

Files X

...

Performing the EDA for Healthcare Providers

```
# prompt: import pandas as pd

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
file_path='/content/Healthcare Providers.csv (1).zip'
df=pd.read_csv(file_path)
print(df.head())
#data Cleaning
print("\nMissing values in each column")
print(df.isnull().sum())
df = df.drop_duplicates()
df = df.drop_duplicates()
print("\nSummary statistics of numerical columns after cleaning")
print(df.describe())
```

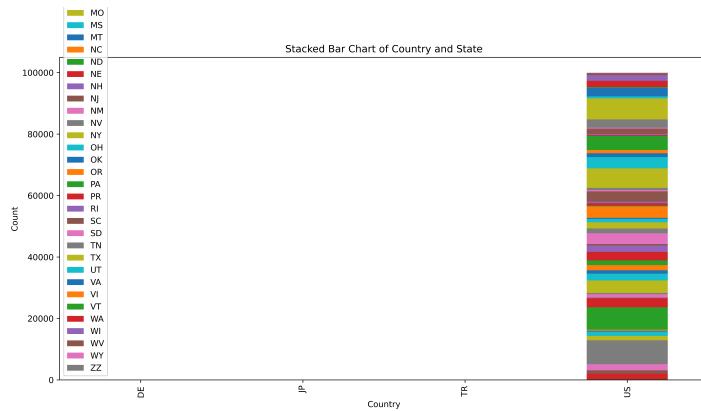


Connecting to a runtime to enable file browsing.

```
Number of Medicare Beneficiaries
Number of Distinct Medicare Beneficiary/Per Day
Average Medicare Allowed Amount
Average Submitted Charge Amount
Average Medicare Payment Amount
Average Medicare Standardized Amount
dtype: int64
```

```
Summary statistics of numerical columns after cleaning
   index  National Provider Identifier
count    1.000000e+05                  1.000000e+05
mean     4.907646e+06                  1.498227e+06
std      2.839633e+06                  2.874125e+06
min      2.090000e+02                  1.003001e+01
25%     2.458791e+06                  1.245669e+01
50%     4.901266e+06                  1.497847e+01
75%     7.349450e+06                  1.740374e+01
max     9.847440e+06                  1.993000e+01
```

histograms, scatter plots and a heatmap for the numerical columns in a dataframe.

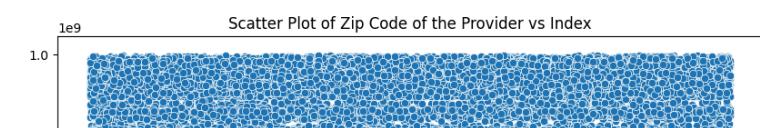
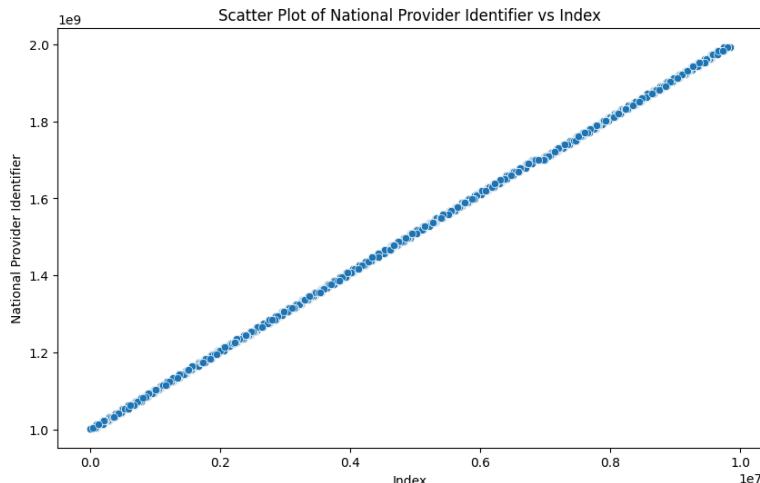
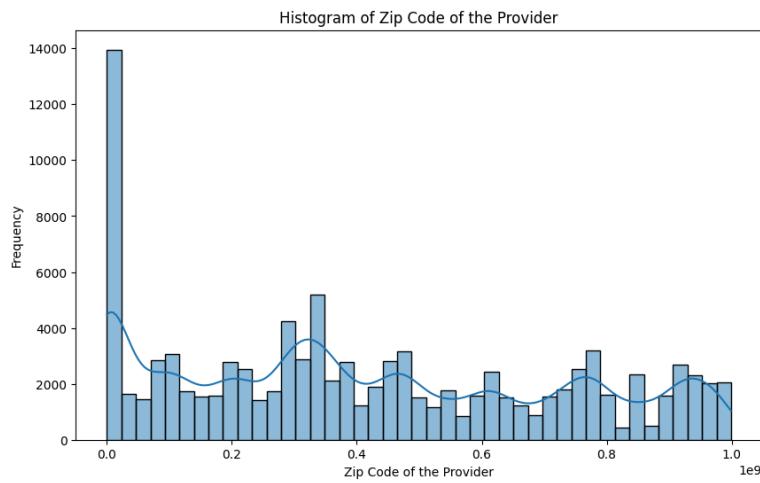
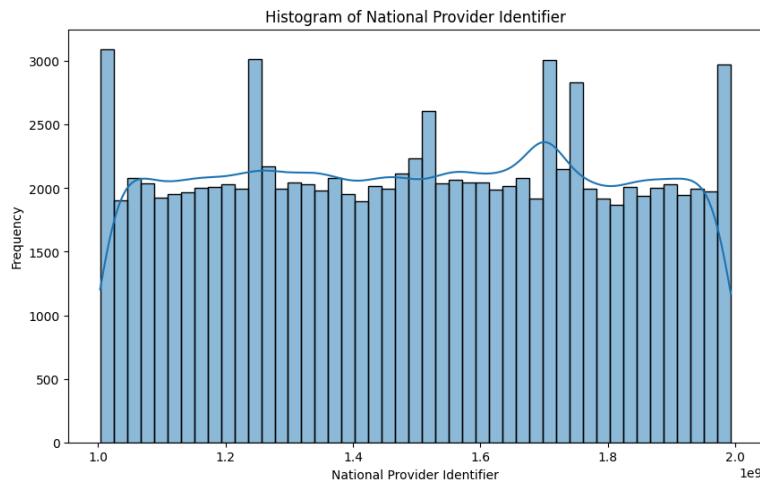
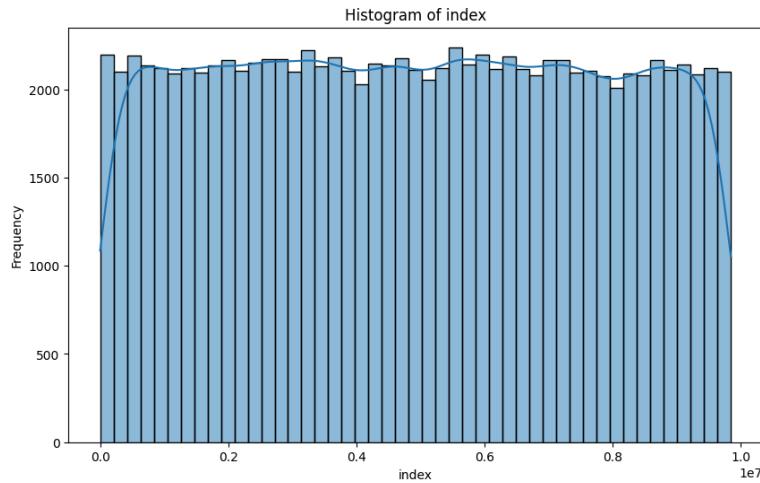


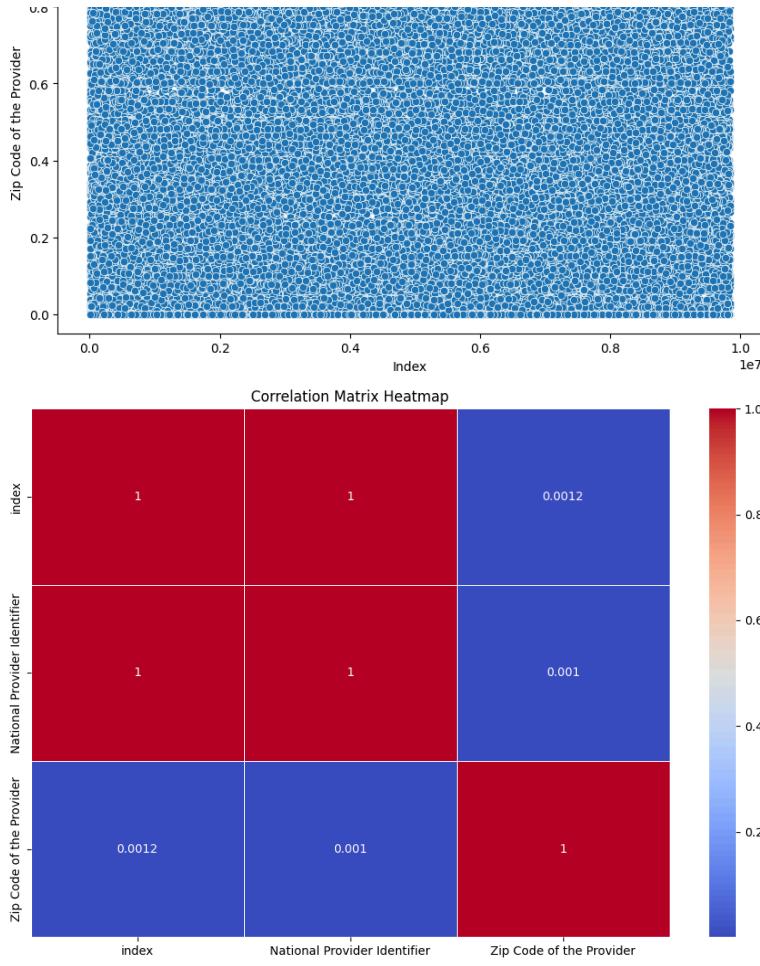
```
# Create histograms for numerical columns
numerical_columns = df.select_dtypes(include=['int64', 'float64'])

for column in numerical_columns:
    plt.figure(figsize=(10, 6))
    sns.histplot(df[column], kde=True)
    plt.title('Histogram of ' + column)
    plt.xlabel(column)
    plt.ylabel('Frequency')
    plt.show()

# Create scatter plots for numerical columns
for column in numerical_columns:
    if column != 'index': # Skip the index column
        plt.figure(figsize=(10, 6))
        sns.scatterplot(x=df['index'], y=df[column])
        plt.title('Scatter Plot of ' + column + ' vs Index')
        plt.xlabel('Index')
        plt.ylabel(column)
        plt.show()

# Create a heatmap for the correlation matrix of numerical columns
plt.figure(figsize=(12, 8))
corr_matrix = df[numerical_columns].corr()
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm',
            square=True, cbar=False)
plt.title('Correlation Matrix Heatmap')
plt.show()
```





UNIVARIATE ANALYSIS: The Histogram, Box Plot, and Density Plot for the top 30 numerical columns:

Number of Services Number of Medicare Beneficiaries
Average Medicare Allowed Amount

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load the data
file_path = 'Healthcare Providers.csv (1).zip'
df = pd.read_csv(file_path)

# Select the top 30 rows
selected_df = df.head(30)

# Convert relevant columns to numeric, coercing errors
selected_columns = ['Number of Services', 'Number of M
for column in selected_columns:
    selected_df[column] = pd.to_numeric(selected_df[co

# Drop rows with NaN values in the selected columns
cleaned_df = selected_df.dropna(subset=selected_column

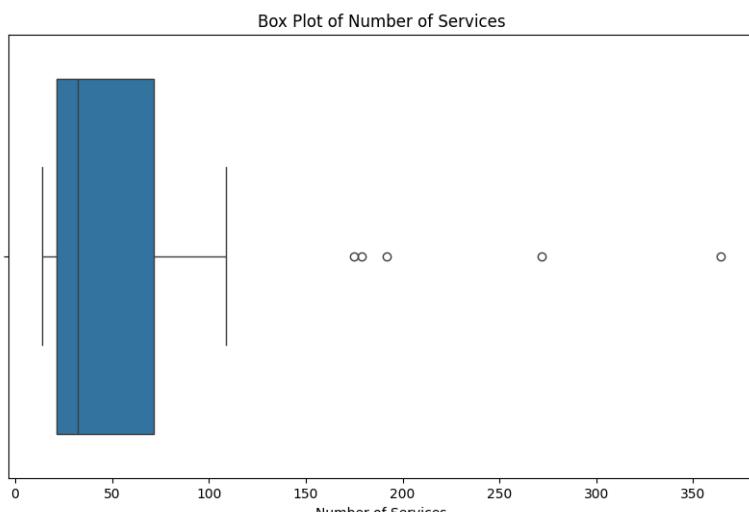
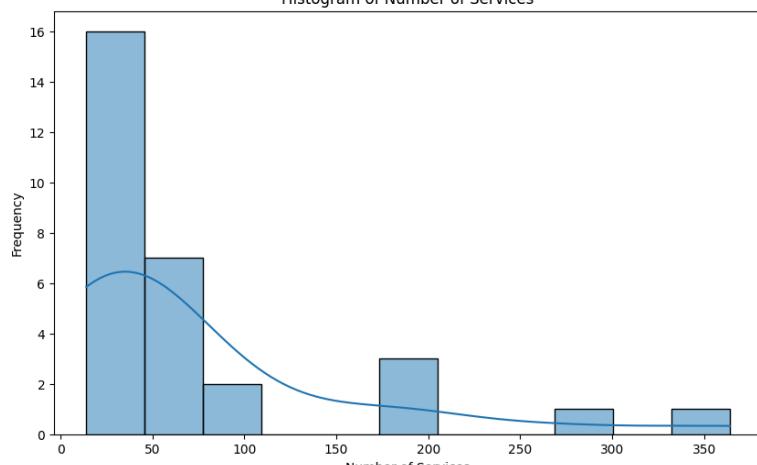
# Plot the visualizations for the selected columns
for column in selected_columns:
    plt.figure(figsize=(10, 6))
    sns.histplot(cleaned_df[column], kde=True)
    plt.title('Histogram of ' + column)
    plt.xlabel(column)
    plt.ylabel('Frequency')
    plt.show()

    plt.figure(figsize=(10, 6))
    sns.boxplot(x=cleaned_df[column])
    plt.title('Box Plot of ' + column)
    plt.xlabel(column)
    plt.show()

    plt.figure(figsize=(10, 6))
    sns.kdeplot(cleaned_df[column], shade=True)
    plt.title('Density Plot of ' + column)
    plt.xlabel(column)
    plt.ylabel('Density')
    plt.show()
```

→ <ipython-input-6-6091b6237e1b>:15: SettingWithCopy
A value is trying to be set on a copy of a slice f
Try using .loc[row_indexer,col_indexer] = value in

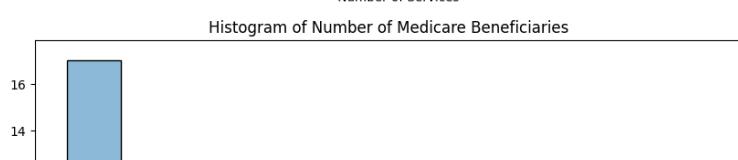
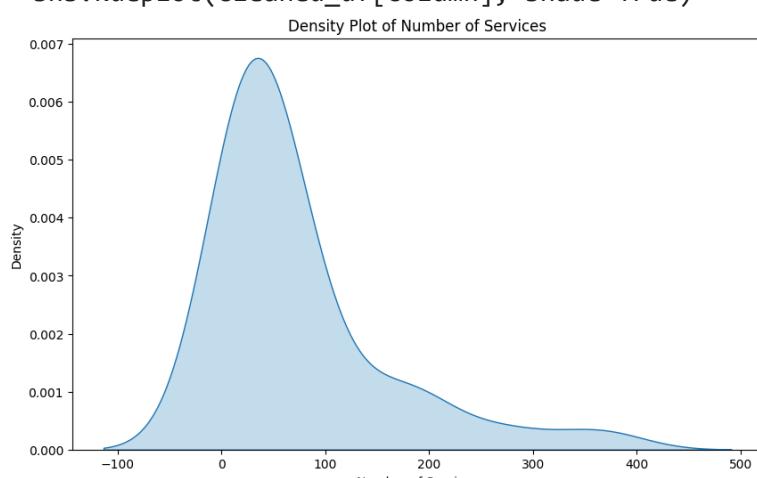
See the caveats in the documentation: <https://pand>
selected_df[column] = pd.to_numeric(selected_df[

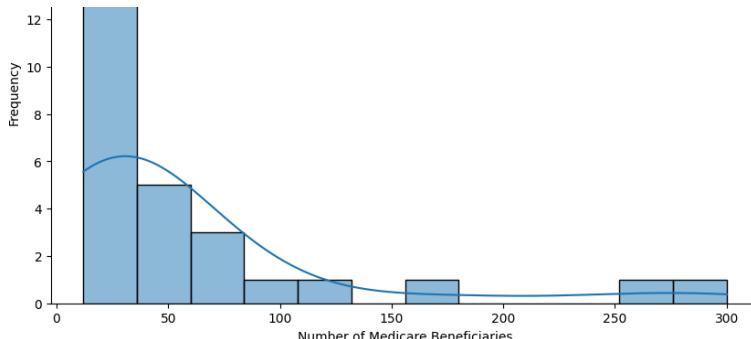


<ipython-input-6-6091b6237e1b>:36: FutureWarning:

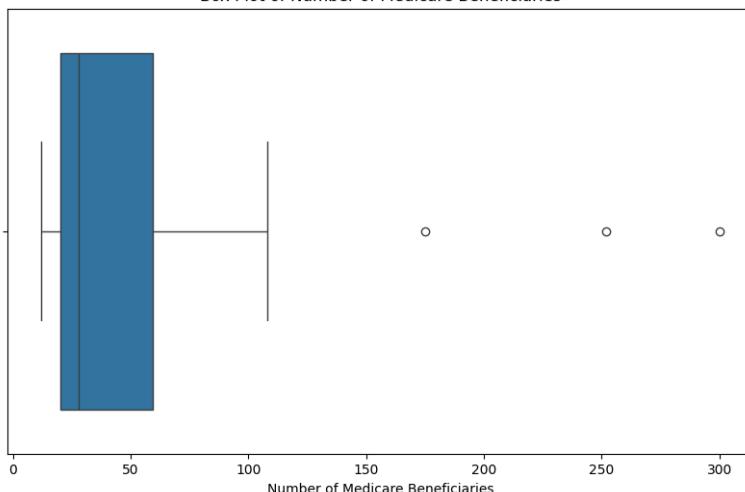
`shade` is now deprecated in favor of `fill`; sett
This will become an error in seaborn v0.14.0; plea

```
sns.kdeplot(cleaned_df[column], shade=True)
```





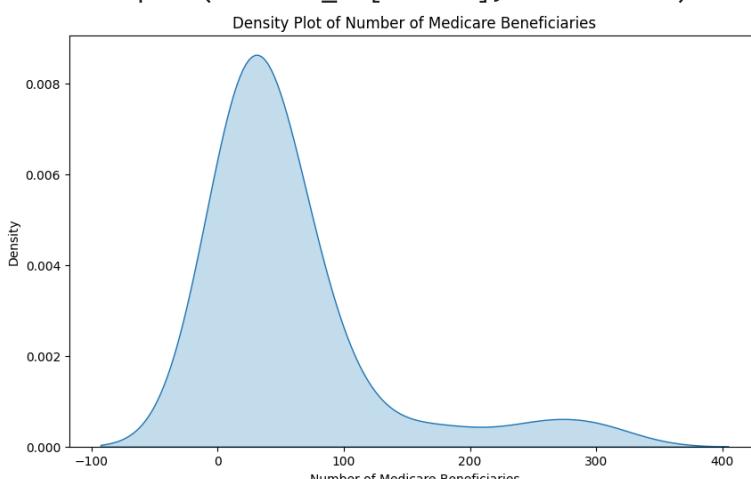
Box Plot of Number of Medicare Beneficiaries



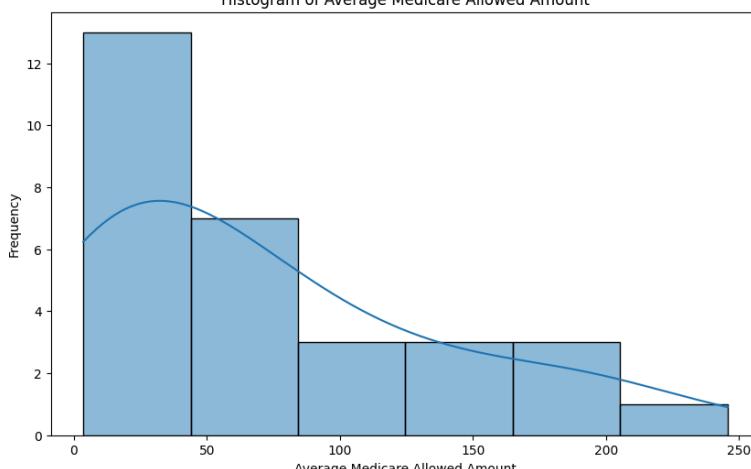
```
<ipython-input-6-6091b6237e1b>:36: FutureWarning:
```

```
`shade` is now deprecated in favor of `fill`; sett
This will become an error in seaborn v0.14.0; plea
```

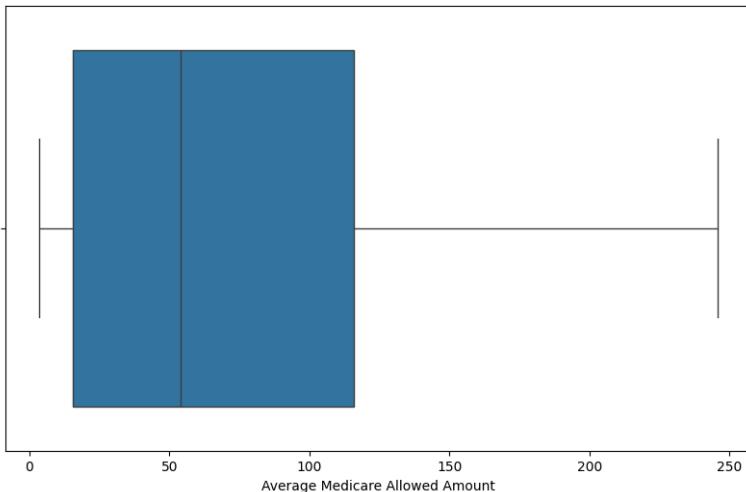
```
sns.kdeplot(cleaned_df[column], shade=True)
```



Histogram of Average Medicare Allowed Amount



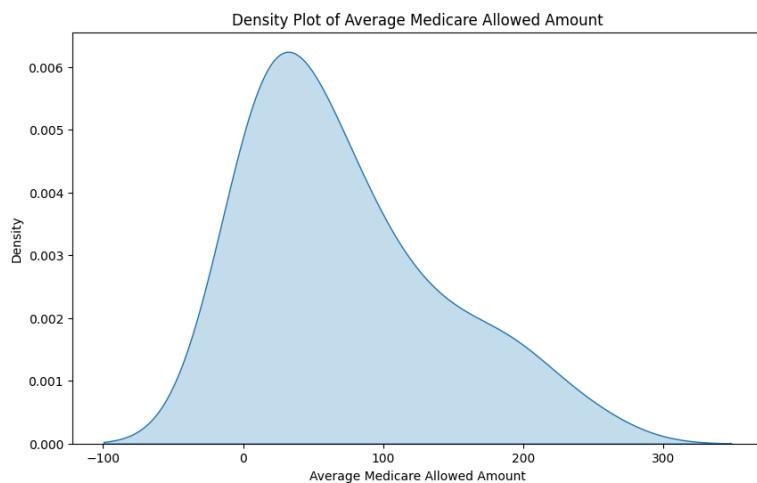
Box Plot of Average Medicare Allowed Amount



```
<ipython-input-6-6091b6237e1b>:36: FutureWarning:
```

```
`shade` is now deprecated in favor of `fill`; sett  
This will become an error in seaborn v0.14.0; plea
```

```
sns.kdeplot(cleaned_df[column], shade=True)
```



Univariate analysis for categorical columns

```
import seaborn as sns
import matplotlib.pyplot as plt

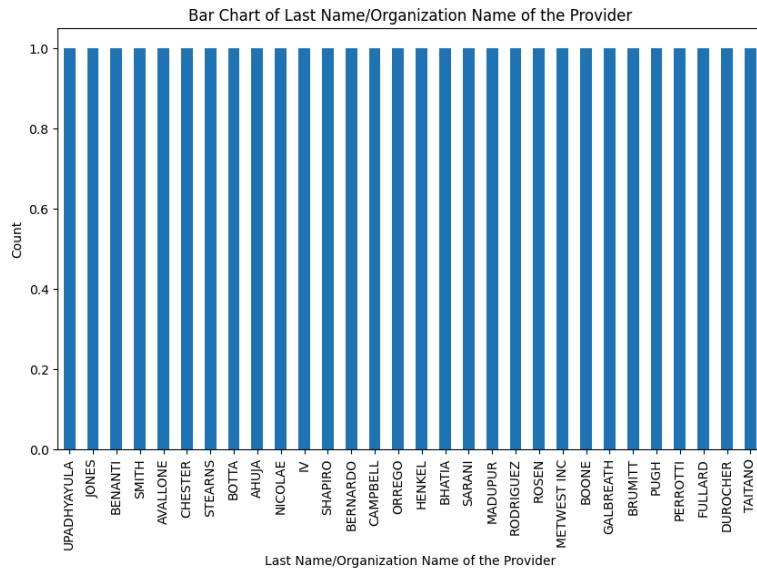
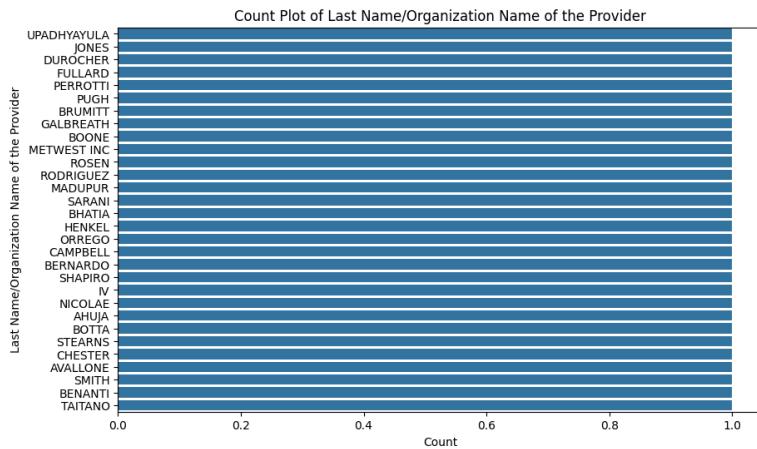
# Select the top 30 rows
selected_df = df.head(30)

# Identify categorical columns
categorical_columns = selected_df.select_dtypes(include=[object]).columns

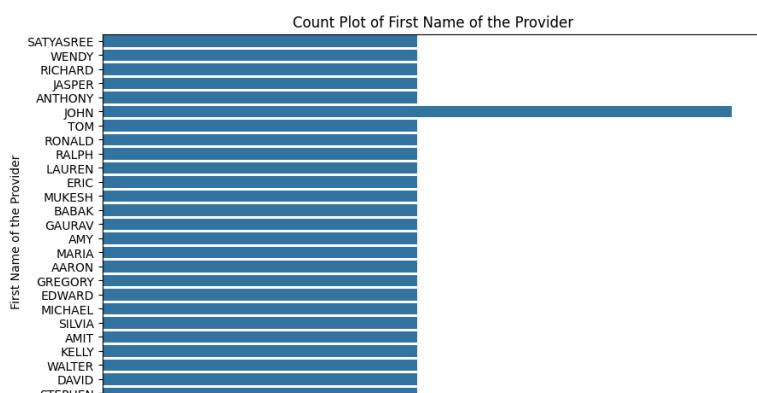
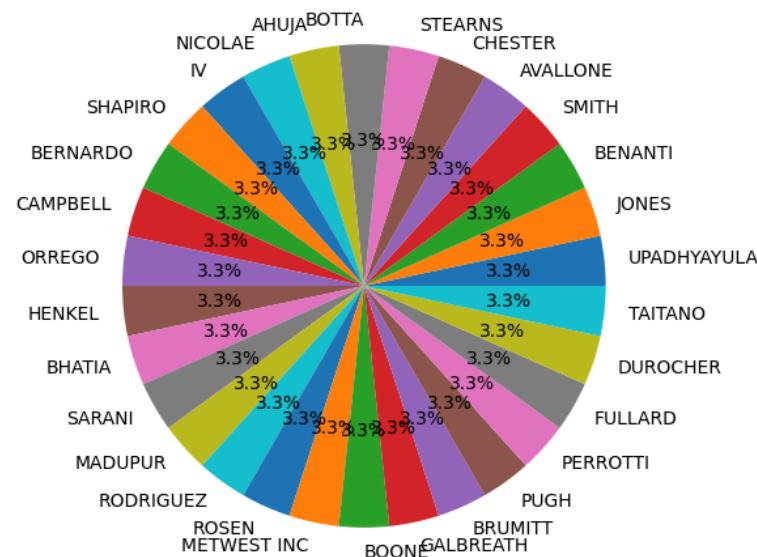
# Plot the visualizations for the selected categorical columns
for column in categorical_columns:
    plt.figure(figsize=(10, 6))
    sns.countplot(y=selected_df[column])
    plt.title('Count Plot of ' + column)
    plt.xlabel('Count')
    plt.ylabel(column)
    plt.show()

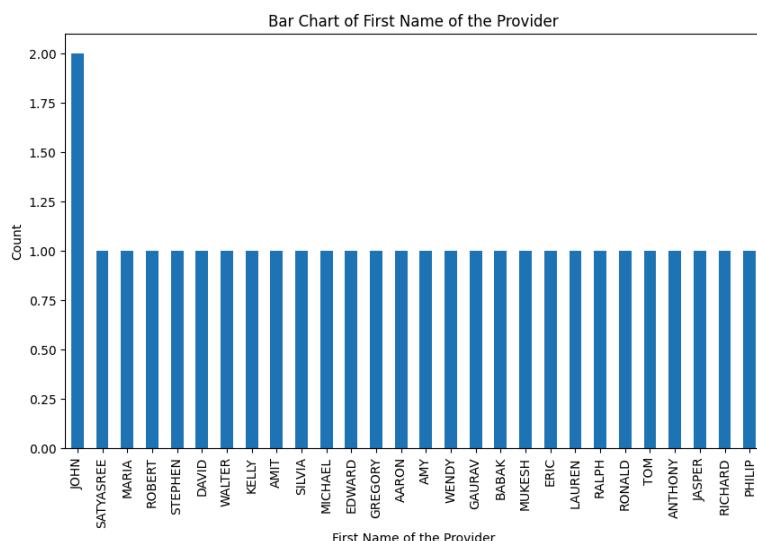
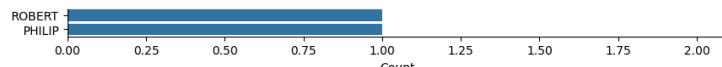
    plt.figure(figsize=(10, 6))
    selected_df[column].value_counts().plot(kind='bar')
    plt.title('Bar Chart of ' + column)
    plt.xlabel(column)
    plt.ylabel('Count')
    plt.show()

    plt.figure(figsize=(10, 6))
    selected_df[column].value_counts().plot(kind='pie')
    plt.title('Pie Chart of ' + column)
    plt.ylabel('')
    plt.show()
```



Pie Chart of Last Name/Organization Name of the Provider



**Pie Chart of First Name of the Provider**

Name	Percentage
JOHN	6.9%
ROBERT	3.4%
MARIA	3.4%
SATYASREE	3.4%
STEPHEN	3.4%
DAVID	3.4%
EDWARD	3.4%
GREGORY	3.4%
AARON	3.4%
AMY	3.4%
WENDY	3.4%
GAURAV	3.4%
BABAK	3.4%
MICHAEL	3.4%
KELLY	3.4%
AMIT	3.4%
SILVIA	3.4%
RICHARD	3.4%
JASPER	3.4%
ANTHONY	3.4%
TOM	3.4%
RONALD	3.4%
RALPH	3.4%

Count Plot of Middle Initial of the Provider

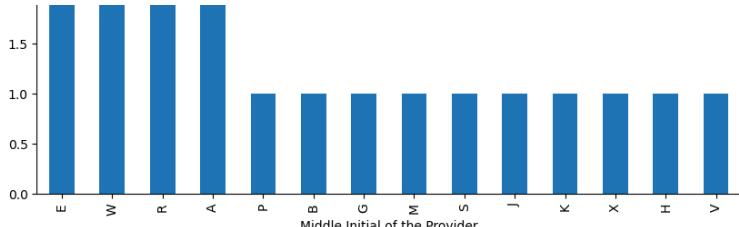
Middle Initial	Count
P	~1.0
W	~2.0
E	~4.0
R	~2.0
B	~1.0
G	~1.0
M	~1.0
S	~1.0
J	~1.0
K	~1.0
X	~1.0
A	~2.0
H	~1.0
V	~1.0

Bar Chart of Middle Initial of the Provider

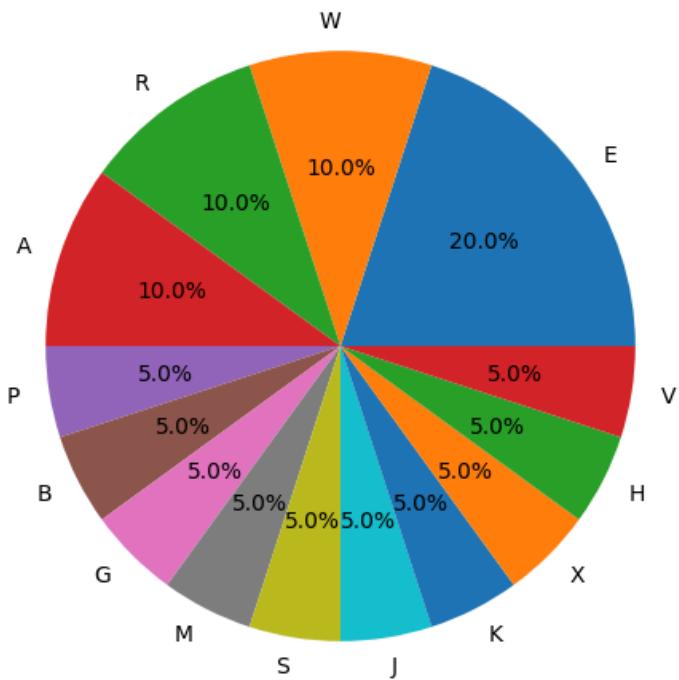
Middle Initial	Count
Initial 1	4.0
Initial 2	2.0
Initial 3	2.0
Initial 4	2.0

https://colab.research.google.com/drive/1cjFmz29vzRYGRZf_zQUxEy40y9lHHp4#scrollTo=_DCOvs3ut0wm&printMode=true

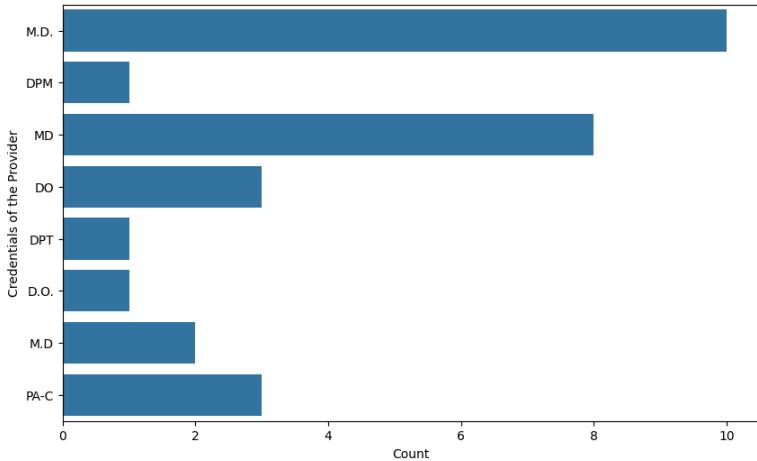
14/41



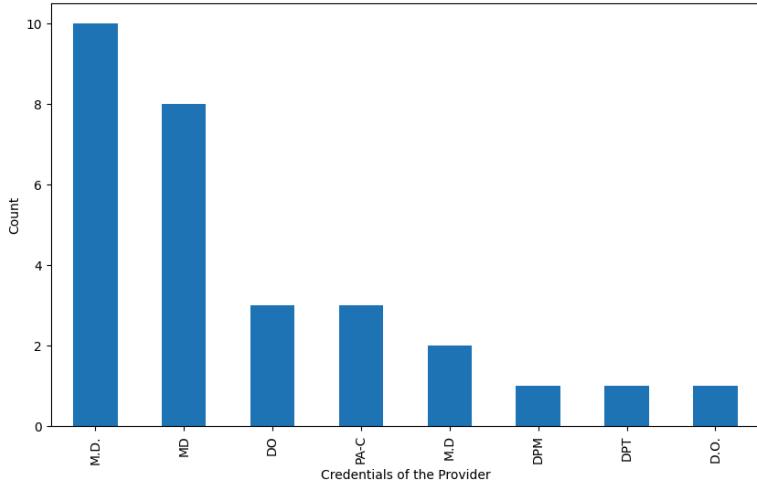
Pie Chart of Middle Initial of the Provider



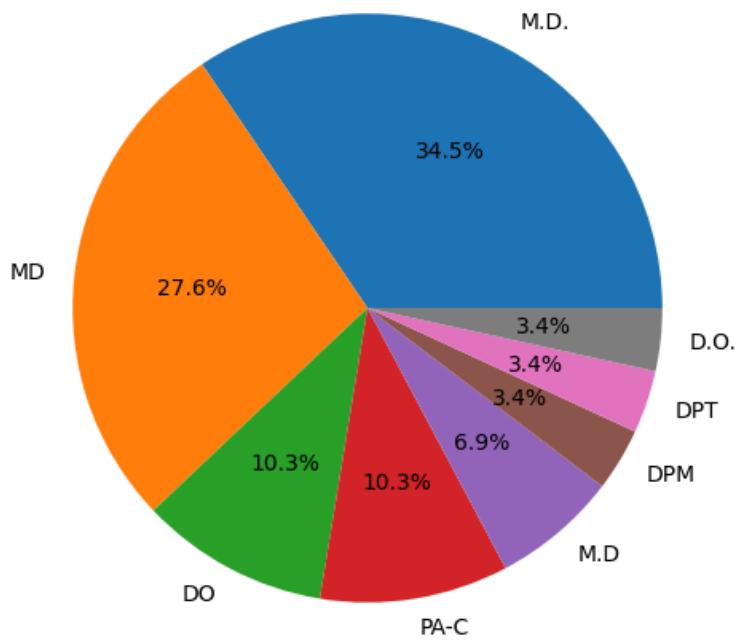
Count Plot of Credentials of the Provider



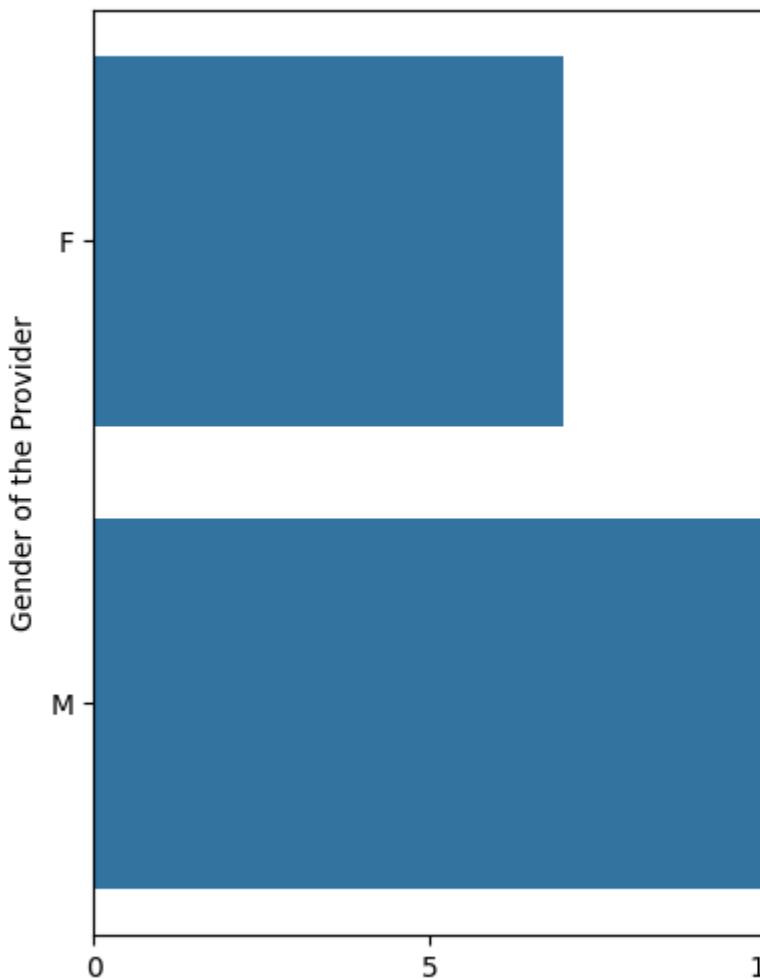
Bar Chart of Credentials of the Provider



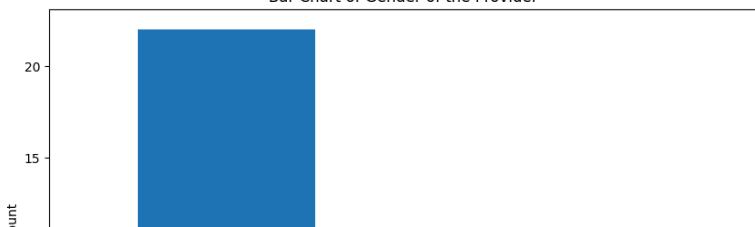
Pie Chart of Credentials of the Provider

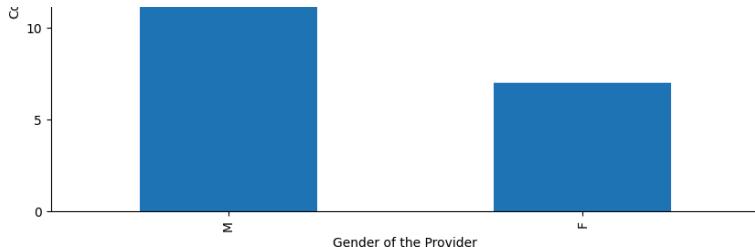


Count Plot c

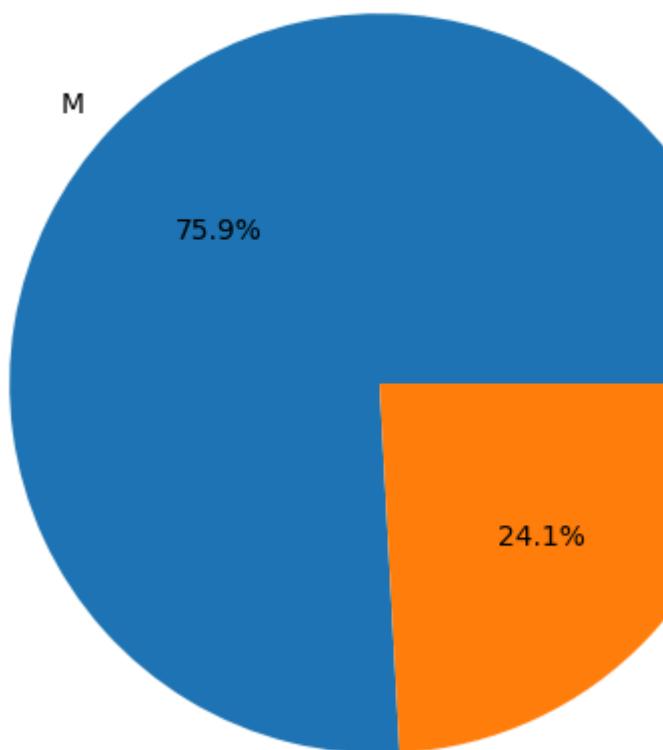


Bar Chart of Gender of the Provider

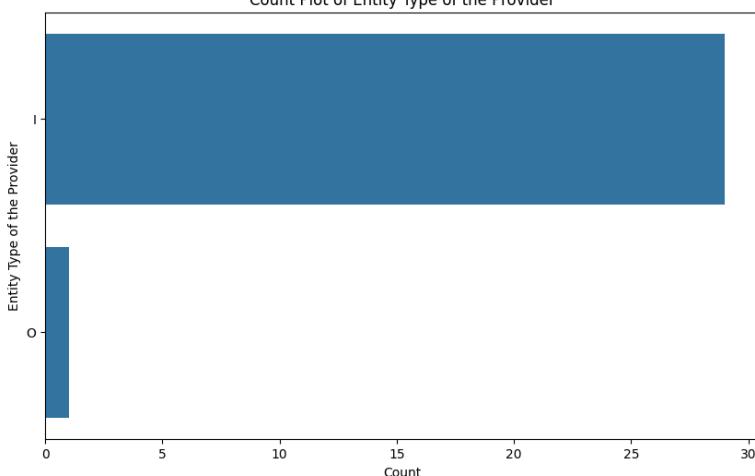




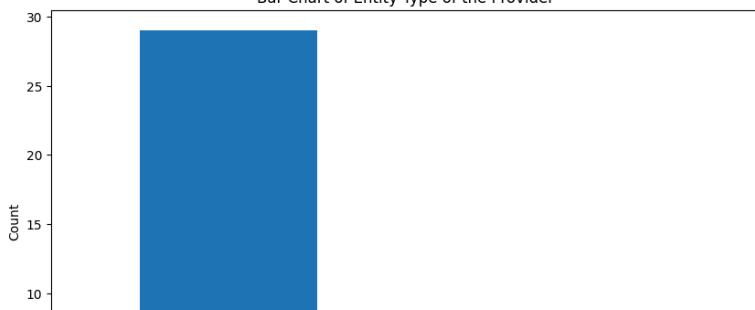
Pie Chart of Gender of the Provider

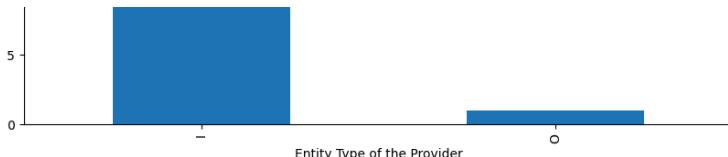


Count Plot of Entity Type of the Provider

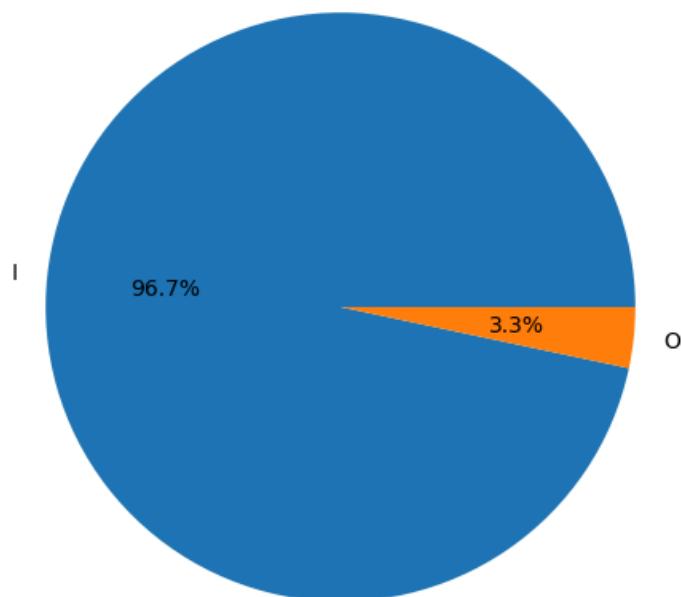


Bar Chart of Entity Type of the Provider

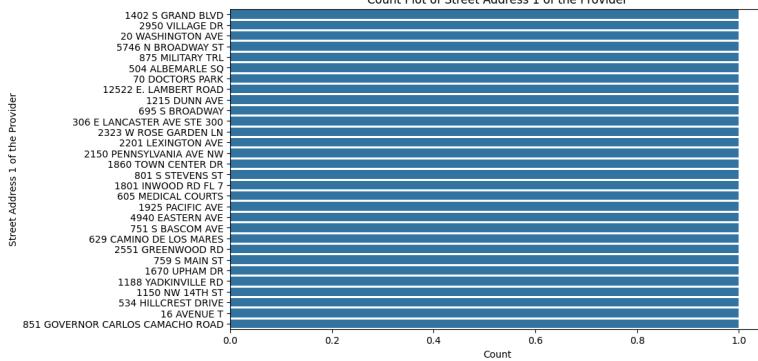




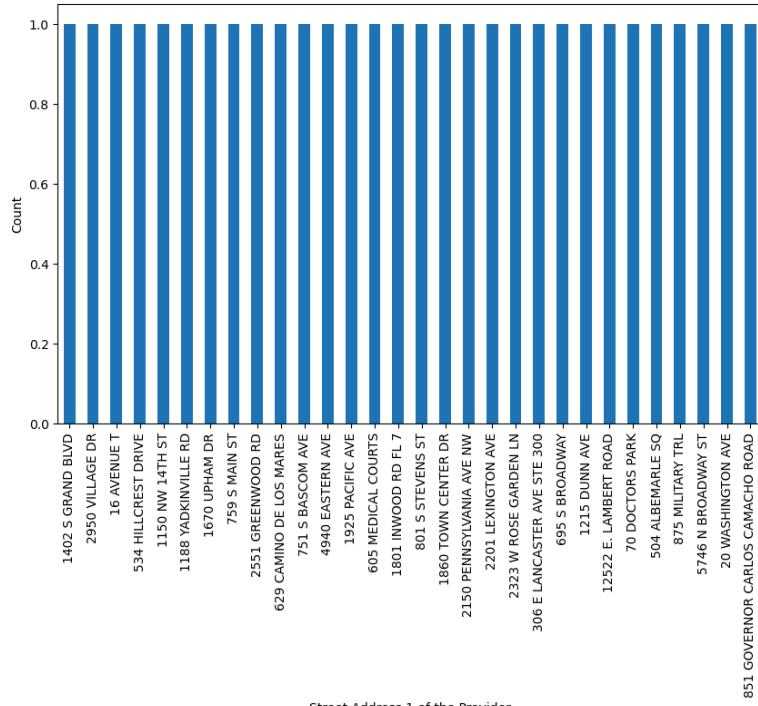
Pie Chart of Entity Type of the Provider



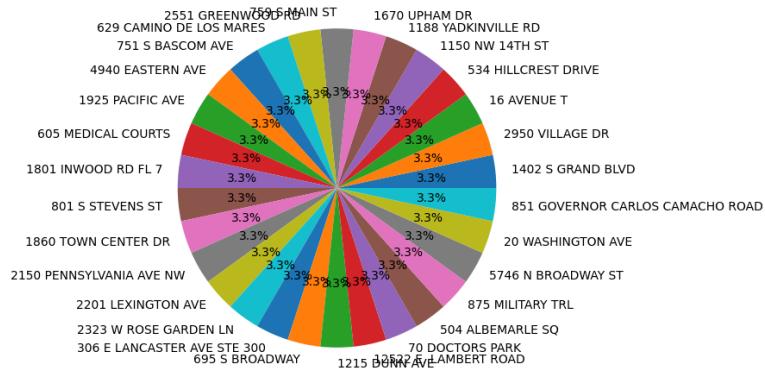
Count Plot of Street Address 1 of the Provider



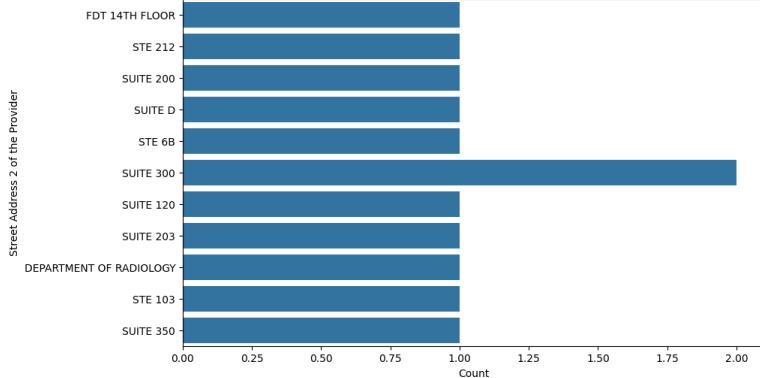
Bar Chart of Street Address 1 of the Provider



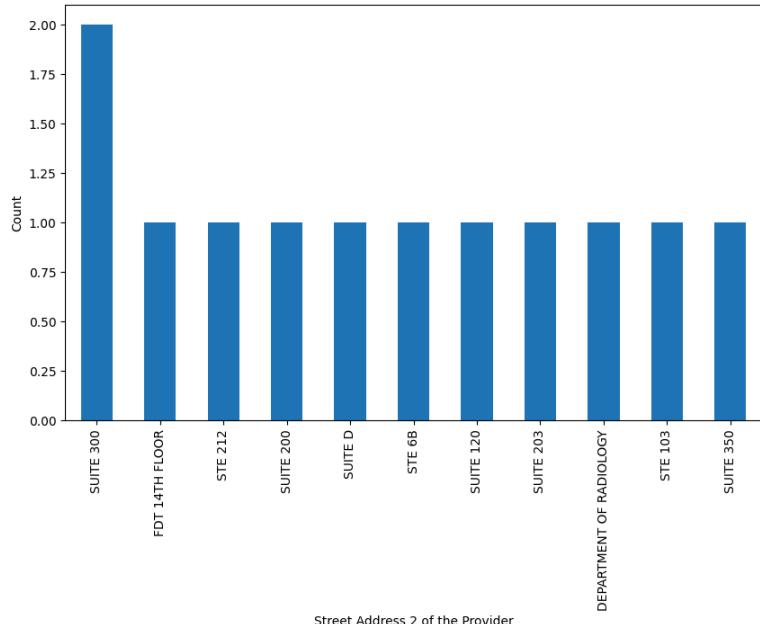
Pie Chart of Street Address 1 of the Provider



Count Plot of Street Address 2 of the Provider

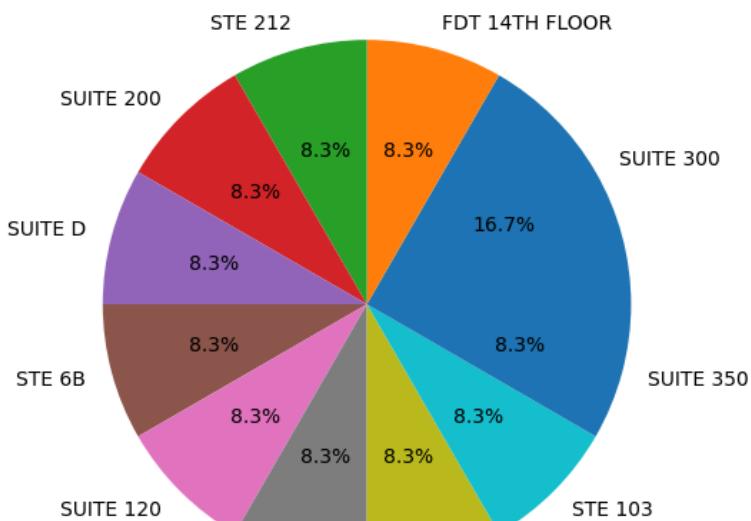


Bar Chart of Street Address 2 of the Provider



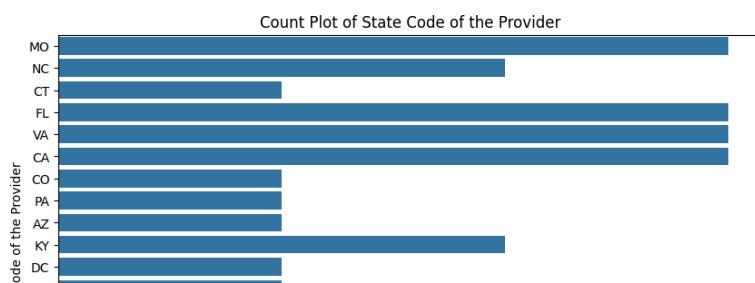
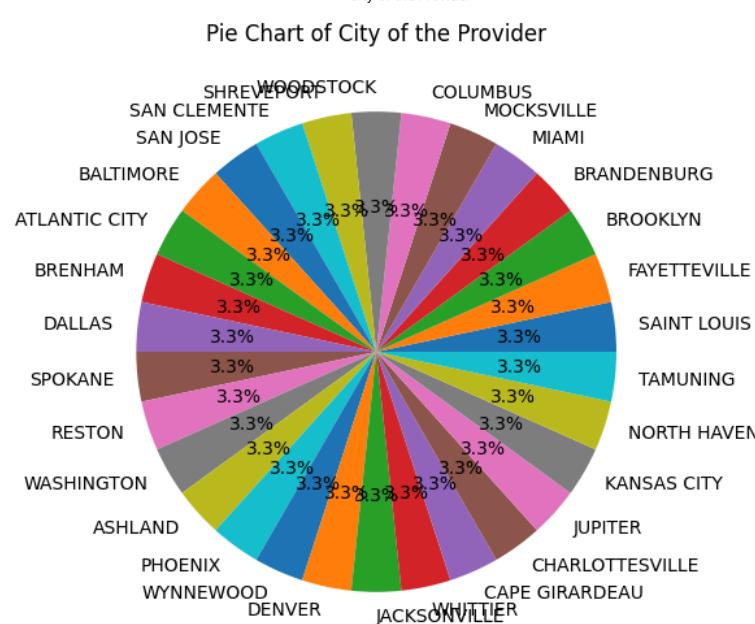
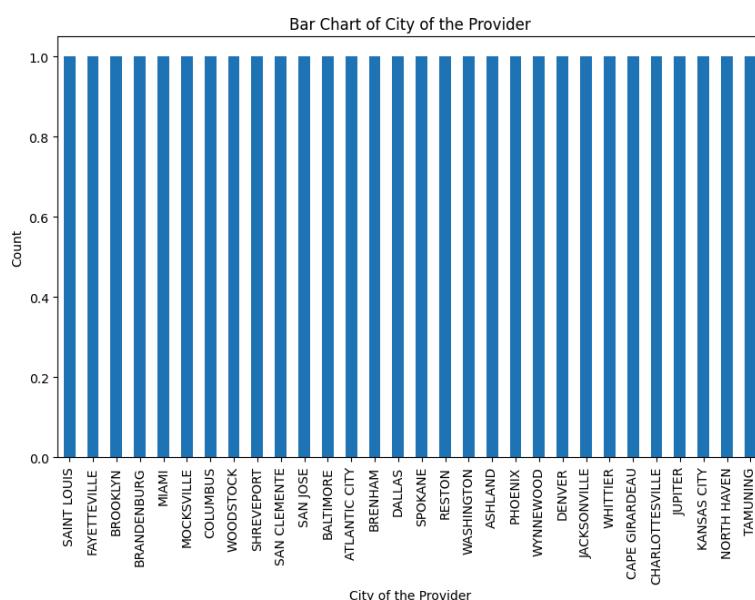
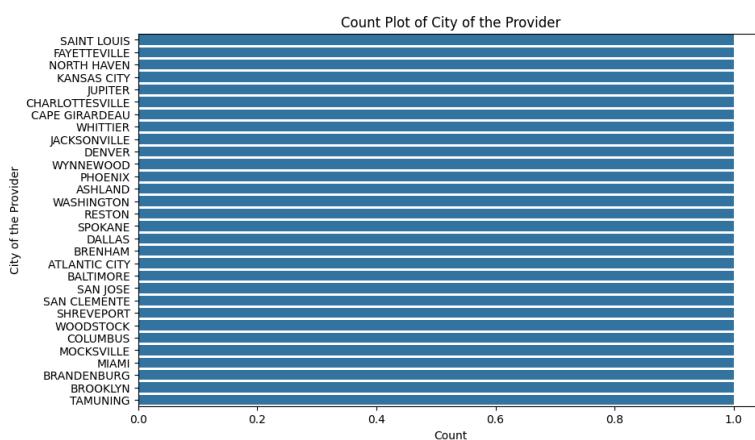
Street Address 2 of the Provider

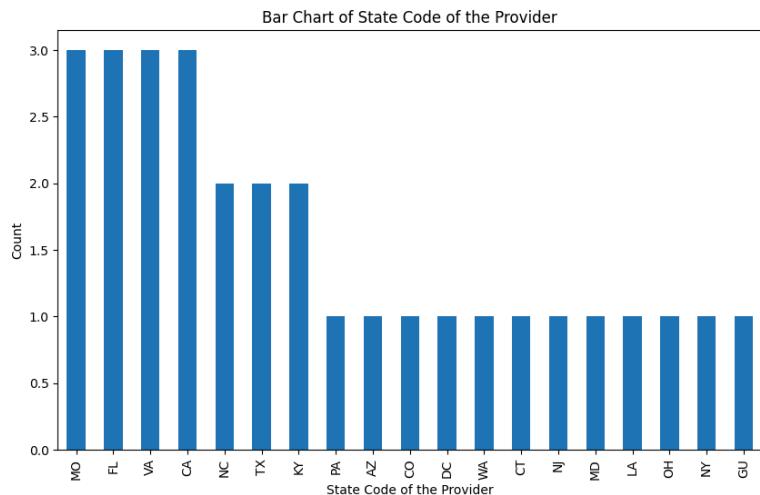
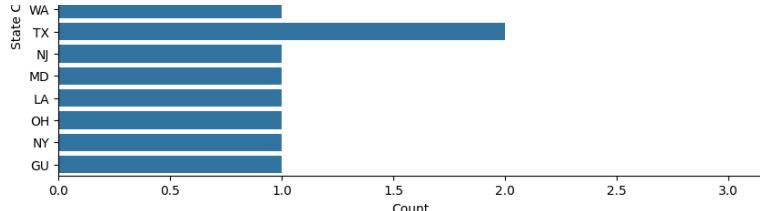
Pie Chart of Street Address 2 of the Provider



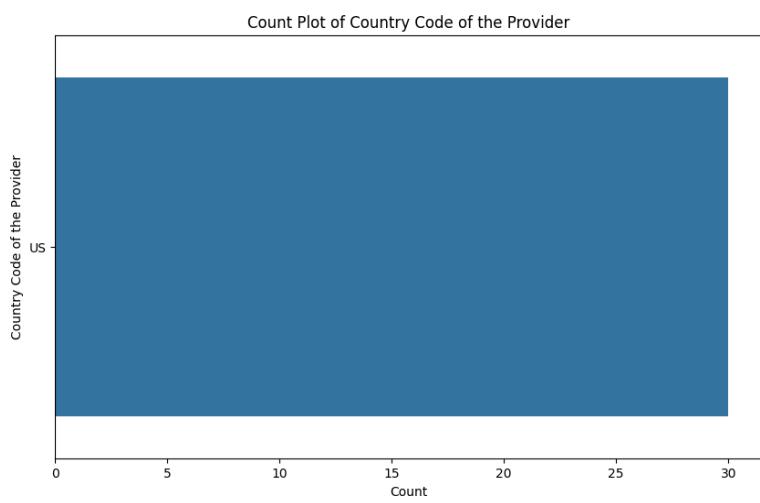
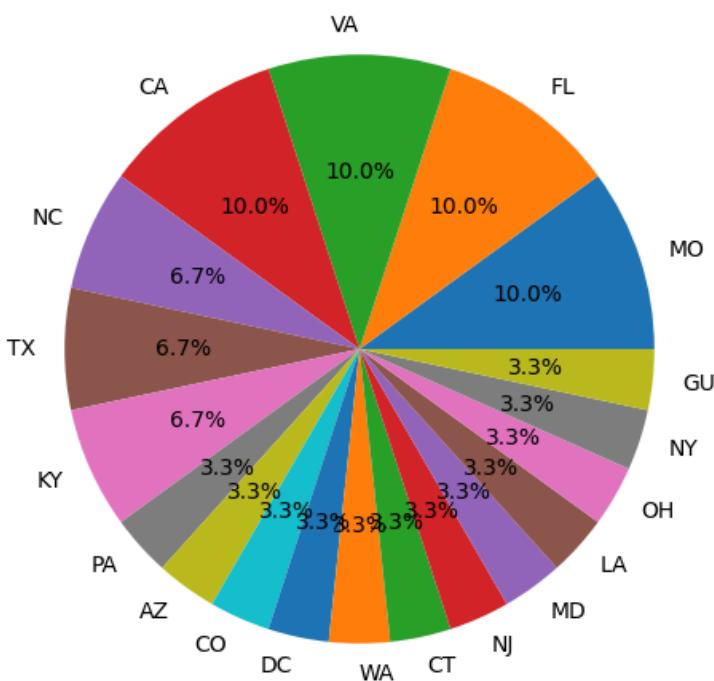
SUITE 203

DEPARTMENT OF RADIOLOGY



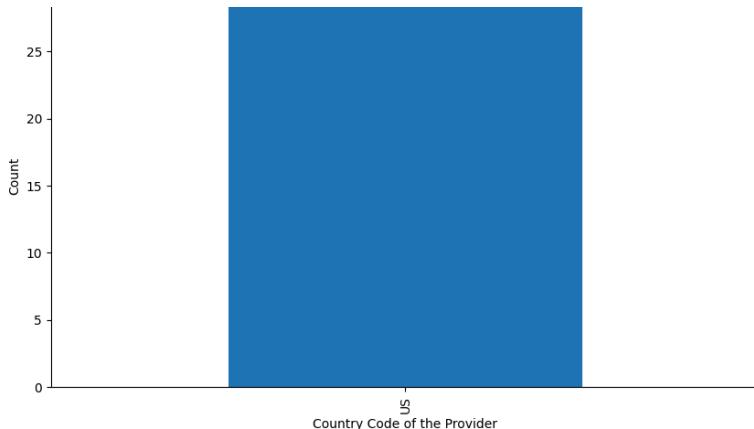


Pie Chart of State Code of the Provider

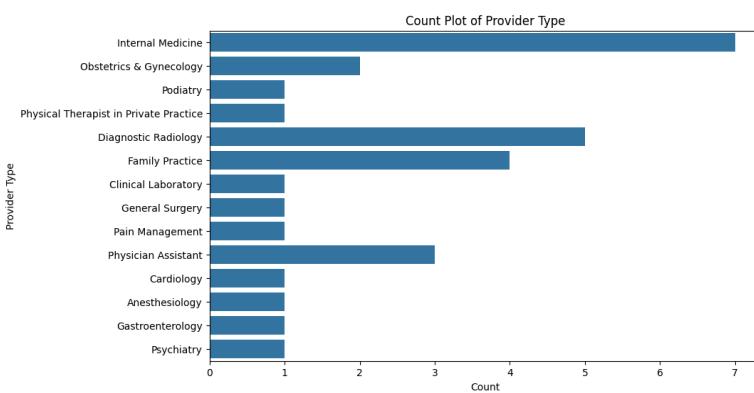
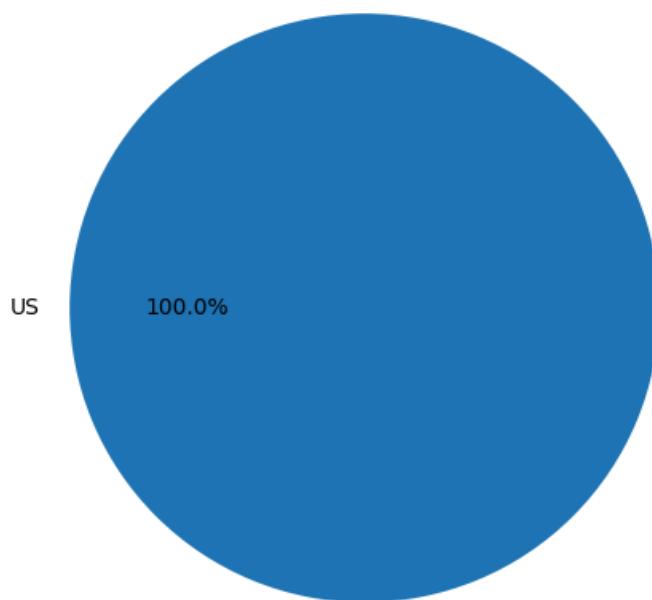


Bar Chart of Country Code of the Provider

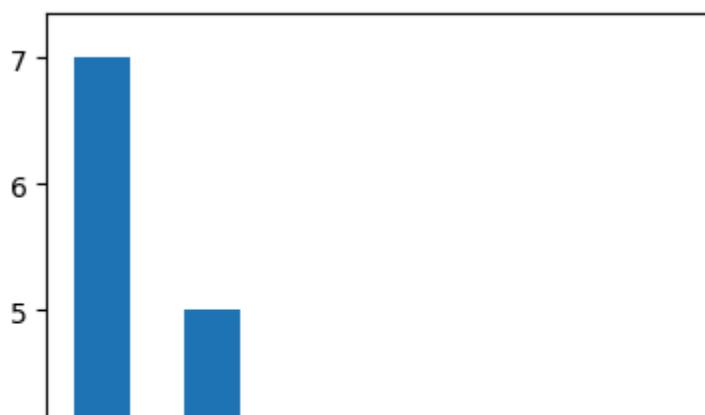


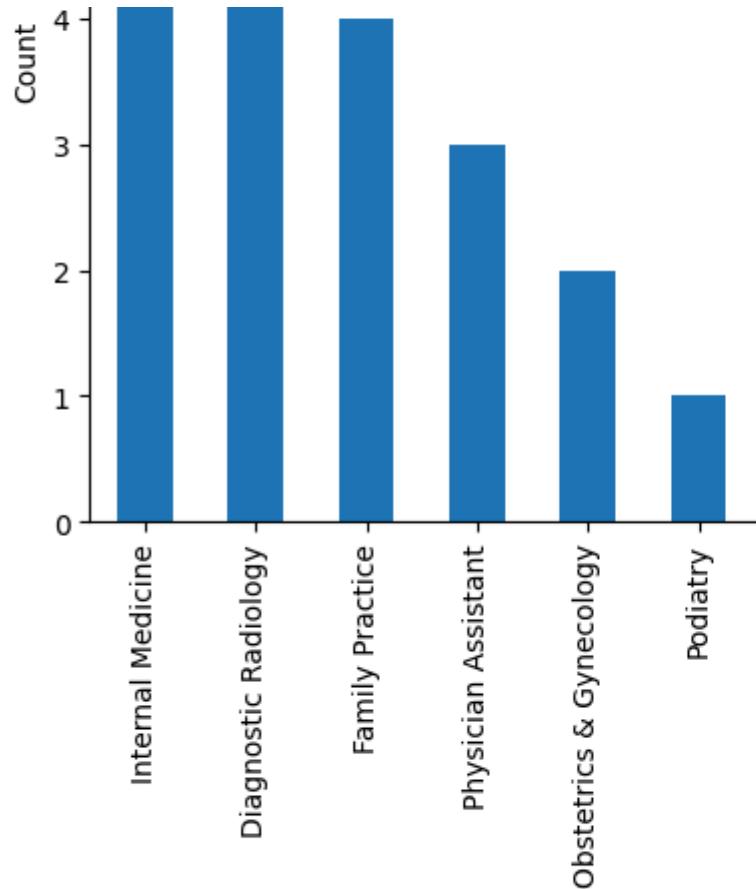


Pie Chart of Country Code of the Provider

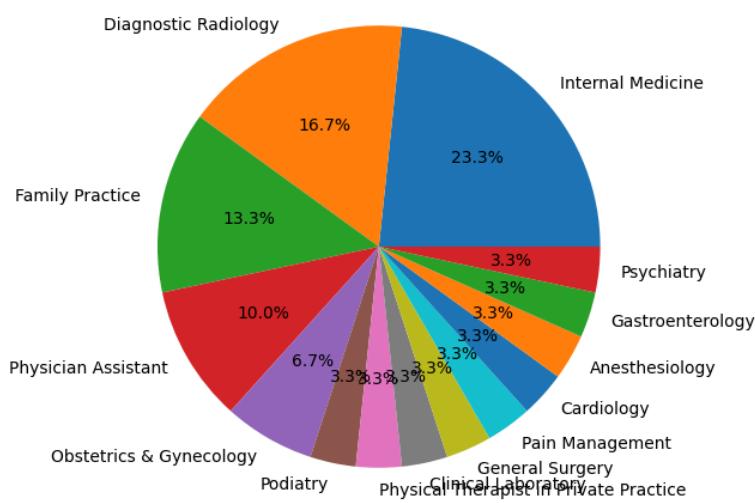


Bar Chart

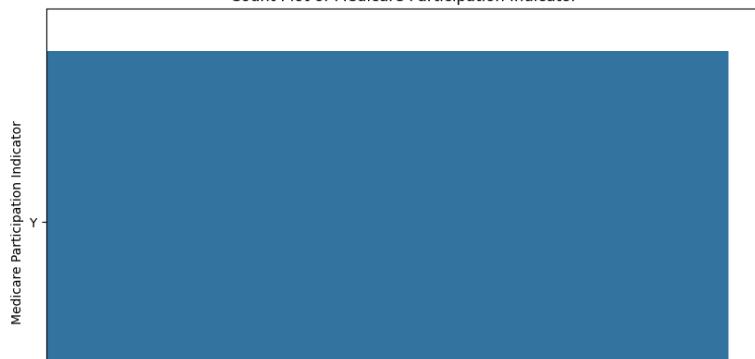


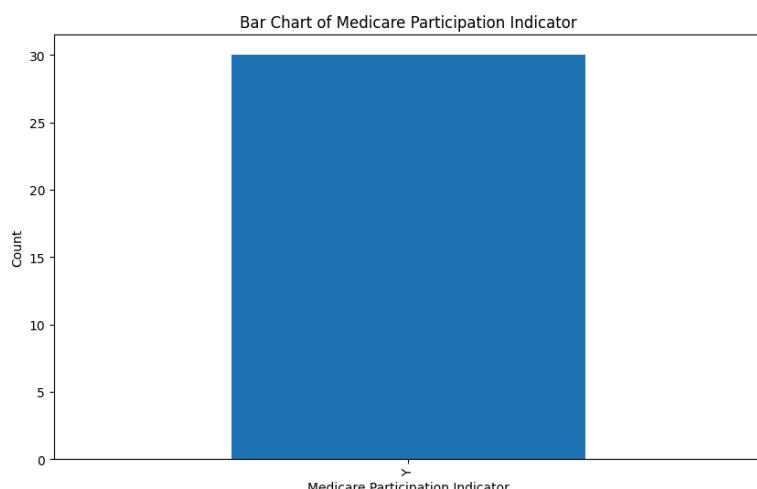
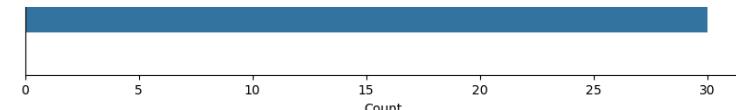


Pie Chart of Provider Type

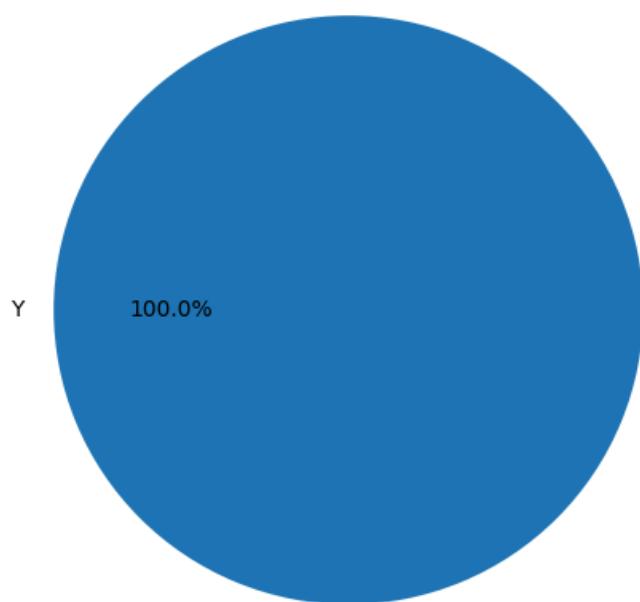


Count Plot of Medicare Participation Indicator

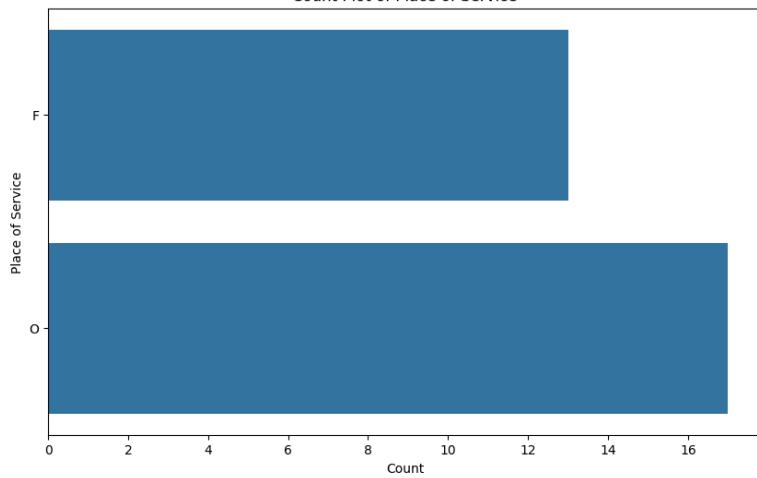




Pie Chart of Medicare Participation Indicator

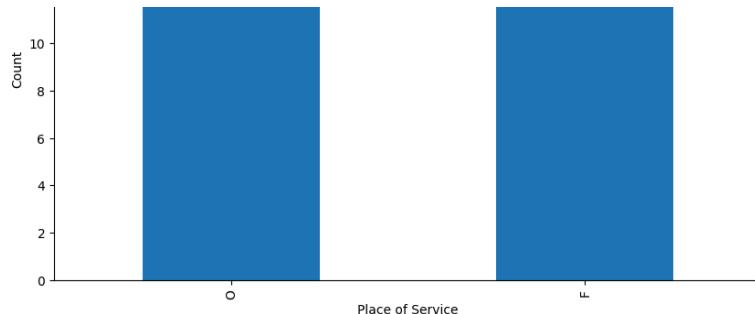


Count Plot of Place of Service

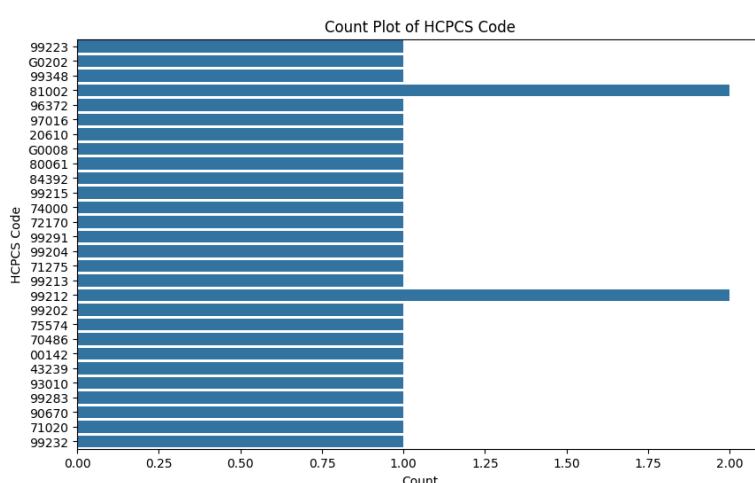
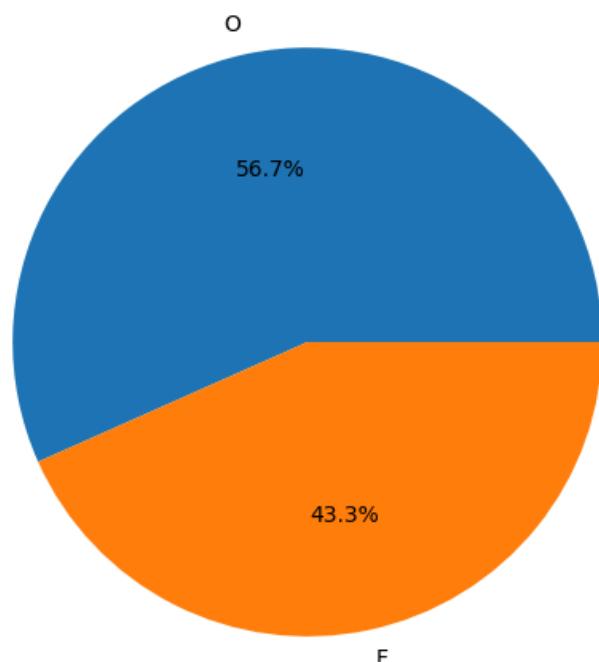


Bar Chart of Place of Service

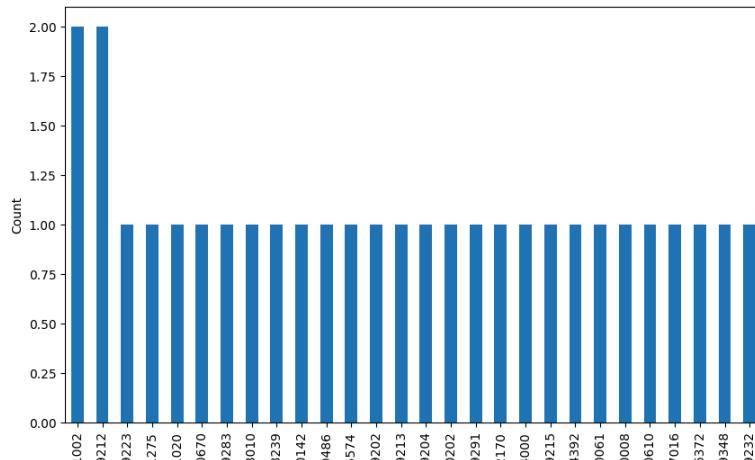


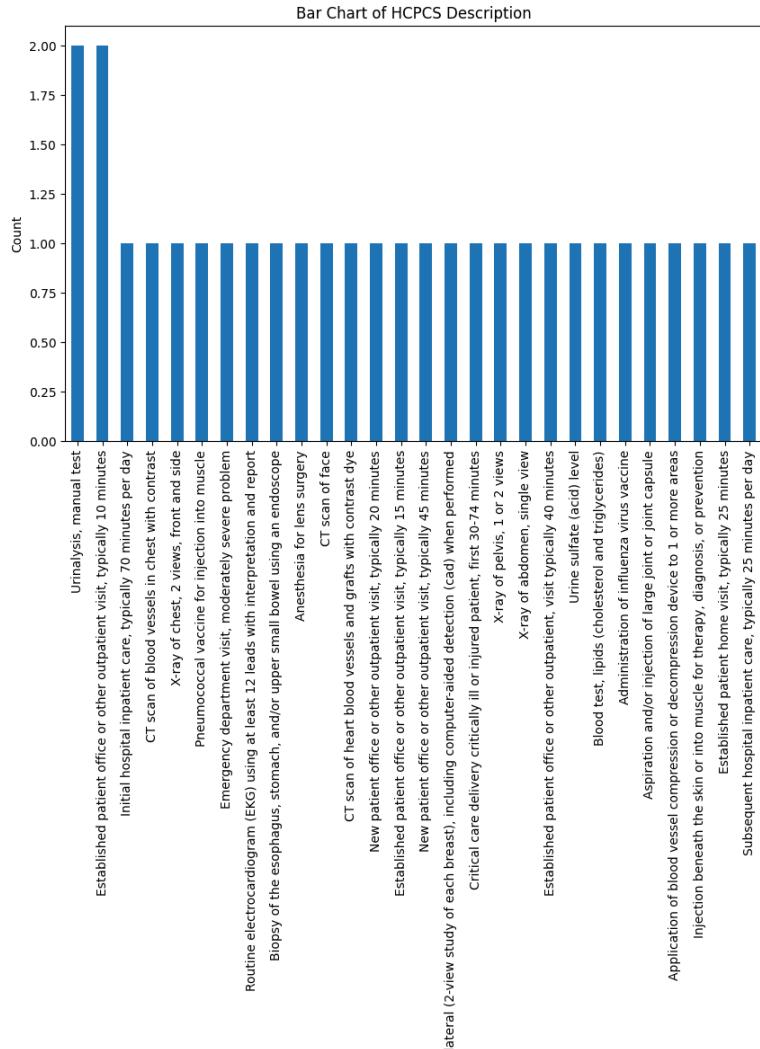
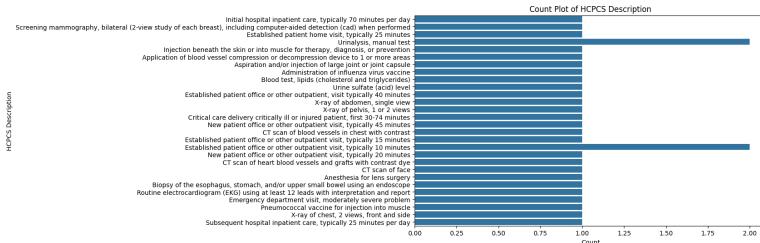
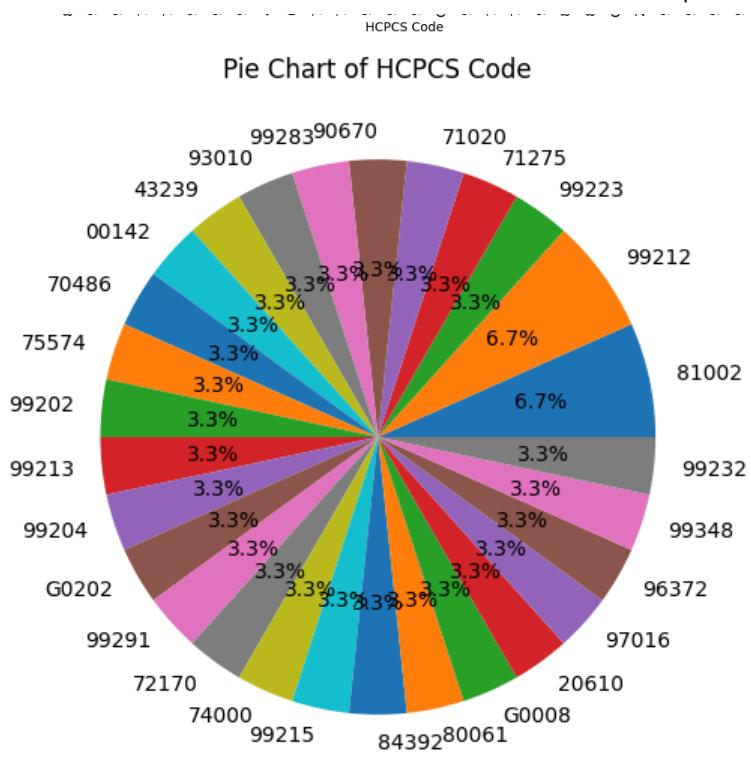


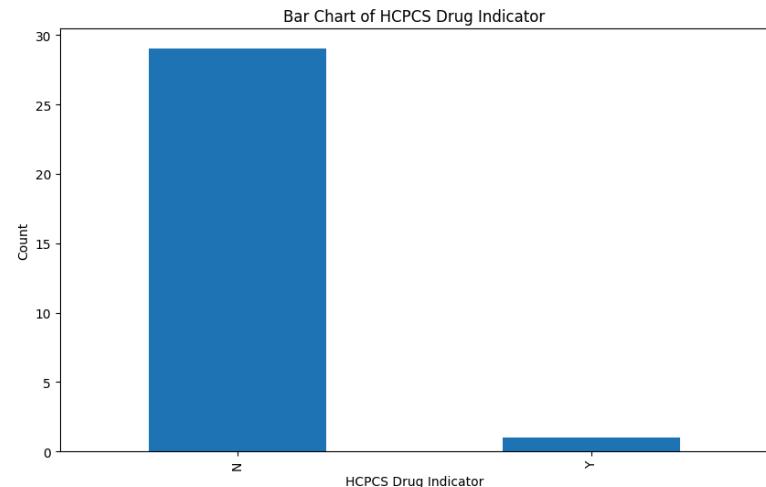
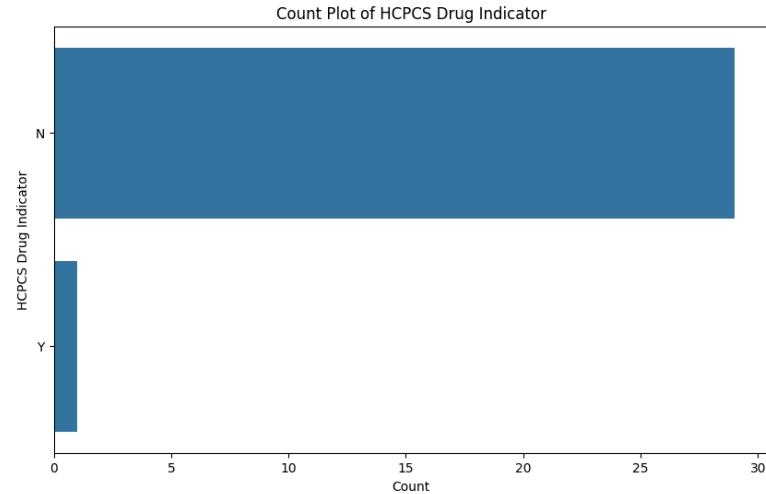
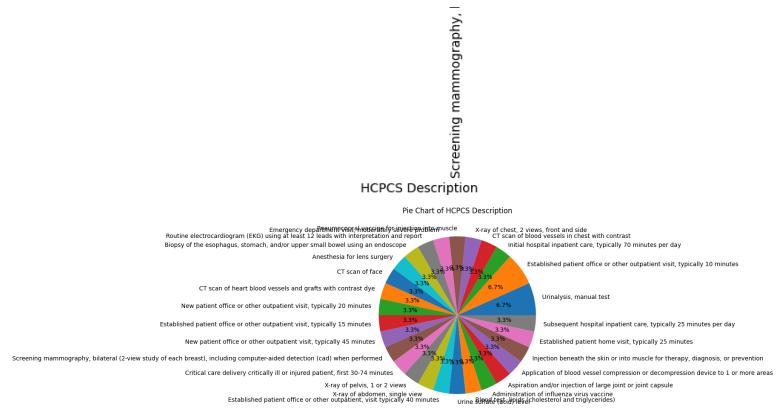
Pie Chart of Place of Service



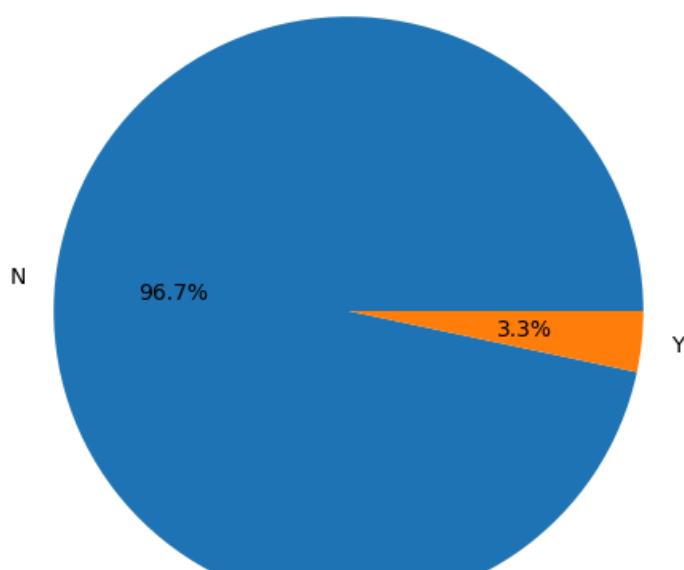
Bar Chart of HCPCS Code

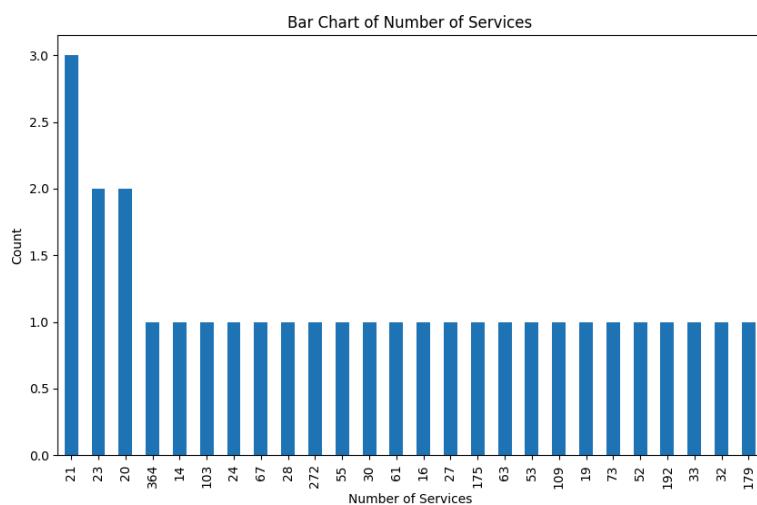
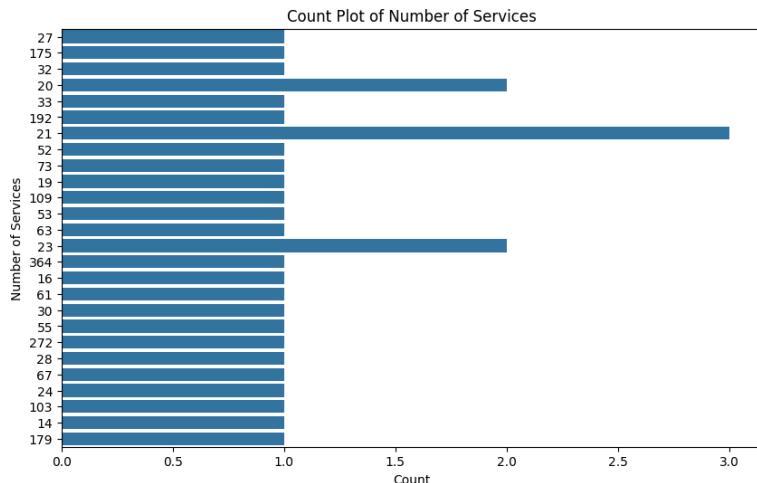




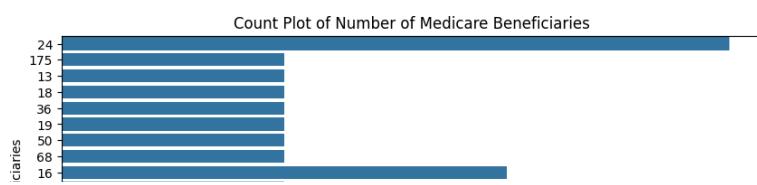
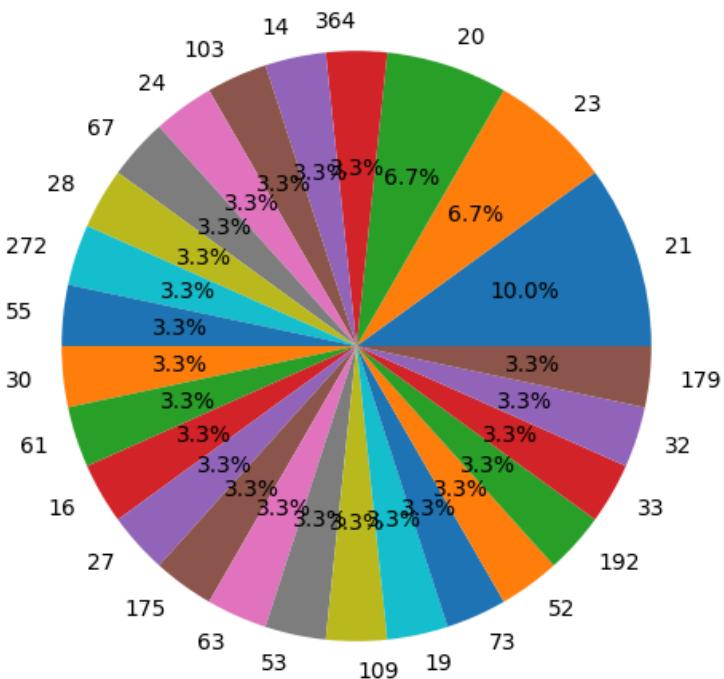


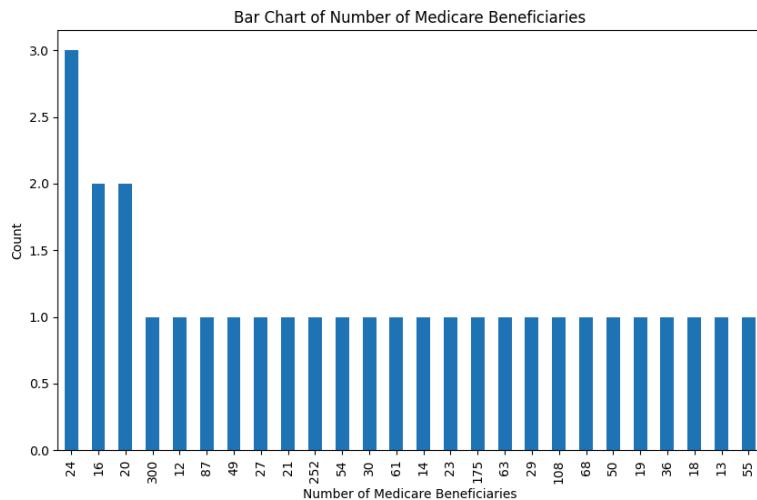
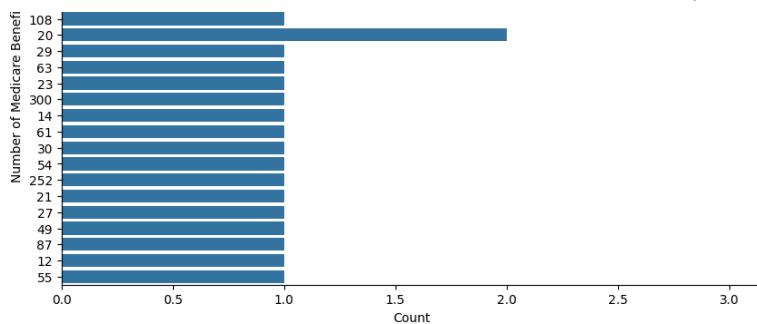
Pie Chart of HCPCS Drug Indicator



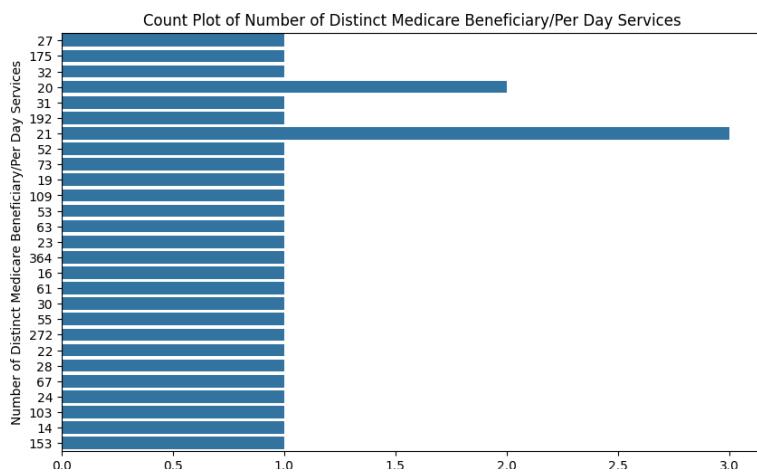
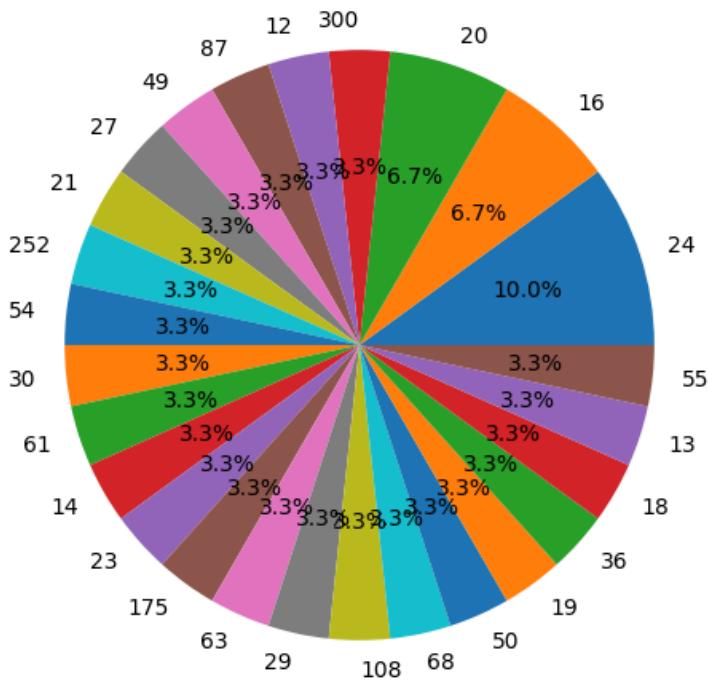


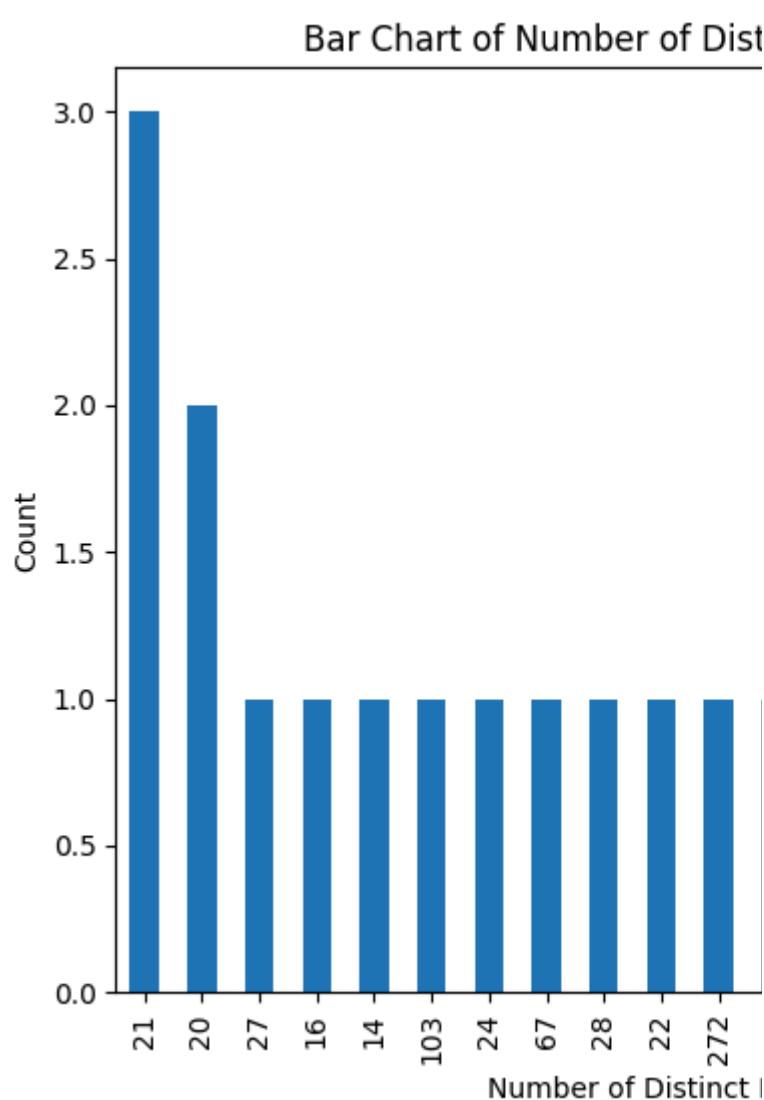
Pie Chart of Number of Services



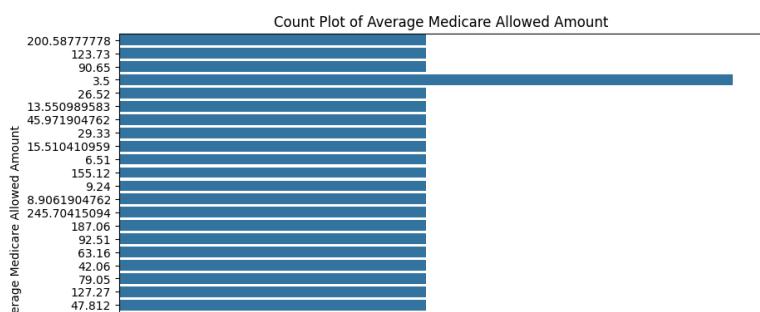
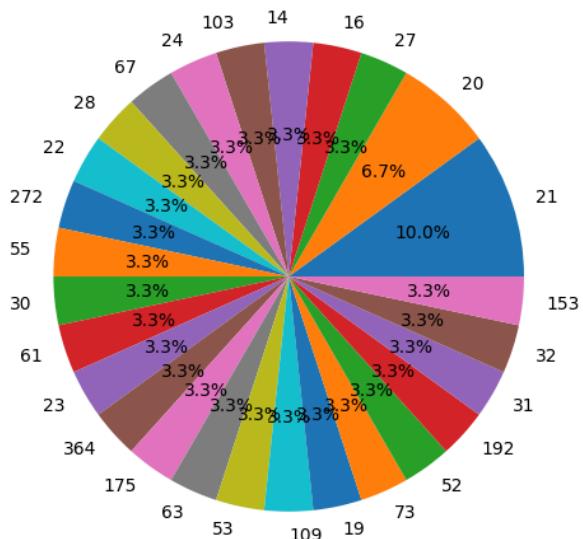


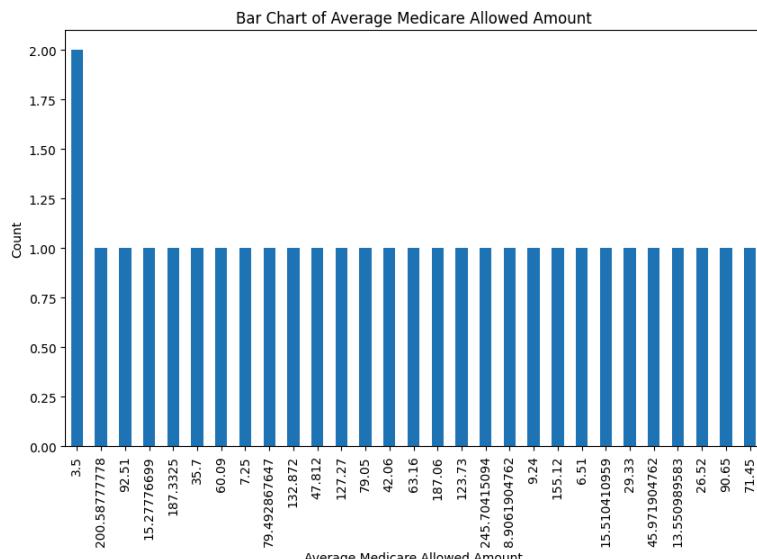
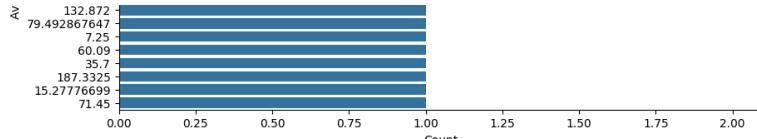
Pie Chart of Number of Medicare Beneficiaries



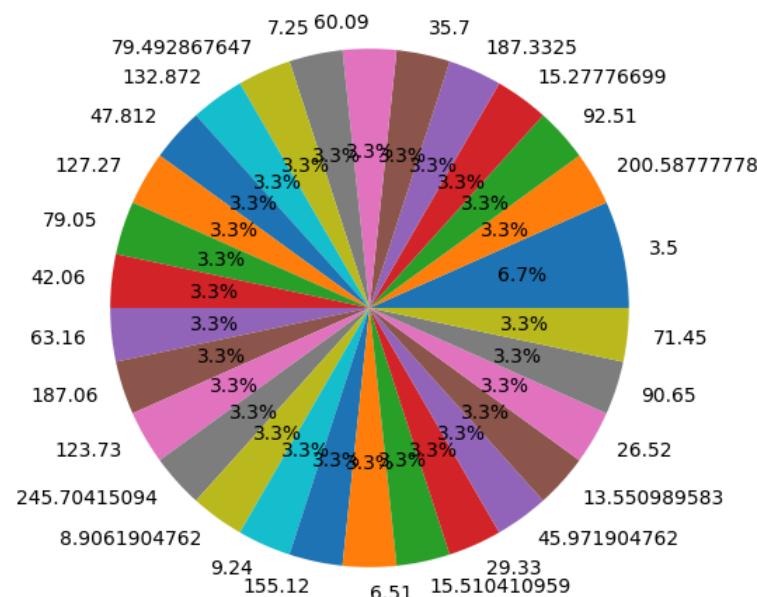


Pie Chart of Number of Distinct Medicare Beneficiary/Per Day Services

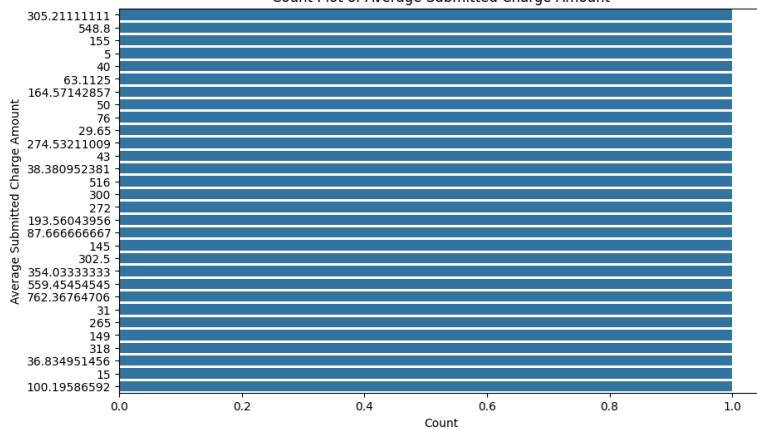




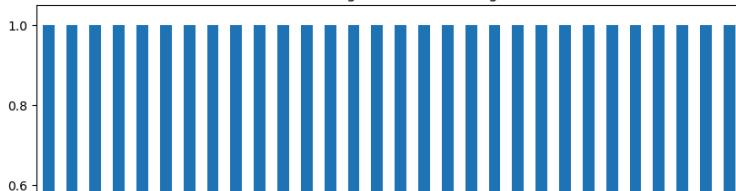
Pie Chart of Average Medicare Allowed Amount

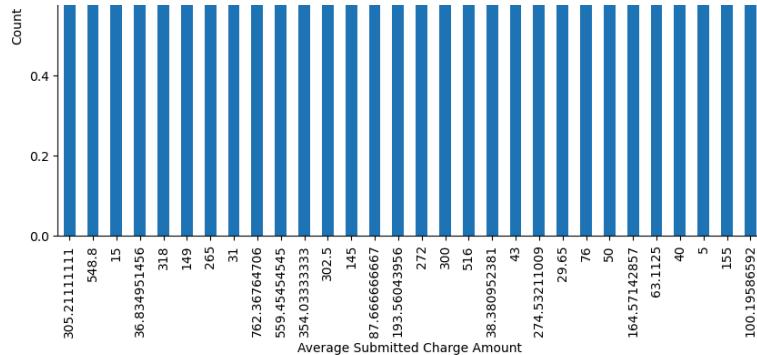


Count Plot of Average Submitted Charge Amount

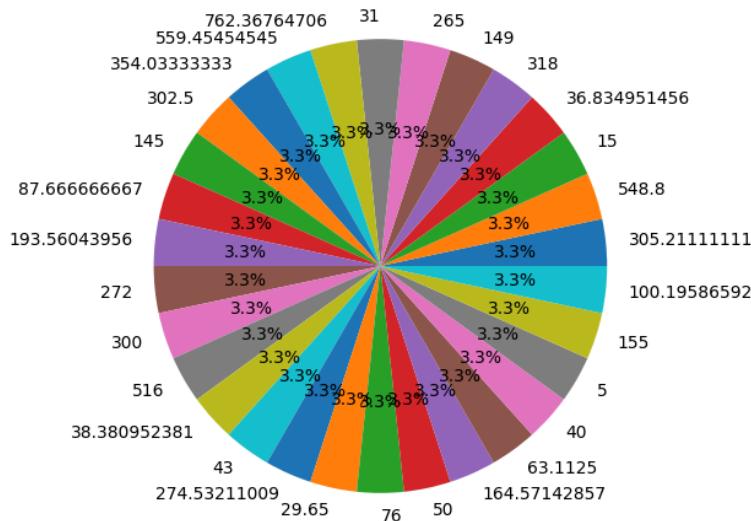


Bar Chart of Average Submitted Charge Amount

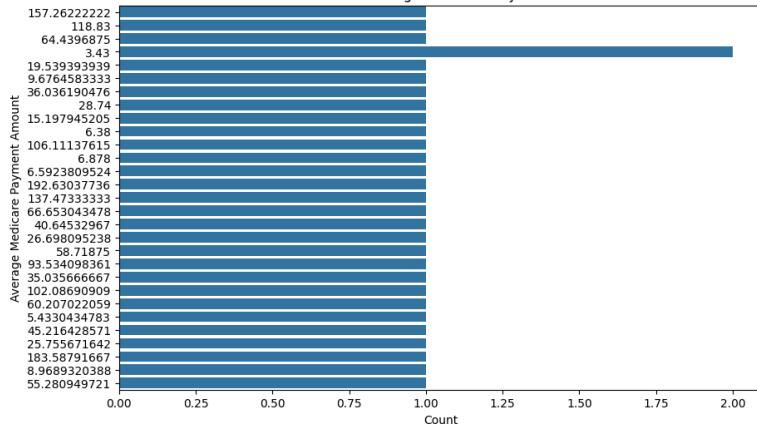




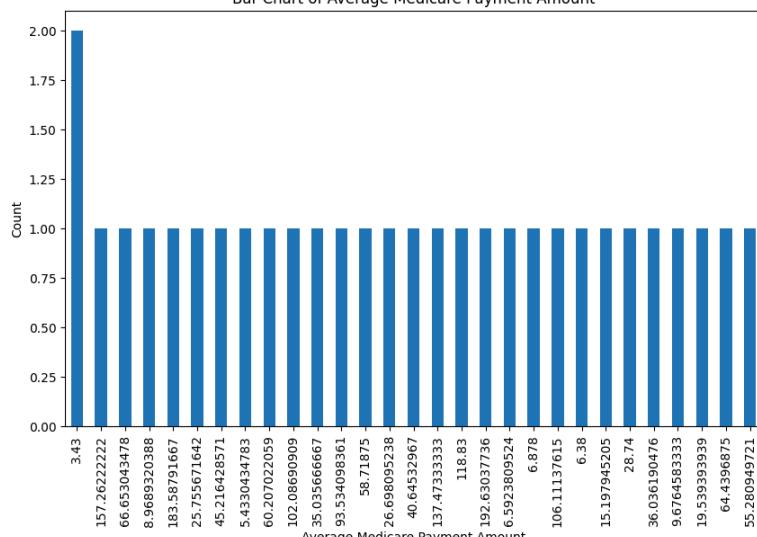
Pie Chart of Average Submitted Charge Amount



Count Plot of Average Medicare Payment Amount

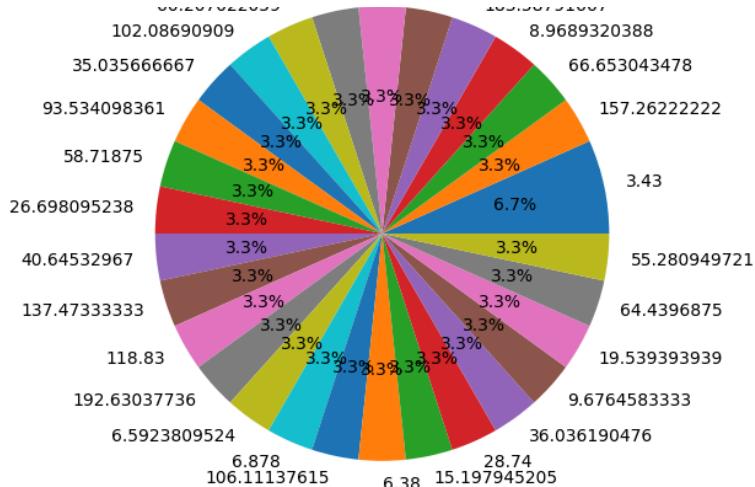


Bar Chart of Average Medicare Payment Amount

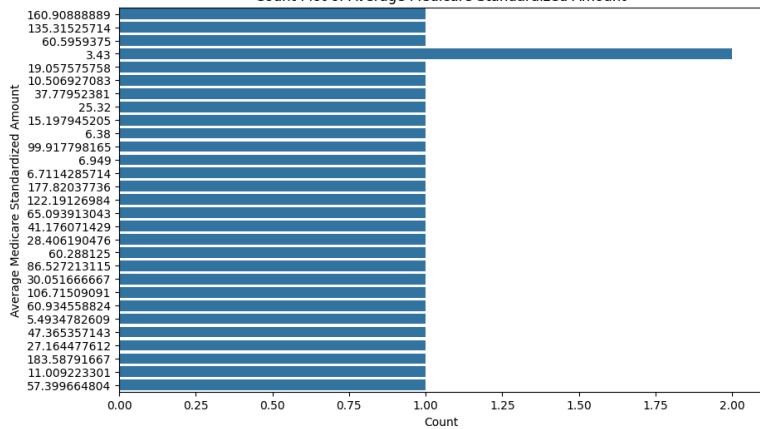


Pie Chart of Average Medicare Payment Amount

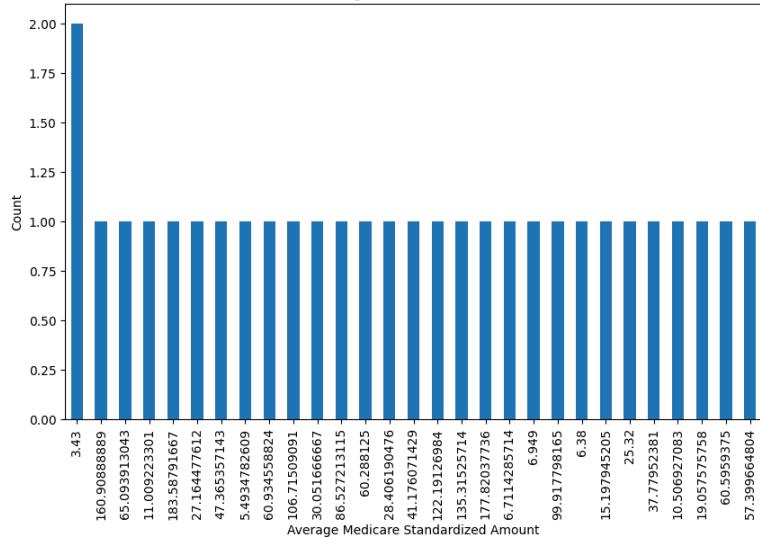
5.4330434783 25.755671642
60.207022059 183.58791667



Count Plot of Average Medicare Standardized Amount



Bar Chart of Average Medicare Standardized Amount



Pie Chart of Average Medicare Standardized Amount

