**Healthcare Providers in US**

**Problem Statement**

The dataset link : [Healthcare Providers.xlsx](https://1drv.ms/x/c/d4d8a221929f9a13/Ee2YqJ5YCaxKvSTlAuZlE5wBcQGeleSLHHLezqzzAVFeIQ?e=GoMbRB)

The problem statement of this dataset is to analyze and improve the quality of healthcare services provided by various healthcare providers in US. This involves identifying any anomalies in the dataset that might indicate issues such as errors in data entry, unusual patterns in healthcare provision, or other irregularities that could affect the quality of service in US Government.

The above Healthcare Providers dataset it containing 26 columns and 100001 rows are present and there are 7 Categorical Columns and 5 Numerical Columns

**Categorical Columns**

**Middle Initial of the Provider**:

* he middle initial of the healthcare provider.
* Helps in uniquely identifying providers, especially when there are common first and last names.

**Gender of the Provider**:

* The gender of the healthcare provider .
* Most of the possible values are Male, Female and none according to the healthcare providers dataset.
* Used for demographic analysis and understanding gender distribution among healthcare providers.

**Entity Type of the Provider**:

* Indicates whether the provider is an individual or an organization.
* Most of the possible values are Individual, Organization.
* It Helps differentiate between individual practitioners and group practices or healthcare facilities.

**State Code of the Provider**:

* The state where the provider is located.

**Medicare Participation Indicator**:

* It Indicates whether the provider participates in Medicare.
* Most of the possible values are Yes, No.
* It Helps in identifying providers who accept Medicare, which can be crucial for patients relying on Medicare coverage.

**Place of Service**:

* Indicates where the service was provided.
* Most of the possible values are Office, Facility.
* It Used to analyze the setting of healthcare services, whether in an office or a facility like a hospital.

**HCPCS Drug Indicator**:

* Indicates if the service involves a drug, as identified by the HCPCS code.
* Y (Yes), N (No).
* It Helps in distinguishing services that include drug administration from other types of services.

**Provider Type and Experience**

* Different provider types might have varying levels of experience, with some specializations potentially requiring more years of experience.

**Specialization**

* Analyzing gender and experience within specializations can provide deeper insights into specific fields.

Through understanding the labels and their associated data, we can gain valuable insights into the composition, trends, and disparities within the healthcare workforce. And Each label provides a different dimension of analysis, helping you understand various aspects like age providers etc of the dataset comprehensively.

Before analyzing the dataset we need do some pre-processing steps

**Pre-Processing steps**

1 . Load the Data

import pandas as pd

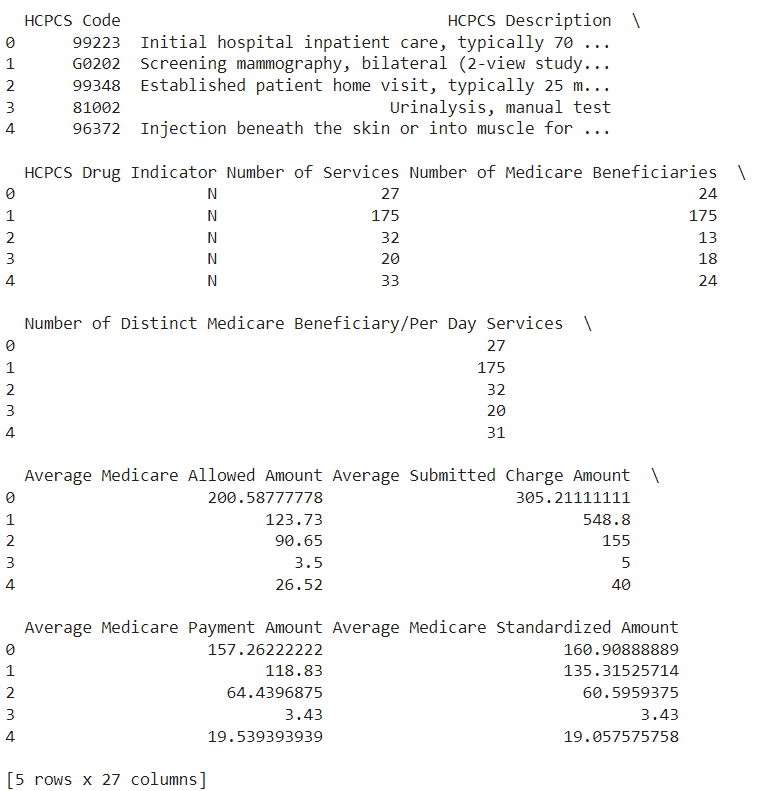
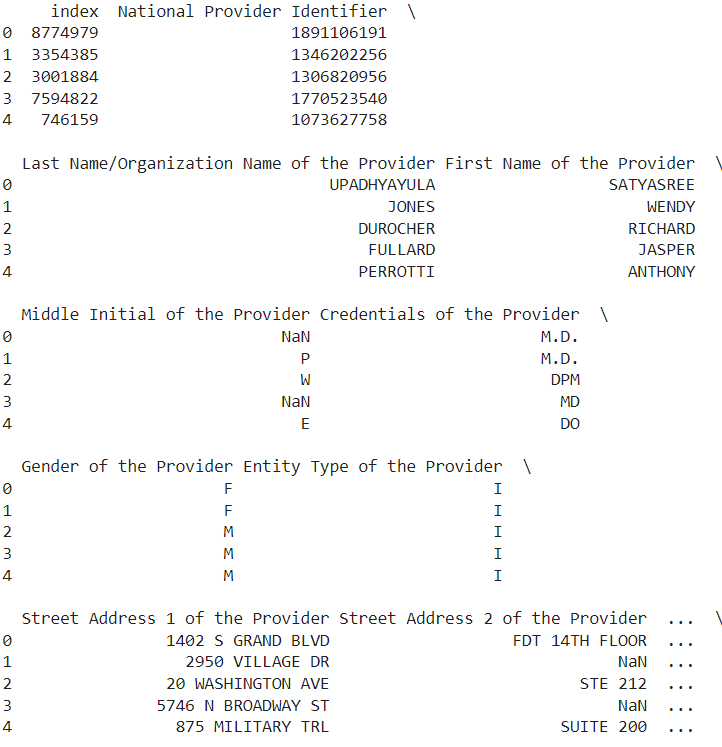
import matplotlib.pyplot as plt

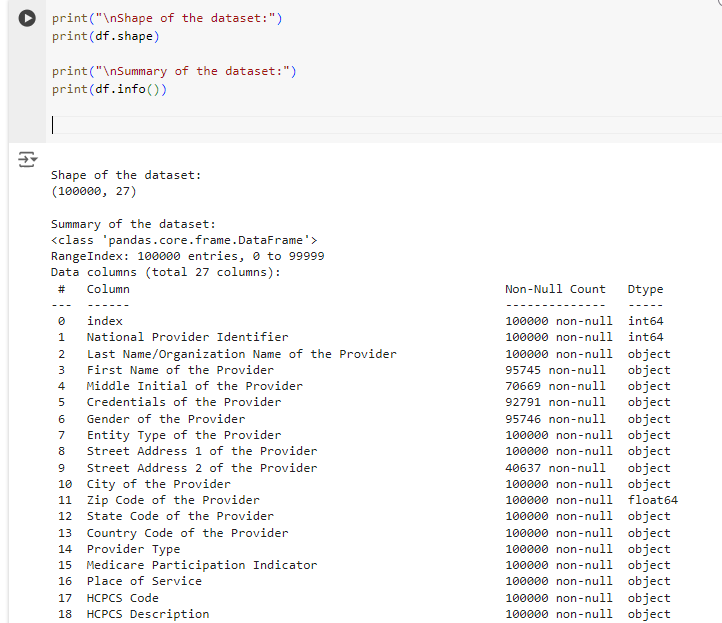
import seaborn as sns

file\_path = '/content/Healthcare Providers.csv'

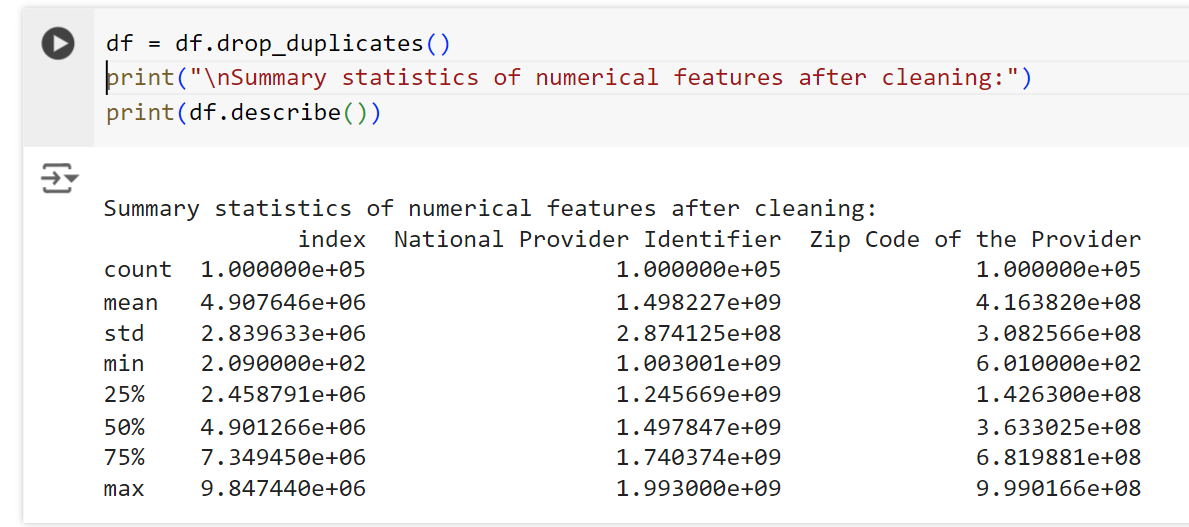
df = pd.read\_csv(file\_path)

print(df.head())



2 . Handle Missing Values

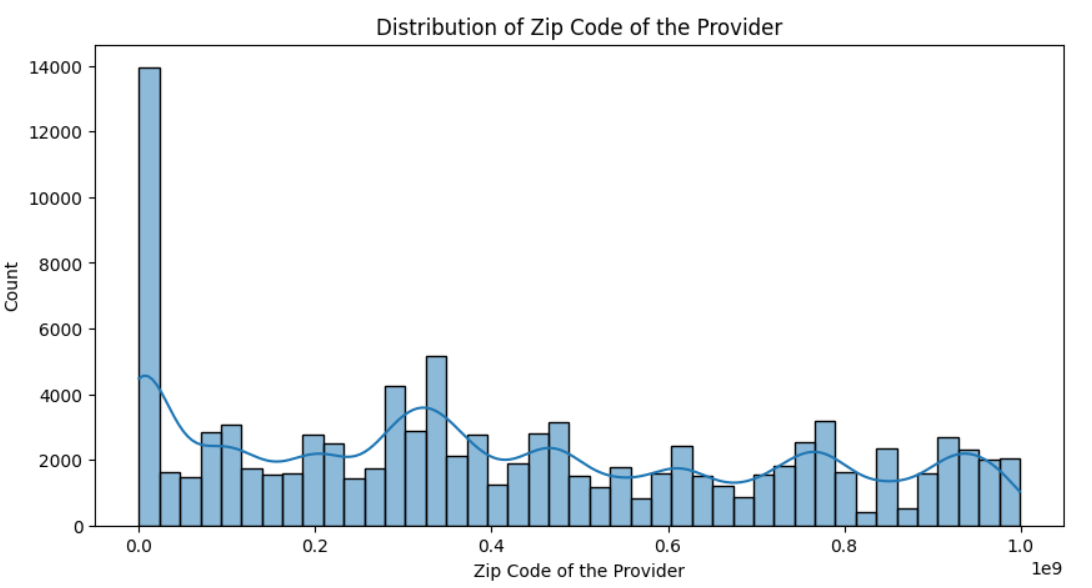
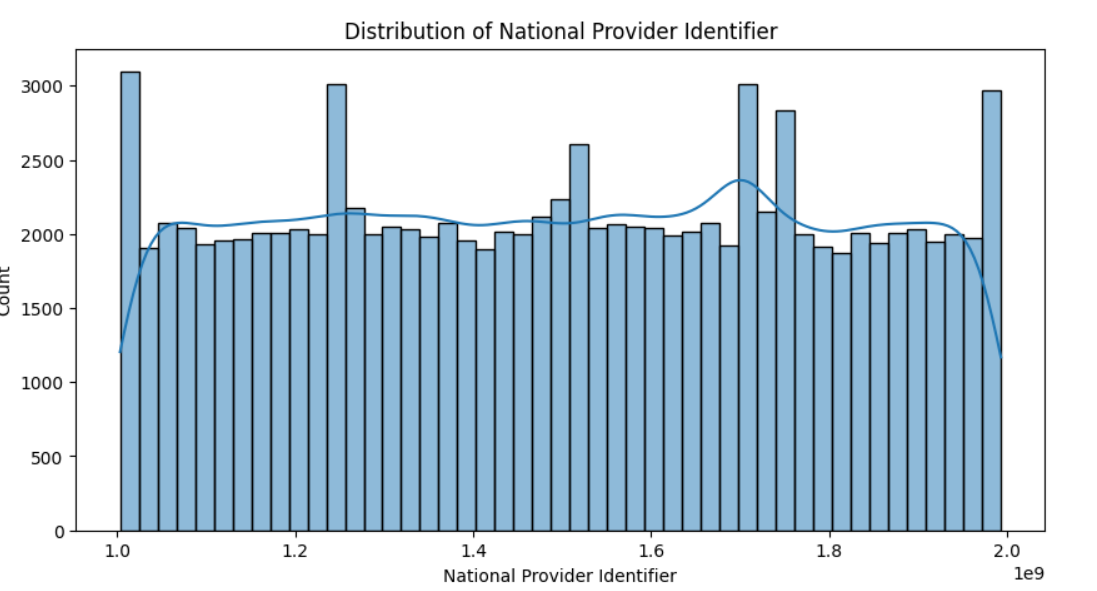
3 . Data Cleaning



4. Feature Engineering(for the above dataset there no use of to perform Feature Engineering)

**Exploratory Data Analysis (EDA)**

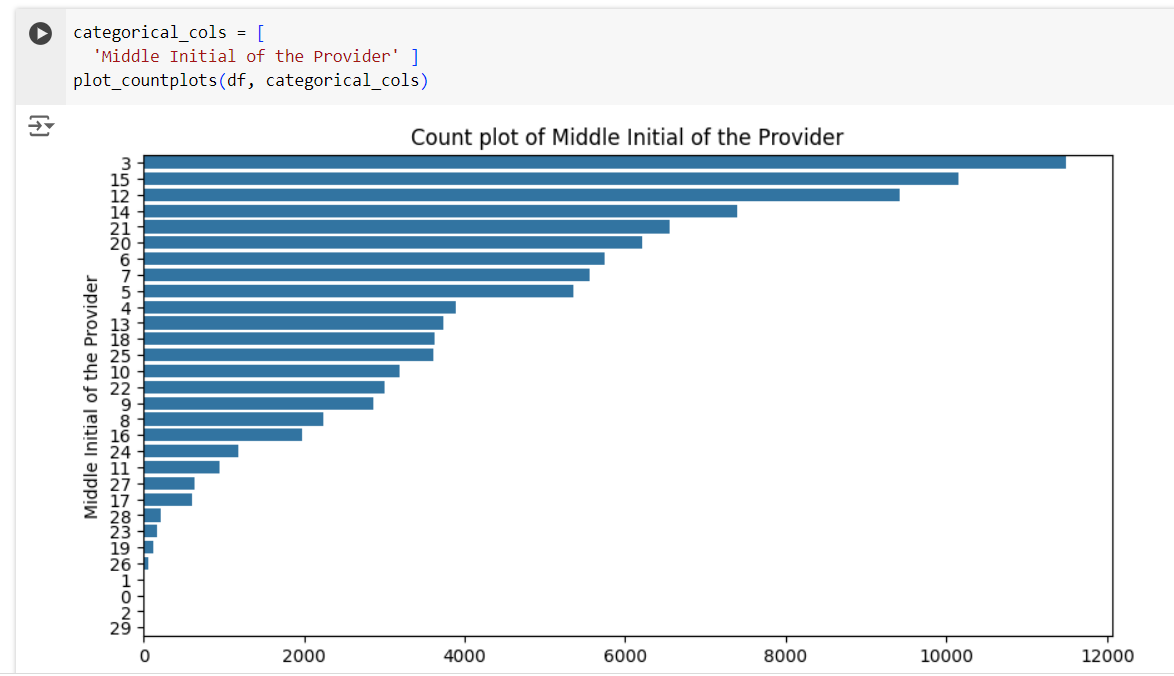
Exploratory Data Analysis (EDA) is an analysis approach that identifies general patterns in the data. These patterns include outliers and features of the data that might be unexpected. Exploratory Data Analysis (EDA) is a crucial initial step in data science projects. It involves analyzing and visualizing data to understand its key characteristics, uncover patterns, and identify relationships between variables refers to the method of studying and exploring record sets to apprehend their predominant traits, discover patterns, locate outliers, and identify relationships between variables.



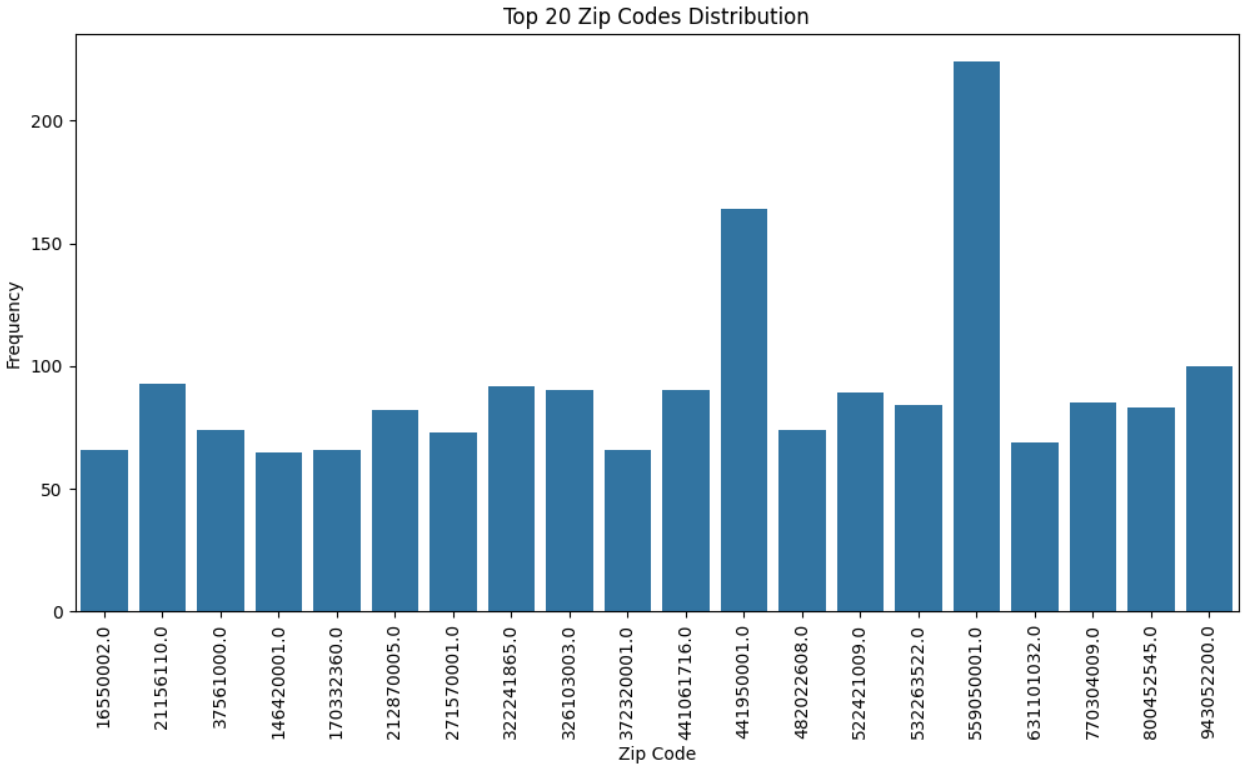
* In this above graph zipcode intially considered as numerical values, so that we need to this to a catagerical value because zipcode should not be considered as a numerical value.

**Count Plot of Middle Initial of the Provider**

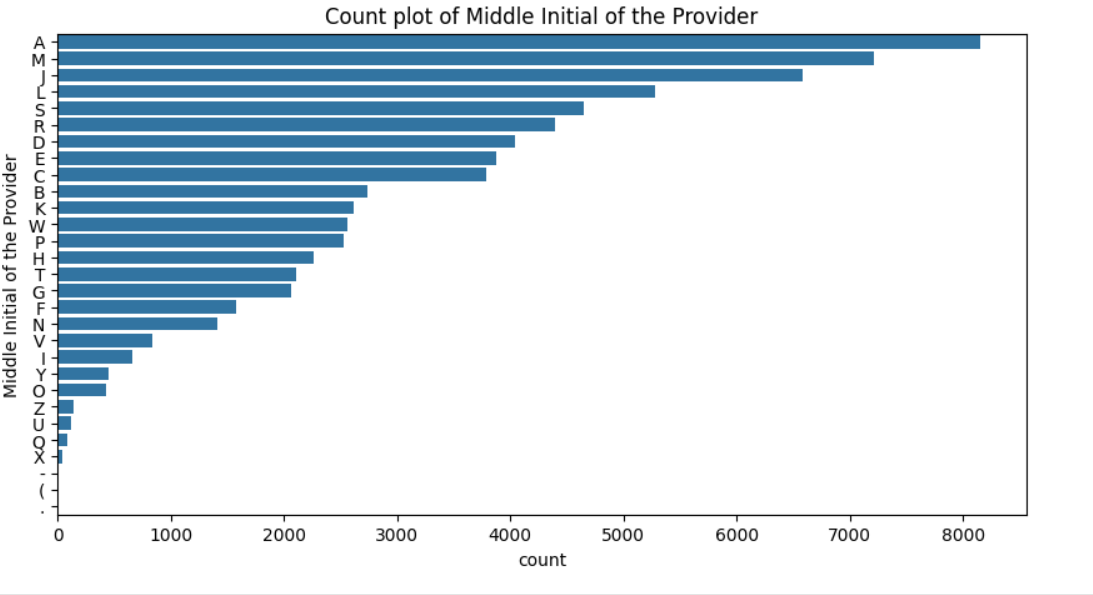
* Many healthcare providers do not have a middle initial name in the dataset that I have choosen.
* The most common middle initial name for the healthcare providers is “m”.



**Zip codes Distribution**

****

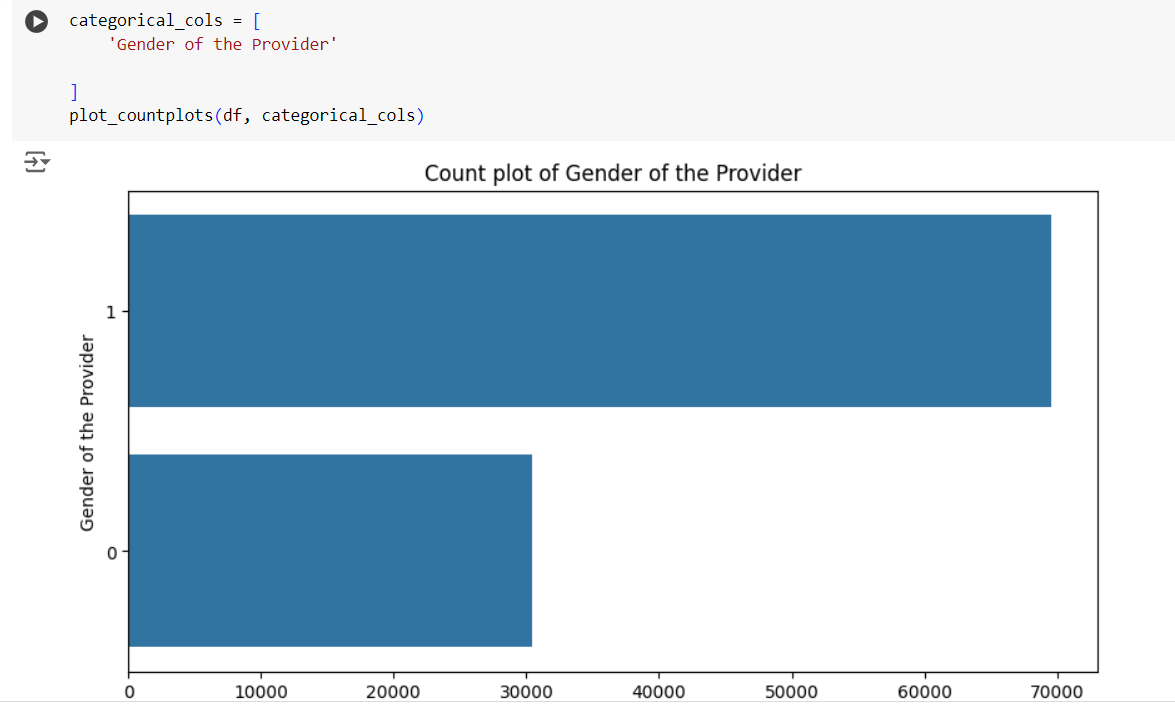
* In this graph the zipcode has been considered as a catagerical value and then its only distributed as top 20 rows because there are so many unique values

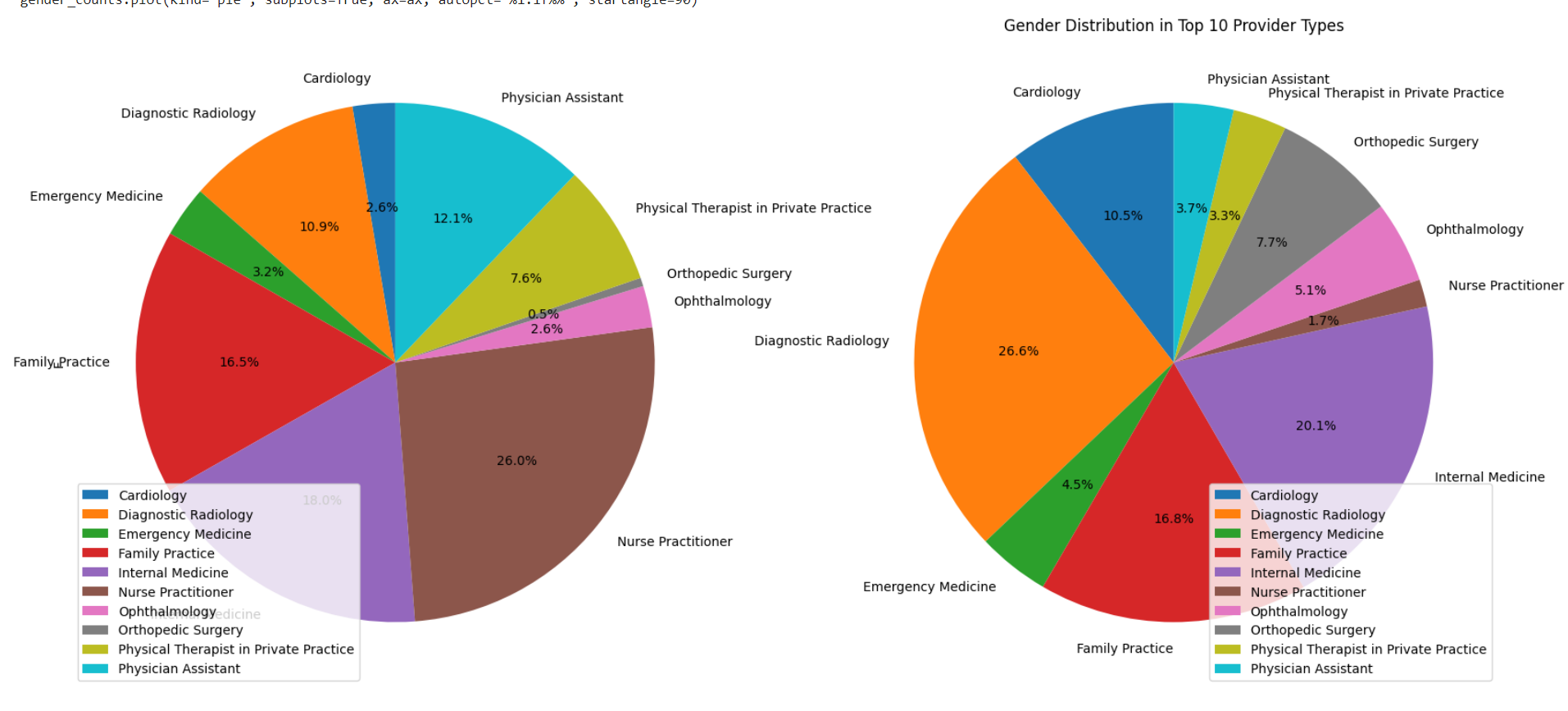
**Middle Initial provider **

* There are so many health care providers who have middle initial as ‘A ‘.
* There are least number of health care providers who has middle initial as ‘x’.

**Count Plot of Gender of the Provider**

* This dataset has more male healthcare providers compared to female healthcare providers.

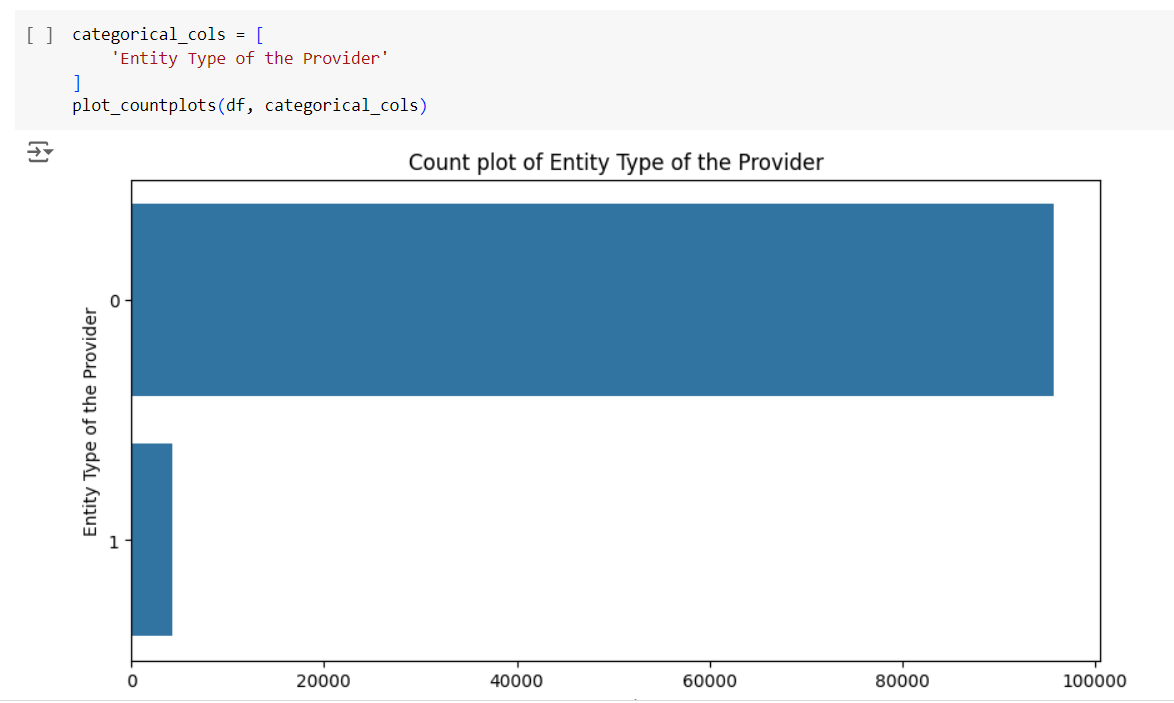




* Through the above pi chat we can understang that hoe many male percentage and female percentage for the each provider type but for the clear understanding I have only conserdier the top 10 provider types.

**Count Plot of Entity Type of the Provider**

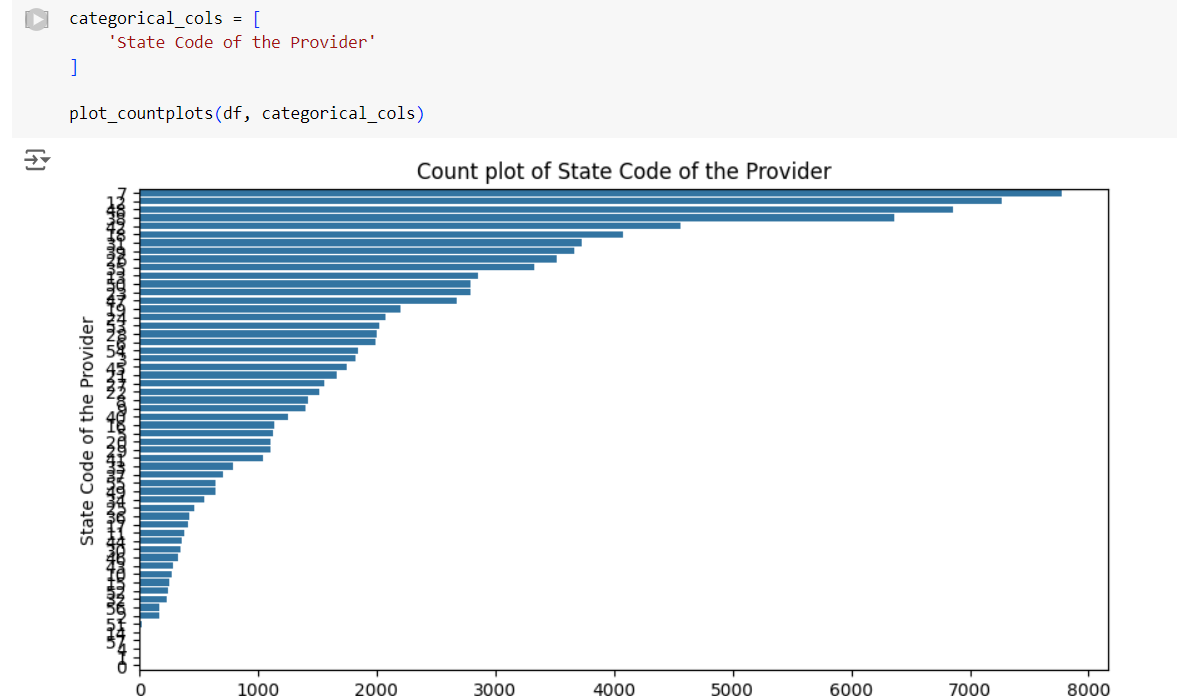
* Here the most of the individual providers are higher than the organizations.



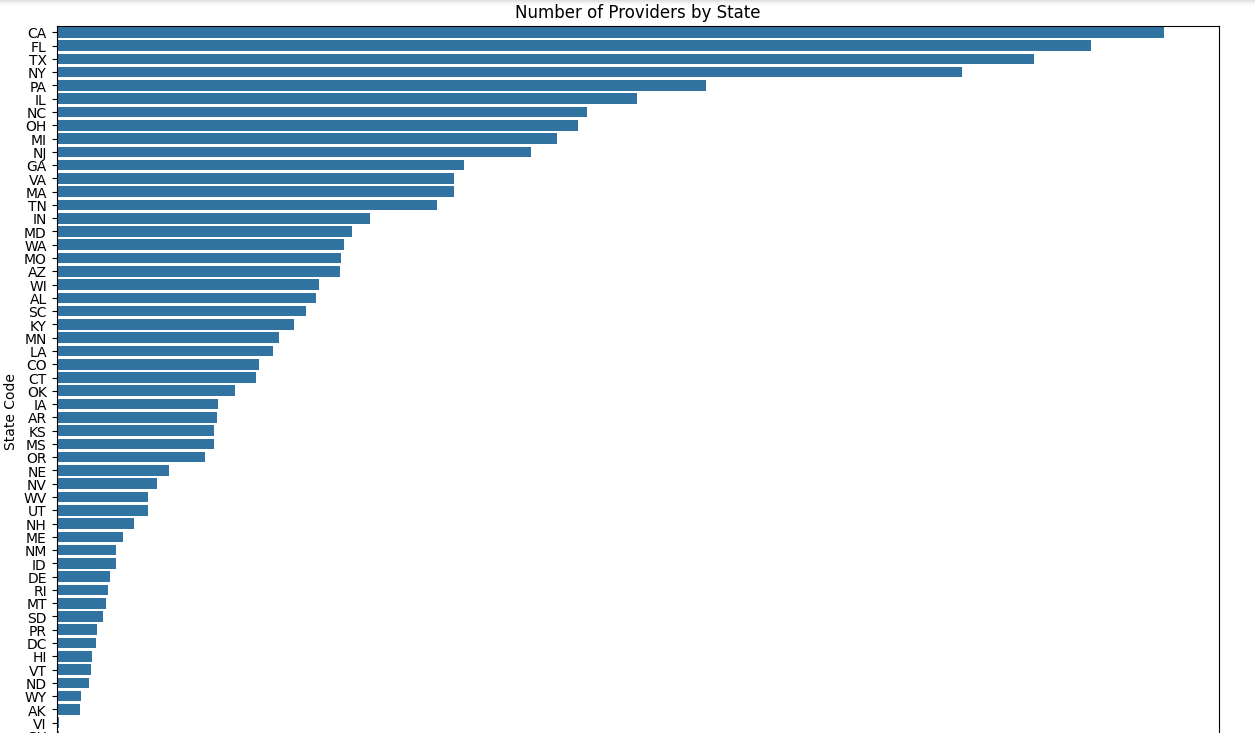
* Through the graph we can observe that their are more number of individual providers and less number of organization who provide health care services.

**Count Plot of State Code of the Provider**

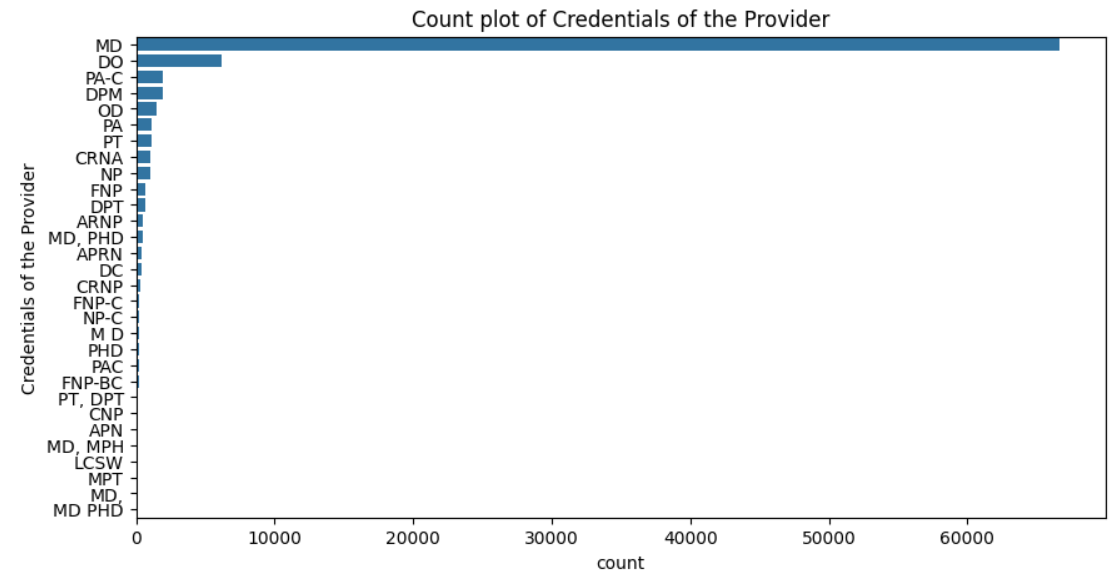
* Certain states have a significantly higher number of providers and the highest is over 7775 for stete CA.
* Some states have very few providers listed as and the least number of providers are in the state AA over 234.



* In this graph its showing that how many providers are there for each state.

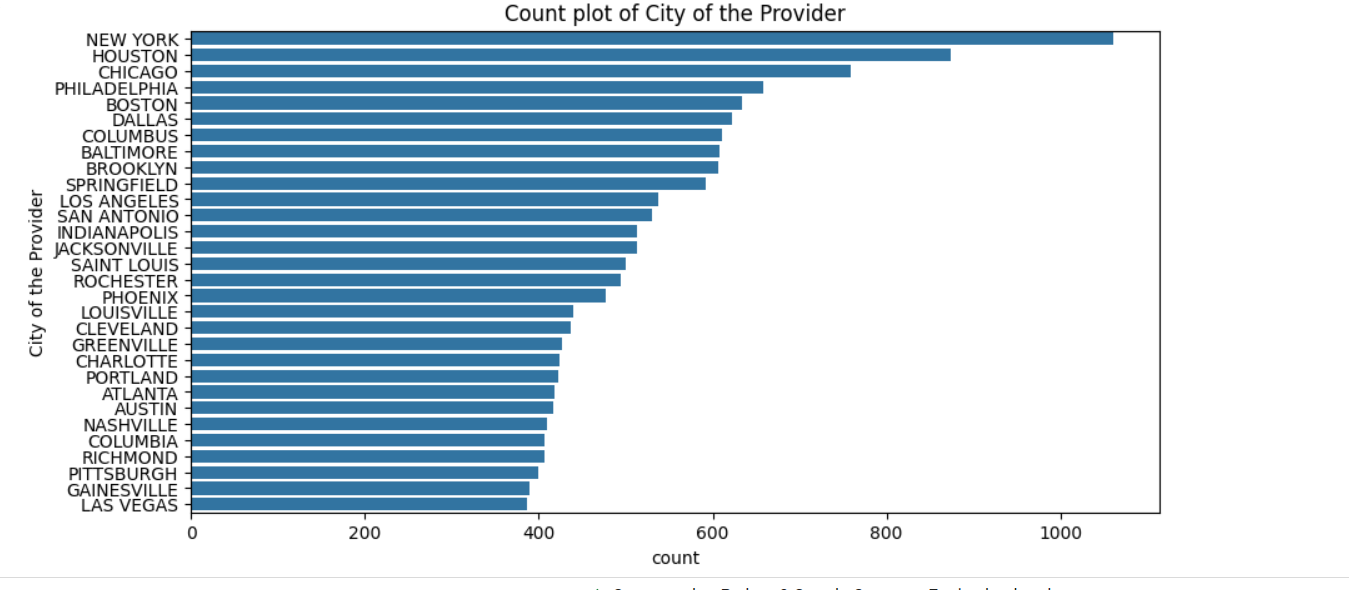


**Credential of the provider**

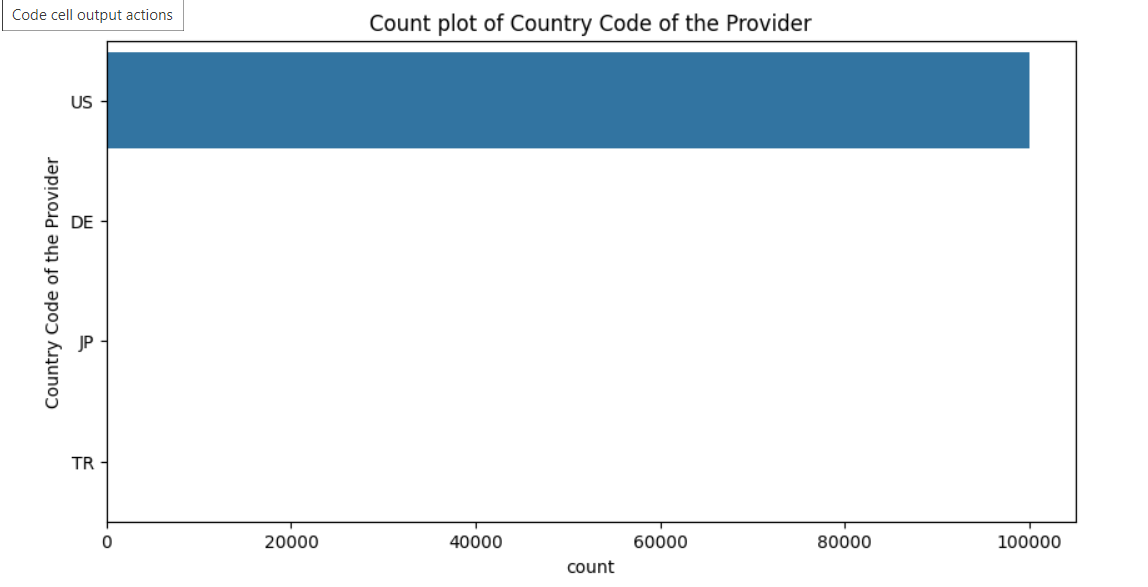
****

* We can actually see through the graph the most of the providers who have done their credentials as MD.
* There are least number of providers who have done their credentials as FNP-BC.

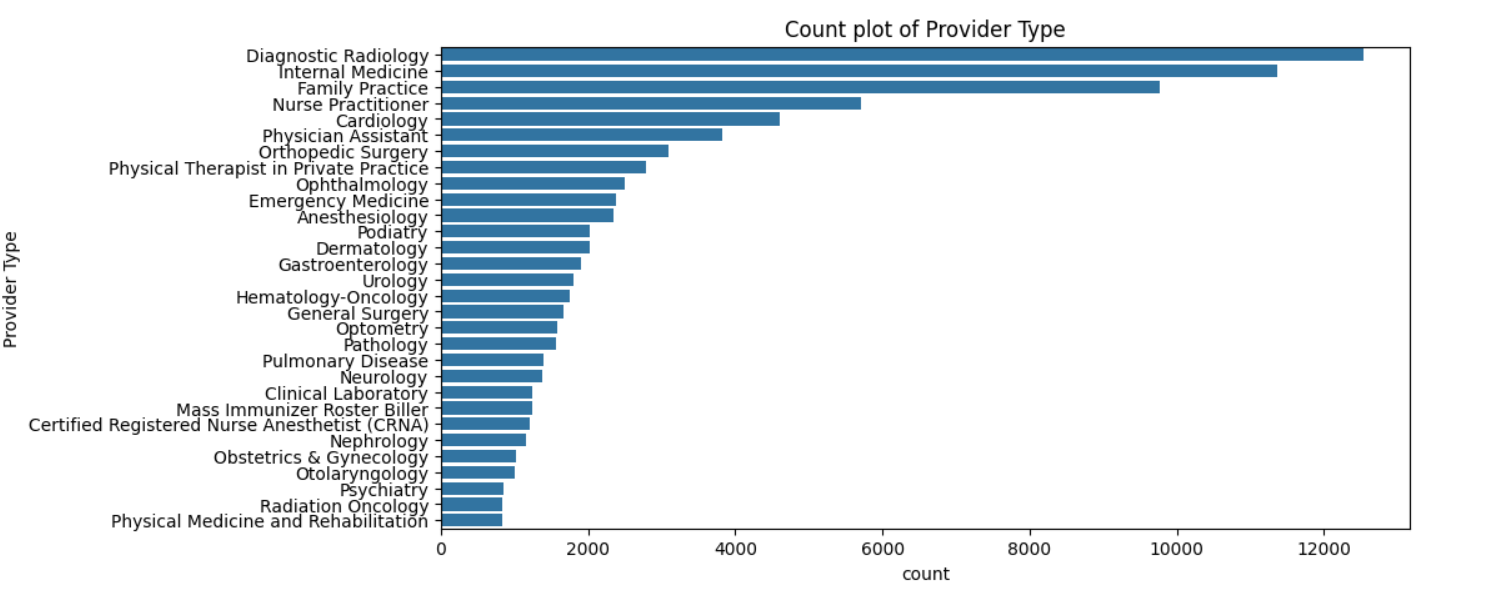
**City of the provider**

****

* These are the top 20 healthcare providers New york has taken position throughout the US.

**Country code of the provider**

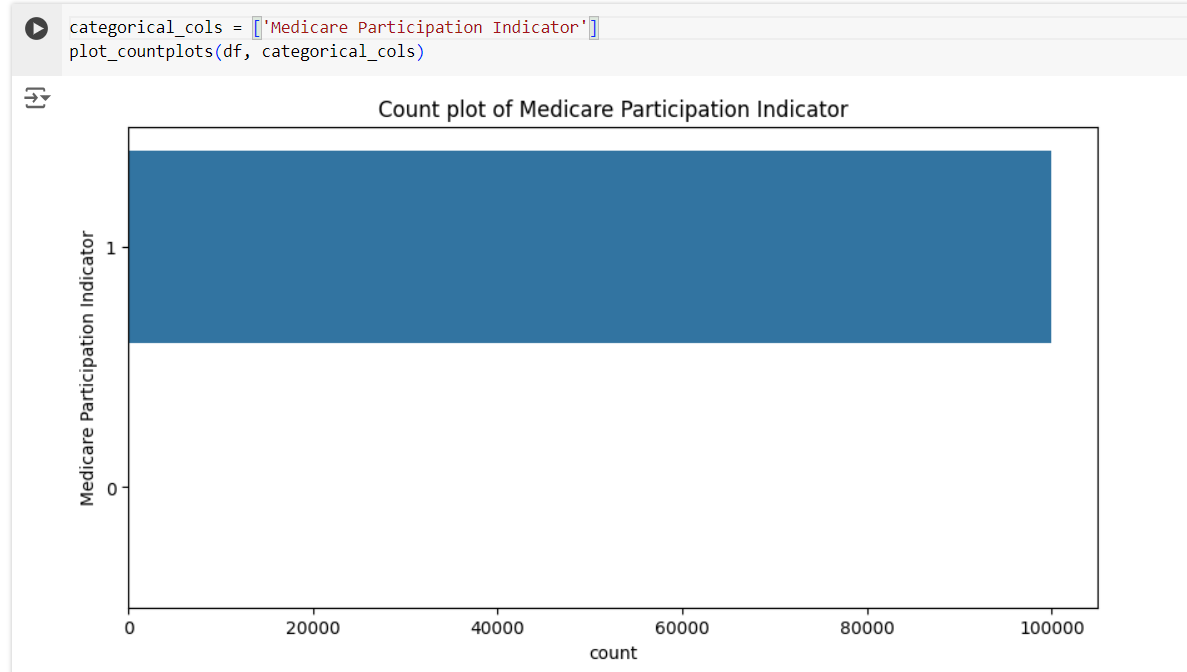
* We can simply say that this data set is only considered from US healthcare providers.

**Count plot of provider type**

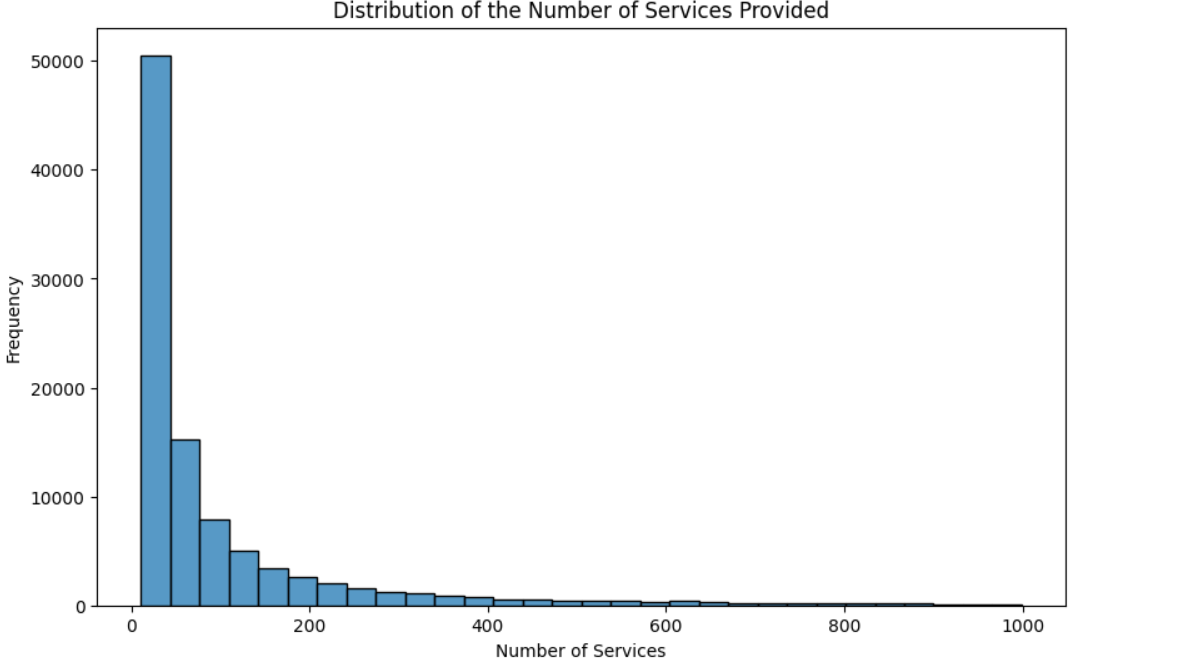
* The observations that I have been observed from the graphs is most of the providers providing Diagnostic Radiology.

**Count Plot of Medicare Participation Indicator**

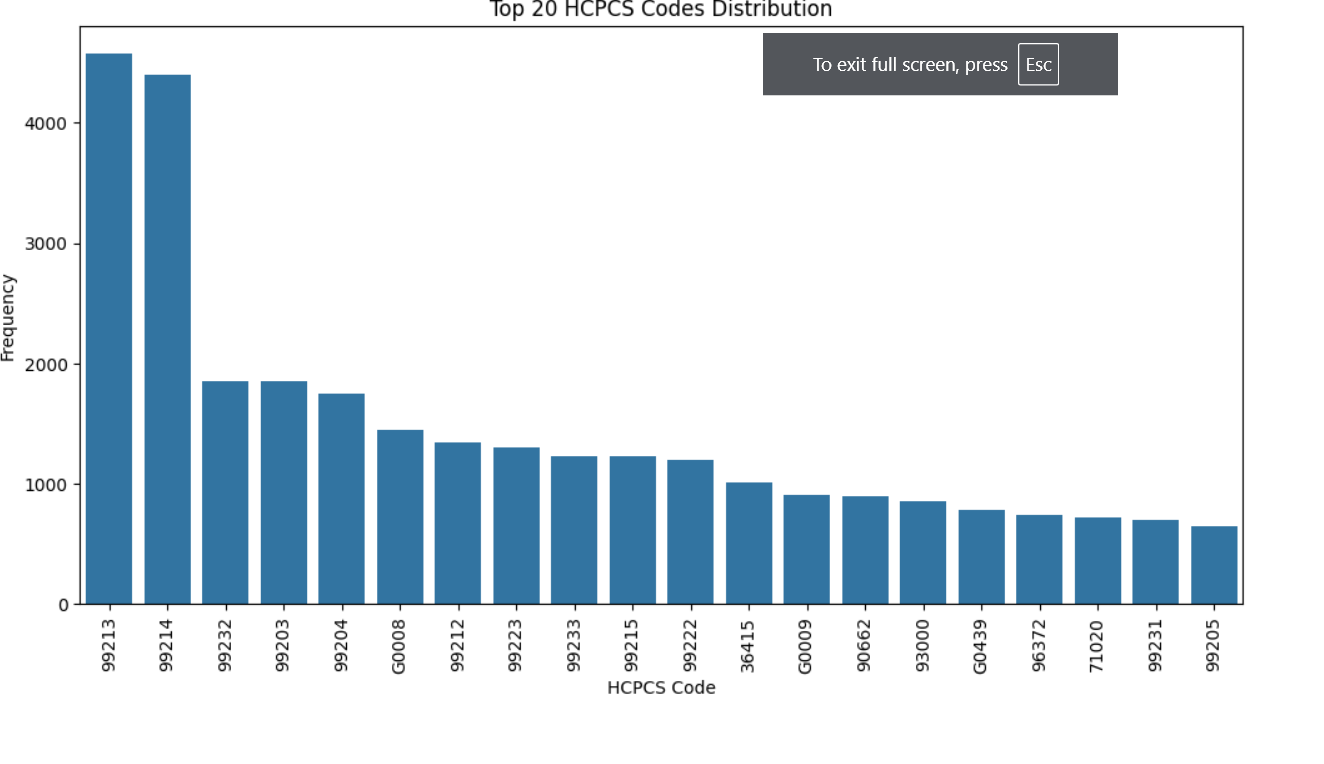
* Almost all the providers were participate in the medicare, which is indicated through the below



**Number of services provided**

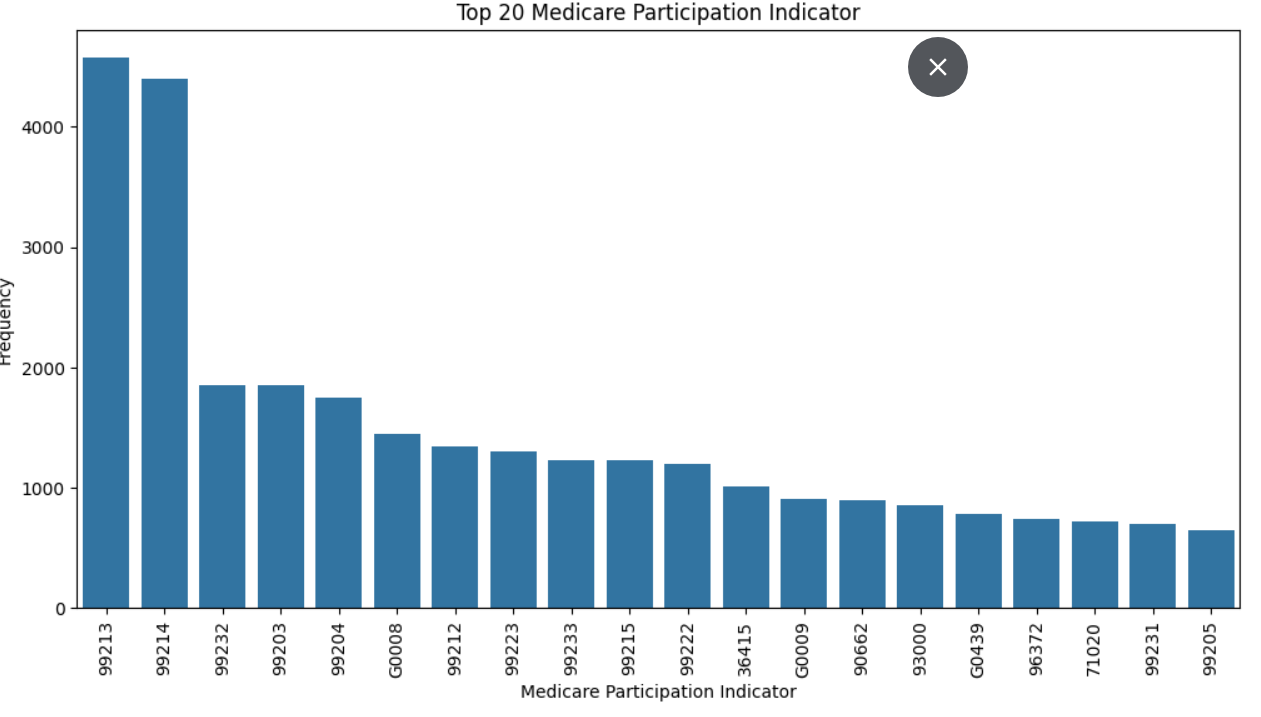


**Top 20 HCPCS Codes Distribution**

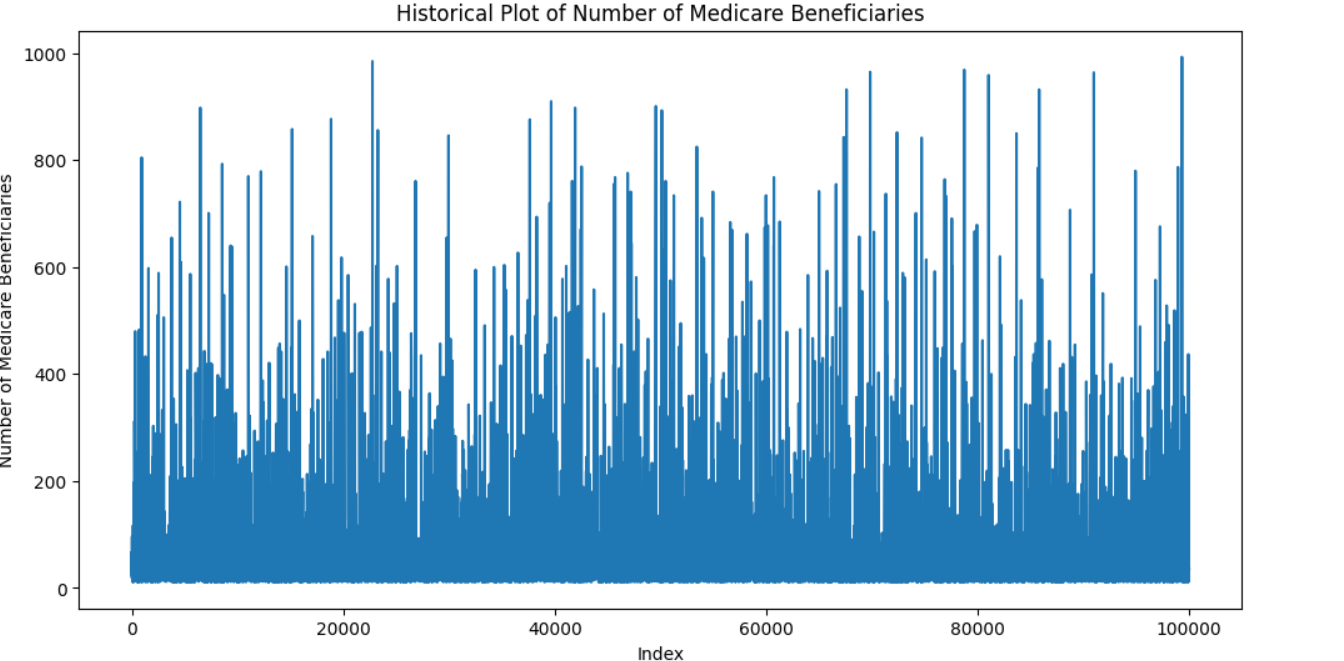


* This codes are essential for facilitating communication between healthcare providers and insures for the types of care provided, so that this graph represents the top 20 distribution code.

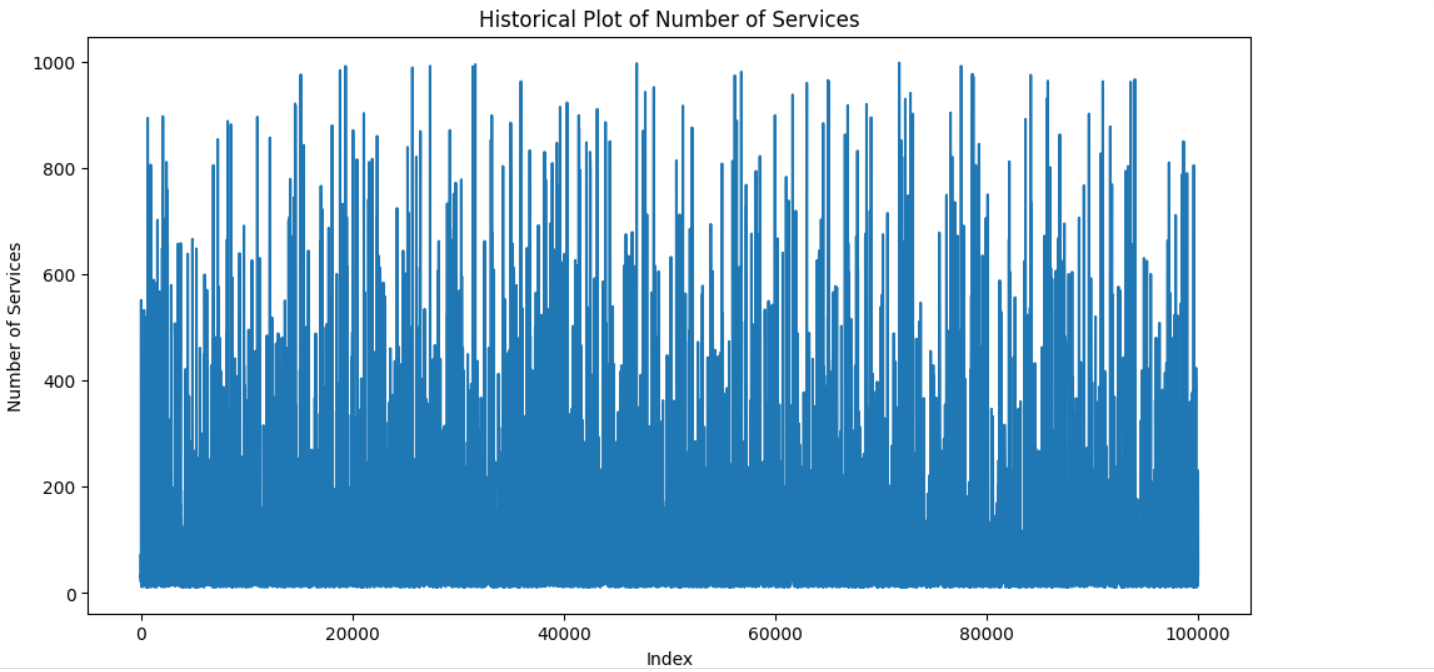
**Top 20 Medicare participation Indicator**



**Number of Medicare Beneficiaries**

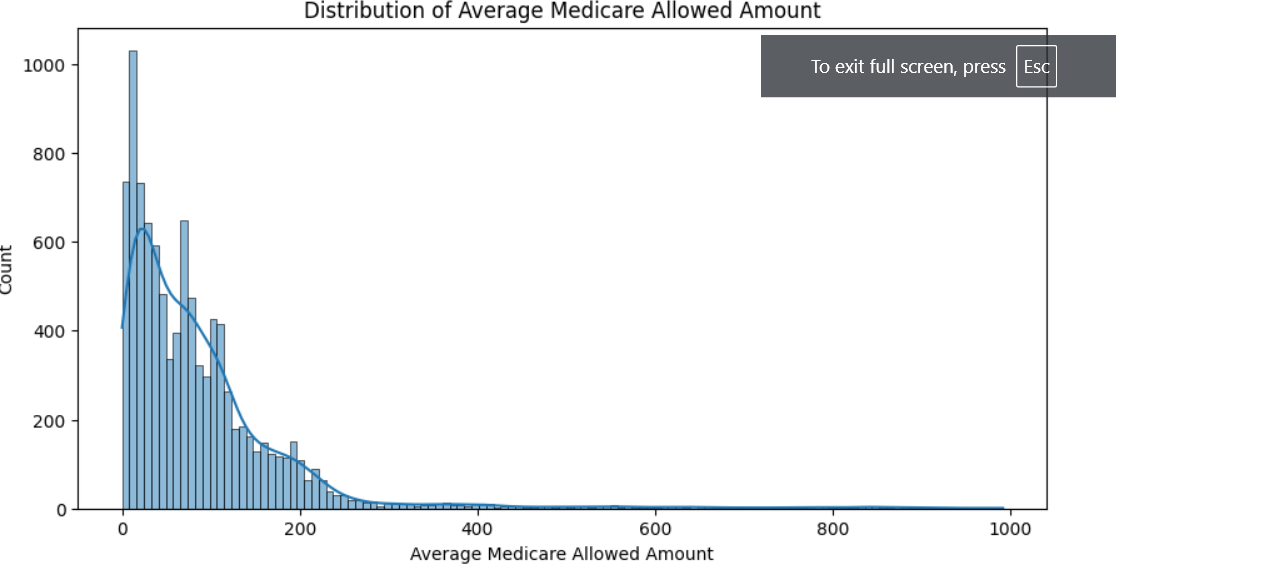


* This graph simply indicates the historical data of Medicare Beneficiaries.

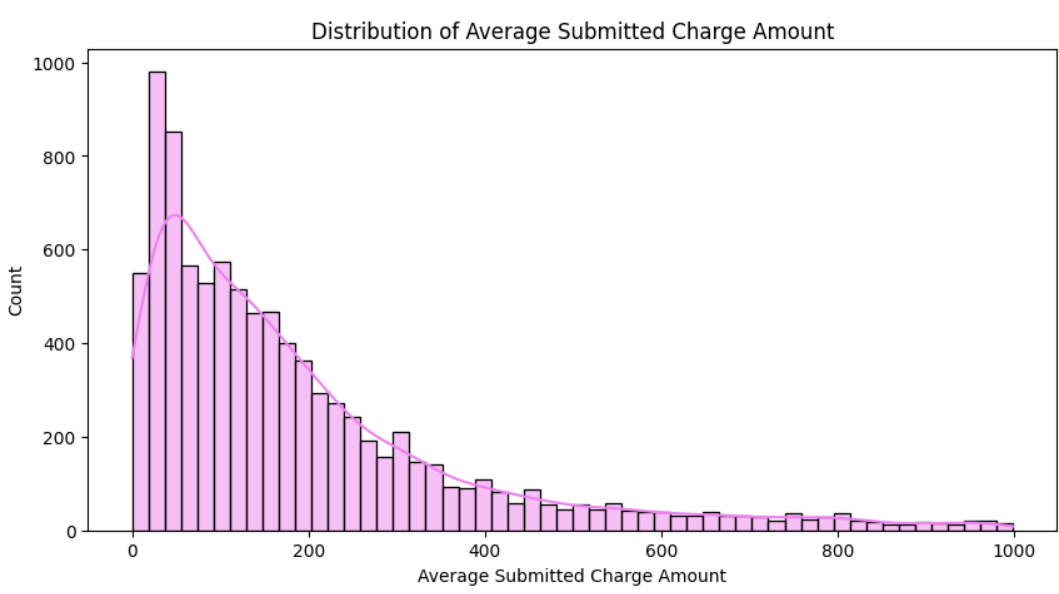
**Number of services** 

* This graph simply indicates the historical data of Medicare Beneficiaries.

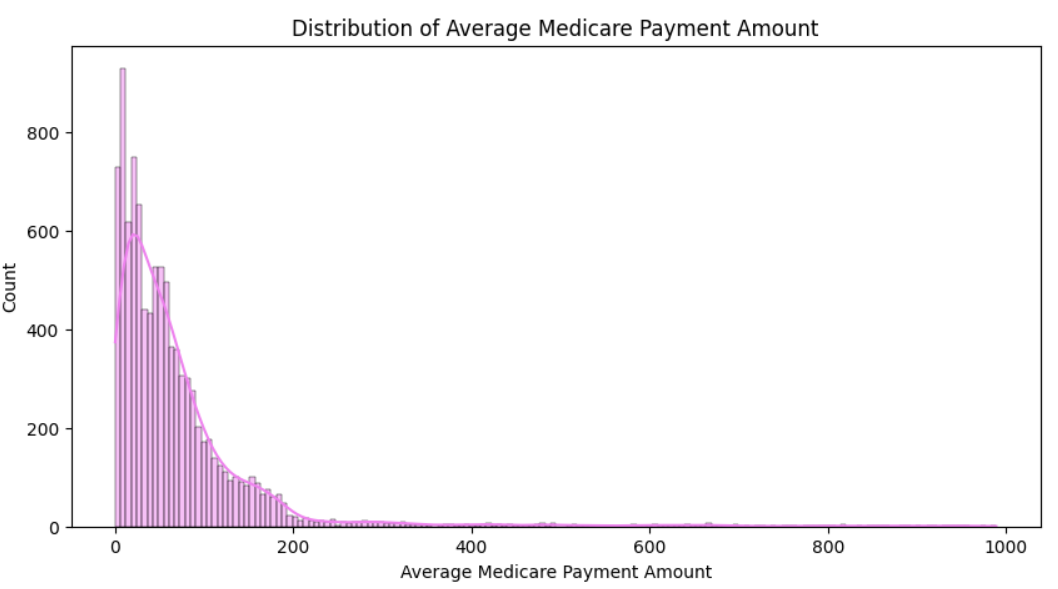
**Distribution Of Average Medicare Allowed**



* This graph represents the numerical column values and those varies from 0-1000 in Average Medicare Allowed Amount.

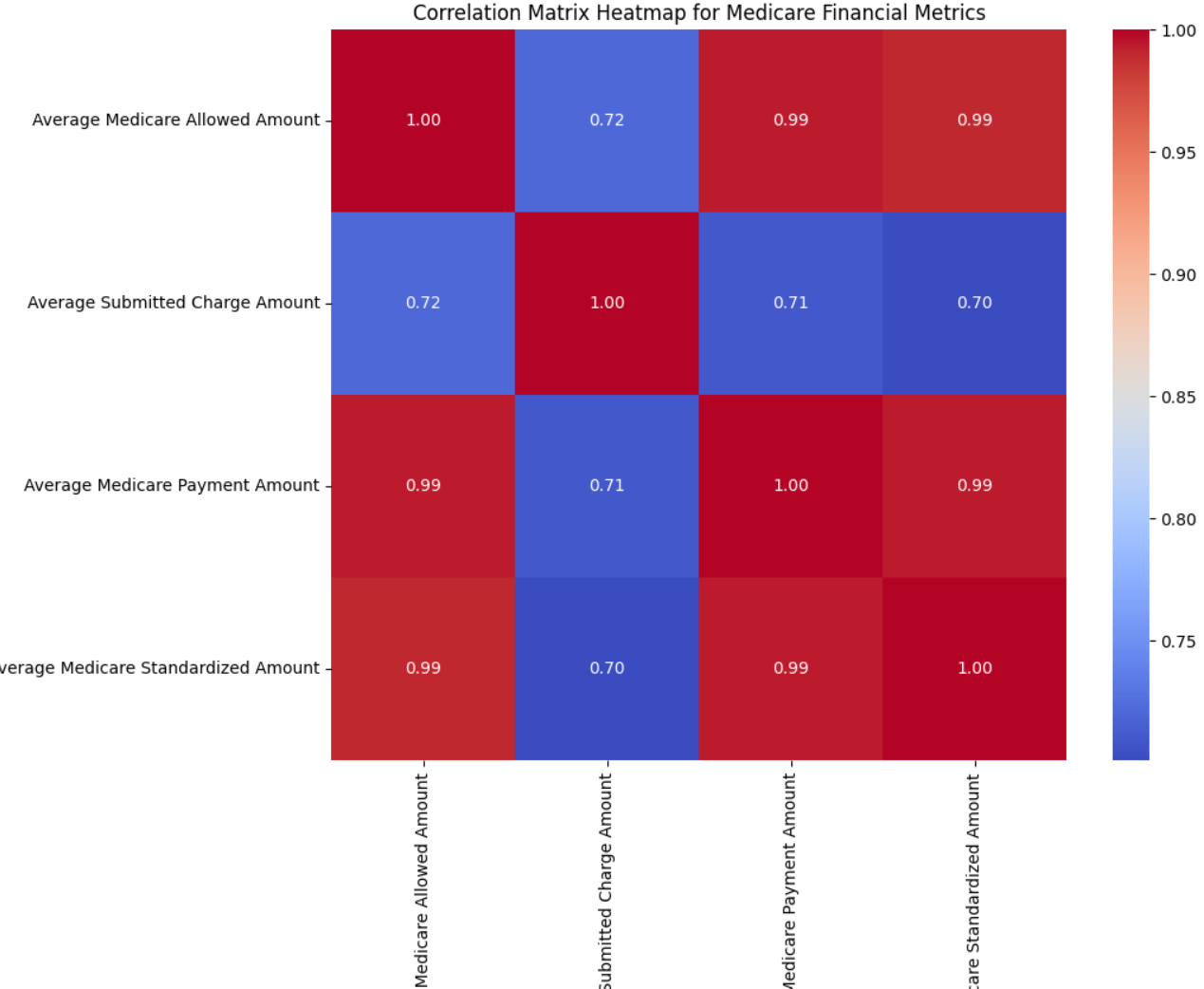
**Average Subbmitted Charge Amount**

**Average Medicare Payment Amount**



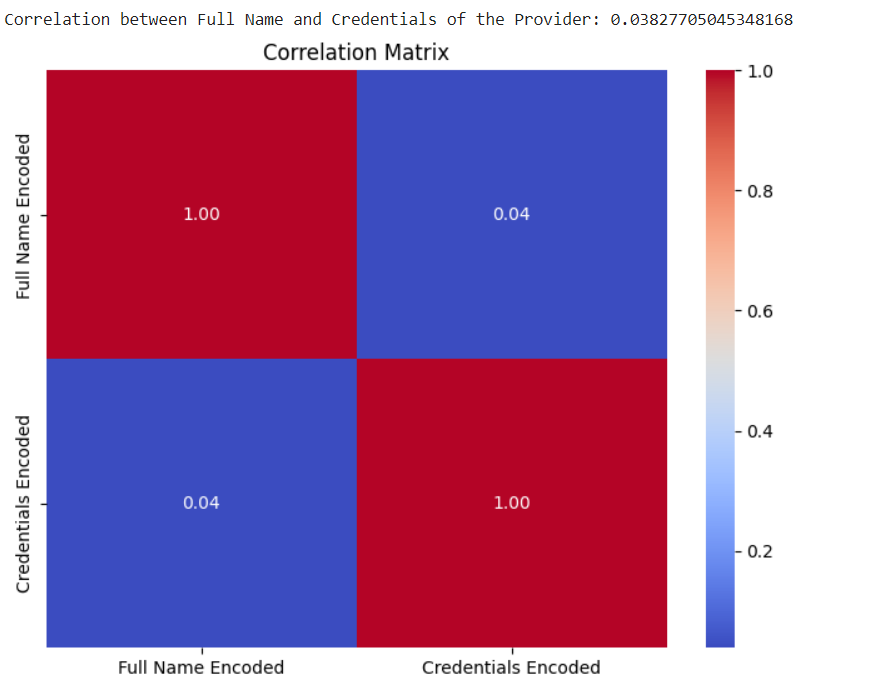
It shows the amount that the payments done by the visited peoples.

**Correlation marix heatmap**



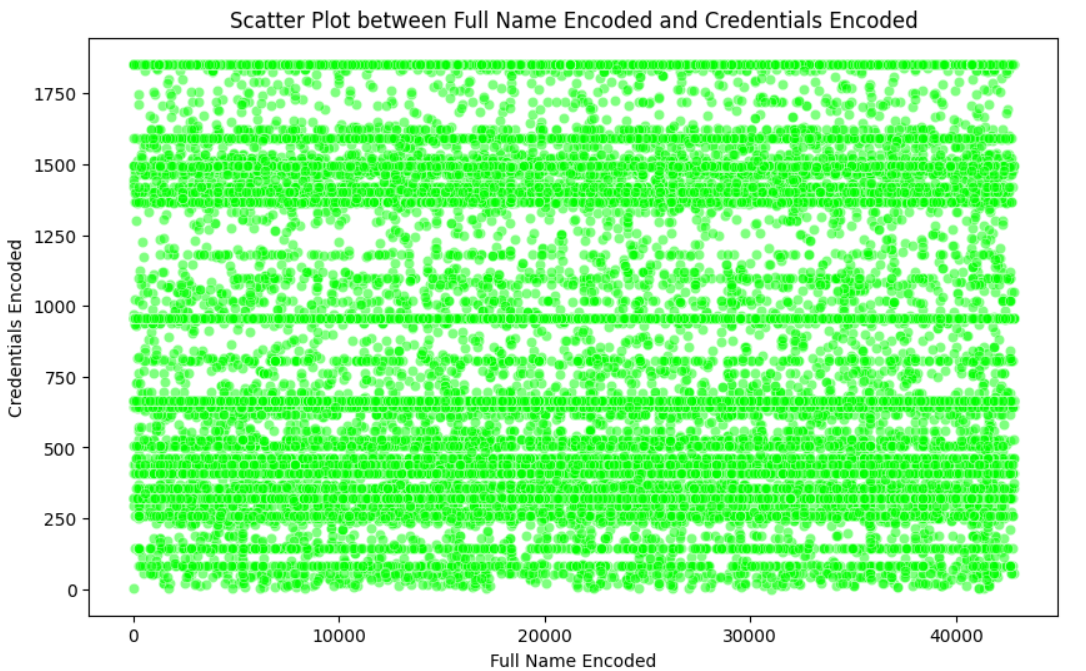
* Here the heatmap show the correlation for the the numerical columns in the dataset.
* Most of the numerical columns have the higest correlations as 1.0 and the least value id 0.70.

**Correlation marix heatmap**

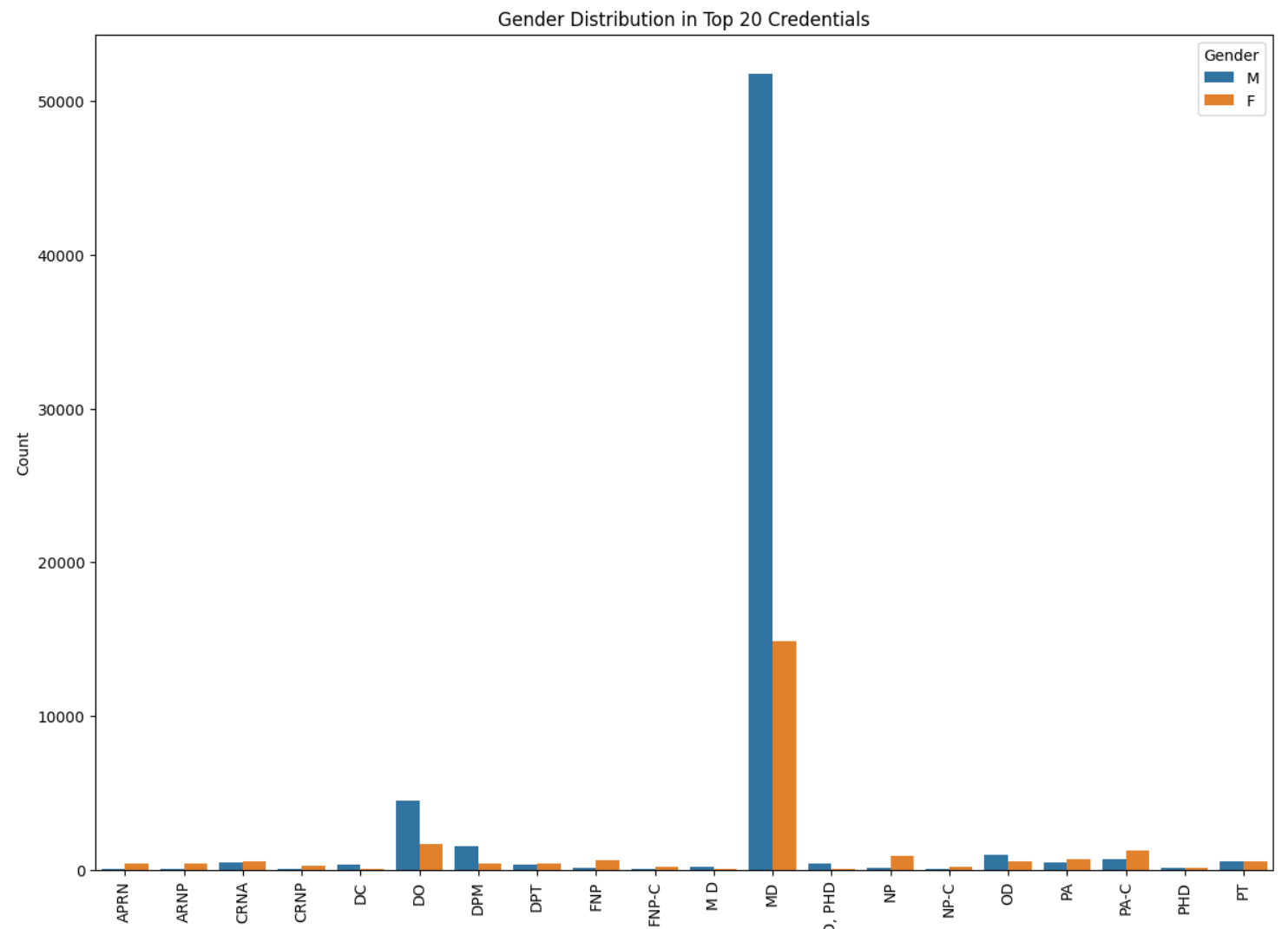


* Correlation marix heatmap for the provider and the credentials of the provider and the correlation is : 0.038277050453.

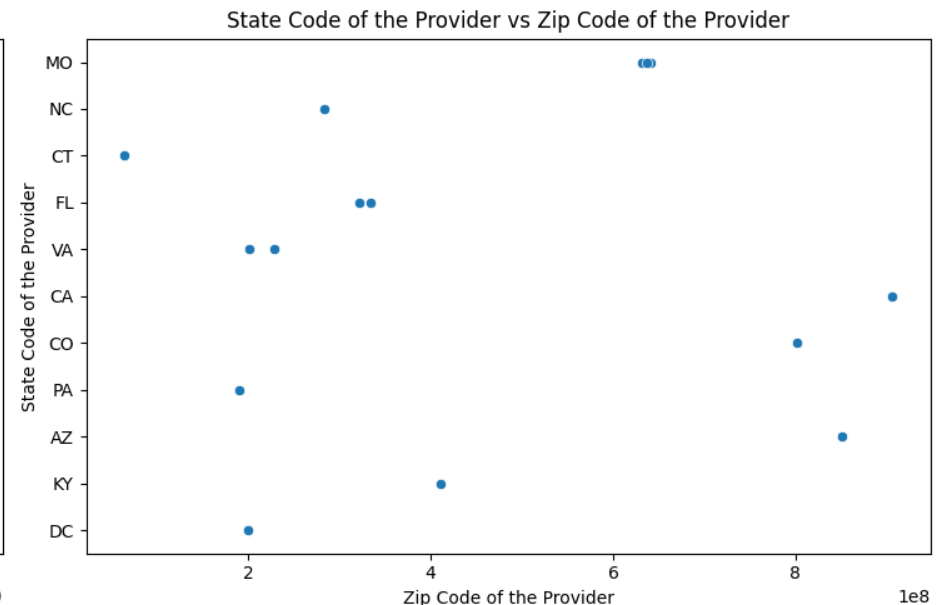
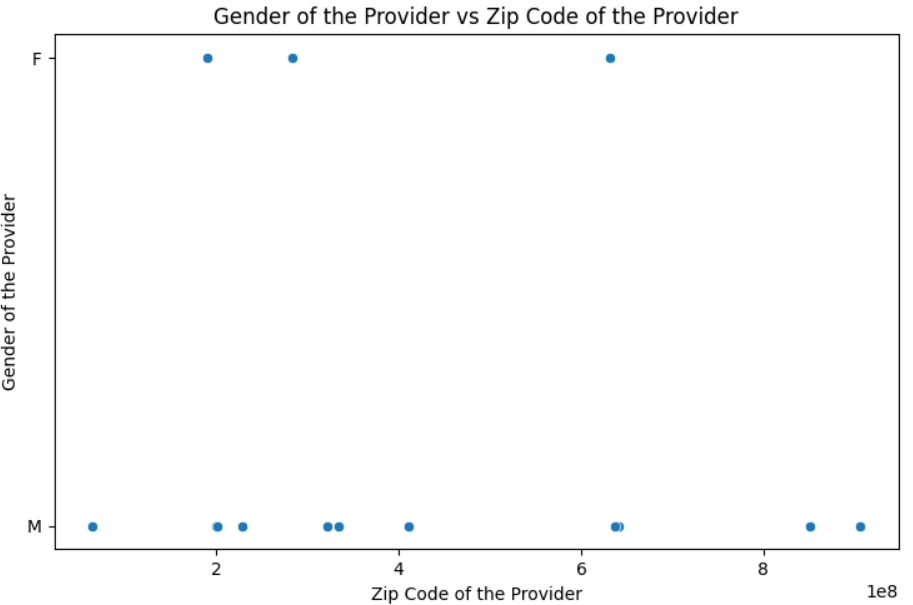
**Scatter Plot for the person and the Credentials**

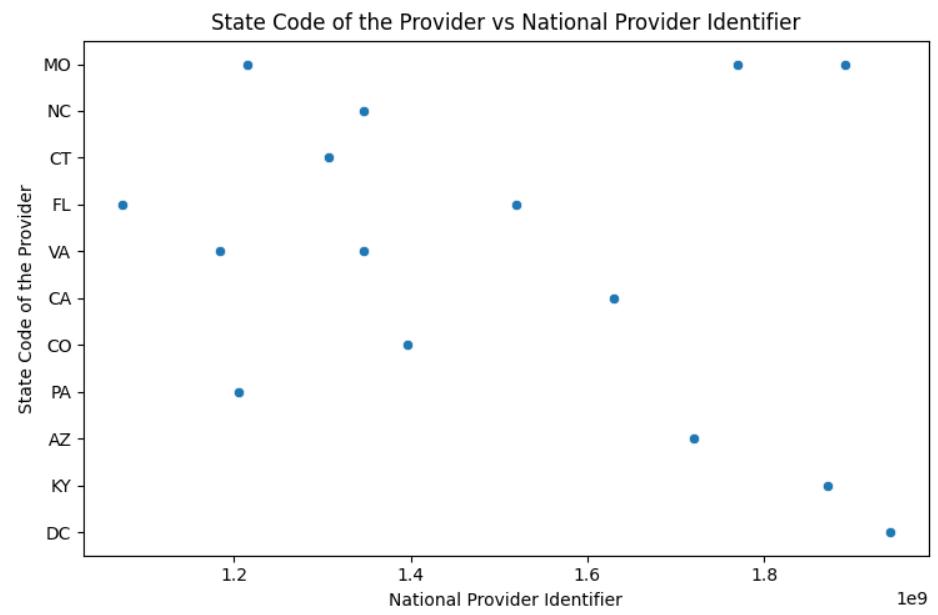


**Relation Ship for the Gender and the** **Credentials**

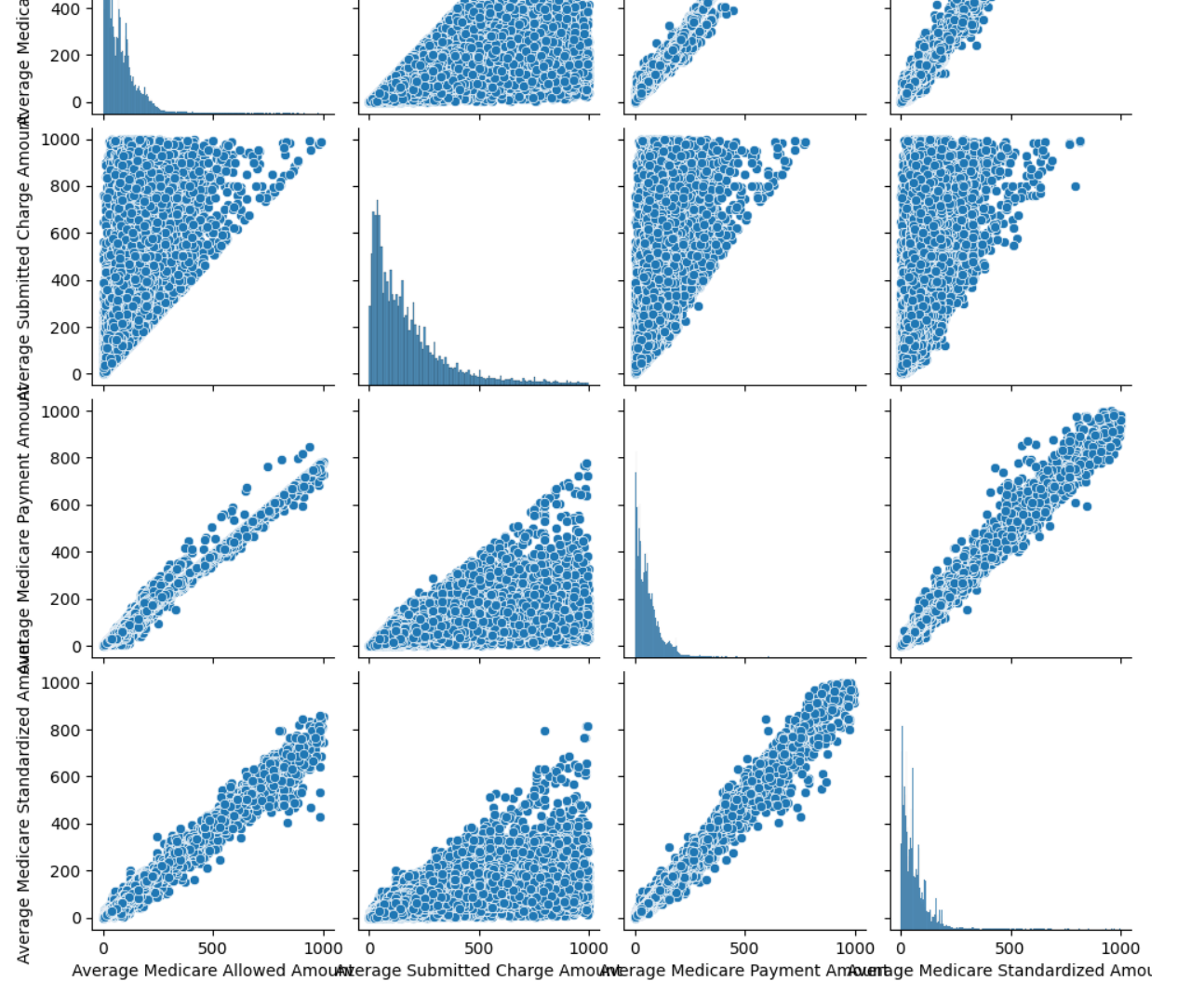


* through the relation ship most of the male providers and the female providers have done there

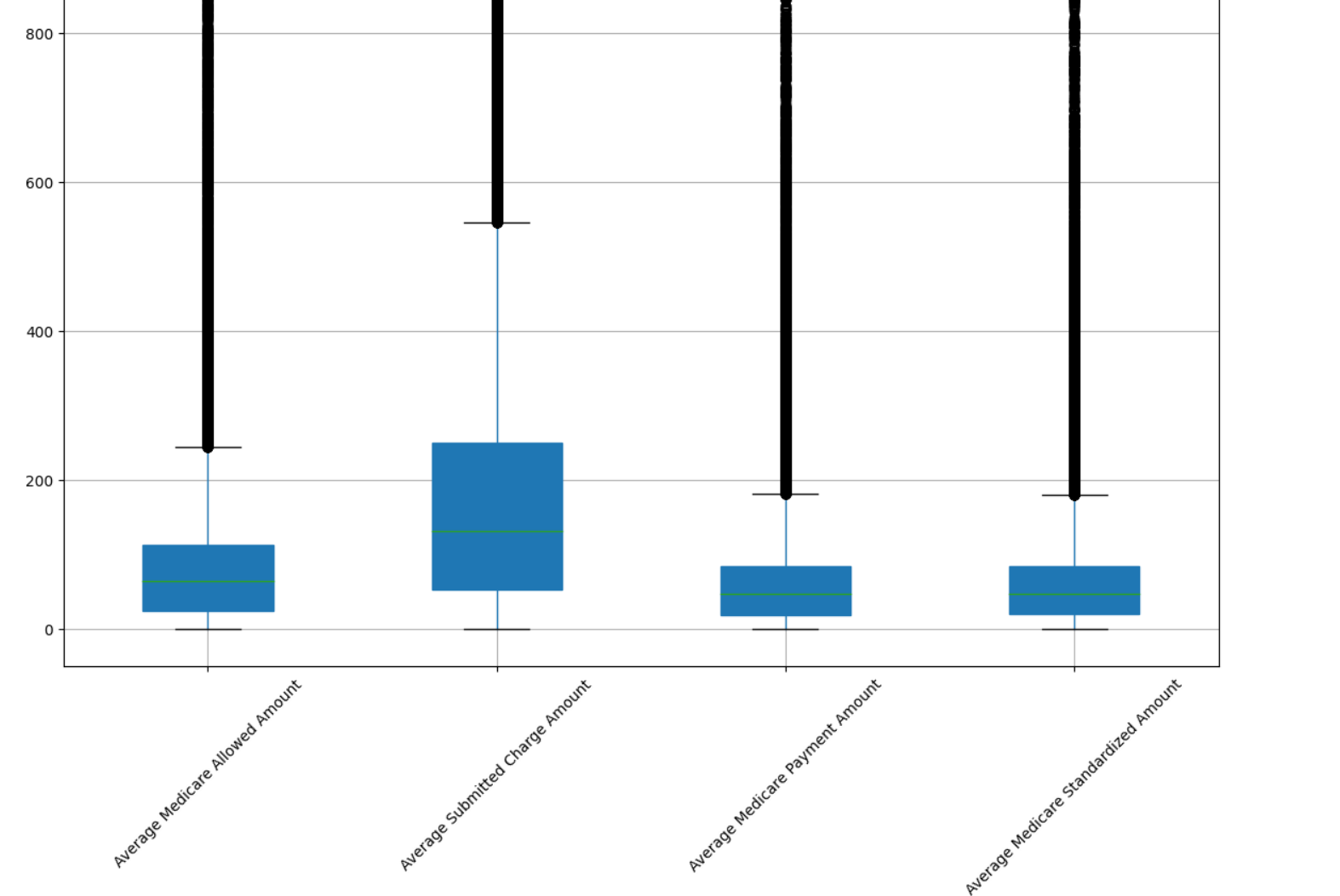




**Pair Plots for the Numerical Columns**



**Box Plot to find out the Outliers**



* There are several outliers in the "Number of Services" column, indicating that some providers have performed a significantly higher number of services compared to the majority.
* Similar to the "Number of Services" column, there are outliers indicating that some providers serve a significantly higher number of Medicare beneficiaries.
* There are a few outliers with higher allowed amounts, but the majority of the data points are clustered within a narrow range.
* The median value is relatively low compared to the outliers.

**Anomaly Detection**

This documentation outlines the steps taken to perform anomaly detection on the healthcare providers dataset. Anomaly detection is a technique used to identify unusual patterns that do not conform to expected behavior. In this context, it helps in identifying anomalies in the healthcare providers' data that could indicate errors, fraud, or other significant deviations.

Anomaly Detection is most of datasets having some error values even through if we remove null values and the some other values like if we consider an IPL cricket match dataset and its contain a team score 15 runs in 1st over, 13 runs in 2nd over, 16 runs in 3rd over and 100 runs in 4th over this can’t be removed in the in pre-processing technique. To remove these type of weird values in the dataset we need to understand the connection of these values in this dataset a team can score 36runs in an over this is nothing but an anomaly values and if we don’t find these values then it cause a huge affect to the service.

For this dataset we use Isolation Forest Method

**Isolation Forest Model**

Through this method we can identify anomalies in the healthcare providers dataset that might indicate unusual patterns, errors, or irregularities in healthcare service providers.

So we need to implement

model = IsolationForest(n\_estimators=100, contamination=0.01, random\_state=42)

model.fit(df\_scaled)

df['anomaly'] = model.predict(df\_scaled)

df['anomaly'] = df['anomaly'].apply(lambda x: 1 if x == -1 else 0)

print("Number of anomalies detected:", df['anomaly'].sum())

plt.scatter(df['feature1'], df['feature2'], c=df['anomaly'])

plt.xlabel('Feature 1')

plt.ylabel('Feature 2')

plt.title('Anomaly Detection using Isolation Forest')

plt.show()