to 22444
Data columns (total 27 columns):
 #   Column                                                         Non-Null Count
Dtype
---  ------                                                         --------------
-----
 0   index                                                          22445 non-null
int64
 1   National Provider Identifier                                   22445 non-null
int64
 2   Last Name/Organization Name of the Provider                    22445 non-null
object
 3   First Name of the Provider                                     21460 non-null
object
 4   Middle Initial of the Provider                                 15886 non-null
object
```

```
 5   Credentials of the Provider                               20804 non-null
object
 6   Gender of the Provider                                    21460 non-null
object
 7   Entity Type of the Provider                               22445 non-null
object
 8   Street Address 1 of the Provider                          22445 non-null
object
 9   Street Address 2 of the Provider                          9134 non-null
object
 10  City of the Provider                                      22445 non-null
object
 11  Zip Code of the Provider                                  22445 non-null
float64
 12  State Code of the Provider                                22445 non-null
object
 13  Country Code of the Provider                              22445 non-null
object
 14  Provider Type                                             22445 non-null
object
 15  Medicare Participation Indicator                          22445 non-null
object
 16  Place of Service                                          22445 non-null
object
 17  HCPCS Code                                                22445 non-null
object
 18  HCPCS Description                                         22445 non-null
object
 19  HCPCS Drug Indicator                                      22445 non-null
object
 20  Number of Services                                        22445 non-null
object
 21  Number of Medicare Beneficiaries                          22445 non-null
object
 22  Number of Distinct Medicare Beneficiary/Per Day Services  22445 non-null
object
 23  Average Medicare Allowed Amount                           22445 non-null
object
 24  Average Submitted Charge Amount                           22445 non-null
object
 25  Average Medicare Payment Amount                           22444 non-null
object
 26  Average Medicare Standardized Amount                      22444 non-null
object
dtypes: float64(1), int64(2), object(24)
memory usage: 4.6+ MB
```

[7]:

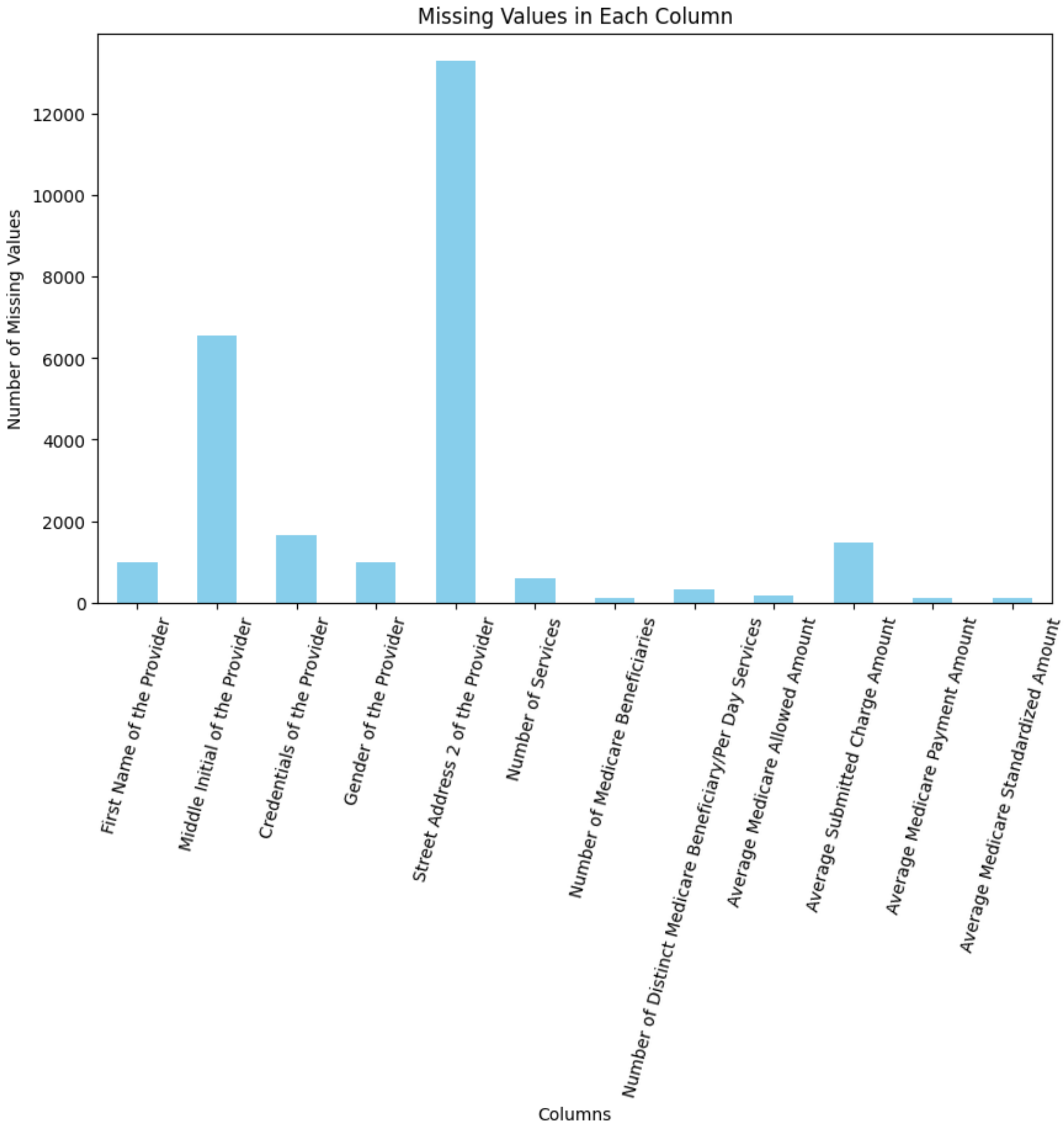
```
numeric_columns = [
 'Number of Services',
 'Number of Medicare Beneficiaries',
 'Number of Distinct Medicare Beneficiary/Per Day Services',
 'Average Medicare Allowed Amount',
 'Average Submitted Charge Amount',
 'Average Medicare Payment Amount',
 'Average Medicare Standardized Amount'
]
for column in numeric_columns:
 data[column] = pd.to_numeric(data[column], errors='coerce')


data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 22445 entries, 0 to 22444
Data columns (total 27 columns):
 #   Column                                                         Non-Null Count
Dtype
---  ------                                                         --------------
-----
 0   index                                                          22445 non-null
int64
 1   National Provider Identifier                                   22445 non-null
int64
 2   Last Name/Organization Name of the Provider                    22445 non-null
object
 3   First Name of the Provider                                     21460 non-null
object
 4   Middle Initial of the Provider                                 15886 non-null
object
 5   Credentials of the Provider                                    20804 non-null
object
 6   Gender of the Provider                                         21460 non-null
object
 7   Entity Type of the Provider                                    22445 non-null
object
 8   Street Address 1 of the Provider                               22445 non-null
object
 9   Street Address 2 of the Provider                               9134 non-null
object
 10  City of the Provider                                           22445 non-null
object
 11  Zip Code of the Provider                                       22445 non-null
float64
 12  State Code of the Provider                                     22445 non-null
```

```
object
 13  Country Code of the Provider                              22445 non-null
object
 14  Provider Type                                             22445 non-null
object
 15  Medicare Participation Indicator                          22445 non-null
object
 16  Place of Service                                          22445 non-null
object
 17  HCPCS Code                                                22445 non-null
object
 18  HCPCS Description                                         22445 non-null
object
 19  HCPCS Drug Indicator                                      22445 non-null
object
 20  Number of Services                                        21860 non-null
float64
 21  Number of Medicare Beneficiaries                          22342 non-null
float64
 22  Number of Distinct Medicare Beneficiary/Per Day Services  22120 non-null
float64
 23  Average Medicare Allowed Amount                           22272 non-null
float64
 24  Average Submitted Charge Amount                           20961 non-null
float64
 25  Average Medicare Payment Amount                           22340 non-null
float64
 26  Average Medicare Standardized Amount                      22344 non-null
float64
dtypes: float64(8), int64(2), object(17)
memory usage: 4.6+ MB
```

[12]:

```
# Calculate the number of missing values in each column
missing_values = data.isnull().sum()
# Filter out columns with non-zero missing values
missing_values = missing_values[missing_values > 0]

# Create a bar chart
plt.figure(figsize=(10, 6))
missing_values.plot(kind='bar', color='skyblue')
plt.title('Missing Values in Each Column')
plt.xlabel('Columns')
plt.ylabel('Number of Missing Values')
plt.xticks(rotation=75)
plt.show()
```



[13]:

```
# Imputation of missing values with mean
data[numeric_columns] = data[numeric_columns].fillna(data[numeric_columns].
  ↪mean())
print(data.isnull().sum())
```

```
index                                                         0
National Provider Identifier                                  0
Last Name/Organization Name of the Provider                   0
First Name of the Provider                                  985
Middle Initial of the Provider                             6559
Credentials of the Provider                                1641
```

```
Gender of the Provider                                            985
Entity Type of the Provider                                         0
Street Address 1 of the Provider                                    0
Street Address 2 of the Provider                                13311
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
```

[14]:
```
# Check for duplicates
print(data.duplicated().sum())
```

```
0
```

[15]:
```
#data preeprocessing
# Merging the name columns into a single column
data['Full Name'] = data['First Name of the Provider'].fillna('') + ' ' + \
 data['Middle Initial of the Provider'].fillna('') + ' ' + \
 data['Last Name/Organization Name of the Provider'].fillna('')
data['Full Name'] = data['Full Name'].str.strip()
data = data.drop(columns=['Last Name/Organization Name of the Provider',
 'First Name of the Provider',
'Middle Initial of the Provider'])
data.head()
```

[15]:
```
     index  National Provider Identifier Credentials of the Provider  \
0  8774979                    1891106191                        M.D.
1  3354385                    1346202256                        M.D.
2  3001884                    1306820956                         DPM
3  7594822                    1770523540                          MD
4   746159                    1073627758                          DO

  Gender of the Provider Entity Type of the Provider  \
0                      F                           I
1                      F                           I
2                      M                           I
3                      M                           I
4                      M                           I

  Street Address 1 of the Provider Street Address 2 of the Provider  \
0                1402 S GRAND BLVD                   FDT 14TH FLOOR
1                  2950 VILLAGE DR                              NaN
2                20 WASHINGTON AVE                          STE 212
3               5746 N BROADWAY ST                              NaN
4                 875 MILITARY TRL                        SUITE 200

  City of the Provider  Zip Code of the Provider State Code of the Provider  \
0          SAINT LOUIS               631041004.0                         MO
1         FAYETTEVILLE               283043815.0                         NC
2          NORTH HAVEN                64732343.0                         CT
3          KANSAS CITY               641183998.0                         MO
4              JUPITER               334585700.0                         FL

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

  HCPCS Drug Indicator Number of Services Number of Medicare Beneficiaries  \
0                    N               27.0                             24.0
1                    N              175.0                            175.0
2                    N               32.0                             13.0
3                    N               20.0                             18.0
4                    N               33.0                             24.0

  Number of Distinct Medicare Beneficiary/Per Day Services  \
0                                              27.0
1                                             175.0
2                                              32.0
3                                              20.0
4                                              31.0

  Average Medicare Allowed Amount Average Submitted Charge Amount  \
0                      200.587778                      305.211111
1                      123.730000                      548.800000
2                       90.650000                      155.000000
3                        3.500000                        5.000000
4                       26.520000                       40.000000
```

```
   Average Medicare Payment Amount  Average Medicare Standardized Amount  \
0                       157.262222                            160.908889
1                       118.830000                            135.315257
2                        64.439688                             60.595937
3                         3.430000                              3.430000
4                        19.539394                             19.057576

                Full Name
0  SATYASREE  UPADHYAYULA
1           WENDY P JONES
2      RICHARD W DUROCHER
3         JASPER  FULLARD
4      ANTHONY E PERROTTI

[5 rows x 25 columns]
```

[16]:

```
# Merging the address columns
data['Full Address'] = data['Street Address 1 of the Provider'].fillna('') + '␣
 ↪' + \
 data['Street Address 2 of the Provider'].fillna('')
data['Full Address'] = data['Full Address'].str.strip()
data = data.drop(columns=['Street Address 1 of the Provider',
 'Street Address 2 of the Provider'])
data.head()
```

[16]:

```
     index  National Provider Identifier Credentials of the Provider  \
0  8774979                    1891106191                        M.D.
1  3354385                    1346202256                        M.D.
2  3001884                    1306820956                         DPM
3  7594822                    1770523540                          MD
4   746159                    1073627758                          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
```

```
  Country Code of the Provider            Provider Type  …  \
0                           US        Internal Medicine  …
1                           US  Obstetrics & Gynecology  …
2                           US                 Podiatry  …
3                           US        Internal Medicine  …
4                           US        Internal Medicine  …

  HCPCS Drug Indicator Number of Services Number of Medicare Beneficiaries  \
0                    N               27.0                             24.0
1                    N              175.0                            175.0
2                    N               32.0                             13.0
3                    N               20.0                             18.0
4                    N               33.0                             24.0

  Number of Distinct Medicare Beneficiary/Per Day Services  \
0                                               27.0
1                                              175.0
2                                               32.0
3                                               20.0
4                                               31.0

  Average Medicare Allowed Amount  Average Submitted Charge Amount  \
0                      200.587778                       305.211111
1                      123.730000                       548.800000
2                       90.650000                       155.000000
3                        3.500000                         5.000000
4                       26.520000                        40.000000

   Average Medicare Payment Amount  Average Medicare Standardized Amount  \
0                       157.262222                            160.908889
1                       118.830000                            135.315257
2                        64.439688                             60.595937
3                         3.430000                              3.430000
4                        19.539394                             19.057576

                Full Name                      Full Address
0  SATYASREE  UPADHYAYULA  1402 S GRAND BLVD FDT 14TH FLOOR
1           WENDY P JONES                   2950 VILLAGE DR
2      RICHARD W DUROCHER         20 WASHINGTON AVE STE 212
3         JASPER  FULLARD                5746 N BROADWAY ST
4      ANTHONY E PERROTTI        875 MILITARY TRL SUITE 200

[5 rows x 24 columns]
```

[20]:

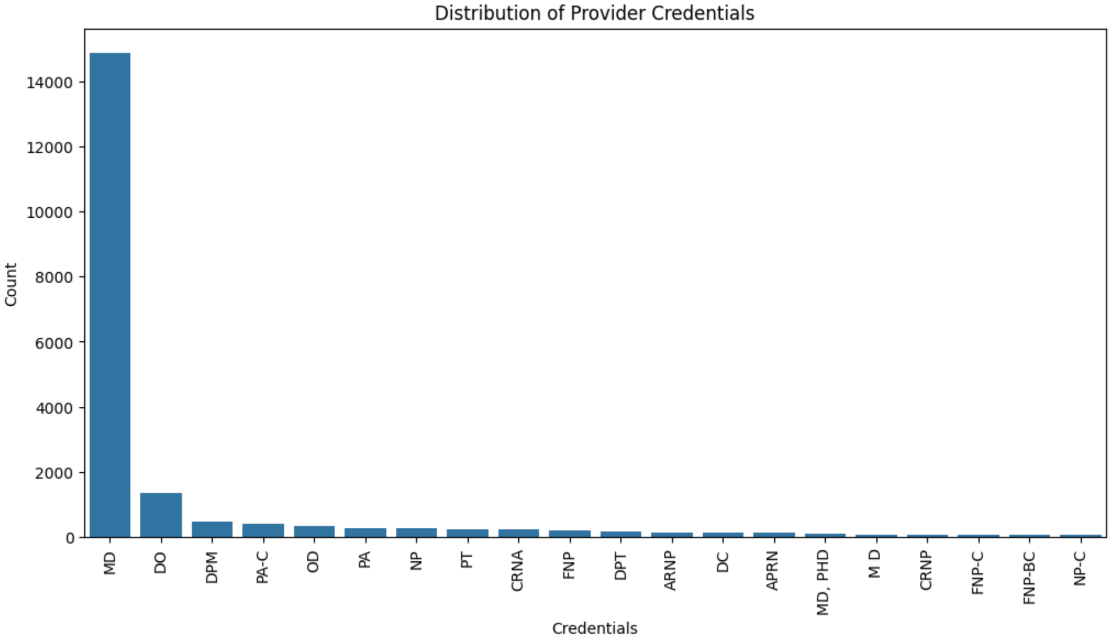
```python
# Standardize credentials
data['Credentials of the Provider'] = data['Credentials of the Provider'].str.replace(r'\.', '', regex=True).str.upper()
```

```
data.head()
```

```
[20]:      index  National Provider Identifier Credentials of the Provider  \
      0  8774979                    1891106191                          MD   
      1  3354385                    1346202256                          MD   
      2  3001884                    1306820956                         DPM   
      3  7594822                    1770523540                          MD   
      4   746159                    1073627758                          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   

        Country Code of the Provider            Provider Type  ...  \
      0                           US        Internal Medicine  ...   
      1                           US  Obstetrics & Gynecology  ...   
      2                           US                 Podiatry  ...   
      3                           US        Internal Medicine  ...   
      4                           US        Internal Medicine  ...   

        HCPCS Drug Indicator Number of Services Number of Medicare Beneficiaries  \
      0                    N               27.0                             24.0   
      1                    N              175.0                            175.0   
      2                    N               32.0                             13.0   
      3                    N               20.0                             18.0   
      4                    N               33.0                             24.0   

        Number of Distinct Medicare Beneficiary/Per Day Services  \
      0                                             27.0           
      1                                            175.0           
      2                                             32.0           
      3                                             20.0           
      4                                             31.0           

        Average Medicare Allowed Amount  Average Submitted Charge Amount  \
      0                      200.587778                       305.211111   
      1                      123.730000                       548.800000   
```

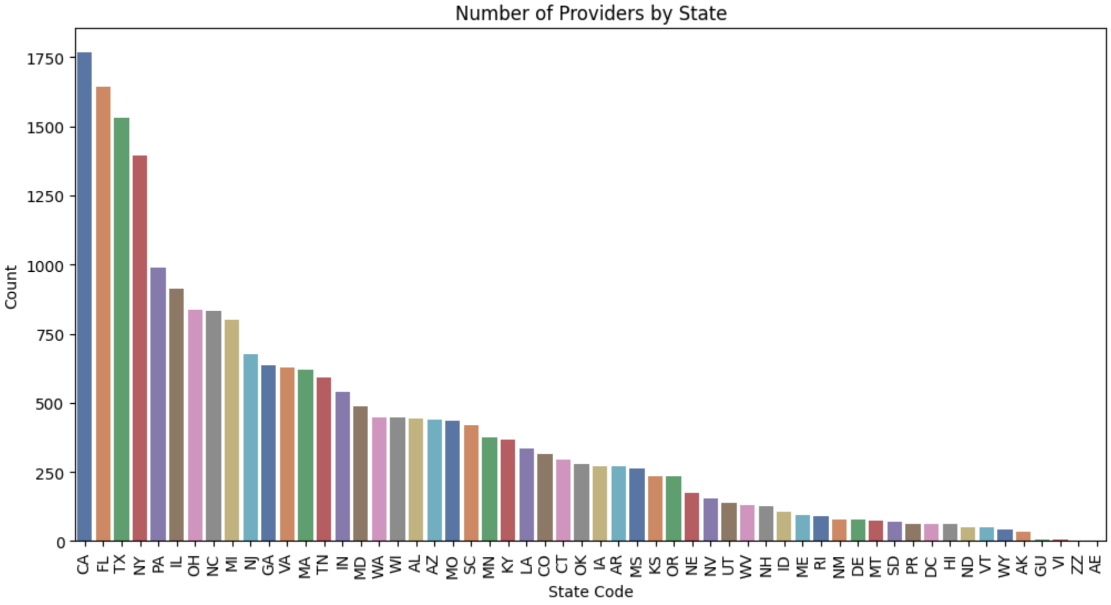
```
2                         90.650000                               155.000000
3                          3.500000                                 5.000000
4                         26.520000                                40.000000

   Average Medicare Payment Amount  Average Medicare Standardized Amount  \
0                       157.262222                            160.908889
1                       118.830000                            135.315257
2                        64.439688                             60.595937
3                         3.430000                              3.430000
4                        19.539394                             19.057576

               Full Name                    Full Address
0  SATYASREE  UPADHYAYULA  1402 S GRAND BLVD FDT 14TH FLOOR
1          WENDY P JONES                   2950 VILLAGE DR
2      RICHARD W DUROCHER        20 WASHINGTON AVE STE 212
3         JASPER  FULLARD               5746 N BROADWAY ST
4      ANTHONY E PERROTTI        875 MILITARY TRL SUITE 200

[5 rows x 24 columns]
```

[21]:
```python
# Plot bar plot for Credentials of the Provider
credentials_counts = data['Credentials of the Provider'].value_counts().head(20)
plt.figure(figsize=(12, 6))
sns.barplot(x=credentials_counts.index, y=credentials_counts.values)
plt.title('Distribution of Provider Credentials')
plt.xlabel('Credentials')
plt.ylabel('Count')
plt.xticks(rotation=90)
plt.show()
```



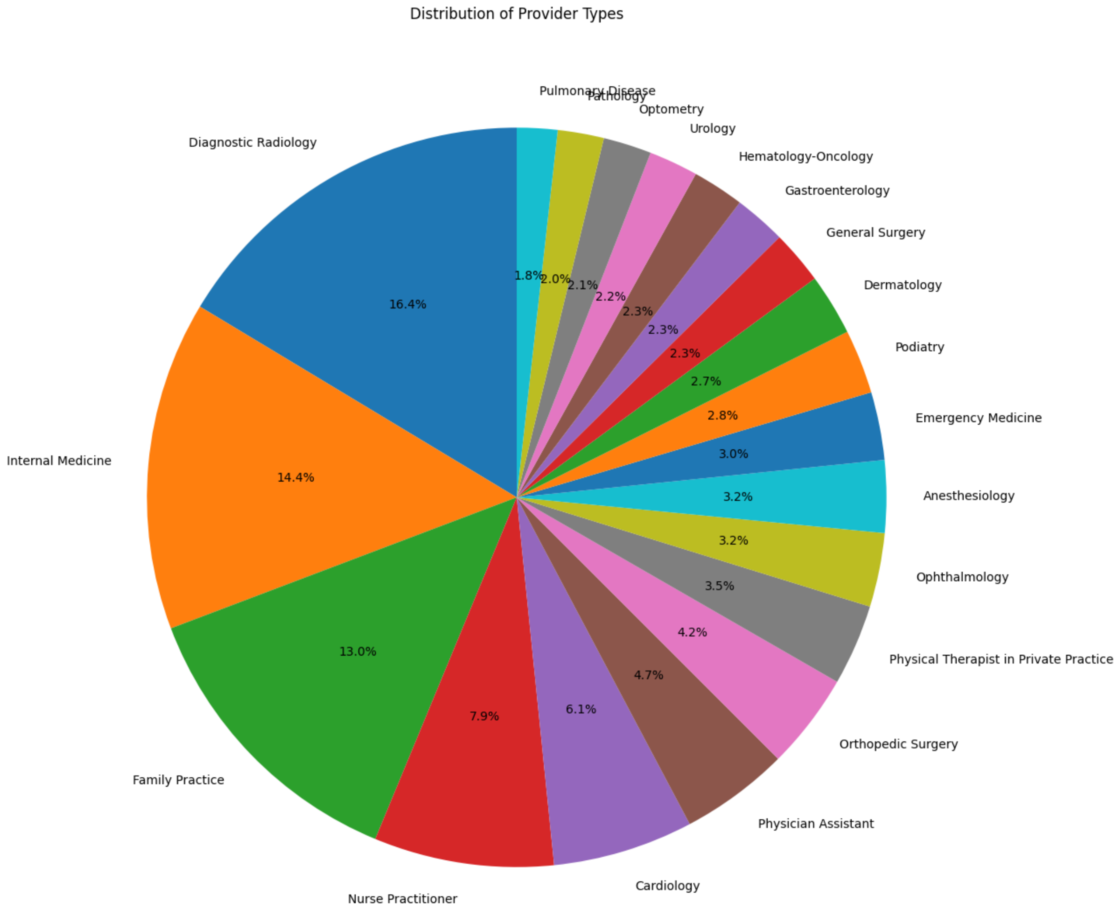
[24]:

```
state_counts = data['State Code of the Provider'].value_counts()
# bar graph for State Code of the Provider
plt.figure(figsize=(12, 6))
sns.barplot(x=state_counts.index, y=state_counts.values, palette='deep')
plt.title('Number of Providers by State')
plt.xlabel('State Code')
plt.ylabel('Count')
plt.xticks(rotation=90)
plt.show()
```



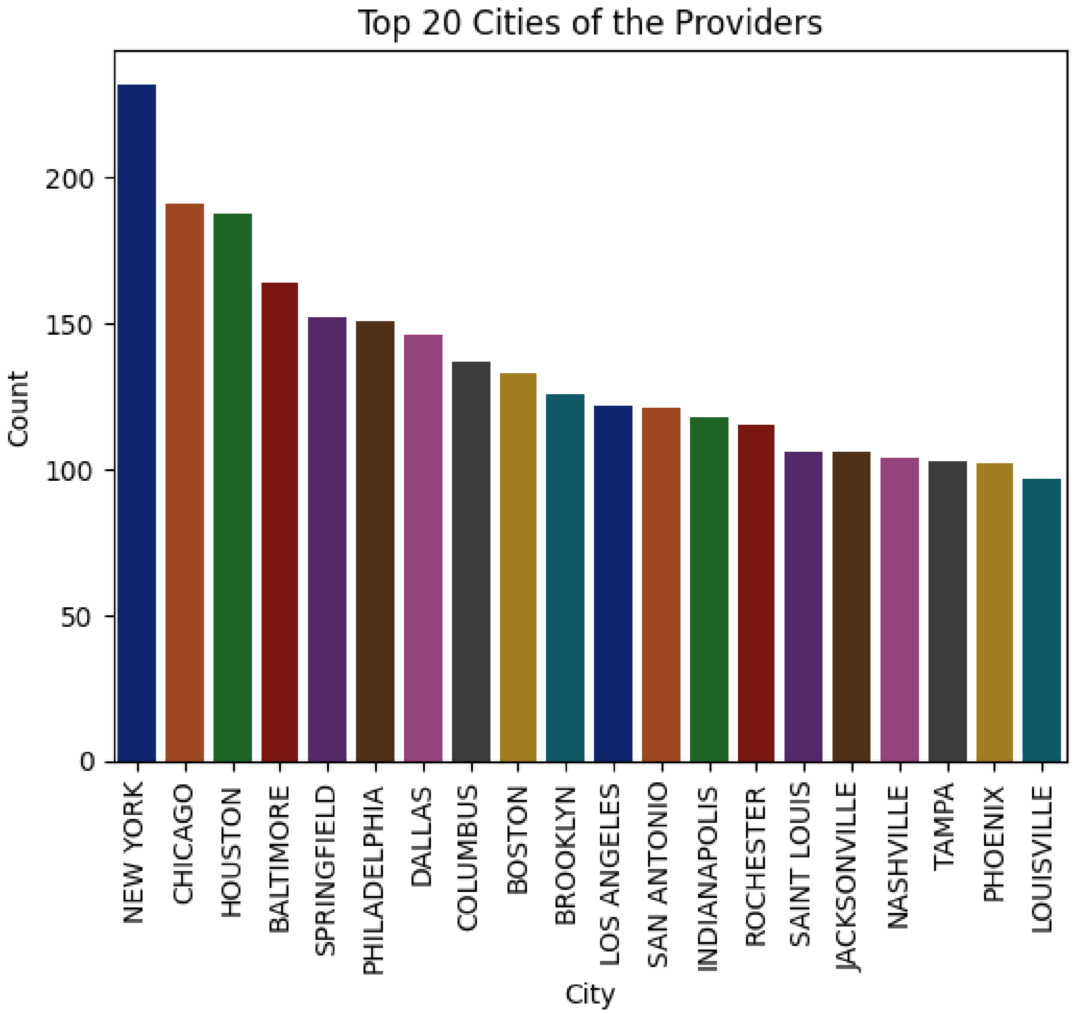
[43]:

```
provider_type_counts = data['Provider Type'].value_counts().head(20)
# pie chart for Provider Types
plt.figure(figsize=(12, 14))
plt.pie(provider_type_counts, labels=provider_type_counts.index, autopct='%1.
 ↪1f%%', startangle=90)
plt.title('Distribution of Provider Types')
plt.axis('equal')
plt.show()
```

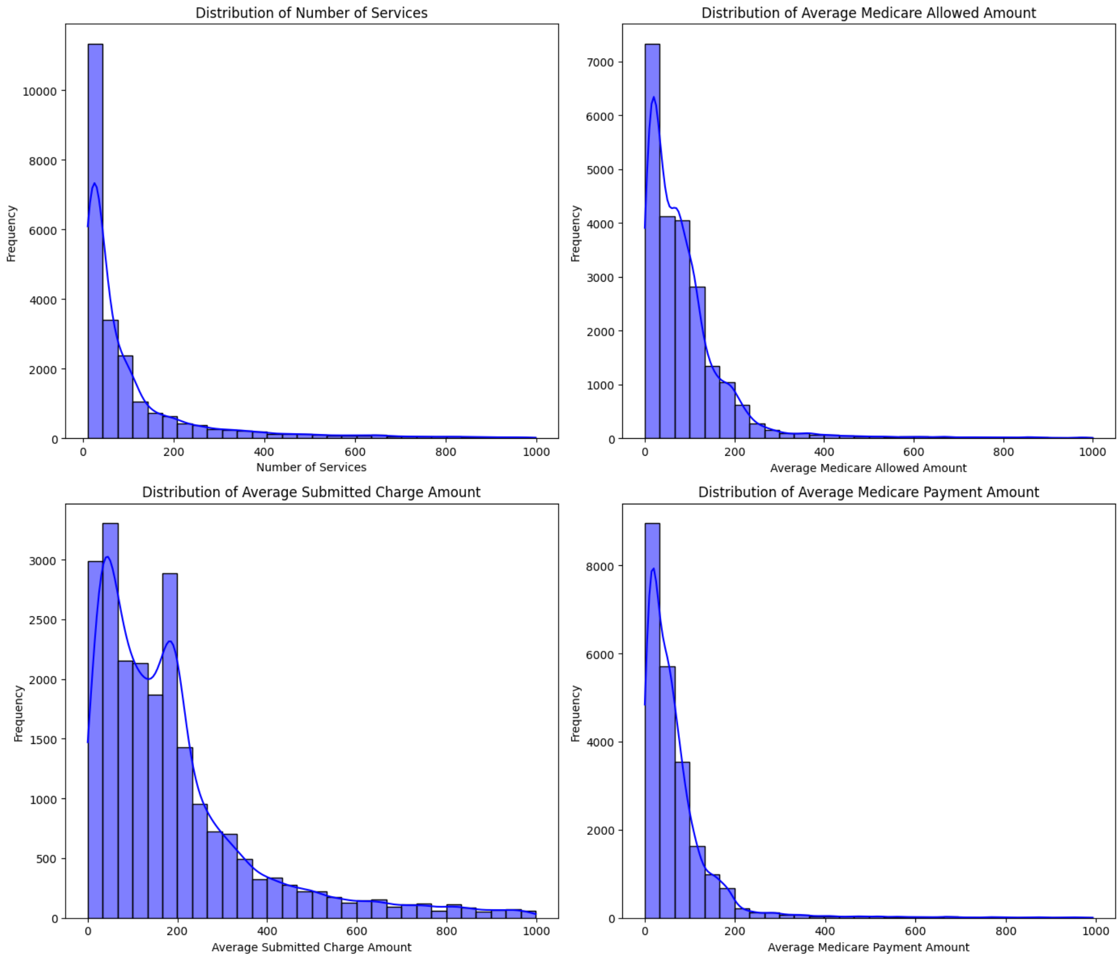


[32]:

```
# occurrences of each city
city_counts = data['City of the Provider'].value_counts().head(20)
# Plot of top 20 cities
sns.barplot(x=city_counts.index, y=city_counts.values, palette='dark')
plt.title('Top 20 Cities of the Providers')
plt.xlabel('City')
plt.ylabel('Count')
plt.xticks(rotation=90)
plt.show()
```

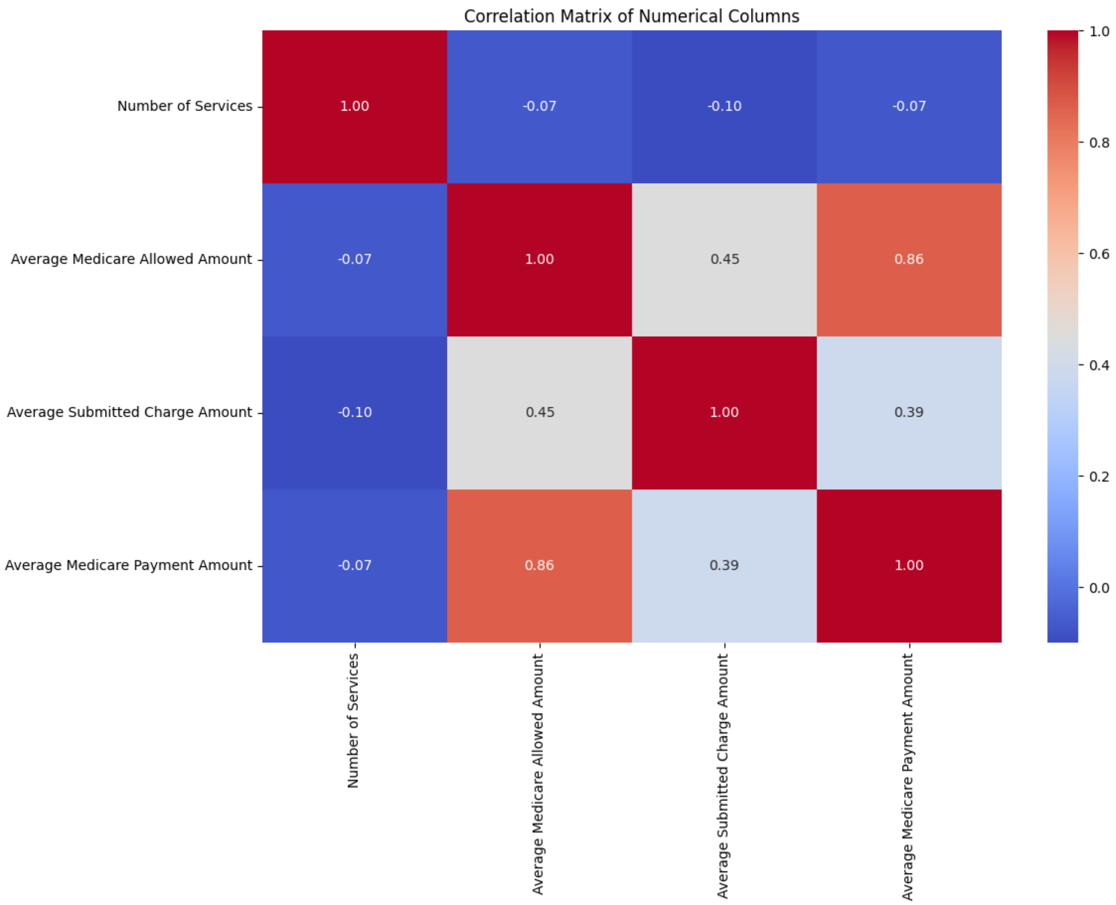


[33]:
```
numeric_columns = [
 'Number of Services',
 'Average Medicare Allowed Amount',
 'Average Submitted Charge Amount',
 'Average Medicare Payment Amount'
]
for column in numeric_columns:
 data[column] = pd.to_numeric(data[column], errors='coerce')
plt.figure(figsize=(14, 12))
for i, column in enumerate(numeric_columns, 1):
 plt.subplot(2, 2, i)
 sns.histplot(data[column].dropna(), bins=30, kde=True, color='blue')
 plt.title(f'Distribution of {column}')
 plt.xlabel(column)
 plt.ylabel('Frequency')
plt.tight_layout()
plt.show()
```



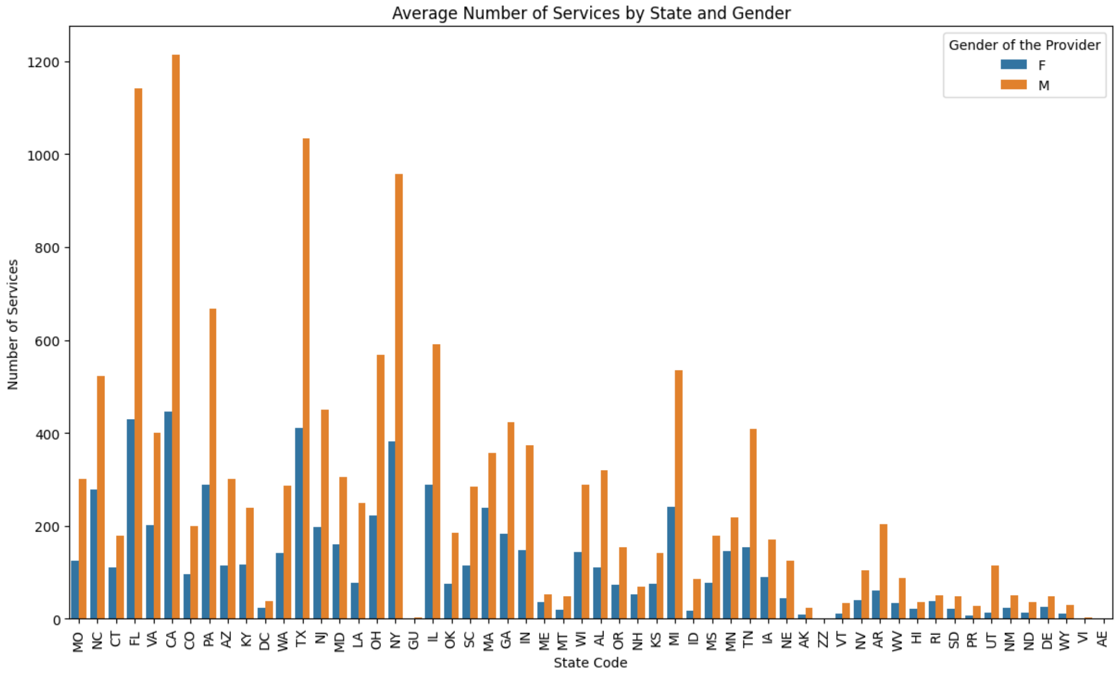
[34]:

```
corr_matrix = data[numeric_columns].corr()
#correlation heatmap
plt.figure(figsize=(12, 8))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Correlation Matrix of Numerical Columns')
plt.show()
```



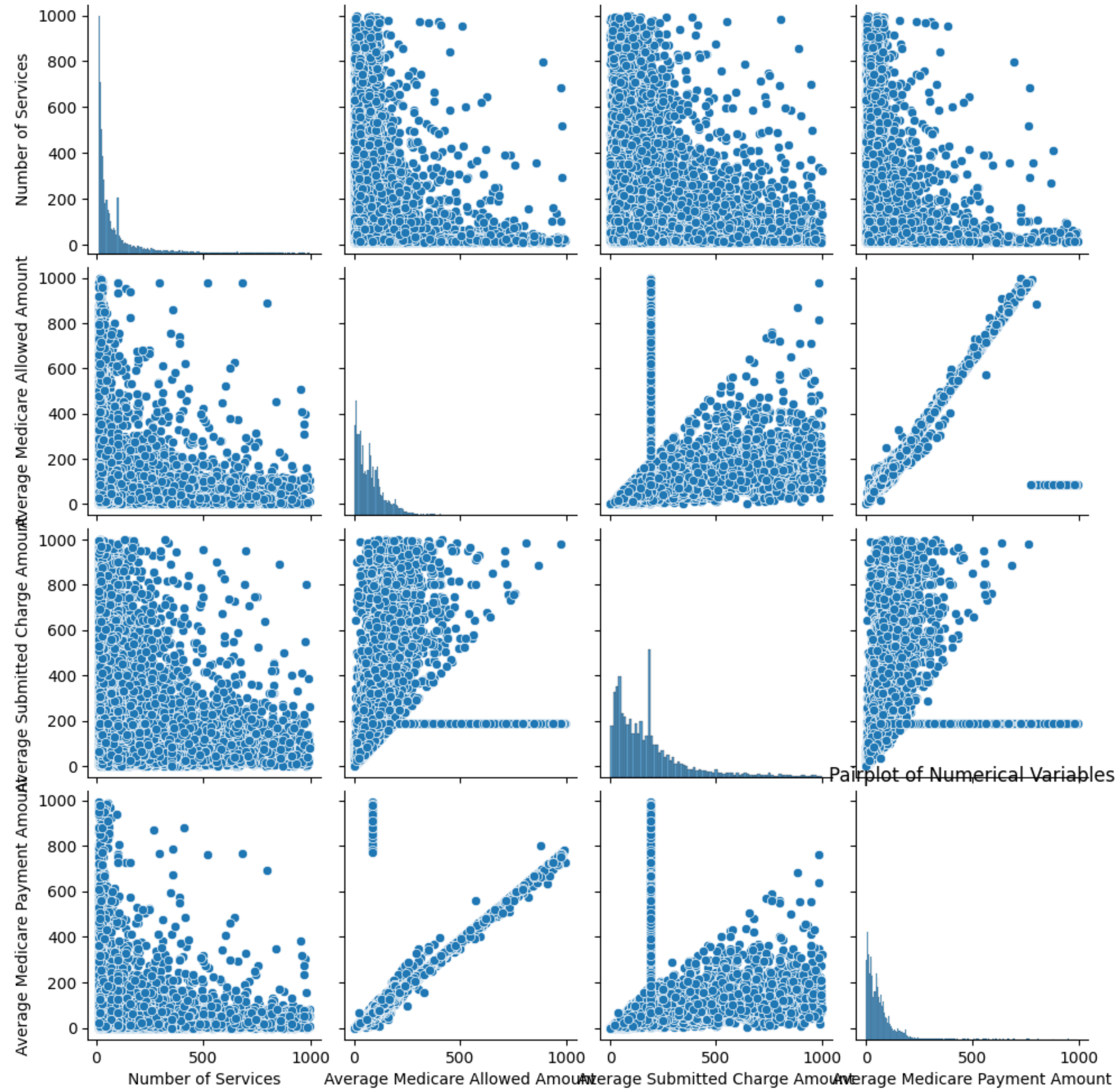
[37]:

```python
# point plot to show the relationship between average Number of Services by State Code of the Provide
plt.figure(figsize=(14, 8))
sns.countplot(x='State Code of the Provider', hue='Gender of the Provider', data=data)
plt.title('Average Number of Services by State and Gender')
plt.xlabel('State Code')
plt.ylabel('Number of Services')
plt.xticks(rotation=90)
plt.show()
```



[45]:

```
#pairplot for numeric values

import matplotlib.pyplot as plt
sns.pairplot(data[numeric_columns])
plt.title('Pairplot of Numerical Variables')
plt.show()
```


Pairplot of Numerical Variables
Number of Services
Average Medicare Allowed Amount
Average Submitted Charge Amount
Average Medicare Payment Amount
0
200
400
600
800
1000
0
500
1000