Anomaly Detection in Healthcare providers Data

Healthcare providers collect and manage vast amounts of data related to patient care, treatment outcomes, billing, and operational metrics. Ensuring the accuracy and integrity of this data is crucial for patient safety, regulatory compliance, and efficient hospital management. Anomalies in healthcare data can indicate a variety of issues, ranging from clerical errors and system malfunctions to potential fraud or misuse. Early detection of these anomalies can help healthcare providers address problems promptly, reduce risks, and maintain high standards of care.

Problem Statement:

The problem statement of this project is to develop an automated anomaly detection system that can identify irregularities in healthcare providers' data. The system should be capable of processing large datasets, detecting various types of anomalies, and providing actionable insights to healthcare administrators and staff.

**Numerical 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**

**Categorical Columns**

* **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**
* **HCPCS Code**
* **HCPCS Description**
* **HCPCS Drug Indicator**
* **Detailed Data File:**
* The following variables are included in the detailed Physician and Other Supplier data file (see Appendix A for a condensed version of variables included)).
* **npi** – National Provider Identifier (NPI) for the performing provider on the claim. The provider NPI is the numeric identifier registered in NPPES.
* **nppes\_provider\_last\_org\_name** – When the provider is registered in NPPES as an individual (entity type code=’I’), this is the provider’s last name. When the provider is registered as an organization (entity type code = ‘O’), this is the organization's name.
* **nppes\_provider\_first\_name** – When the provider is registered in NPPES as an individual (entity type code=’I’), this is the provider’s first name. When the provider is registered as an organization (entity type code = ‘O’), this will be blank.
* **nppes\_provider\_mi** – When the provider is registered in NPPES as an individual (entity type code=’I’), this is the provider’s middle initial. When the provider is registered as an organization (entity type code= ‘O’), this will be blank.
* **nppes\_credentials** – When the provider is registered in NPPES as an individual (entity type code=’I’), these are the provider’s credentials. When the provider is registered as an organization (entity type code = ‘O’), this will be blank.
* **nppes\_provider\_gender** – When the provider is registered in NPPES as an individual (entity type code=’I’), this is the provider’s gender. When the provider is registered as an organization (entity type code = ‘O’), this will be blank.
* **nppes\_entity\_code** – Type of entity reported in NPPES. An entity code of ‘I’ identifies providers registered as individuals and an entity type code of ‘O’ identifies providers registered as organizations.
* **nppes\_provider\_street1** – The first line of the provider’s street address, as reported in NPPES.
* **nppes\_provider\_street** – The second line of the provider’s street address, as reported in NPPES.
* **nppes\_provider\_city** – The city where the provider is located, as reported in NPPES.
* **nppes\_provider\_zip** – The provider’s zip code, as reported in NPPES.
* **nppes\_provider\_state** – The state where the provider is located, as reported in NPPES. The fifty U.S. states and the District of Columbia are reported by the state postal abbreviation. The following values are used for all other areas:
* 'XX' = 'Unknown'  
  'AA' = 'Armed Forces Central/South America'  
  'AE' = 'Armed Forces Europe'  
  'AP' = 'Armed Forces Pacific'  
  'AS' = 'American Samoa'  
  'GU' = 'Guam'  
  'MP' = 'North Mariana Islands'  
  'PR' = 'Puerto Rico'  
  'VI' = 'Virgin Islands'  
  'ZZ' = 'Foreign Country'
* **nppes\_provider\_country** – The country where the provider is located, as reported in NPPES. The country code will be ‘US’ for any state or U.S. possession. For foreign countries (i.e., state values of ‘ZZ’), the provider country values include the following:  
  AE=United Arab Emirates IT=Italy  
  AG=Antigua JO= Jordan  
  AR=Argentina JP=Japan  
  AU=Australia KR=Korea  
  BO=Bolivia KW=Kuwait  
  BR=Brazil KY=Cayman Islands  
  CA=Canada LB=Lebanon  
  CH=Switzerland MX=Mexico  
  CN=China NL=Netherlands  
  CO=Colombia NO=Norway  
  DE= Germany NZ=New Zealand  
  ES= Spain PA=Panama  
  FR=France PK=Pakistan  
  GB=Great Britain RW=Rwanda  
  GR=Greece SA=Saudi Arabia  
  HU= Hungary SY=Syria  
  IL= Israel TH=Thailand  
  IN=India TR=Turkey  
  IS= Iceland VE=Venezuela
* **provider\_type** – Derived from the provider specialty code reported on the claim.
* **medicare\_participation\_indicator** – Identifies whether the provider participates in Medicare and/or accepts the assigned assignment of Medicare allowed amounts.
* **place\_of\_service** – Identifies whether the place of service submitted on the claims is a facility (value of ‘F’) or non-facility (value of ‘O’). Non-facility is generally an office setting; however other entities are included in non-facility.
* **hcpcs\_code** – HCPCS code used to identify the specific medical service furnished by the provider.
* **hcpcs\_description** – Description of the HCPCS code for the specific medical service furnished by the provider.
* **hcpcs\_drug\_indicator** –Identifies whether the HCPCS code for the specific service furnished by the provider is an HCPCS listed on the Medicare Part B Drug Average Sales Price (ASP) File.
* **line\_srvc\_cnt** – Number of services provided; note that the metrics used to count the number provided can vary from service to service.
* **bene\_unique\_cnt** – Number of distinct Medicare beneficiaries receiving the service.
* **bene\_day\_srvc\_cnt** – Number of distinct Medicare beneficiary/per day services.
* **average\_Medicare\_allowed\_amt** – Average of the Medicare allowed amount for the service.
* **stdev\_Medicare\_allowed\_amt** – Standard deviation of the Medicare allowed amounts.
* **average\_submitted\_chrg\_amt** – Average of the charges that the provider submitted for the service.
* **stdev\_submitted\_chrg\_amt** – Standard deviation of the charge amounts submitted by the provider.
* **average\_Medicare\_payment\_amt** – Average amount that Medicare paid after deductible and coinsurance amounts have been deducted for the line item service.

.

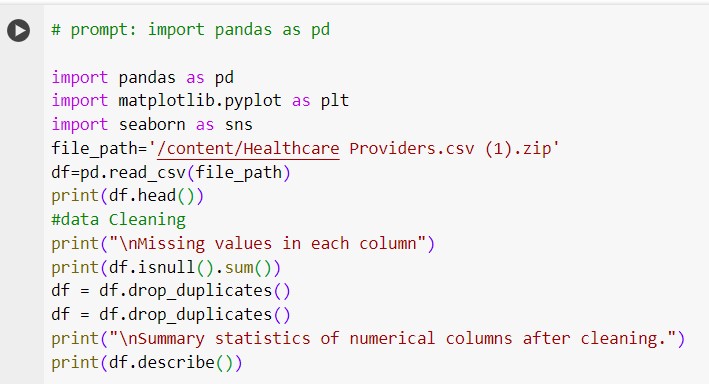
The project will involve:

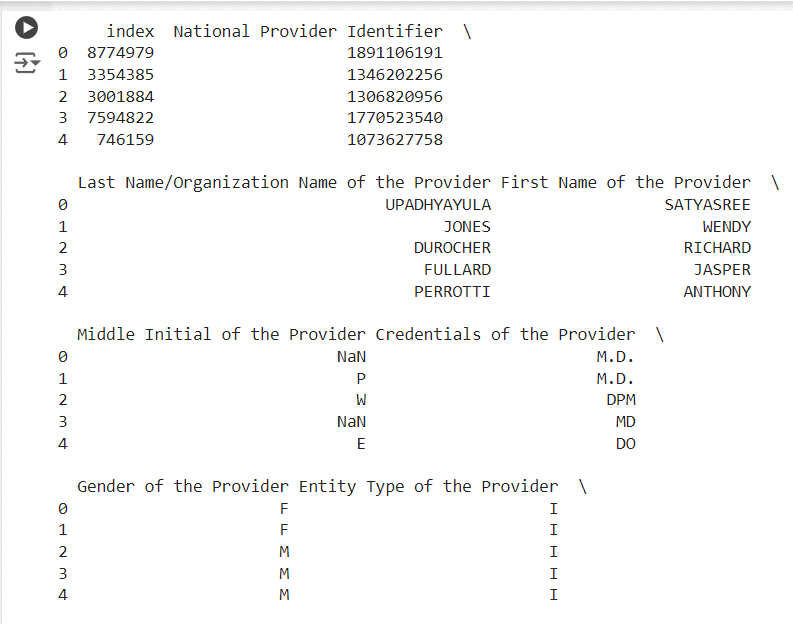
1. **Data Collection and Preprocessing**: Aggregating data from various sources (e.g., Electronic Health Records (EHR), billing systems) and preprocessing it to handle missing values, normalize formats, and anonymize sensitive information.
2. **Anomaly Detection Algorithms**: Implementing and comparing various anomaly detection algorithms, including statistical methods, machine learning models, and deep learning techniques.
3. **Validation and Testing**: Validating the effectiveness of the anomaly detection system using historical data with known anomalies and testing it in a live environment.
4. **User Interface**: Developing an intuitive user interface that allows healthcare providers to visualize anomalies, understand their potential impact, and take corrective actions.
5. **Compliance and Security**: Ensuring the system complies with healthcare regulations and maintains the highest standards of data security.

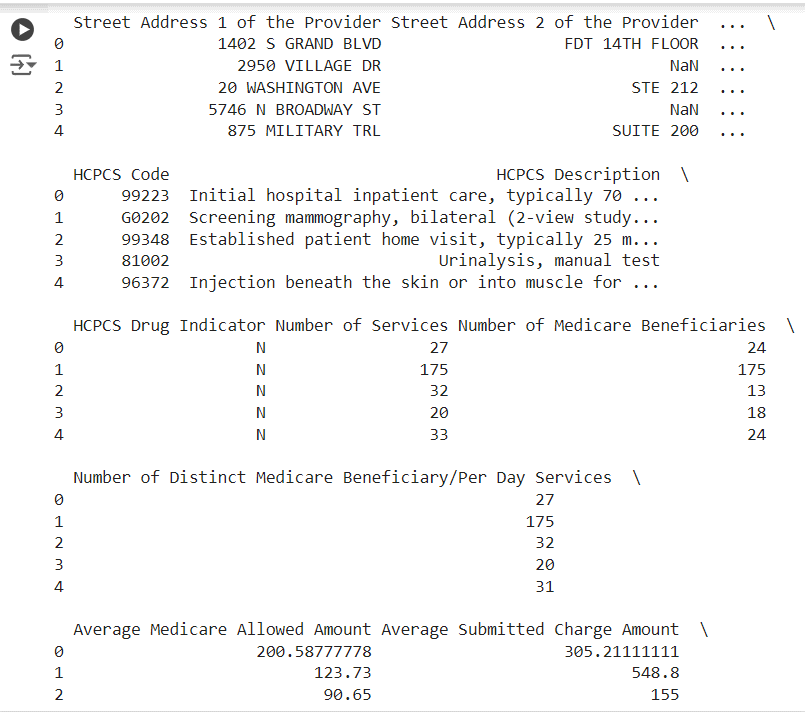
.

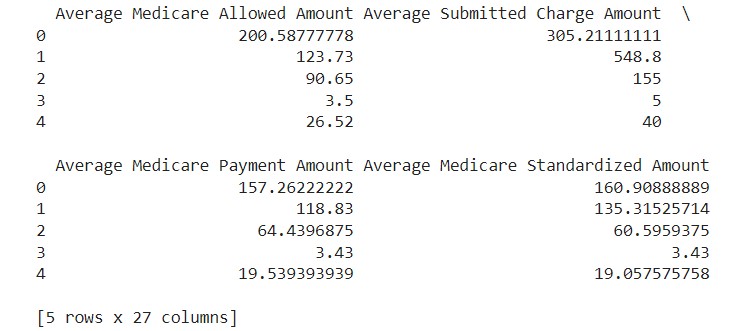
Preprocessing steps

1.load the Data

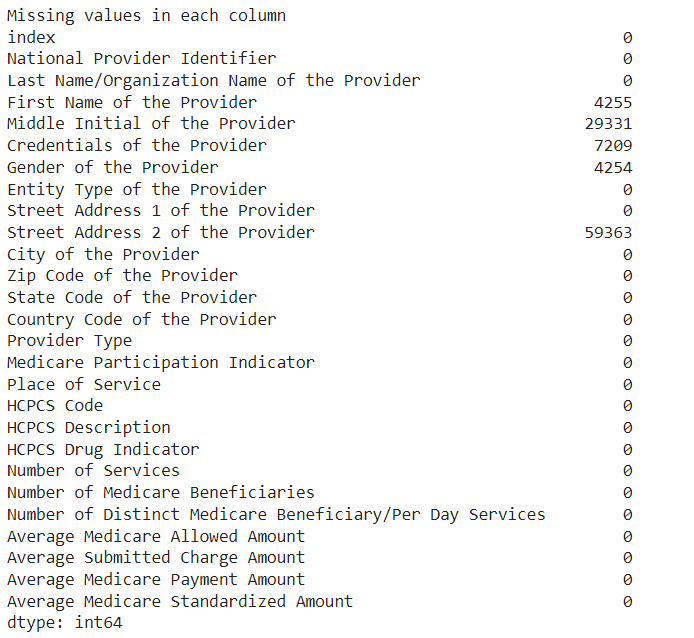




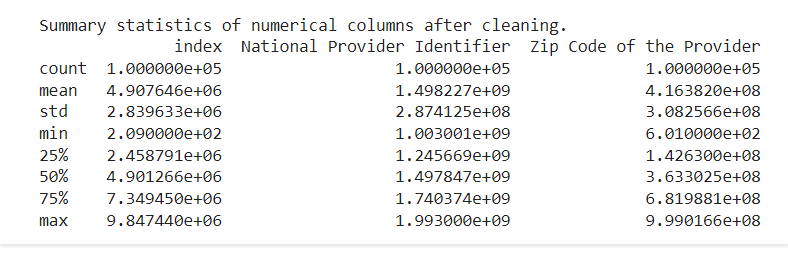




3.Handle missing values :

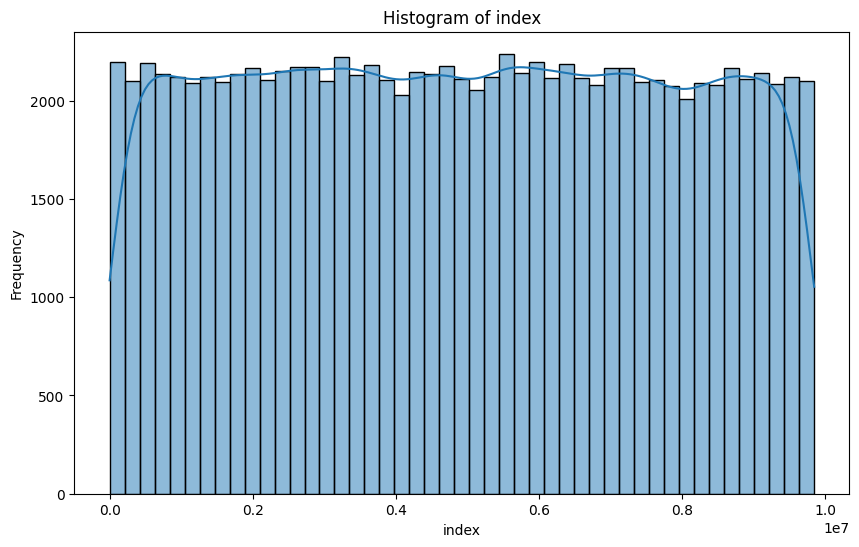


Data cleaning:



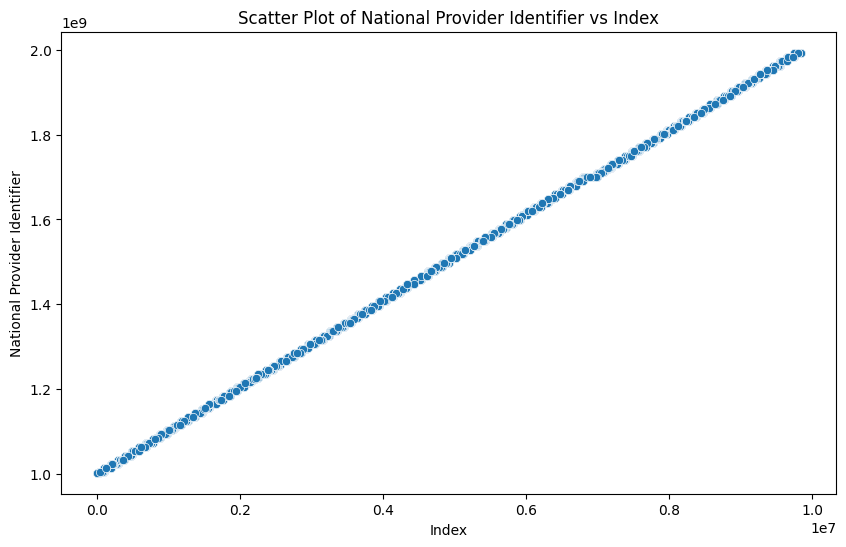
Exploratory Data Analysis(EDA)

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.



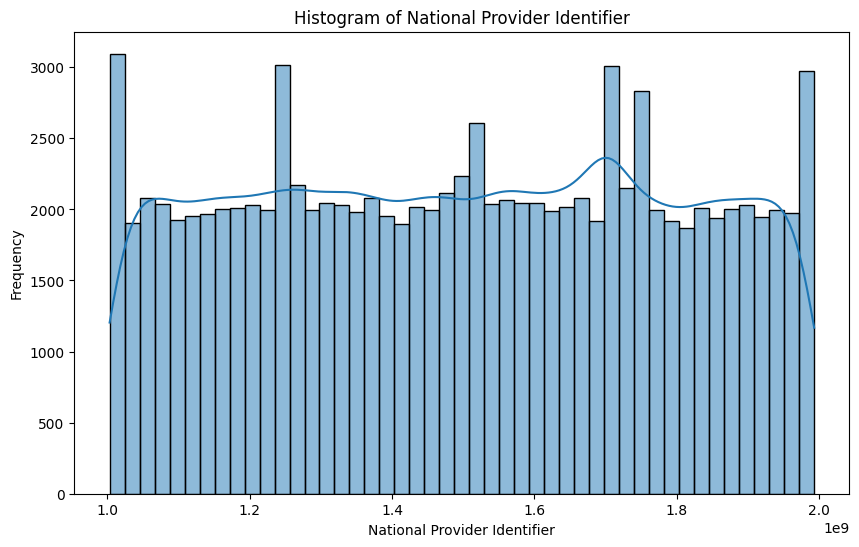
observations

* The distribution is nearly **uniform**, with each interval having a similar frequency count.
* There is no significant peak or skewness, indicating a relatively **even spread** of data points across the index range.



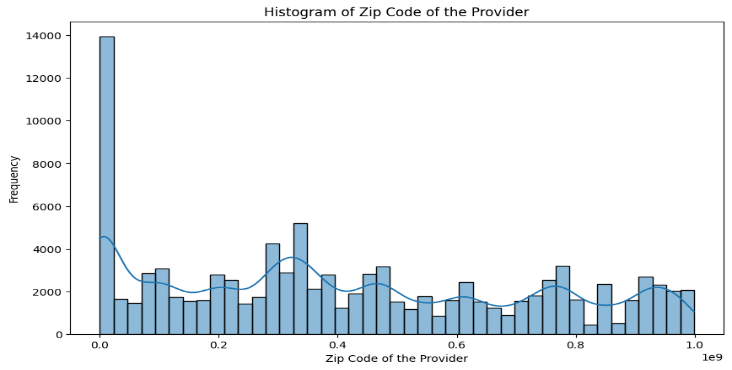
Observation:  The scatter plot depicts a perfect linear relationship, where the National Provider Identifier increases linearly with the index.

 This suggests that the National Provider Identifier may have been assigned in a sequential manner, which is why the plot shows a straight line.

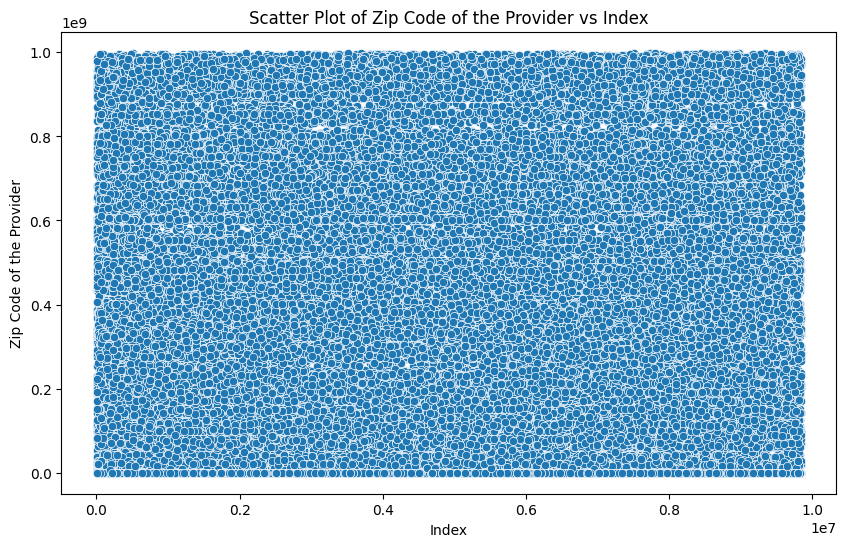


Observation: The histogram of the National Provider Identifier reveals a generally uniform distribution with minor peaks, indicating consistent and systematic assignment of identifiers. The observations from the histogram align with the sequential pattern observed in the scatter plot, reinforcing the understanding of the dataset's structure and integrity.

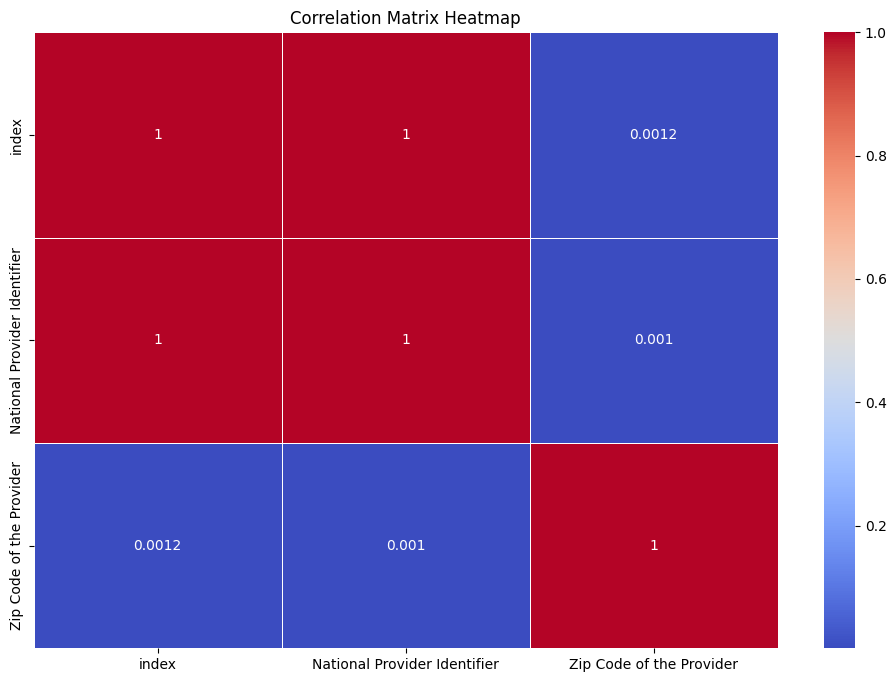
4o



* Observations : **High Frequency at the Beginning**: There is a significant peak near zero on the x-axis, indicating a large number of providers with zip codes in the lower range.
* **Multiple Peaks and Valleys**: The distribution has several peaks and valleys, suggesting that certain zip code ranges are more common than others.
* **Decreasing Frequency**: After the initial peak, there is a noticeable drop in frequency, followed by smaller peaks throughout.
* **Skewed Distribution**: Overall, the histogram is skewed to the left, with most data points falling on the lower end of the zip code range.

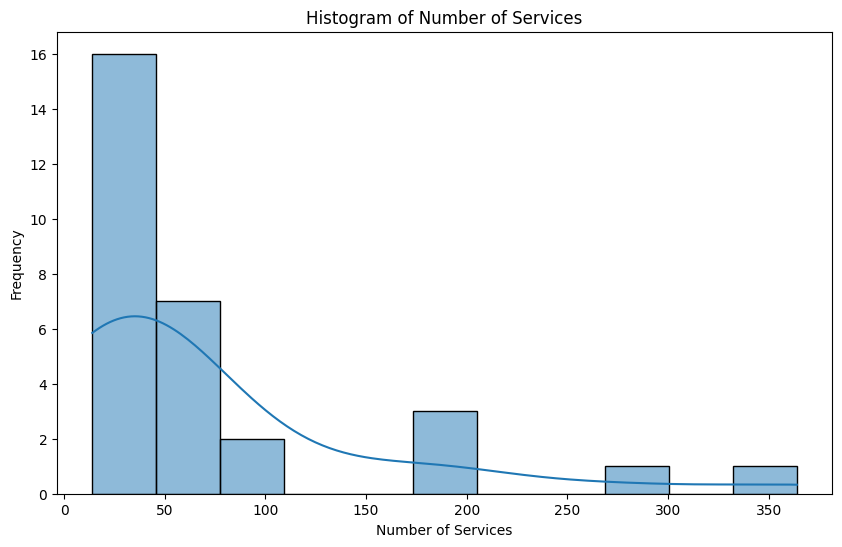


1. Observations: **Uniform Distribution**: The data points are densely spread across the entire plot, indicating a uniform distribution of zip codes across different index values.
2. **No Clear Pattern**: There is no discernible pattern or clustering among the data points. Providers with similar index values have diverse zip codes, and vice versa.



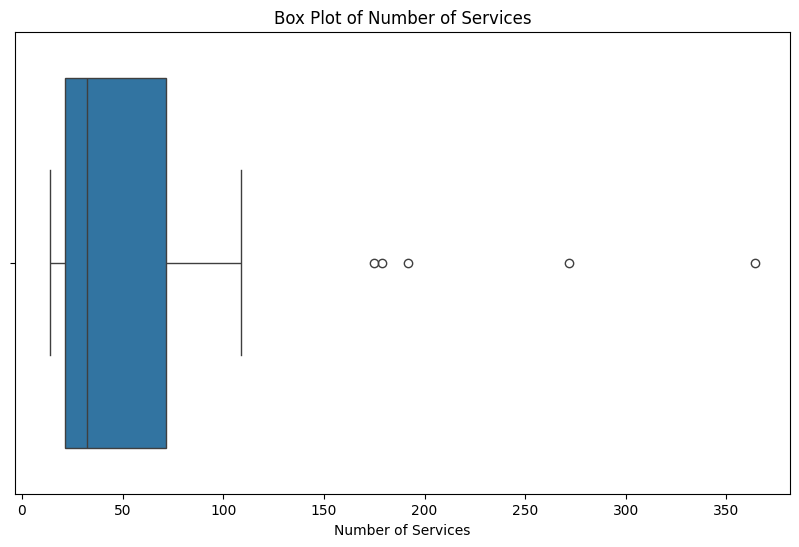
1. Observation ; **Perfect Positive Correlations**:
   * The diagonal elements (self-correlations) are all 1.0, which is expected since any variable is perfectly correlated with itself.
   * Specifically, ‘index’ has a perfect positive correlation with itself, as does ‘National Provider Identifier’ (NPI) with itself.

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



Observations from the graph are as follows:

* + There is a significant peak in the frequency of the number of services in the 0-50 range.
  + This suggests that a large number of entities provide a relatively small number of services.

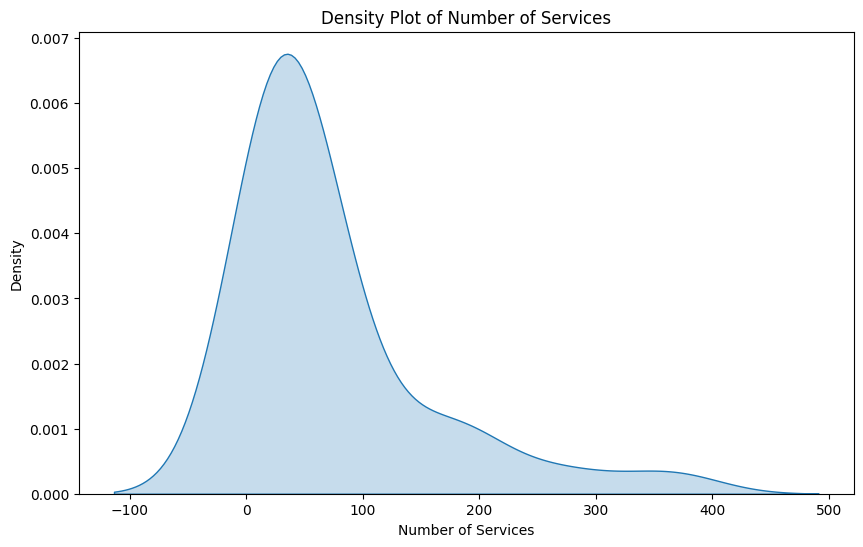


observations:

* + The median (central line inside the box) appears to be around 100 services.
  + This suggests that approximately half of the entities provide services around this median value.
  + The interquartile range (IQR), represented by the height of the box, spans from approximately 25 to 75 services.
  + The middle 50% of entities fall within this range.
  + The whiskers extend from around 25 to 175 services, capturing most of the data.

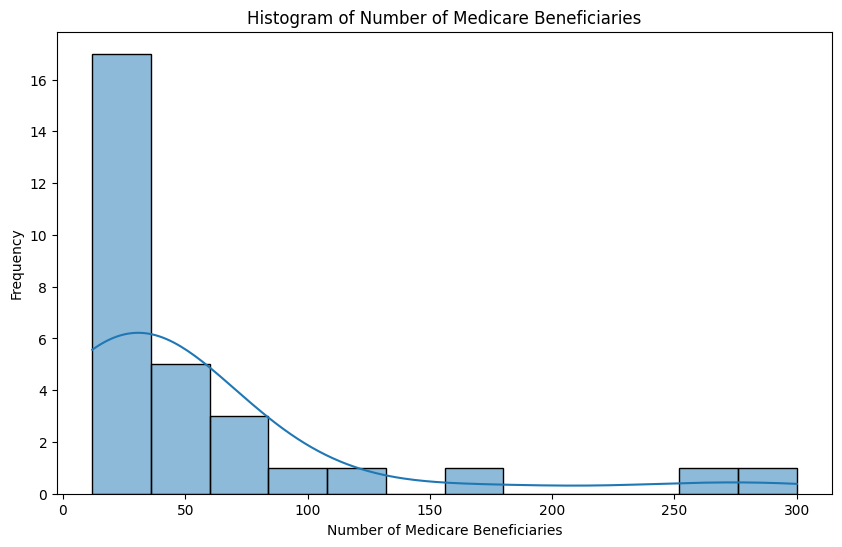
**Outliers**:

* + There are several individual points plotted beyond the upper whisker (values around 200, 225, 250, and 300).
  + These points are likely outliers, indicating a few entities with exceptionally high numbers of services.



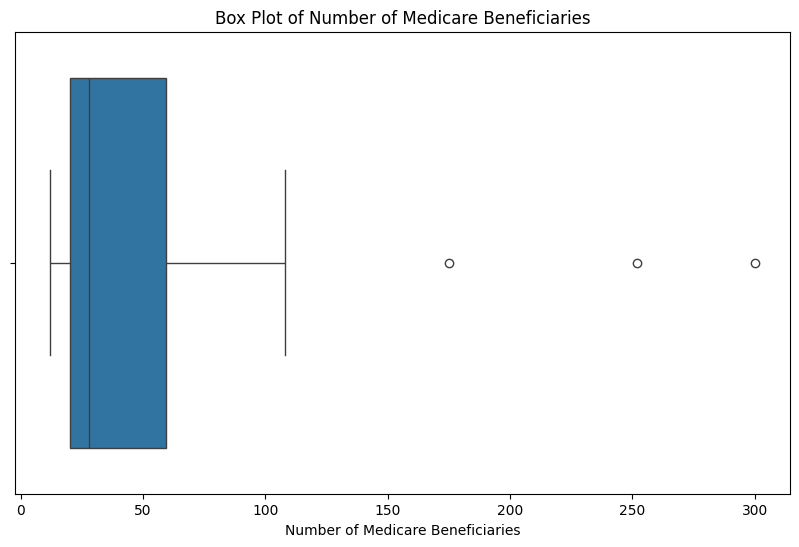
Observations:

* + The density plot shows a sharp peak around the number 0 on the horizontal axis (Number of Services).
  + There are individual points plotted beyond the upper whisker (values around 200, 225, 250, and 300).
  + These points represent outliers—entities with exceptionally high numbers of services.

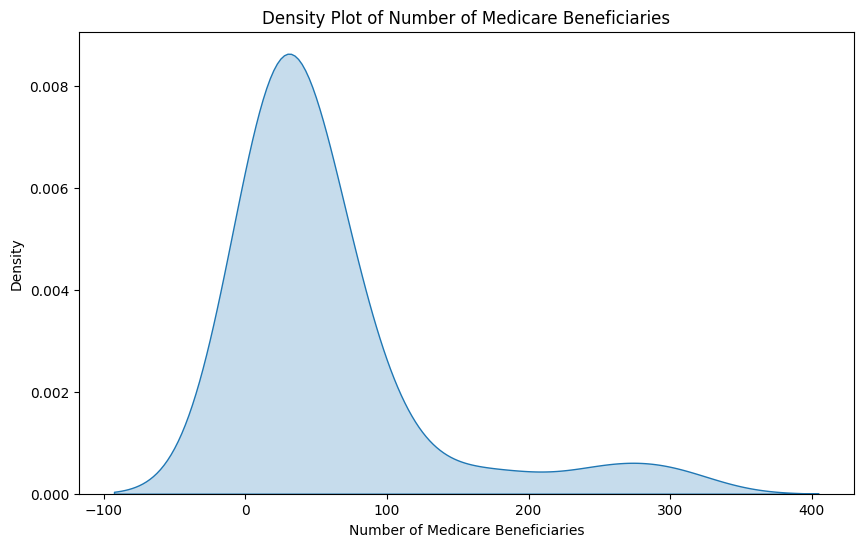


 observations:

* The histogram shows a sharp peak at the leftmost side (around 0 beneficiaries), indicating that a significant number of beneficiaries have received no services or very few services.
  + As the number of beneficiaries increases, the frequency of occurrence decreases sharply.
  + There are significantly fewer beneficiaries in the subsequent ranges (e.g., 50-100, 100-150, etc.).

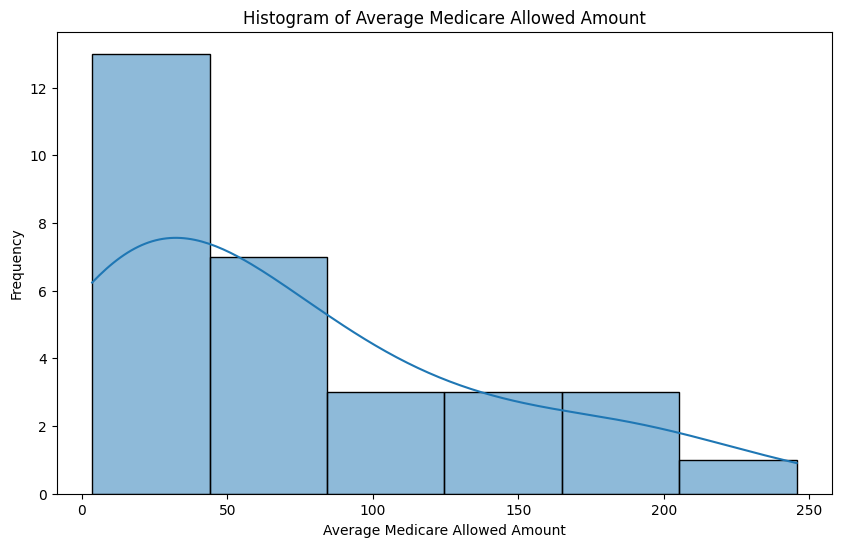


* observations: The **median** (middle value) of the number of beneficiaries is around **100**, as indicated by the line within the box.
* The **interquartile range (IQR)**, which spans from the first quartile (Q1) to the third quartile (Q3), is quite large, suggesting significant variability in the number of beneficiaries.
* The **lower quartile (Q1)** lies around **50 beneficiaries**, while the **upper quartile (Q3)** is close to **150 beneficiaries**.
* There are **outliers** present above the upper whisker, indicating values significantly higher than expected based on Q3 and IQR.
* The ‘whiskers’ extend from approximately **0 to over 200**, suggesting that most data points fall within this range.



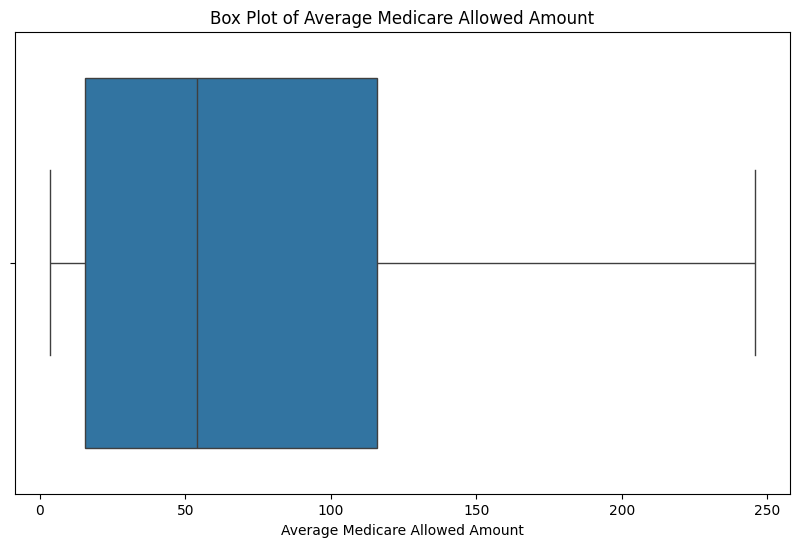
observations:

* + The density plot shows a sharp peak around the number 0 on the horizontal axis (Number of Medicare Beneficiaries).
  + As the number of beneficiaries increases, the frequency of occurrence decreases sharply.
  + There are significantly fewer beneficiaries in the subsequent ranges (e.g., 50-100, 100-150, etc.).
  + The line graph overlaid on the histogram suggests a normal distribution curve but skewed to the left.



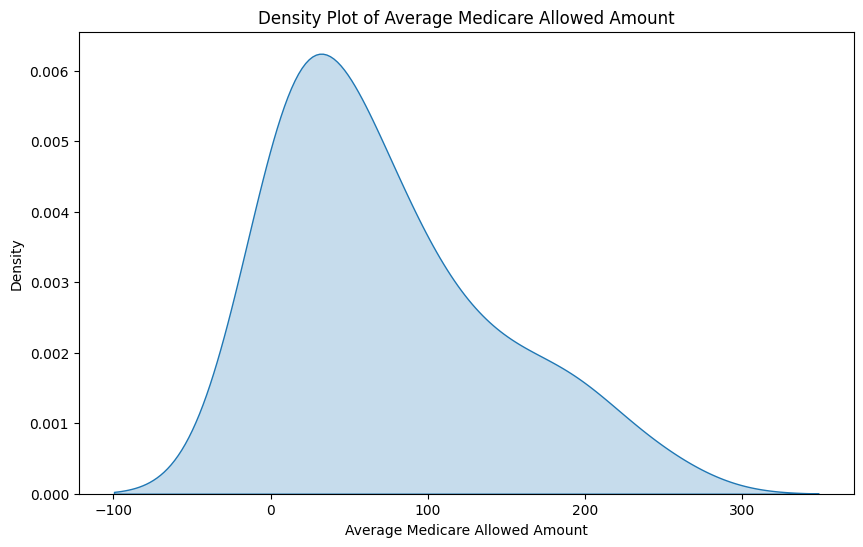
observations:

* + The density plot shows a sharp peak around the number 0 on the horizontal axis (Number of Medicare Beneficiaries).
  + As the number of beneficiaries increases, the frequency of occurrence decreases sharply.
  + There are significantly fewer beneficiaries in the subsequent ranges (e.g., 50-100, 100-150, etc.).
  + This skewness indicates that there is a higher frequency of lower values (fewer beneficiaries) in this dataset.



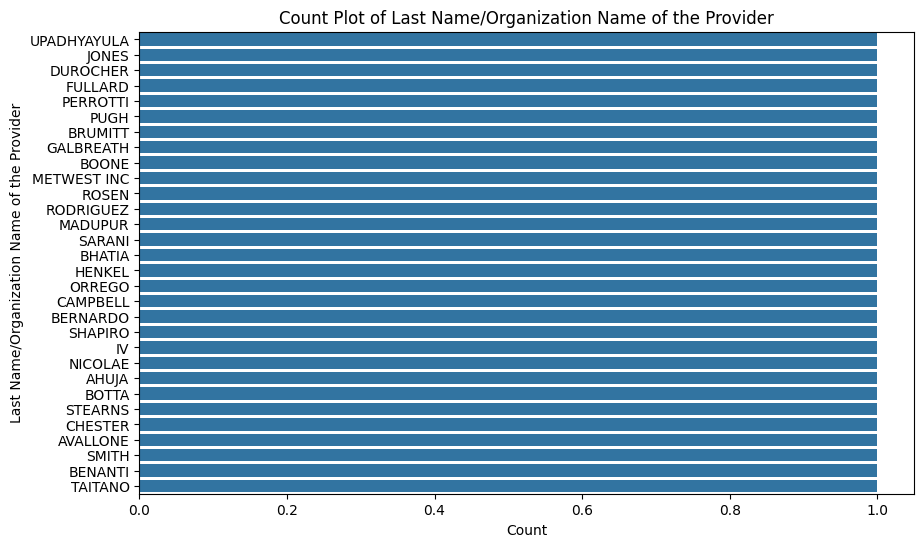
observations:

* **Central Tendency**: The median value (50th percentile) lies around $100, indicating that half of the Medicare payments fall below this amount.
* **Variability**: The interquartile range (IQR) spans from approximately $50 to $200, showing the middle 50% of payments.
* **Outliers**: There are no visible outliers above the upper whisker, suggesting that extreme payment amounts are rare.



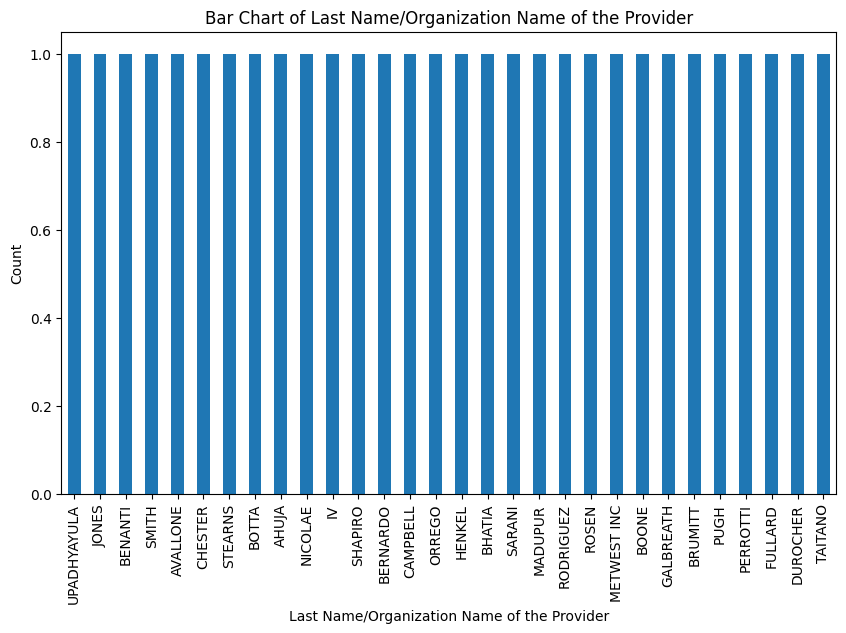
observations:

* **Distribution Shape**: The plot is bell-shaped and skewed to the right, indicating that most average amounts cluster around a mode slightly above zero.
* **Long Tail**: There is a long tail extending toward higher values (up to 300), suggesting some extreme payment amounts.
* **Density**: The density ranges from 0 to approximately 0.006, emphasizing the concentration of values around the mode.
* Univariate analysis for categorical columns



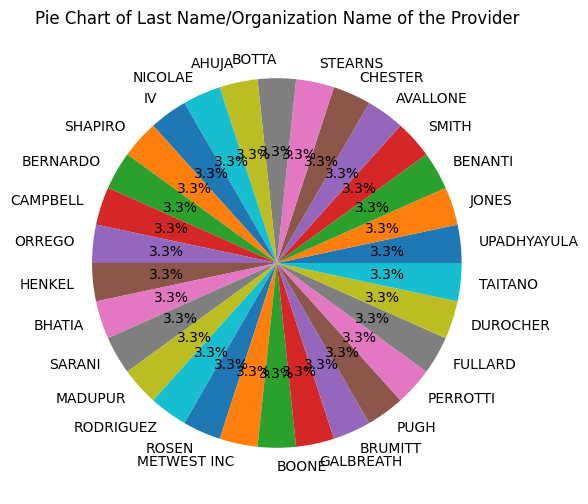
observations:

* **Distribution Shape**: The plot is bell-shaped and skewed to the right, indicating that most average amounts cluster around a mode slightly above zero.
* **Long Tail**: There is a long tail extending toward higher values (up to 300), suggesting some extreme payment amounts.
* **Density**: The density ranges from 0 to approximately 0.006, emphasizing the concentration of values around the mode.



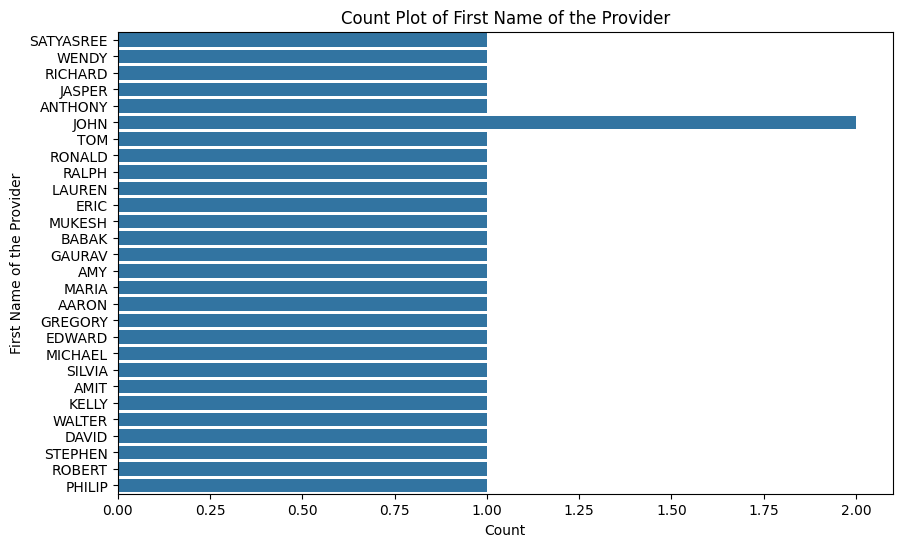
observations:

1. The chart shows the count of providers based on their last names or organization names.
2. The names “INDIVIDUALS,” “BARNETT,” “CALISTI,” and others have equal counts, suggesting that they are common among providers.
3. The chart provides a visual representation of the distribution of provider names.



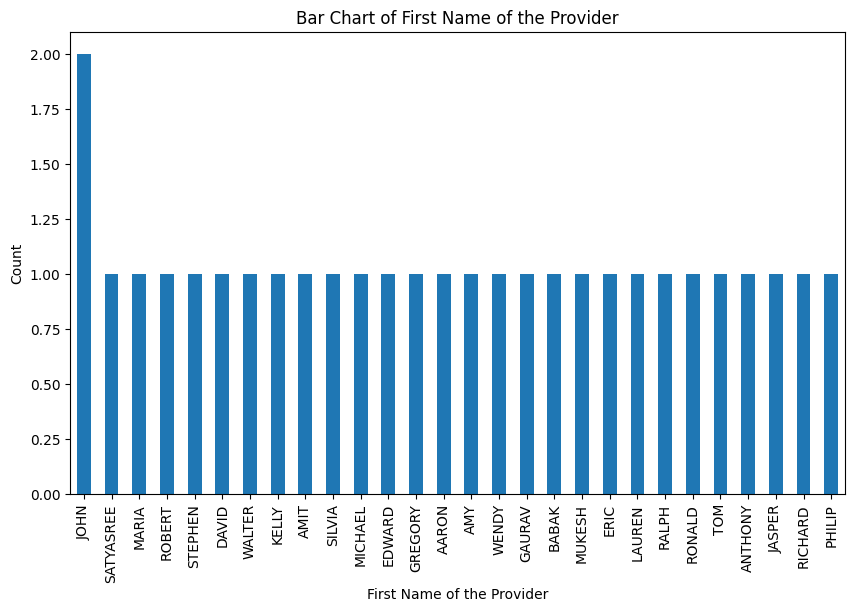
 observations:

* **Distribution Shape**: The plot is bell-shaped and skewed to the right, indicating that most average amounts cluster around a mode slightly above zero.
* **Long Tail**: There is a long tail extending toward higher values (up to 300), suggesting some extreme payment amounts.
* **Density**: The density ranges from 0 to approximately 0.006, emphasizing the concentration of values around the mode.



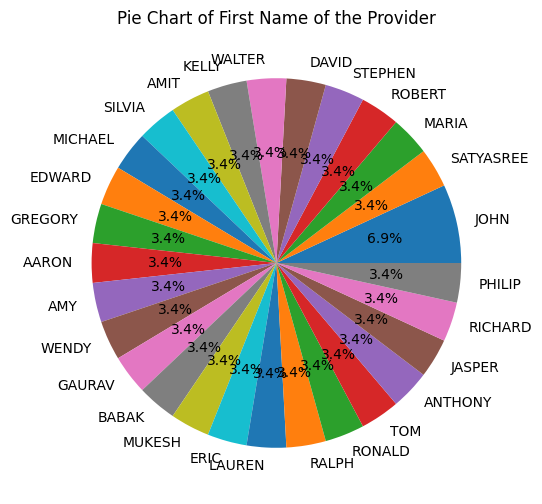
 observations:

* **Distribution Shape**: The plot is bell-shaped and skewed to the right, indicating that most average amounts cluster around a mode slightly above zero.
* **Long Tail**: There is a long tail extending toward higher values (up to 300), suggesting some extreme payment amounts.
* **Density**: The density ranges from 0 to approximately 0.006, emphasizing the concentration of values around the mode.



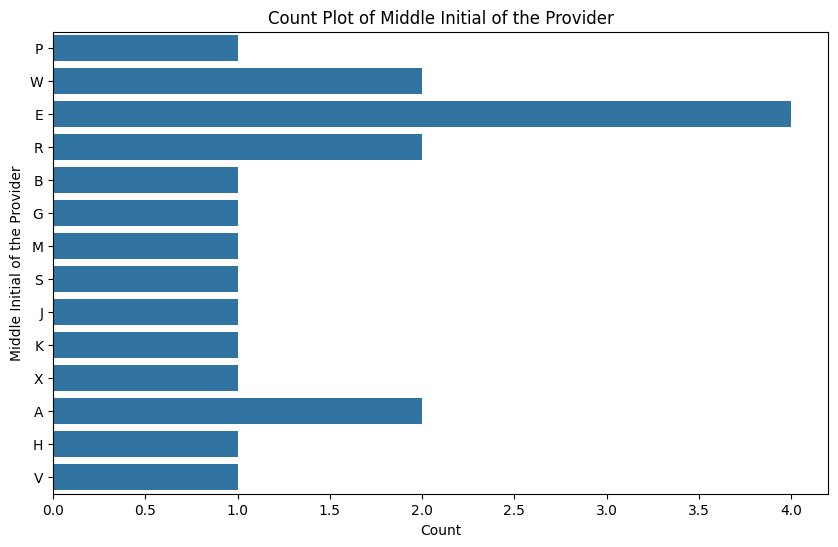
observations:

* **Distribution Shape**: The plot is bell-shaped and skewed to the right, indicating that most average amounts cluster around a mode slightly above zero.
* **Long Tail**: There is a long tail extending toward higher values (up to 300), suggesting some extreme payment amounts.
* **Density**: The density ranges from 0 to approximately 0.006, emphasizing the concentration of values around the mode.



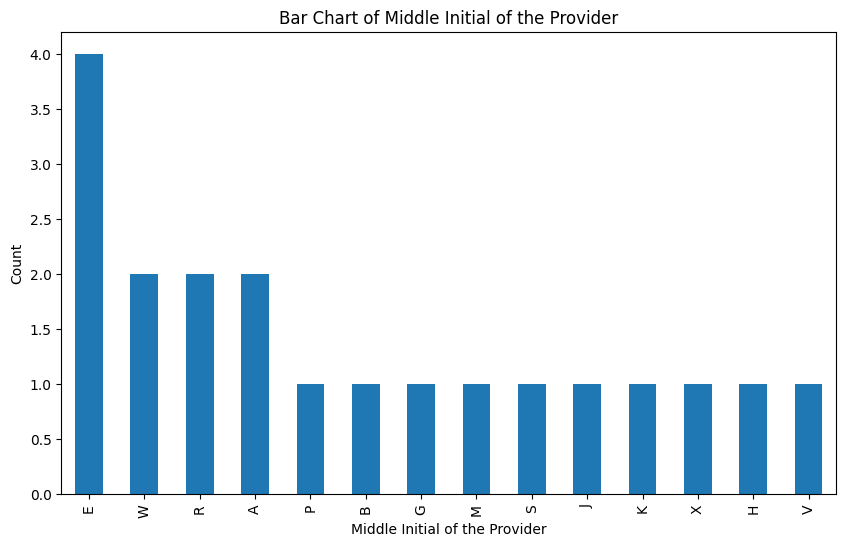
key observations:

* **Distribution Shape**: The plot is bell-shaped and skewed to the right, indicating that most average amounts cluster around a mode slightly above zero.
* **Long Tail**: There is a long tail extending toward higher values (up to 300), suggesting some extreme payment amounts.
* **Density**: The density ranges from 0 to approximately 0.006, emphasizing the concentration of values around the mode.



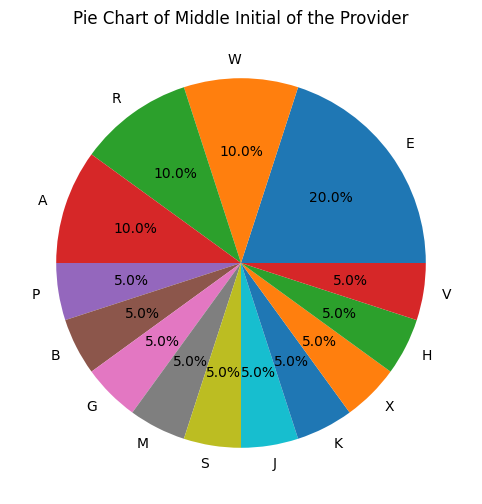
observations:

* **Distribution Shape**: The plot is bell-shaped and skewed to the right, indicating that most average amounts cluster around a mode slightly above zero.
* **Long Tail**: There is a long tail extending toward higher values (up to 300), suggesting some extreme payment amounts.
* **Density**: The density ranges from 0 to approximately 0.006, emphasizing the concentration of values around the mode.



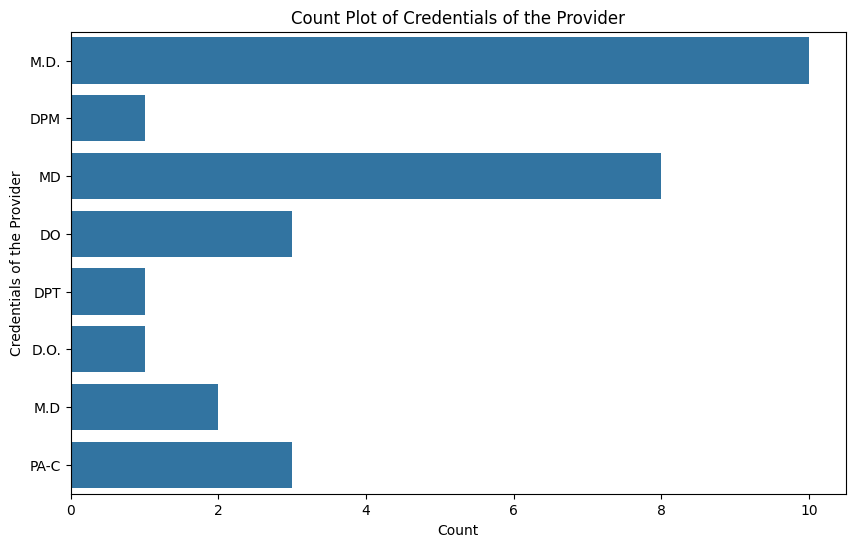
observations:

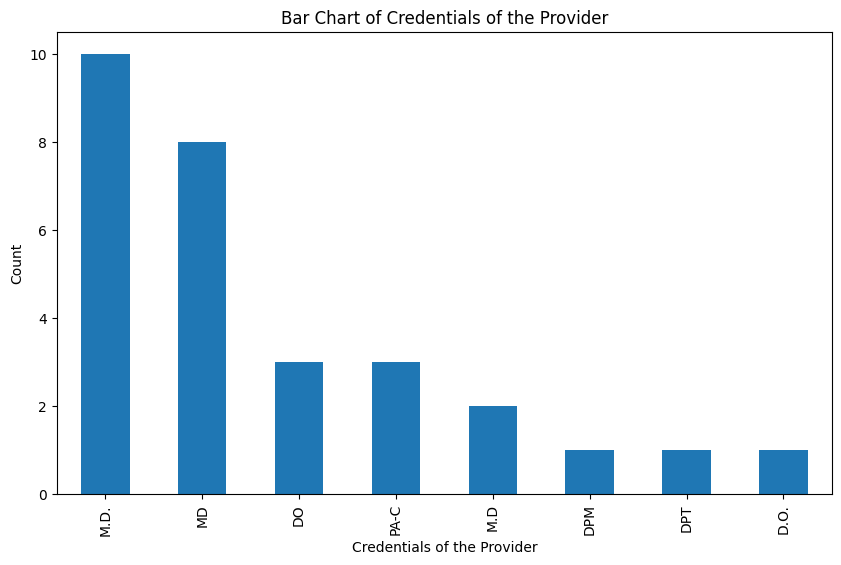
* **Distribution Shape**: The plot is bell-shaped and skewed to the right, indicating that most average amounts cluster around a mode slightly above zero.
* **Long Tail**: There is a long tail extending toward higher values (up to 300), suggesting some extreme payment amounts.
* **Density**: The density ranges from 0 to approximately 0.006, emphasizing the concentration of values around the mode.

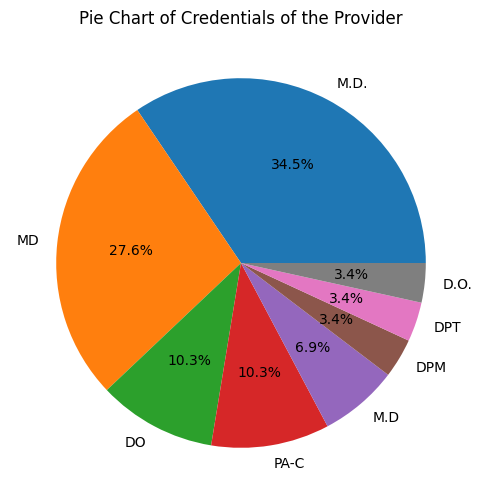


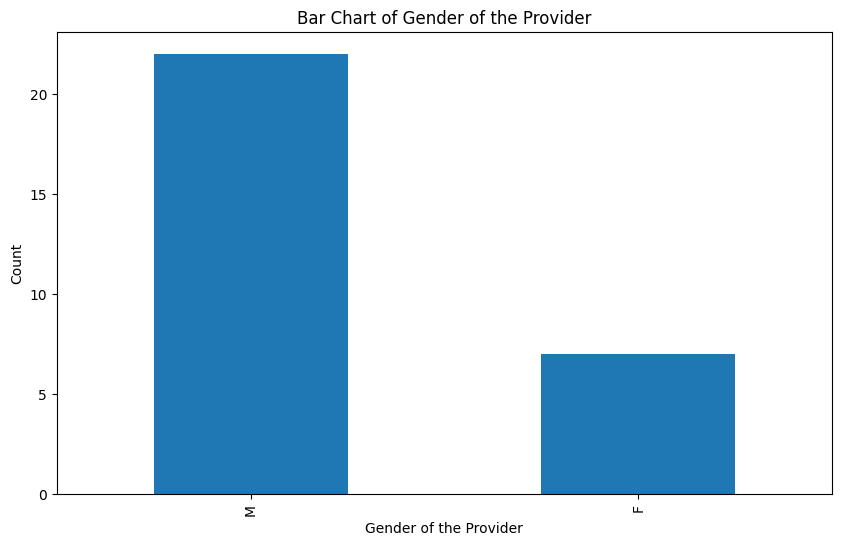
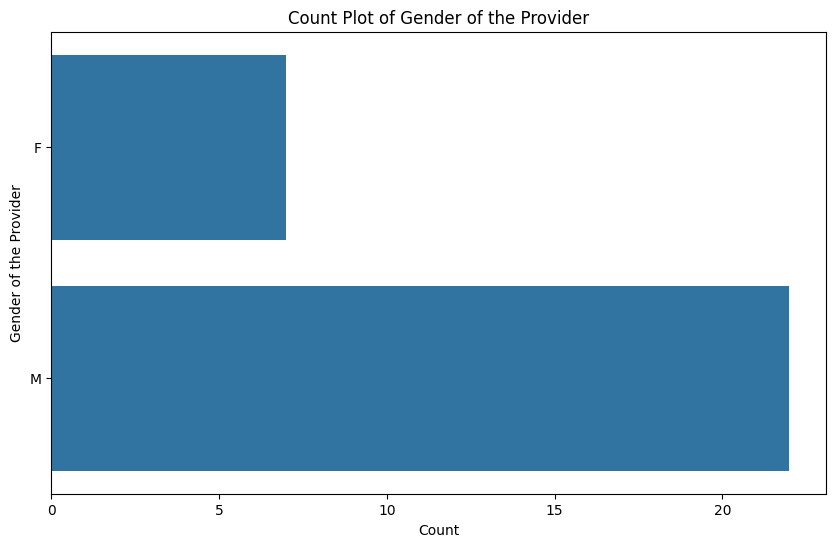
observations:

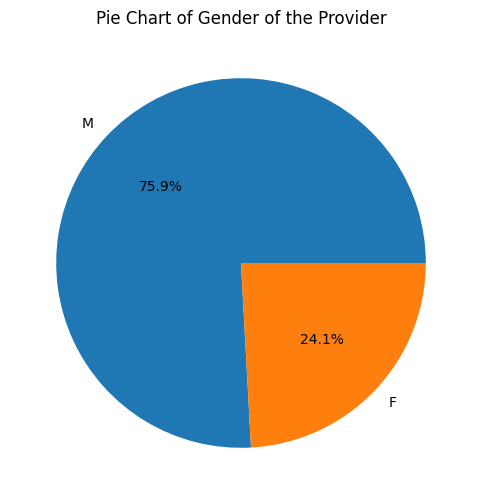
* **Distribution Shape**: The plot is bell-shaped and skewed to the right, indicating that most average amounts cluster around a mode slightly above zero.
* **Long Tail**: There is a long tail extending toward higher values (up to 300), suggesting some extreme payment amounts.
* **Density**: The density ranges from 0 to approximately 0.006, emphasizing the concentration of values around the mode.

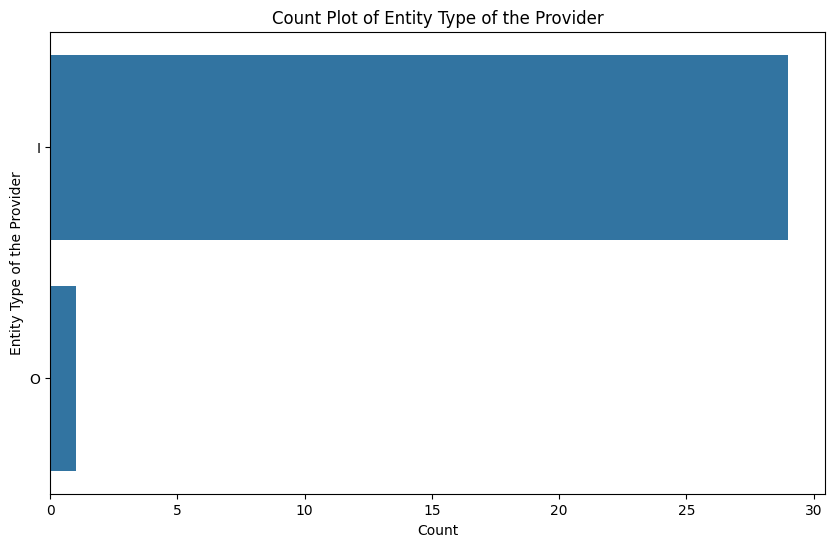


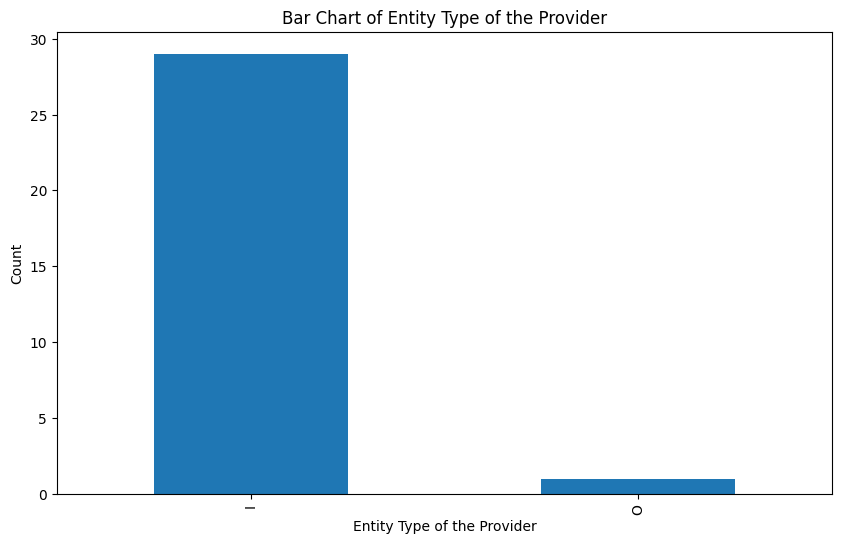


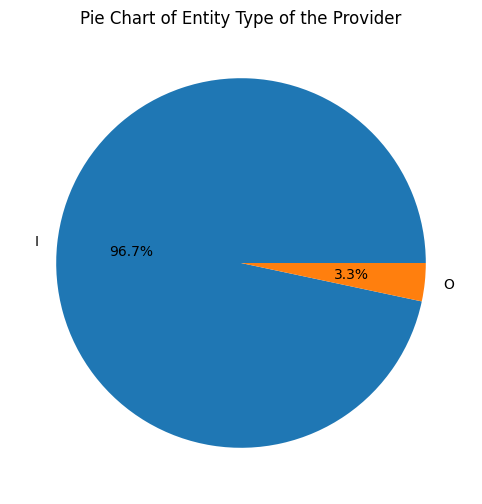


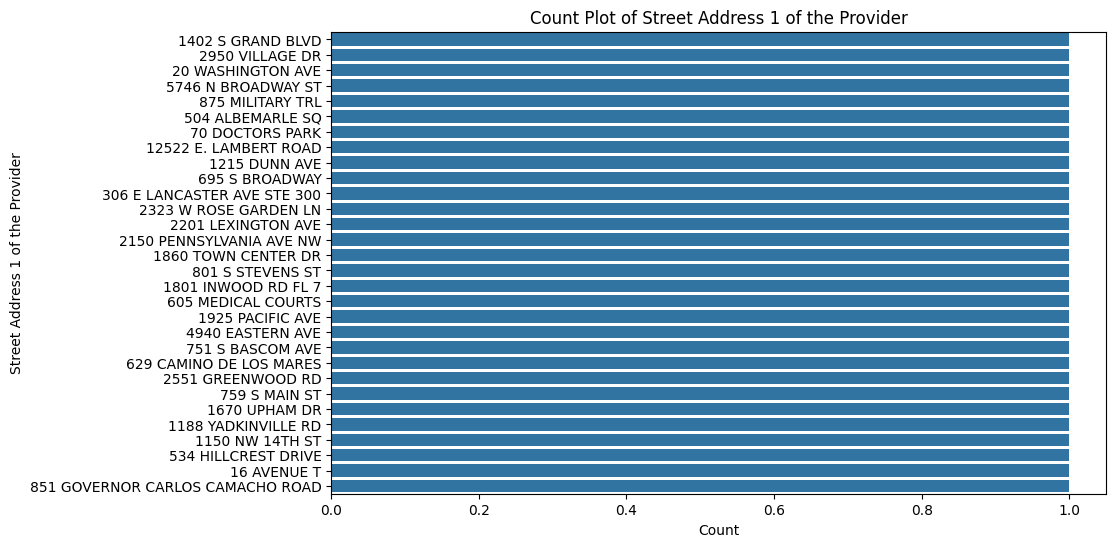


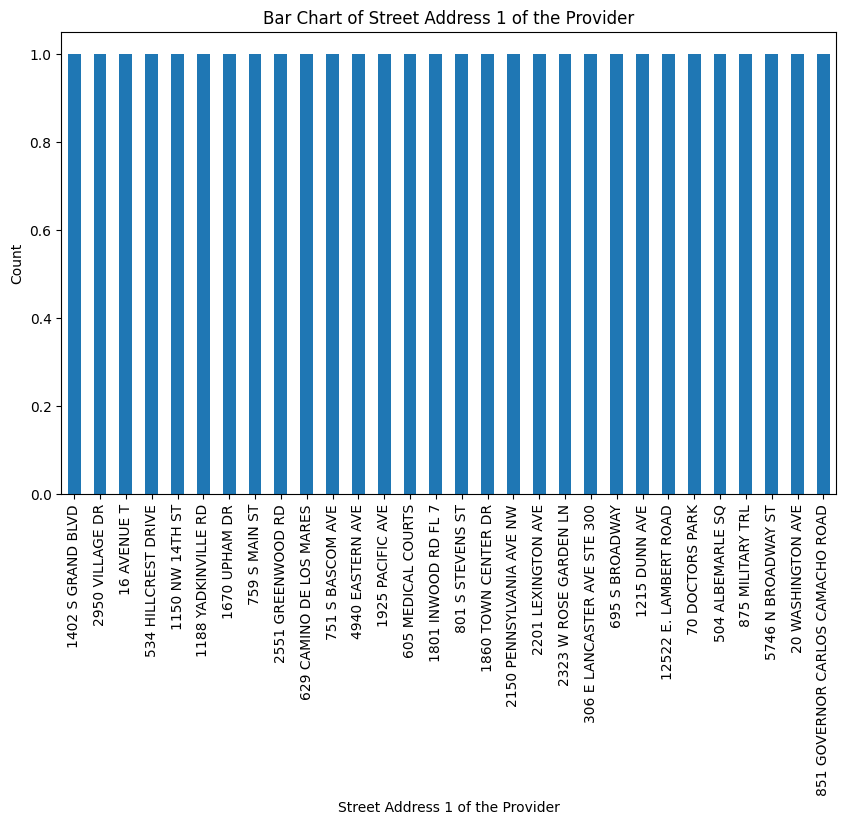


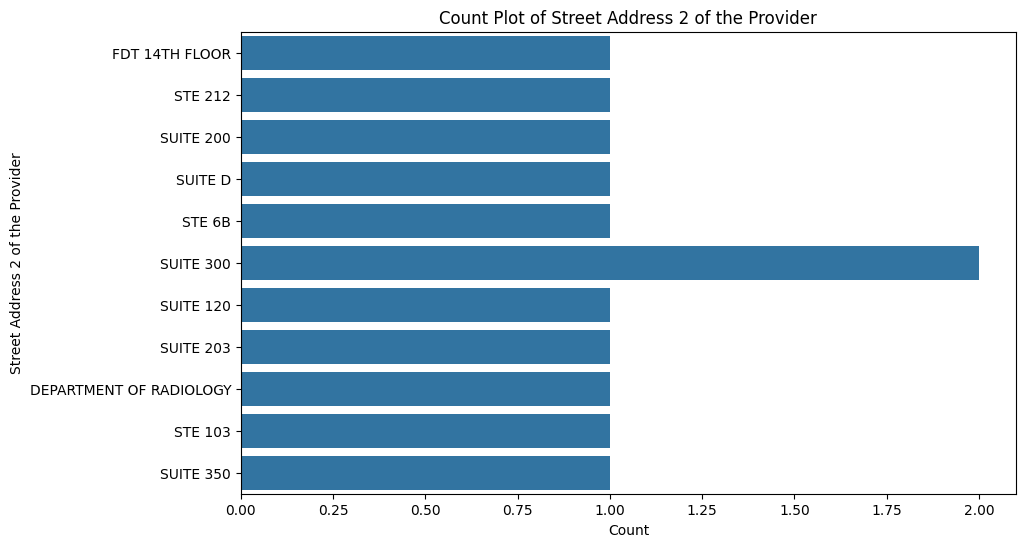


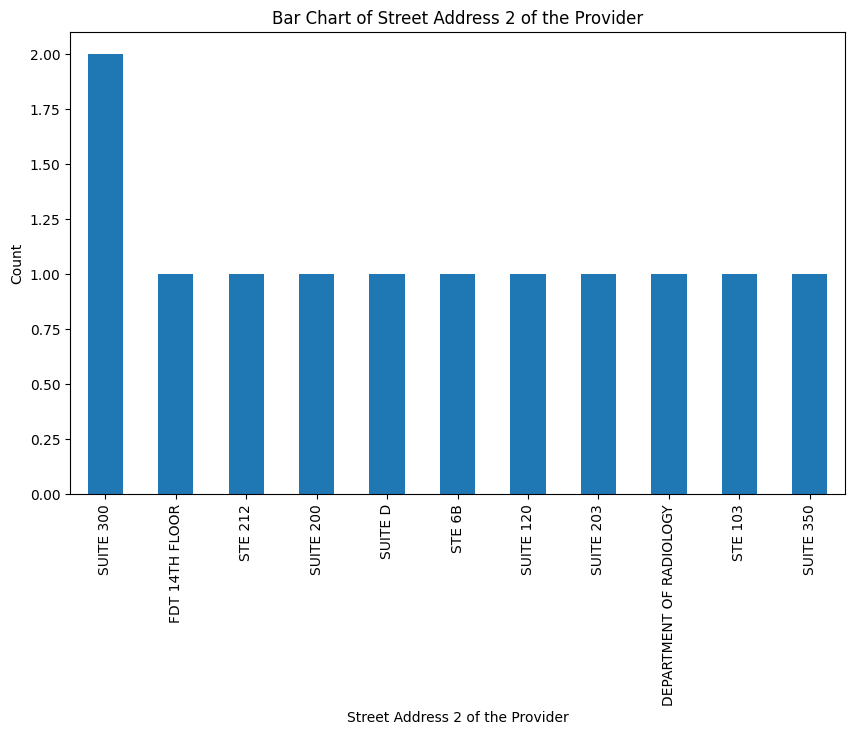


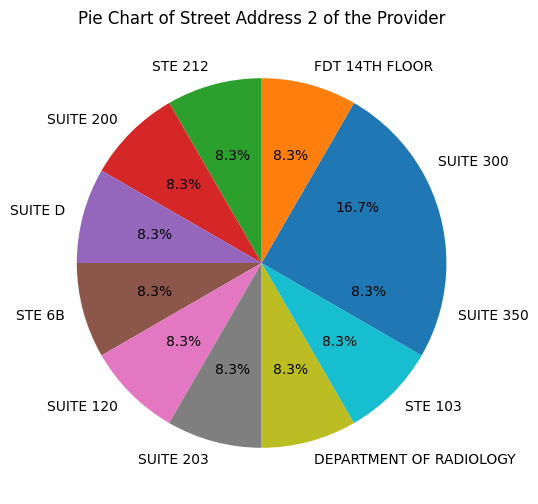


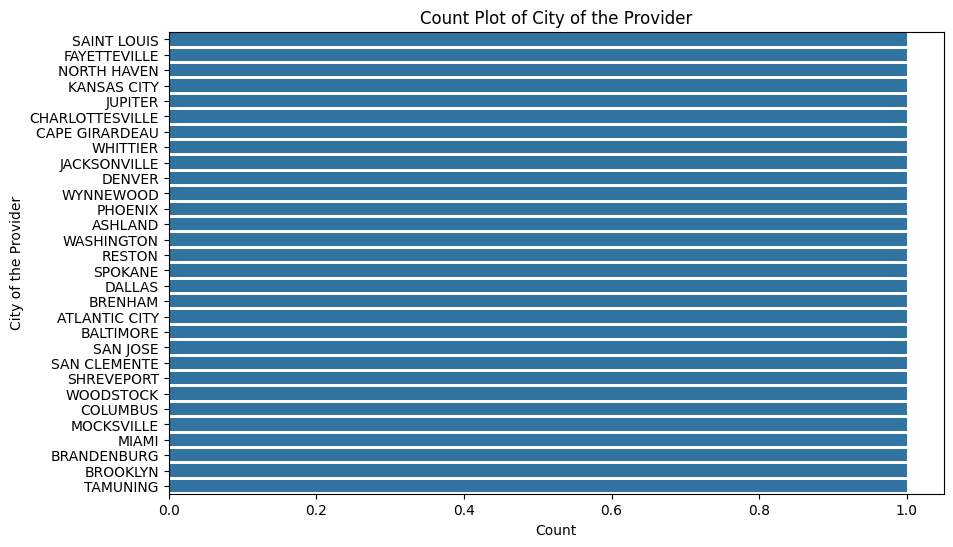


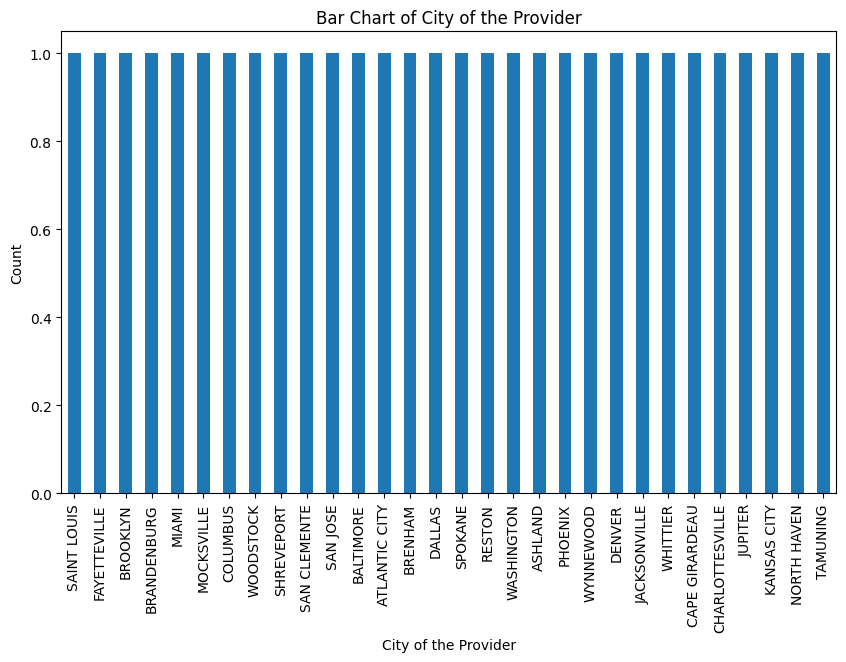


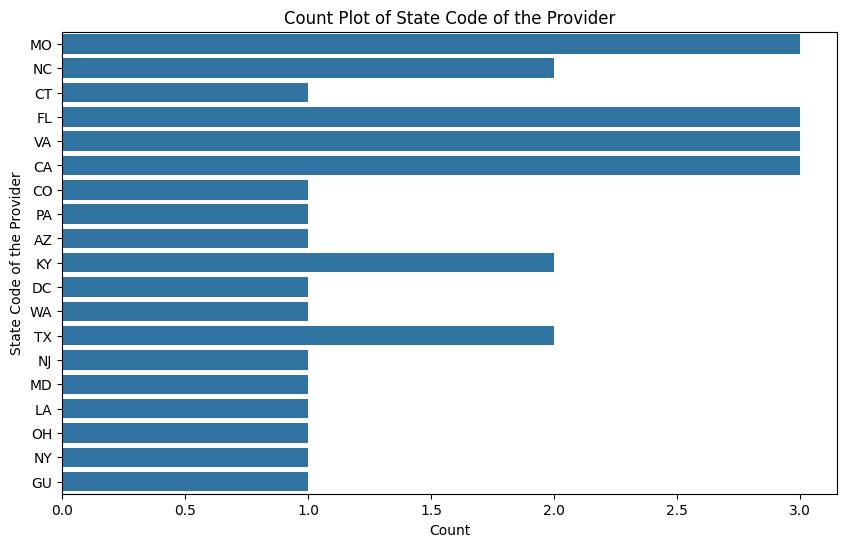


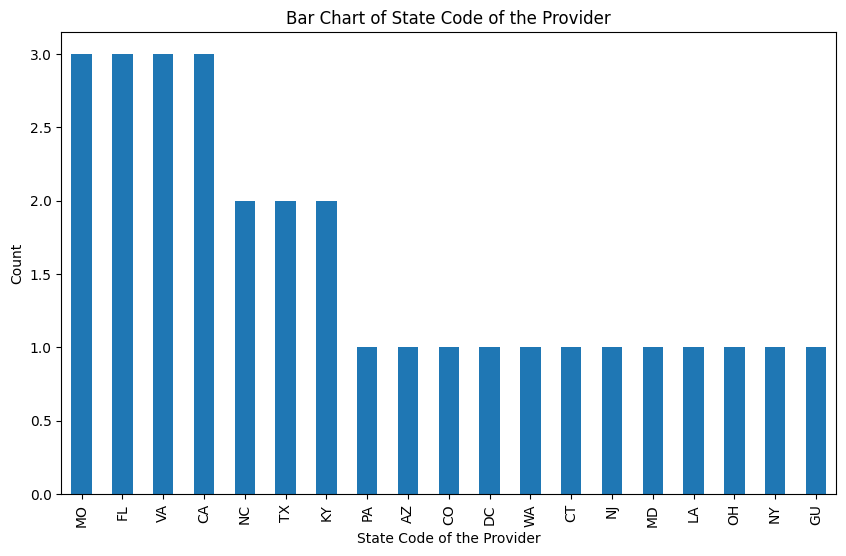


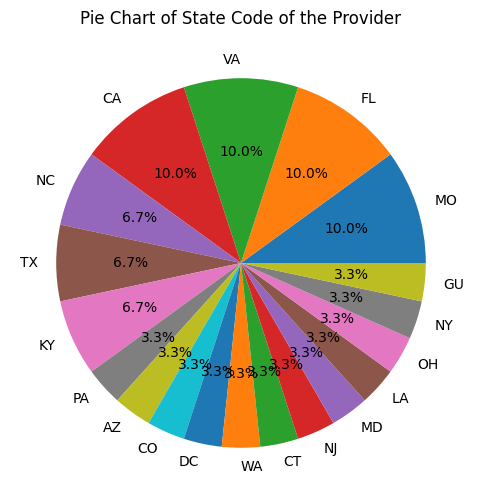


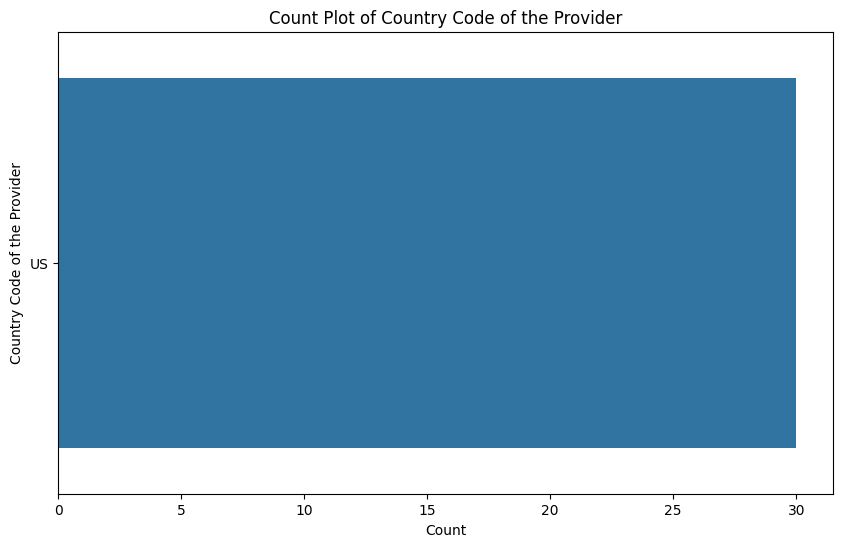


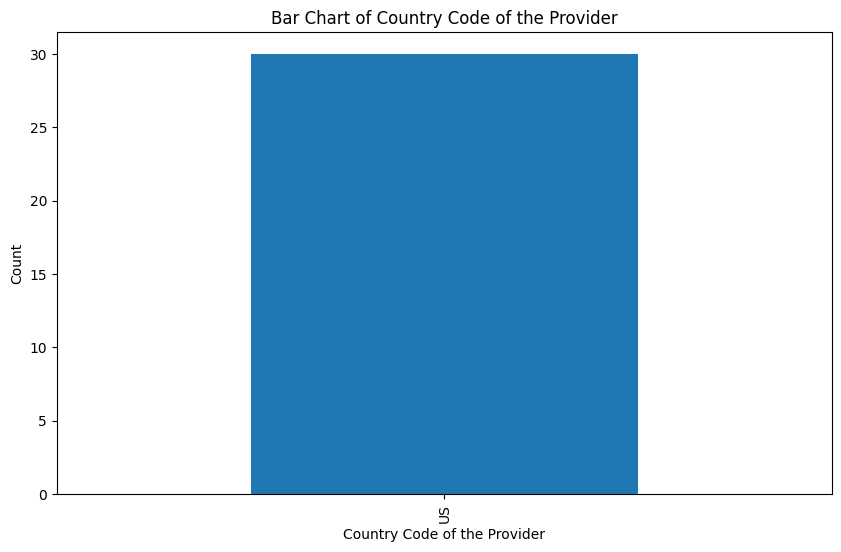


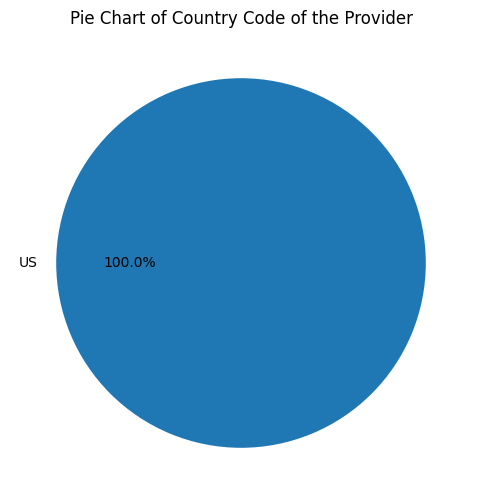


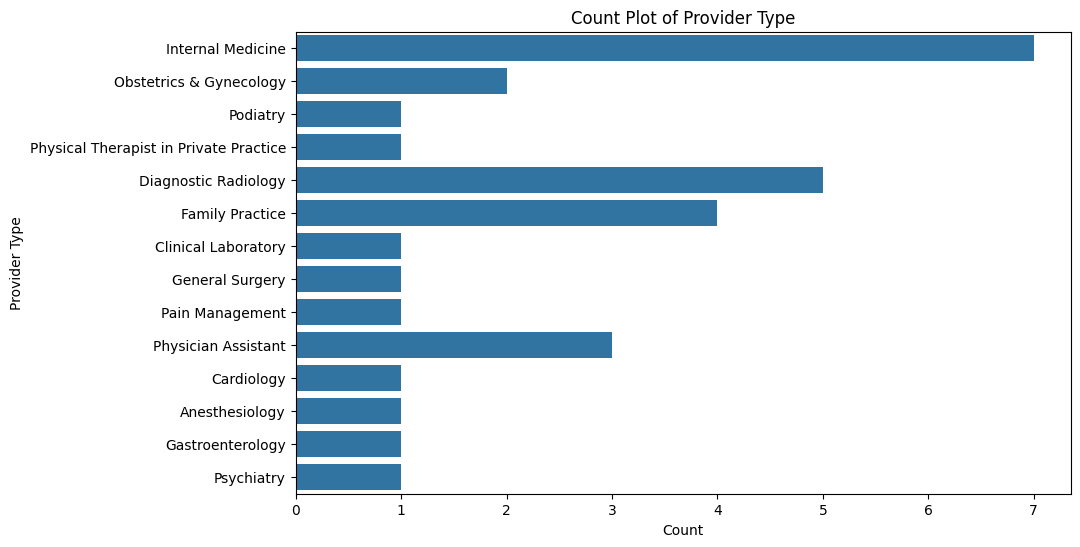


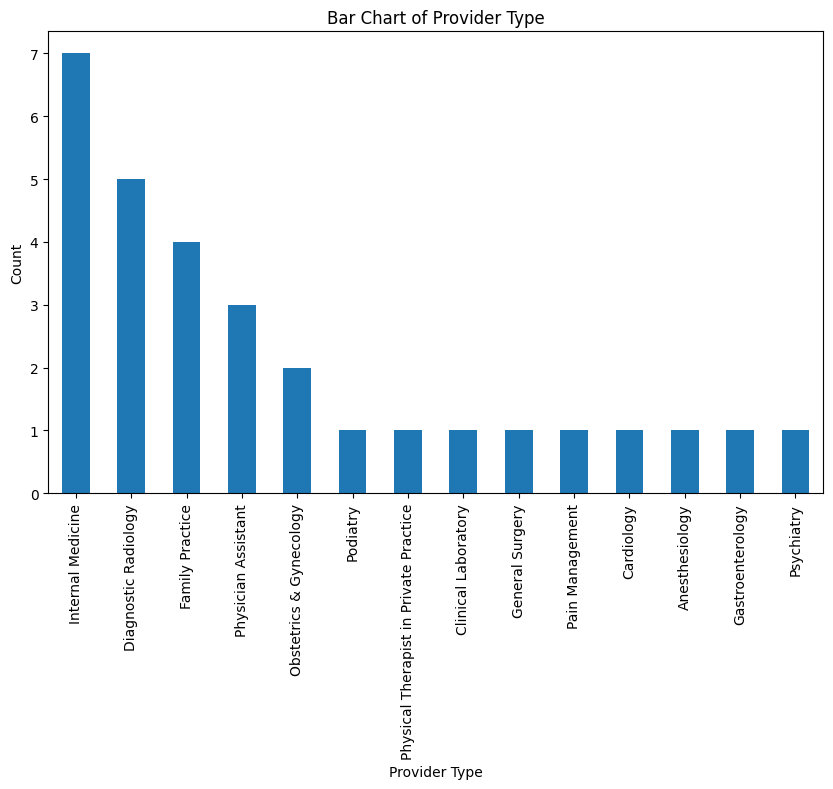


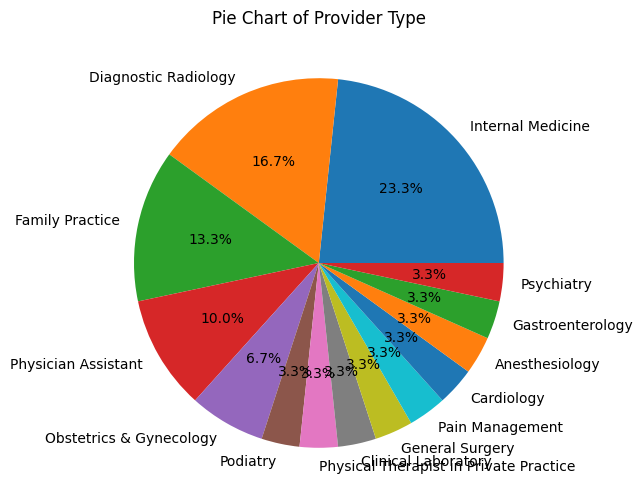


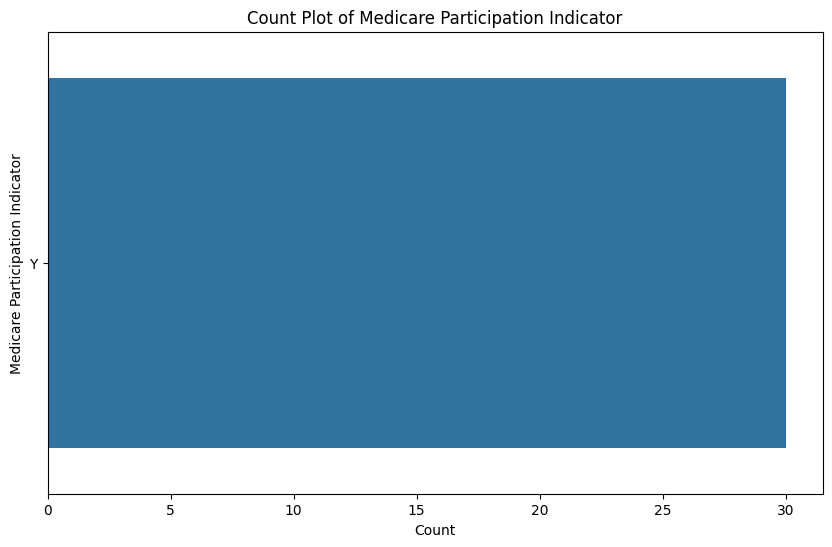


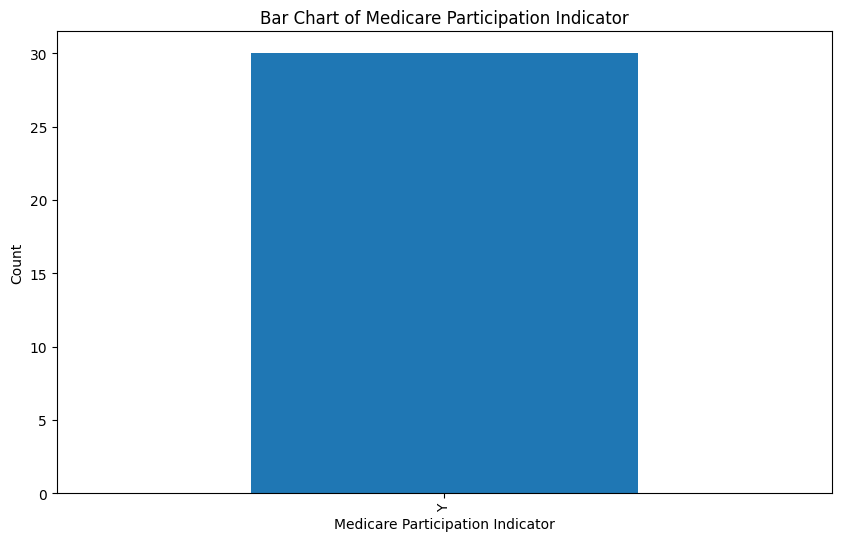


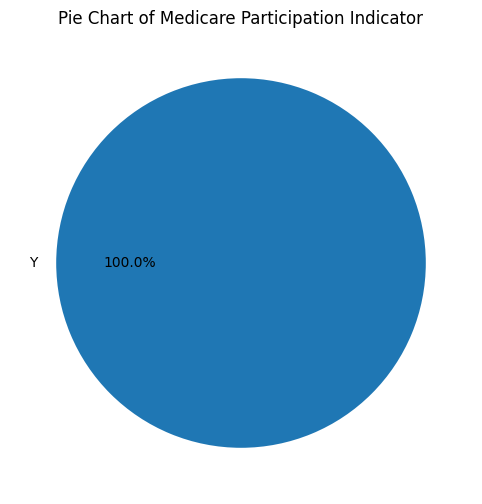


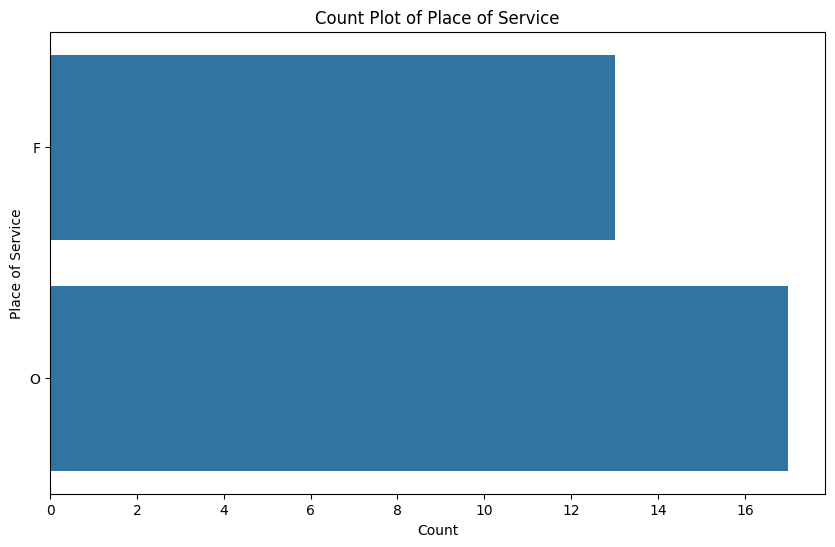


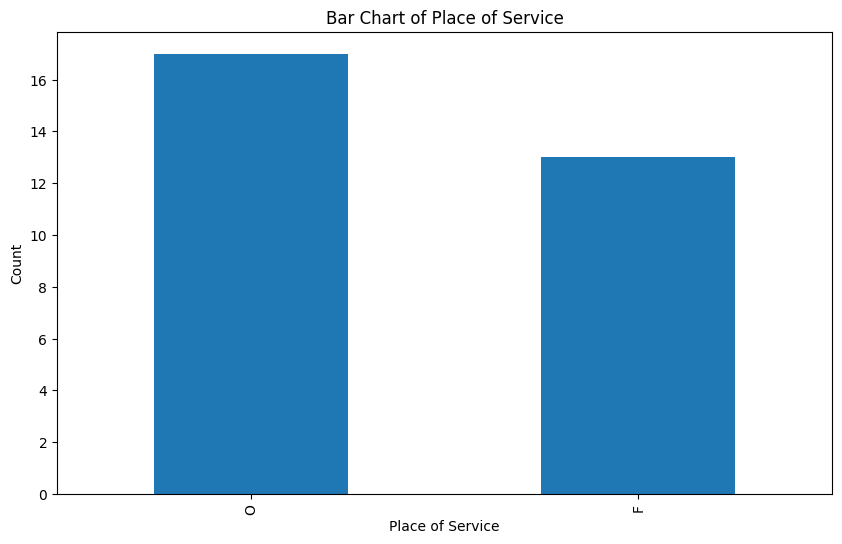


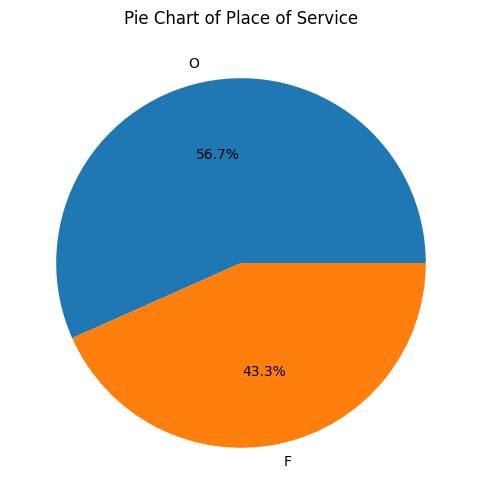


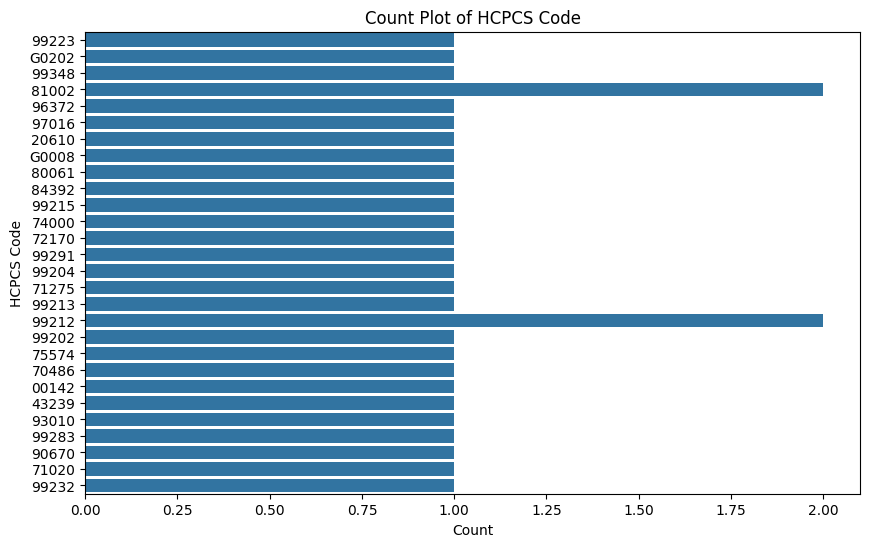


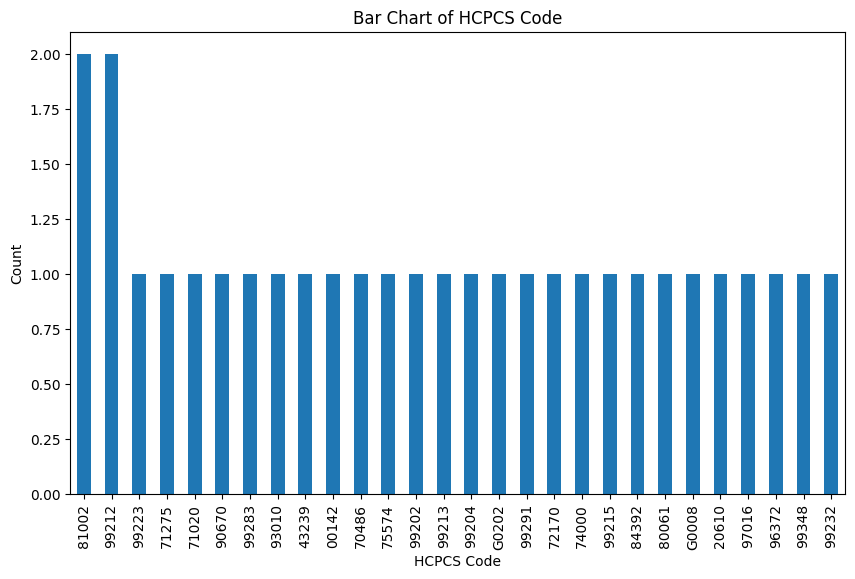


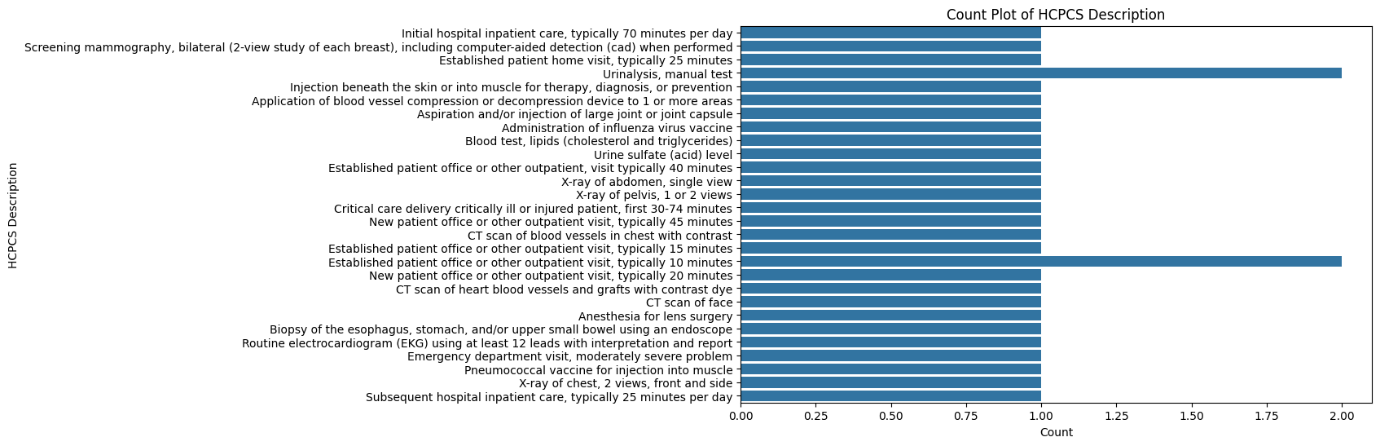


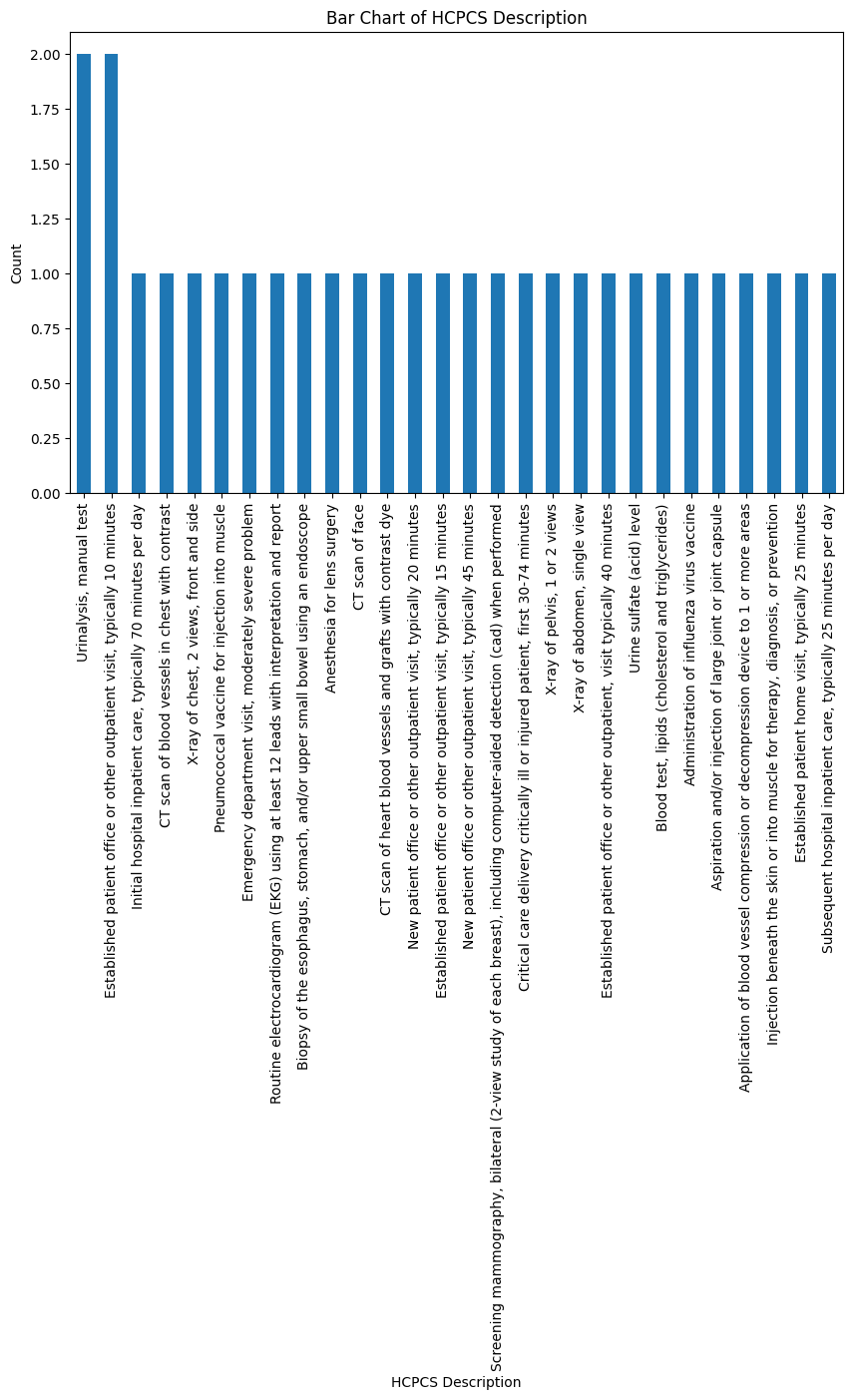


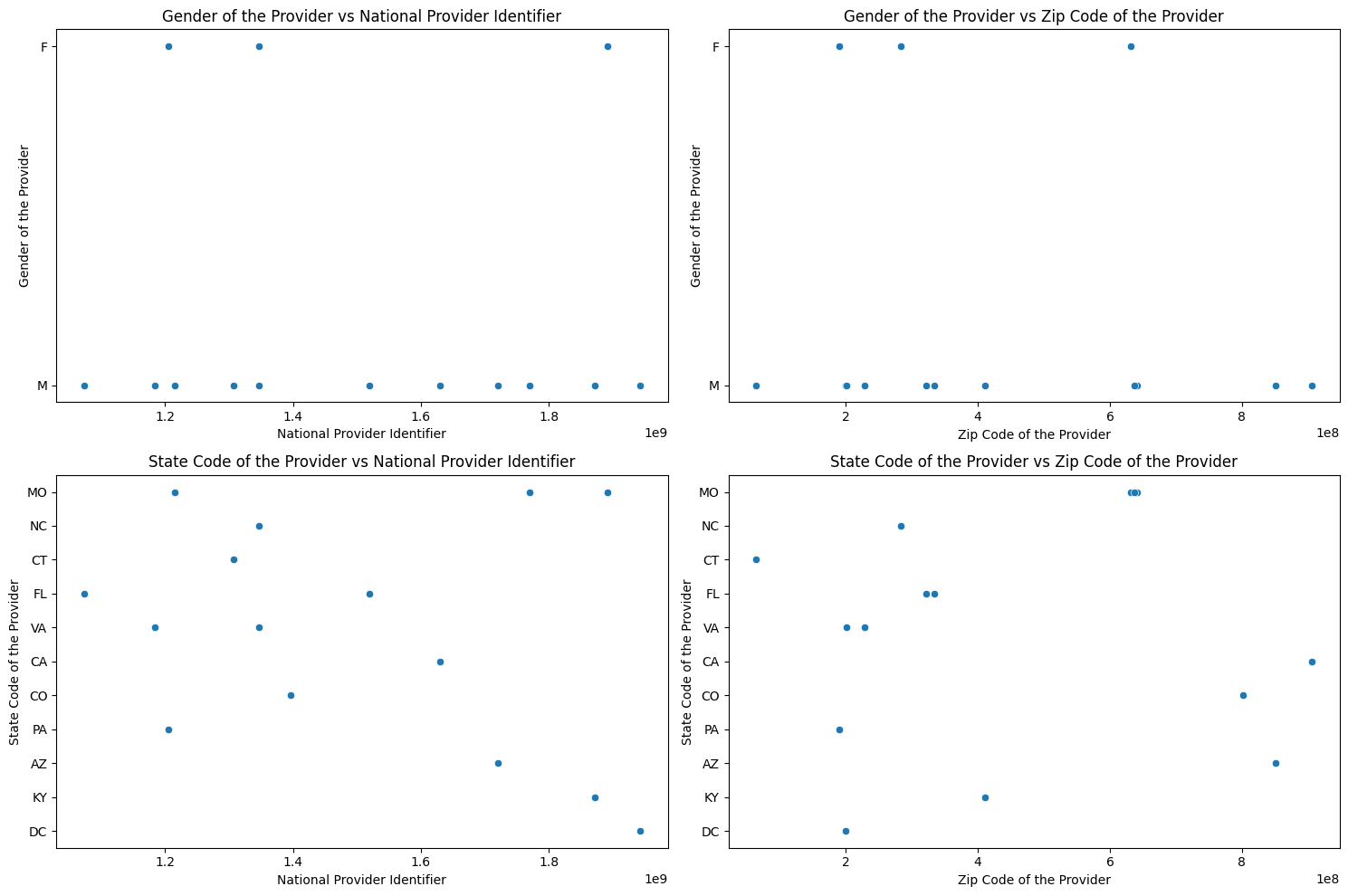




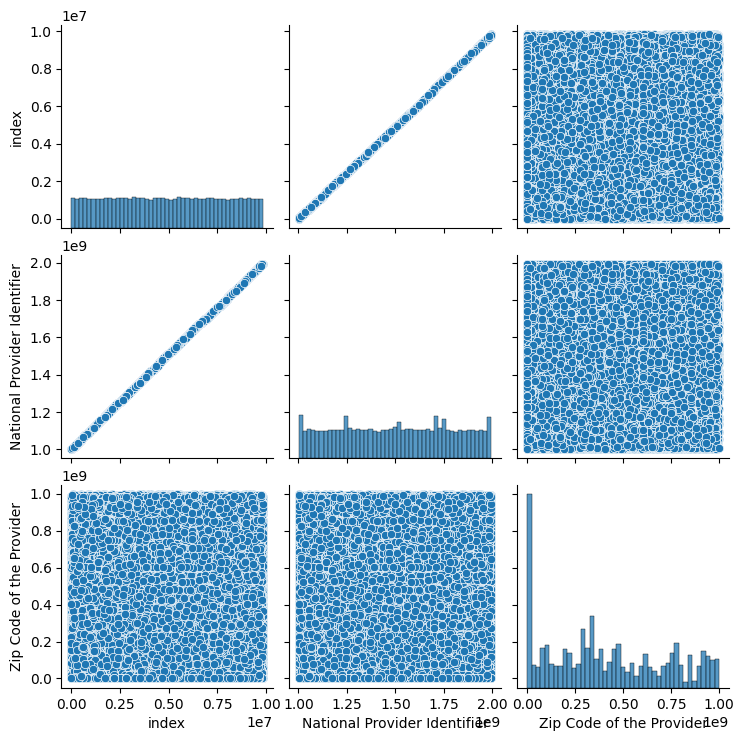








1. **Gender of the Provider vs. National Provider Identifier (NPI)**:
   * No clear correlation or trend between gender and NPI.
   * Data points are randomly distributed.
2. **Gender of the Provider vs. Zip Code of the Provider**:
   * No discernible pattern between gender and zip code.
   * Points are scattered without any clear relationship.
3. **State Code of the Provider vs. NPI**:
   * No significant correlation between state code and NPI.
   * Data points are widely spread across different states.
4. **State Code of the Provider vs. Zip Code of the Provider**:
   * Similar to the previous plot, no strong relationship between state code and zip code.
   * Points are randomly distributed across states and zip codes.



Observations:

1.for the graph index vs index all the x-axis indices are at the same range and their value is in

Between 0.1-0.2.

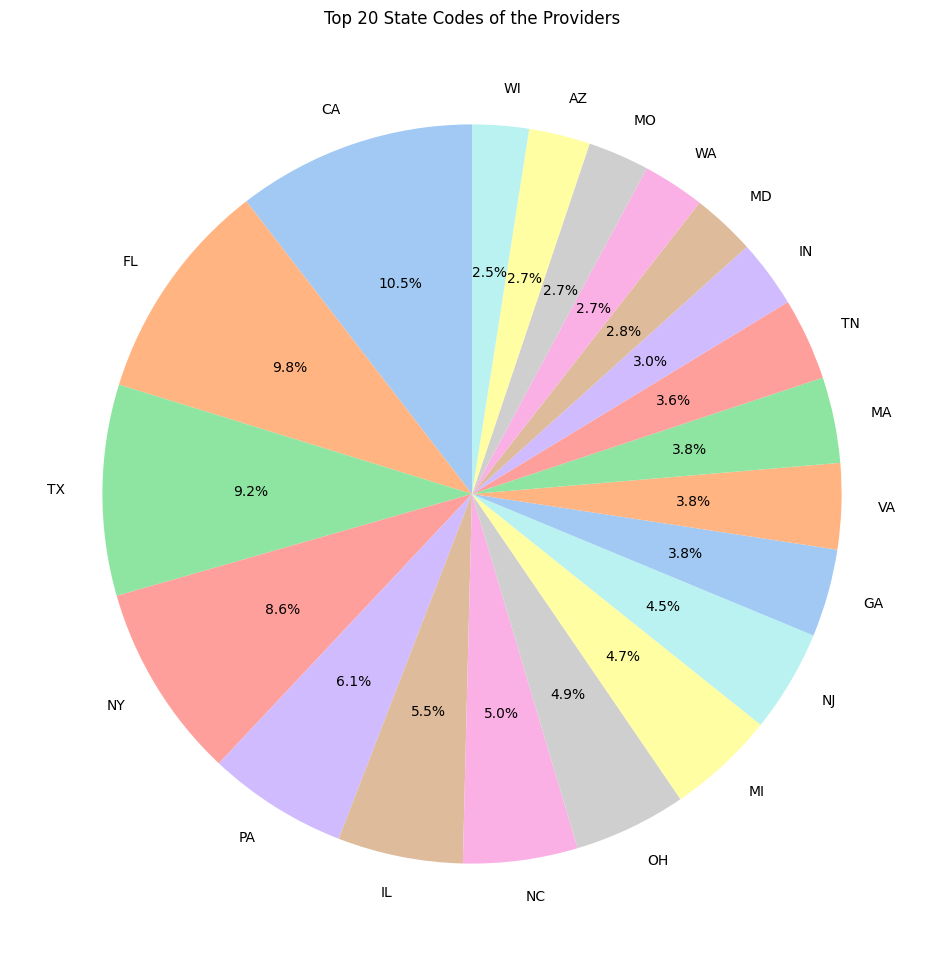
2.for the graph index vs NPI there is a uniform increase of the graph where indices and NPI

Are directly proportional to each other.

3.for the index vs zip code we can see a pair plot which is the unform distribution and we cannot able to see the outliers

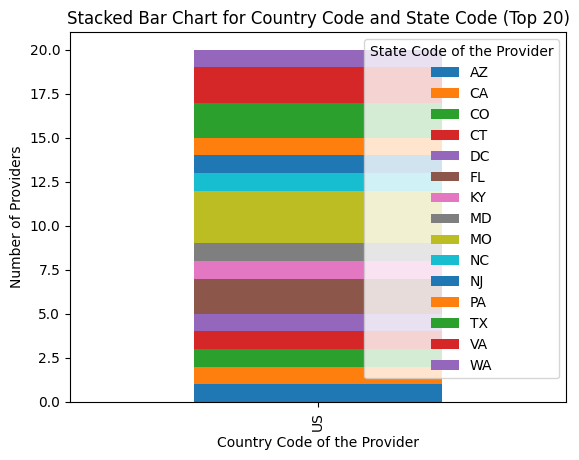
4.for the graph NPI vs ZIP code, Dense clustering of points without a clear pattern or correlation.

No discernible relationship between NPI and zip code.



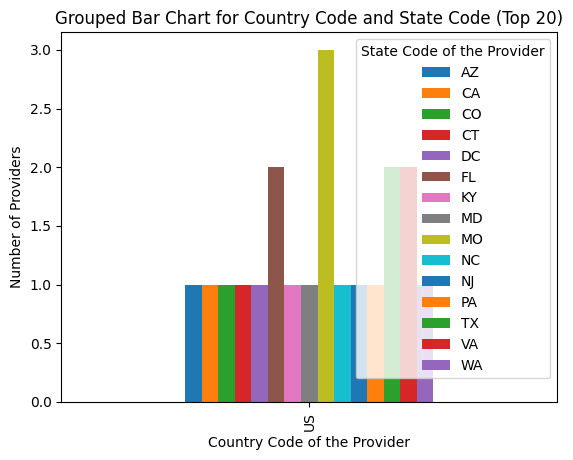
observations:

1. **California (CA)** has the highest percentage of providers at **10.5%**.
2. **Texas (TX)** follows closely with **9.2%** of providers.
3. Other states with significant percentages include **Florida (FL)** at **8.6%**, **New York (NY)**, **Pennsylvania (PA)**, and **Illinois (IL)**.
4. The remaining states contribute smaller percentages, ranging from **2.5%** to **3.8%**.



observations:

* **Distribution Shape**: The plot is bell-shaped and skewed to the right, indicating that most average amounts cluster around a mode slightly above zero.
* **Long Tail**: There is a long tail extending toward higher values (up to 300), suggesting some extreme payment amounts.
* **Density**: The density ranges from 0 to approximately 0.006, emphasizing the concentration of values around the mode.



* **Distribution Shape**: The plot is bell-shaped and skewed to the right, indicating that most average amounts cluster around a mode slightly above zero.
* **Long Tail**: There is a long tail extending toward higher values (up to 300), suggesting some extreme payment amounts.
* **Density**: The density ranges from 0 to approximately 0.006, emphasizing the concentration of values around the mode.