ANOMALY DETECTION USING UNSUPERVISED MACHINE LEARNING

PRESENTED BY:-

PUSHPENDRA KUMAR

INFOSYS SPRINGBOARD INTERN

DOMAIN – AI

GROUP – 122

PROBLEM STATEMENT

- This project aims to develop a system using unsupervised learning to detect anomalies or potential fraud among healthcare providers based on their behavior and transaction data.

- The system integrates diverse datasets for comprehensive analysis, extracts relevant features, and applies clustering and anomaly detection algorithms.

- It aims to differentiate between normal variations and potentially fraudulent activities, establishing real-time monitoring and scoring mechanisms.

HEALTHCARE PROVIDERS

HEALTHCARE PROVIDERS DATA(KAGGLE.COM)

Healthcare fraud represents a significant societal challenge, diverting funds intended for medication, elderly care, and emergency services towards dishonest practitioners or patients. This diversion contributes significantly to the escalating costs of healthcare amid rising expenditures.

This dataset is a collection of healthcare providers records, comprising 100,000 entries. It consists of various columns including index, National Provider Identifier(NPI), Name of Provider, Credentials, Gender, State Code, Number of Services, Average Medicare Amounts, etc.

DATASET ANALYSIS

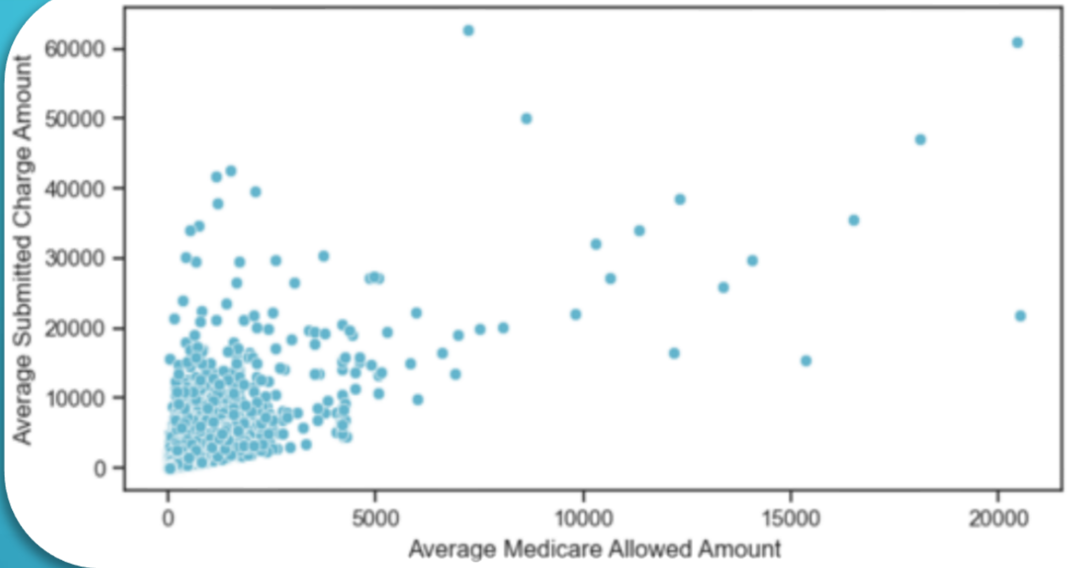
- This dataset consists of total 27 columns having 100K of total entries.
- There are 5 Categorical, 17 Texts, and 3 Numeric and 2 Boolean variables in the dataset.
- It consists of various data types including object, float64, and int64.
- There are total 104412 cells missing in the dataset which is 3.9% of total cells.
- There is no any duplicate rows present in the dataset.

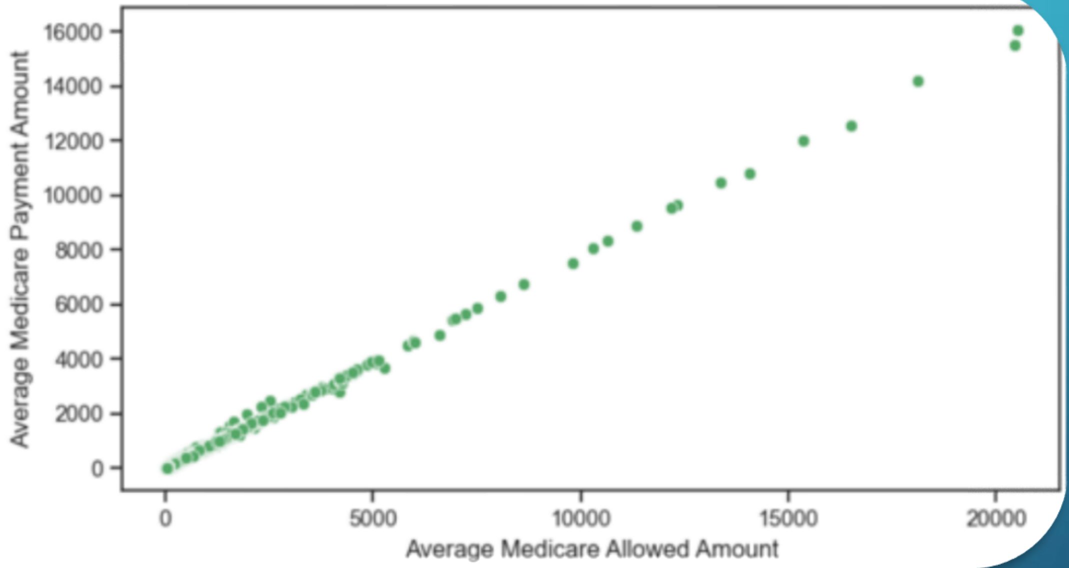
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 27 columns):
 #   Column                                                    Non-Null Count   Dtype
---  ------                                                    --------------   -----
 0   index    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

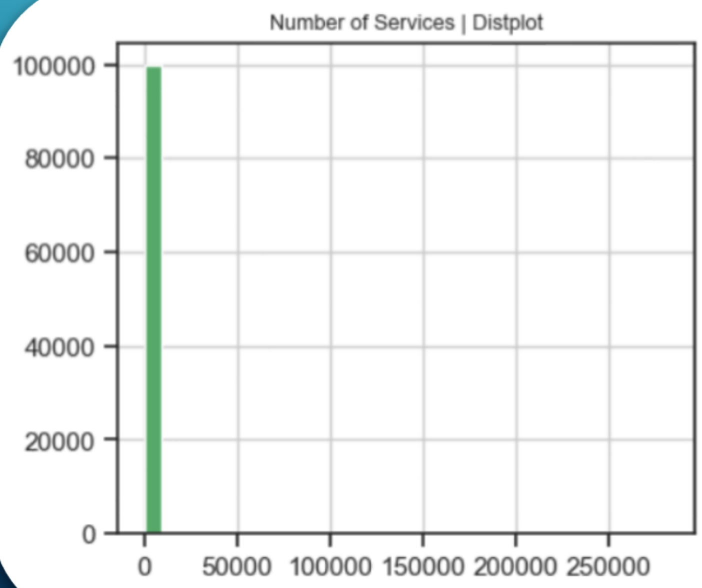
EXPLORATORY DATA ANALYSIS (EDA)

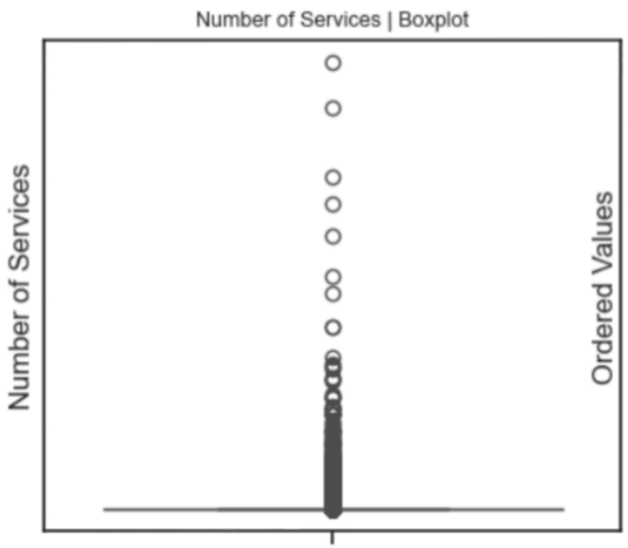
- The healthcare providers dataset is analysed thoroughly, all its features are observed and checked for its usefulness and importance in anomaly detection.
- Pre-processing steps are followed such as: Handling missing values, Dropping unnecessary columns, Checking duplicates, and Filling null values with mode.
- Data Cleaning is done by converting the object columns which are supposed to be numeric, removing any comma separators, and removing periods(.) from “Credentials of the Provider” column.
- Data Visualization using various scatter plots (Univariate and Bivariate Analysis).
- Normalization, Standardization, and Encoding.
- Dimensionality Reduction using Principal Component Analysis (PCA).

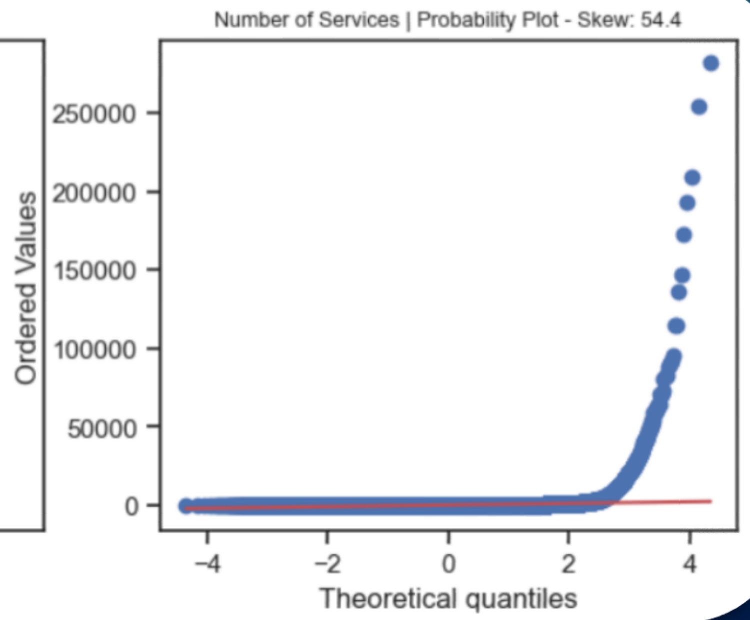
EDA VISUALISATION



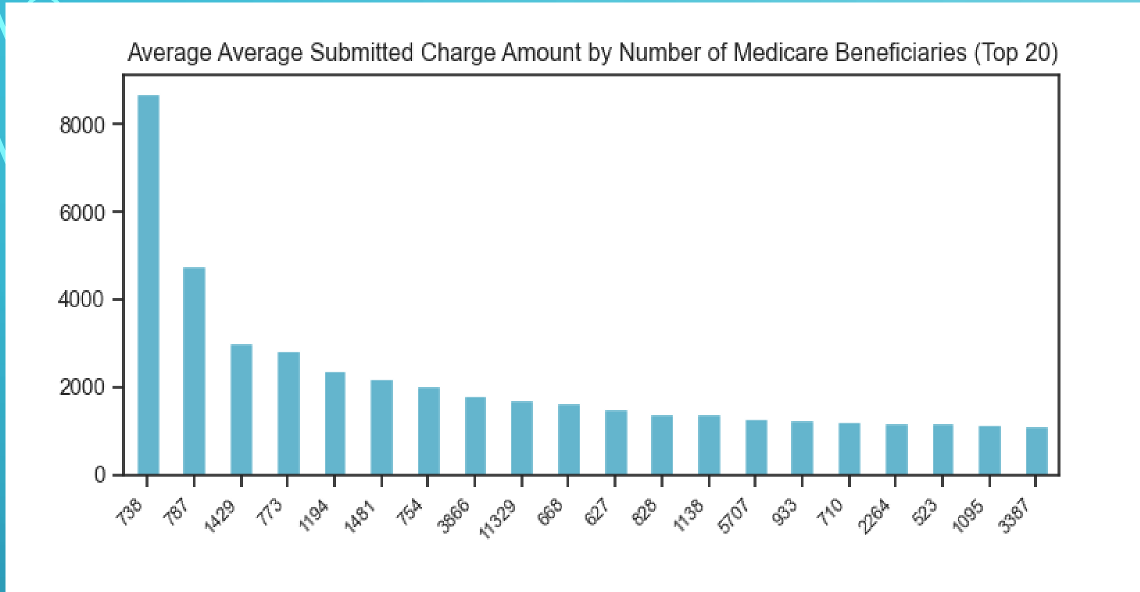


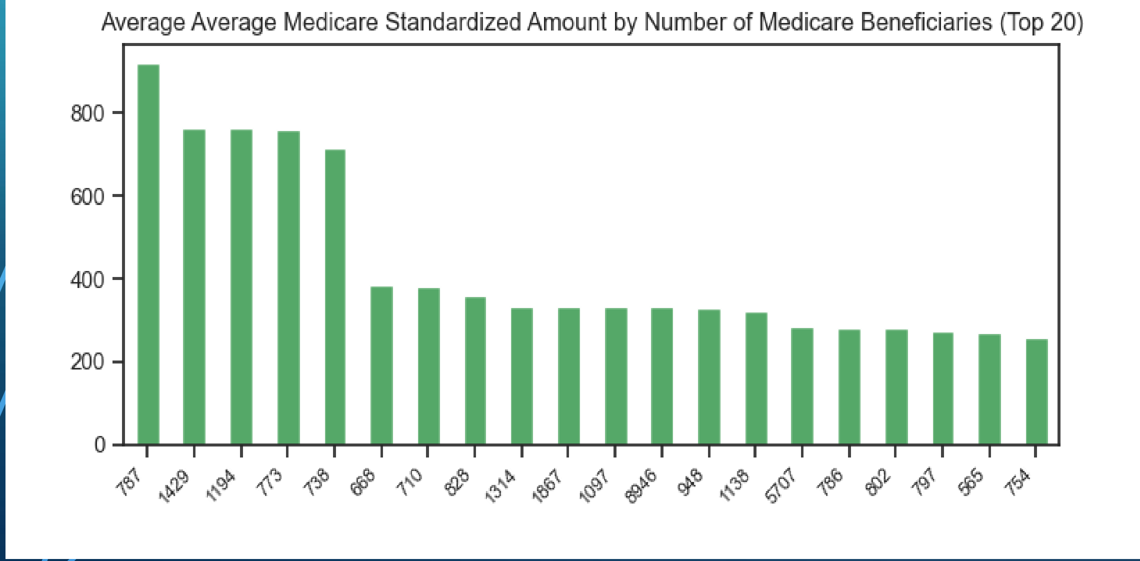


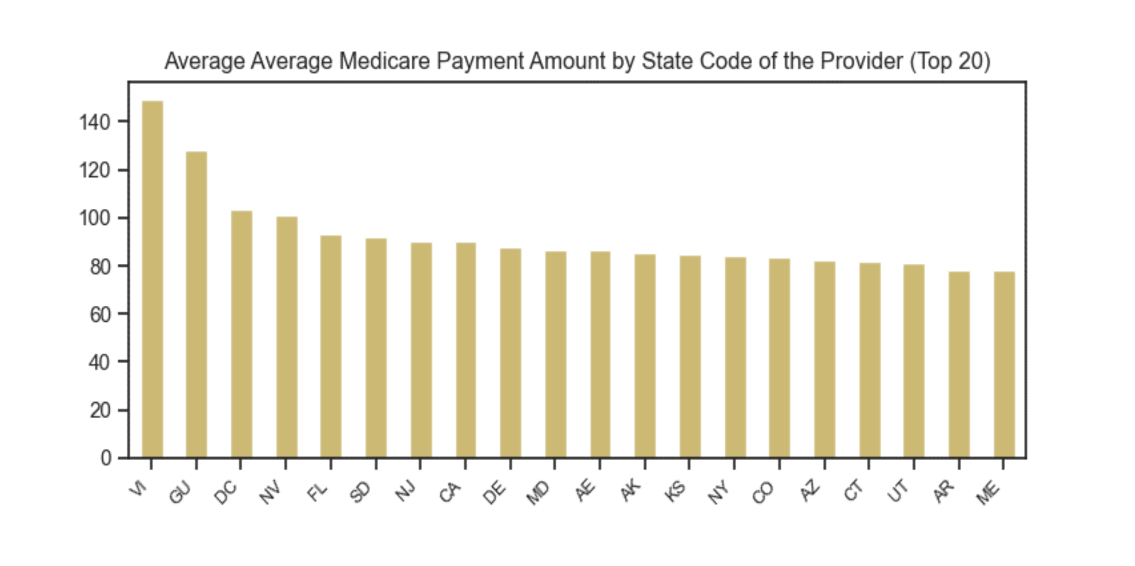


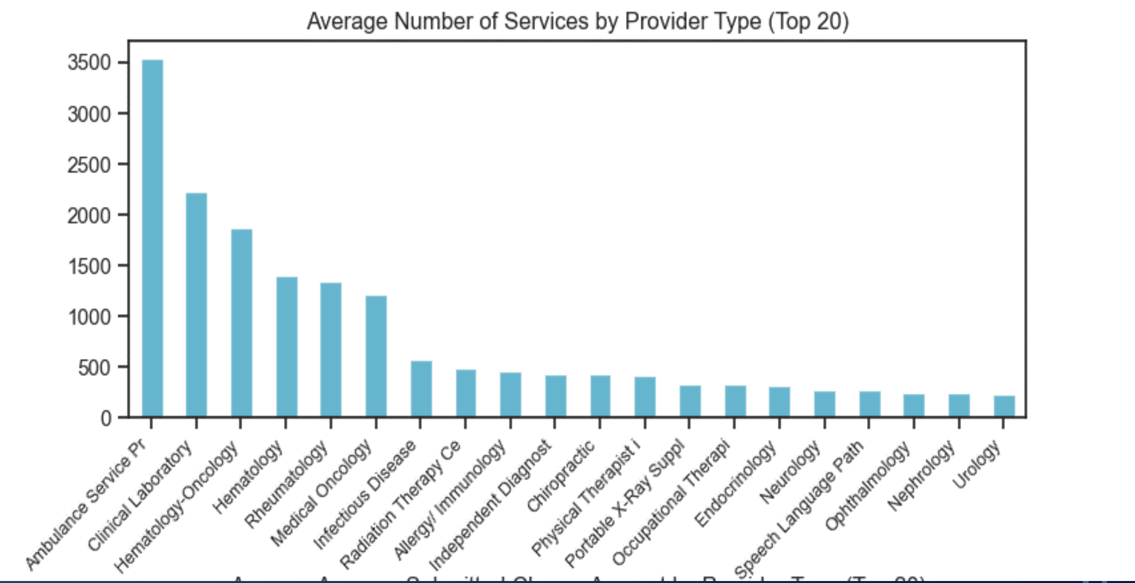


EDA VISUALISATION (CONTD.)

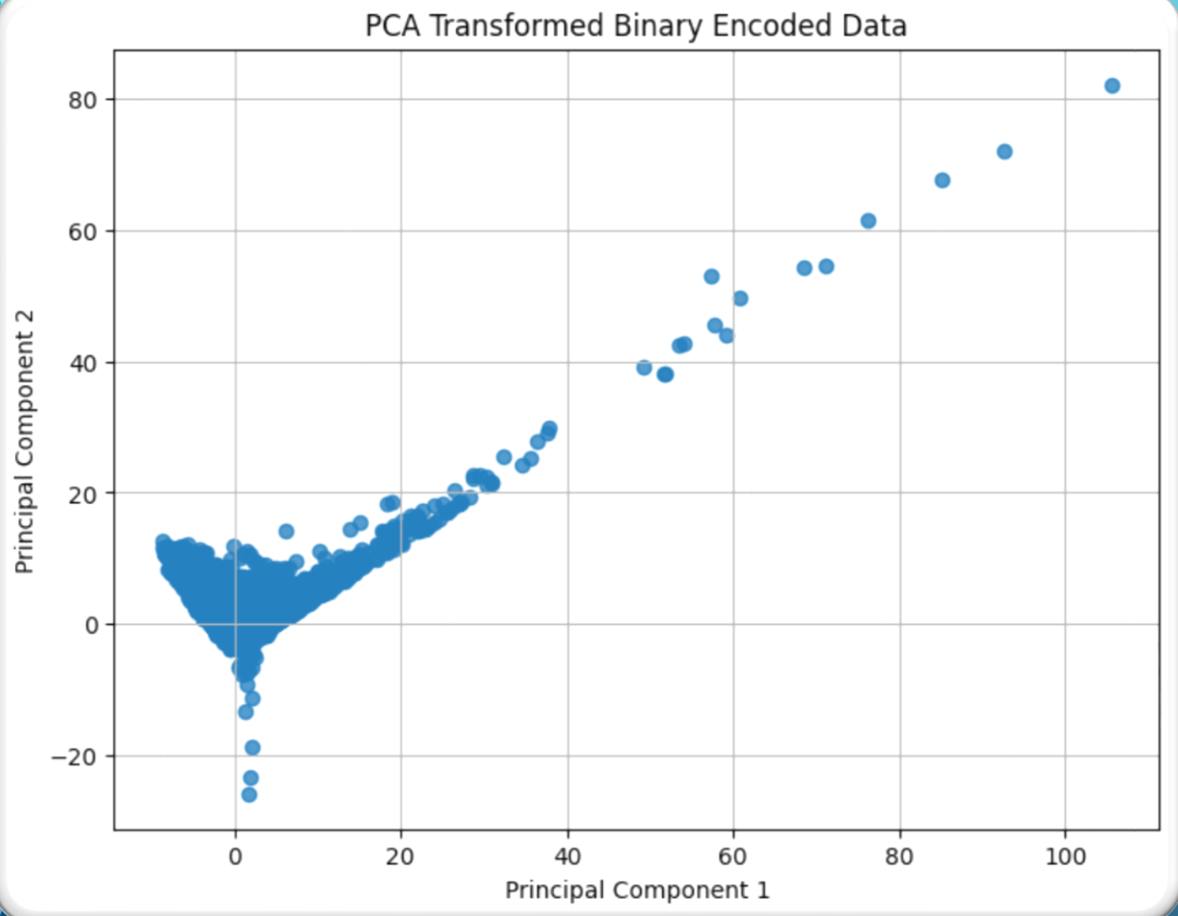


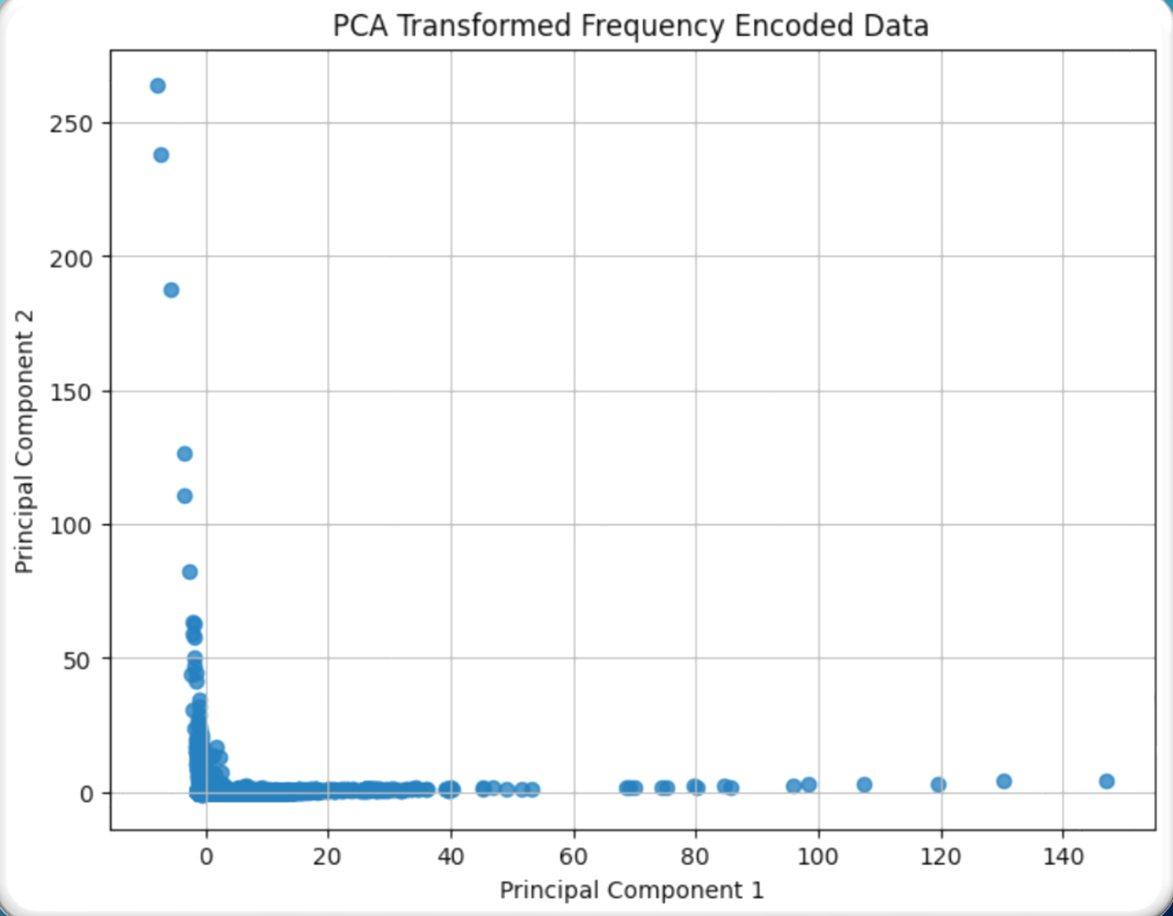






PCA VISUALISATION

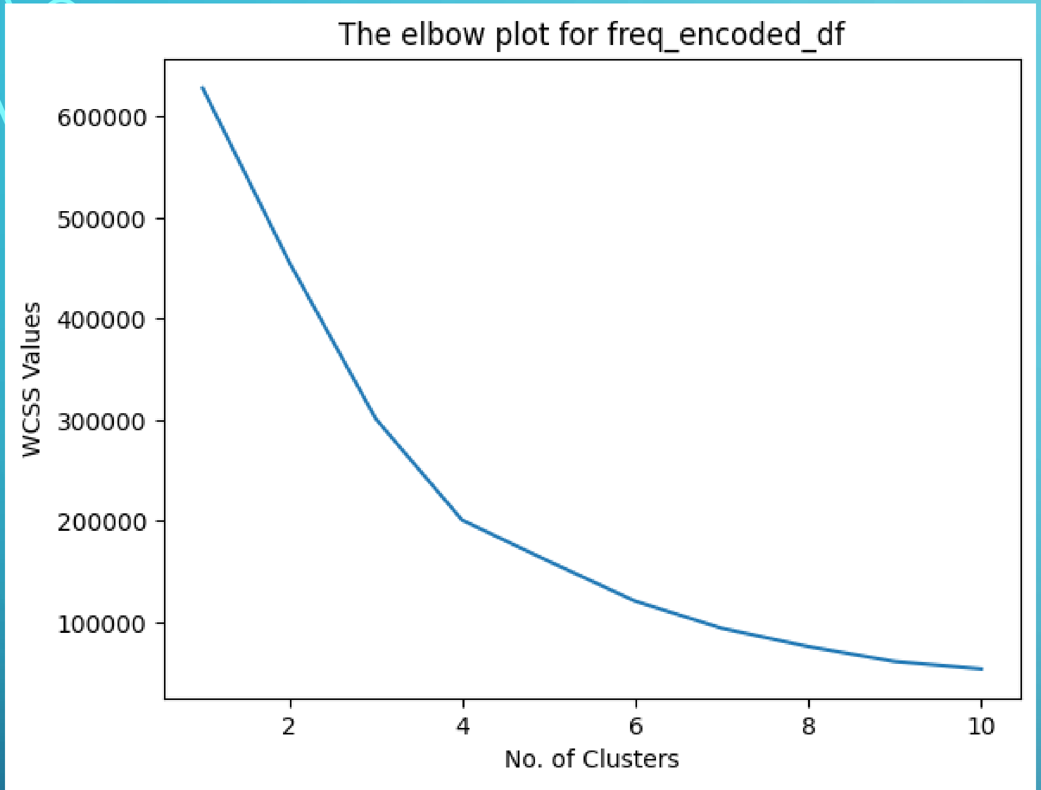




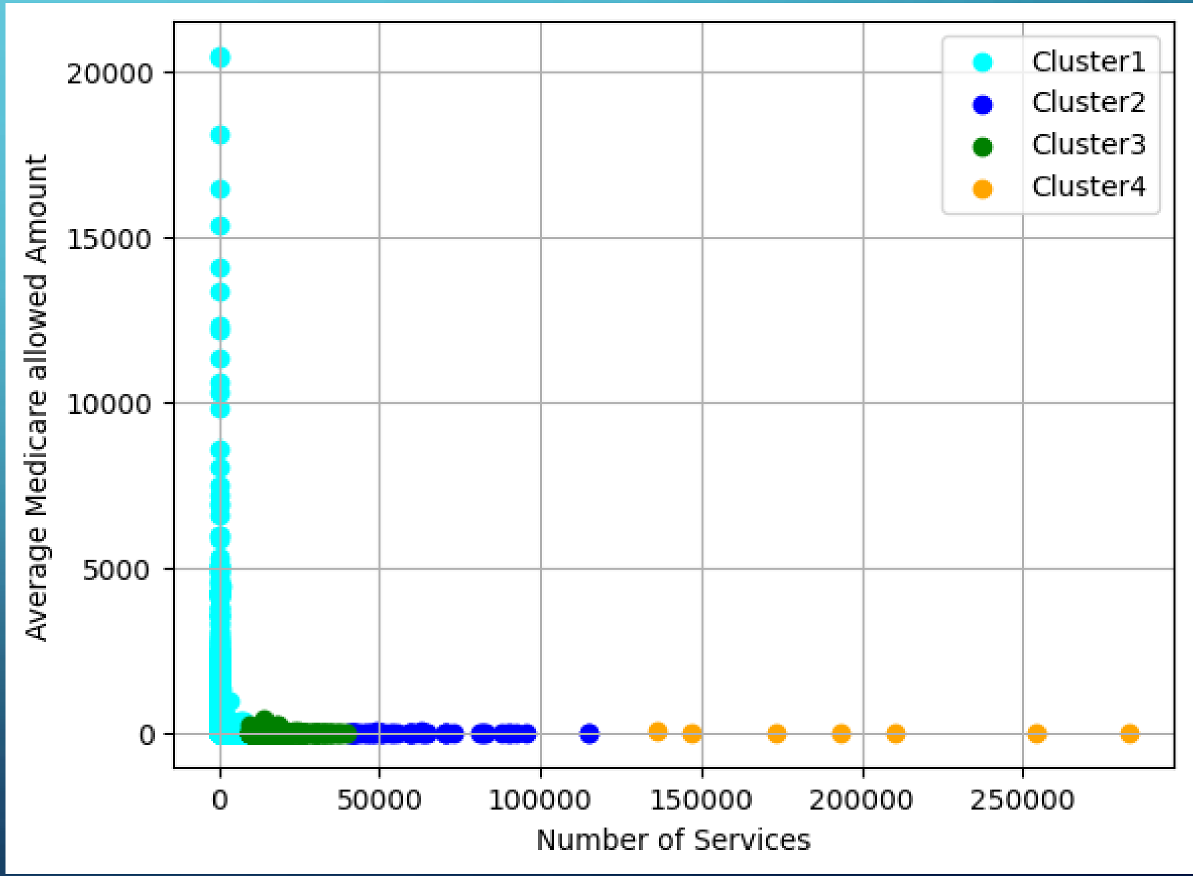
CLUSTERING ALGORITHMS

- K-means Clustering
- DBSCAN Clustering

K-MEANS CLUSTERING

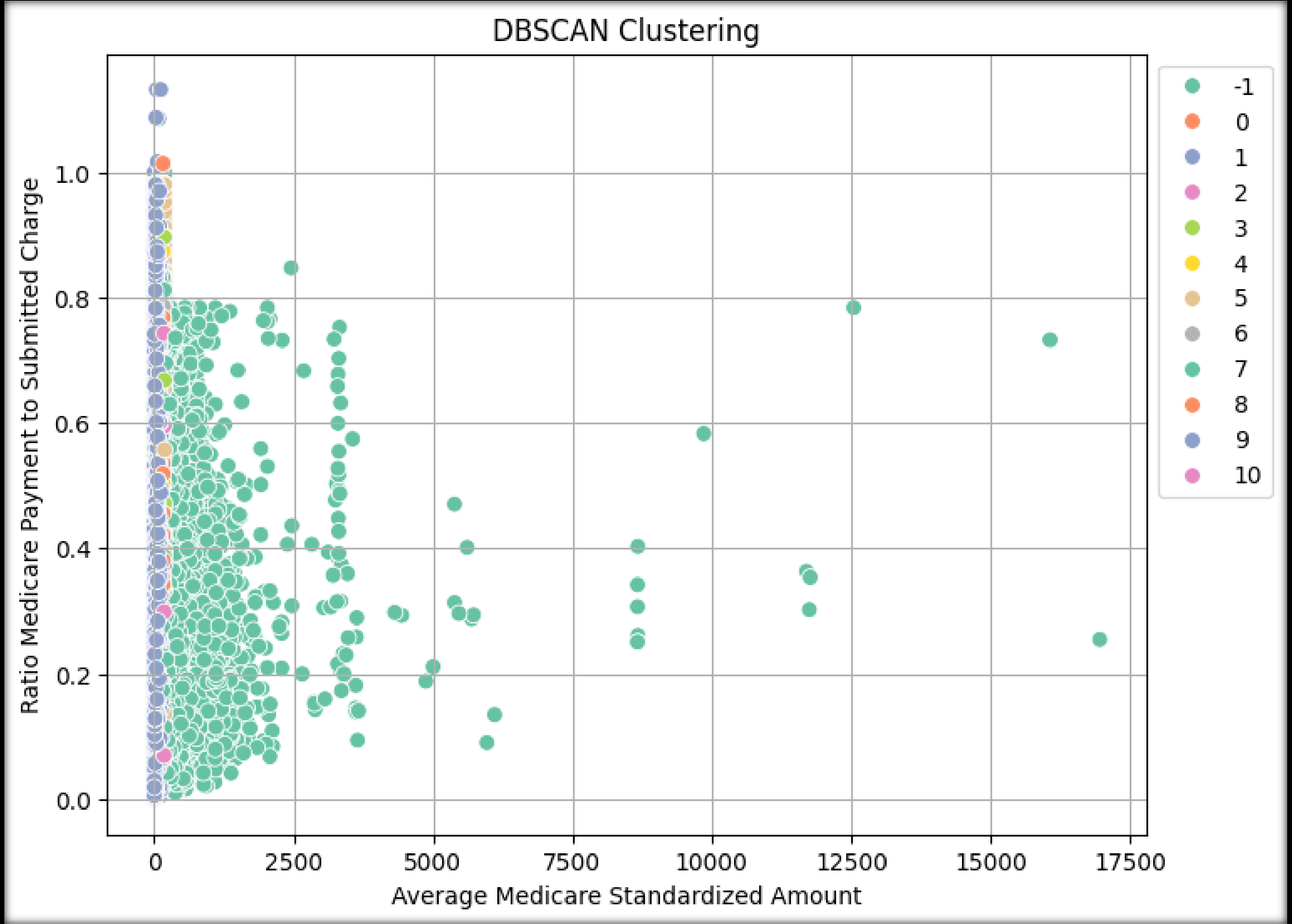


The 4 clusters formed by the distribution of values of "Number of Service" vs "Avg. Medicare Allowed Amount".



The elbow plot above tells us the ideal number of clusters to be formed.

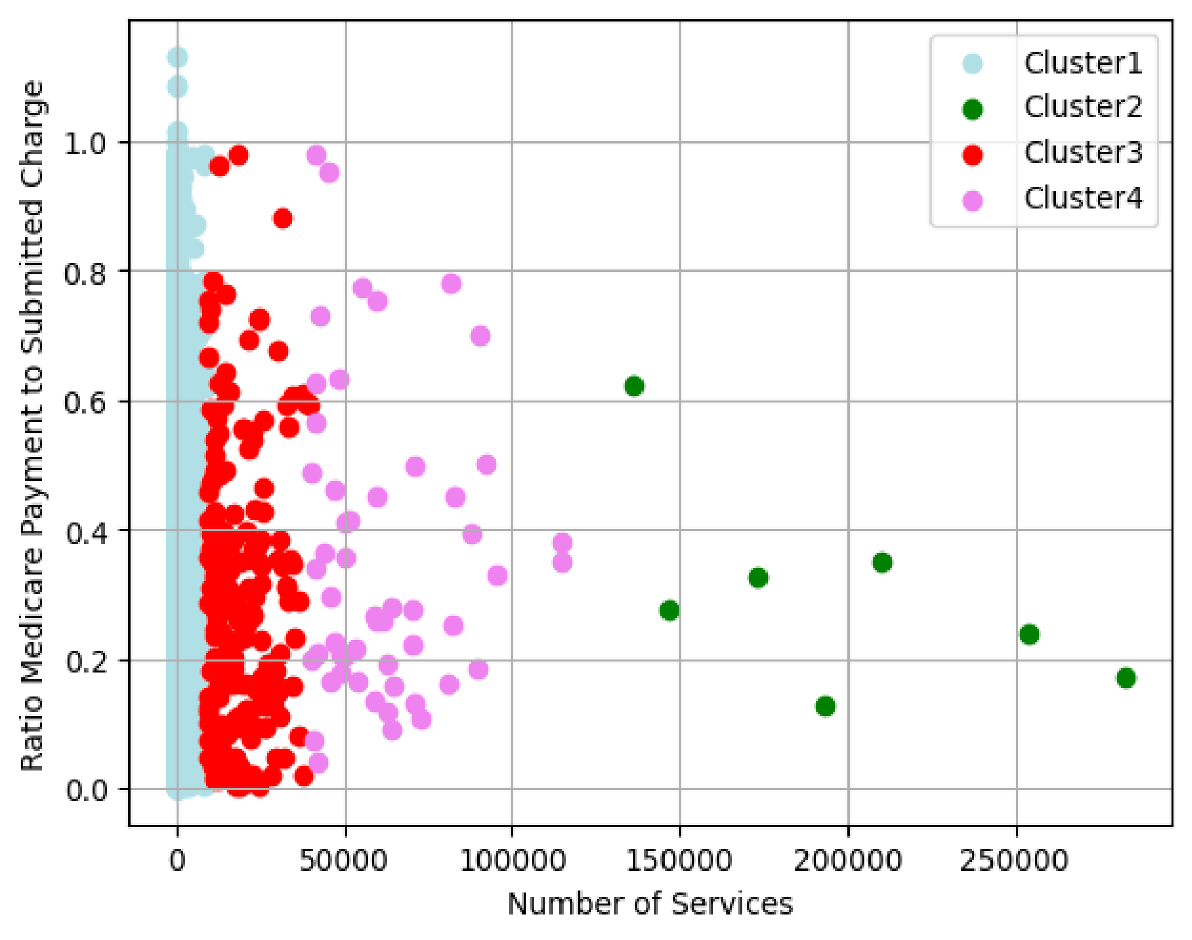
DBSCAN CLUSTERING



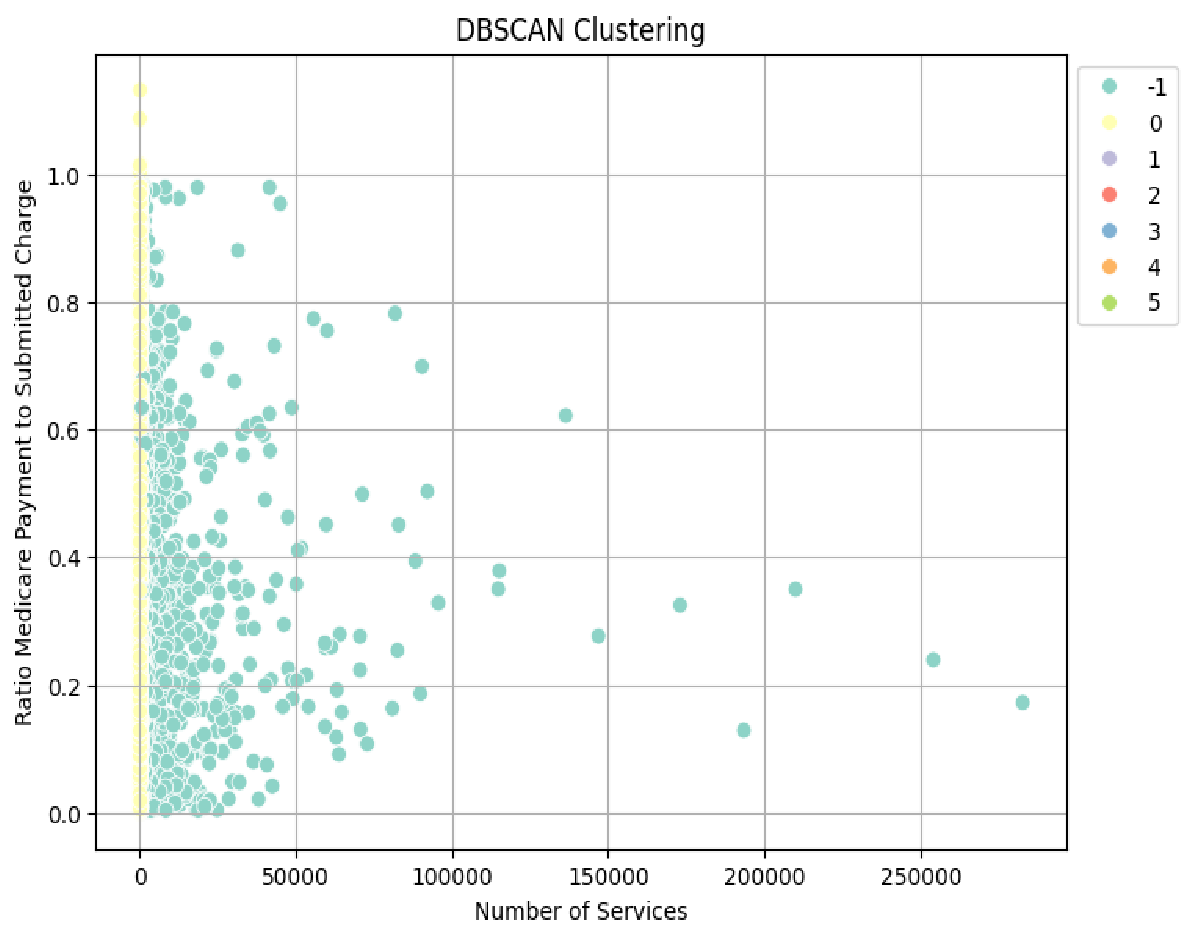
Total 10 clusters are formed using DBSCAN clustering algorithm.

This plot shows the clusters formed by two features “Avg. Medicare Standardized Amount” vs “Ratio Medicare Payment to Submitted Charge”.

CLUSTERING COMPARISON



K-means Clustering



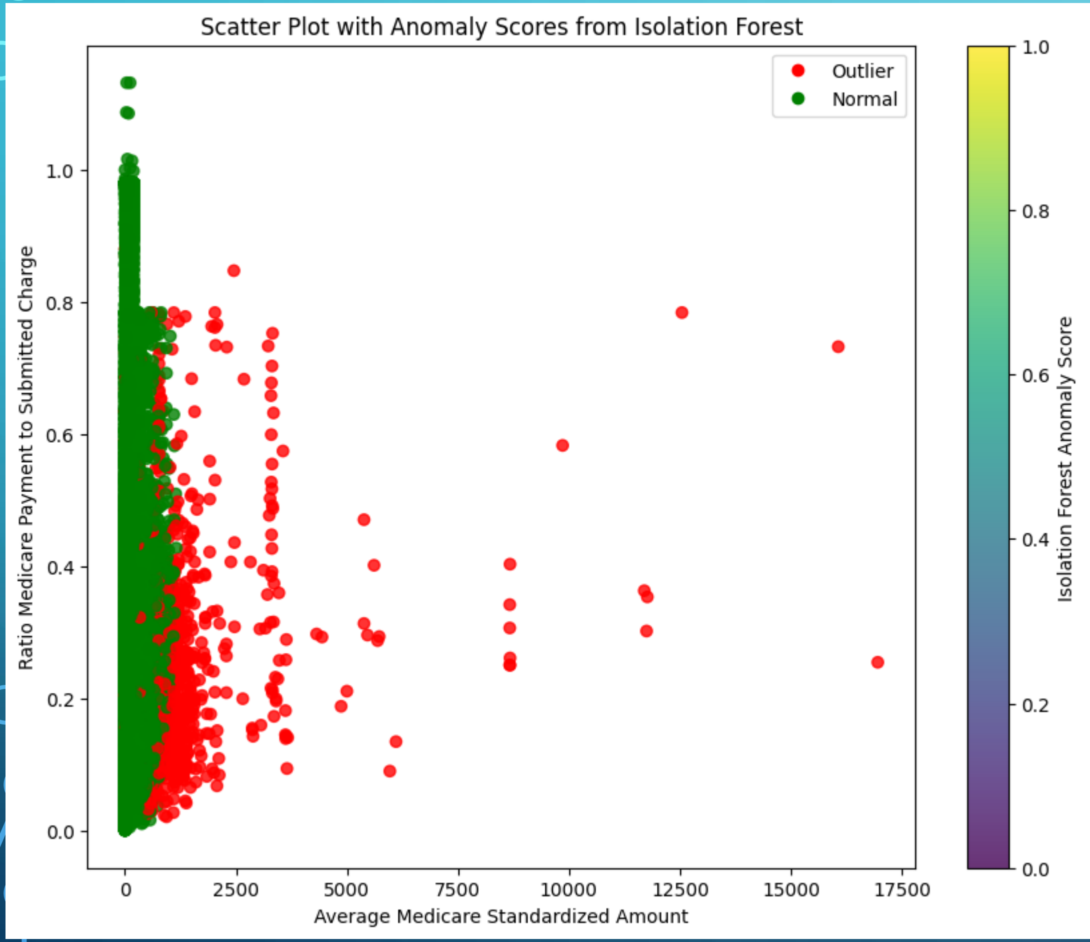
DBSCAN Clustering

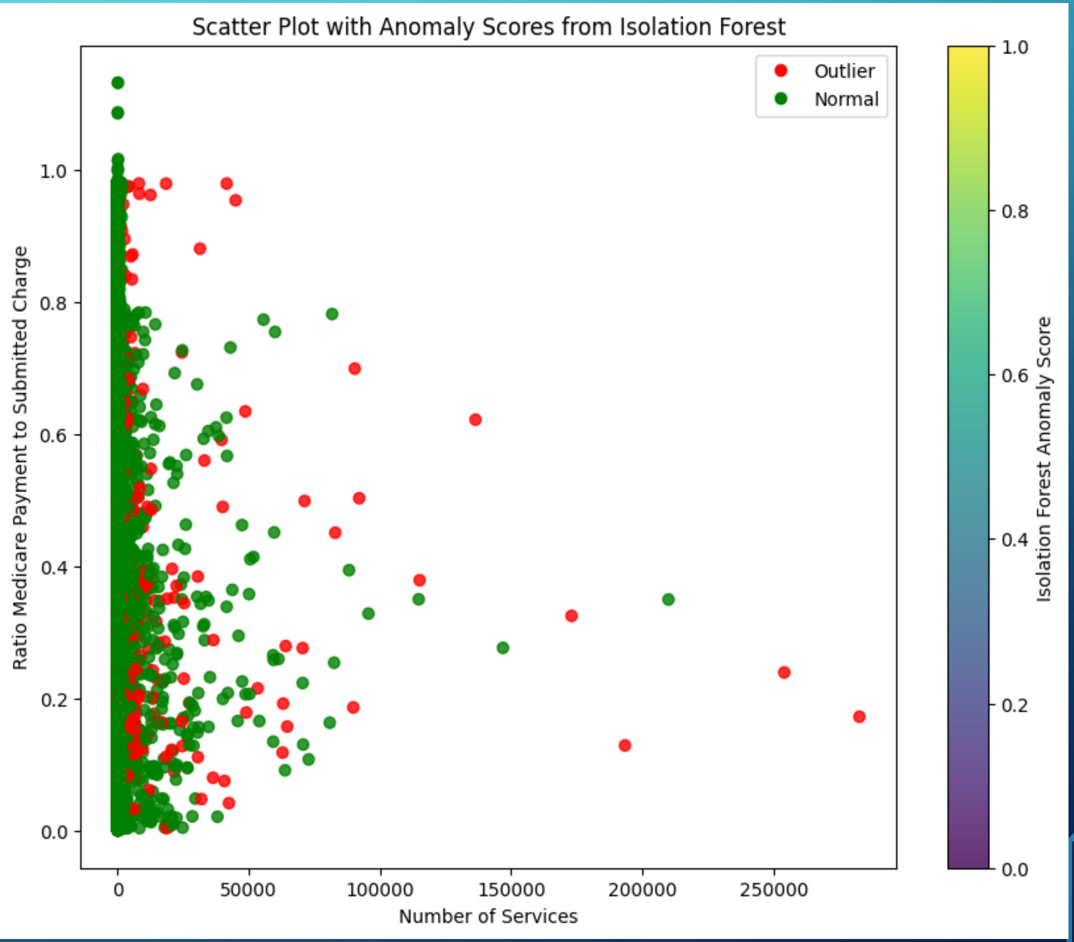
MACHINE LEARNING ALGORITHMS

- Isolation Forest
- Elliptic Envelope
- One Class SVM

ML ALGORITHM (ISOLATION FOREST)

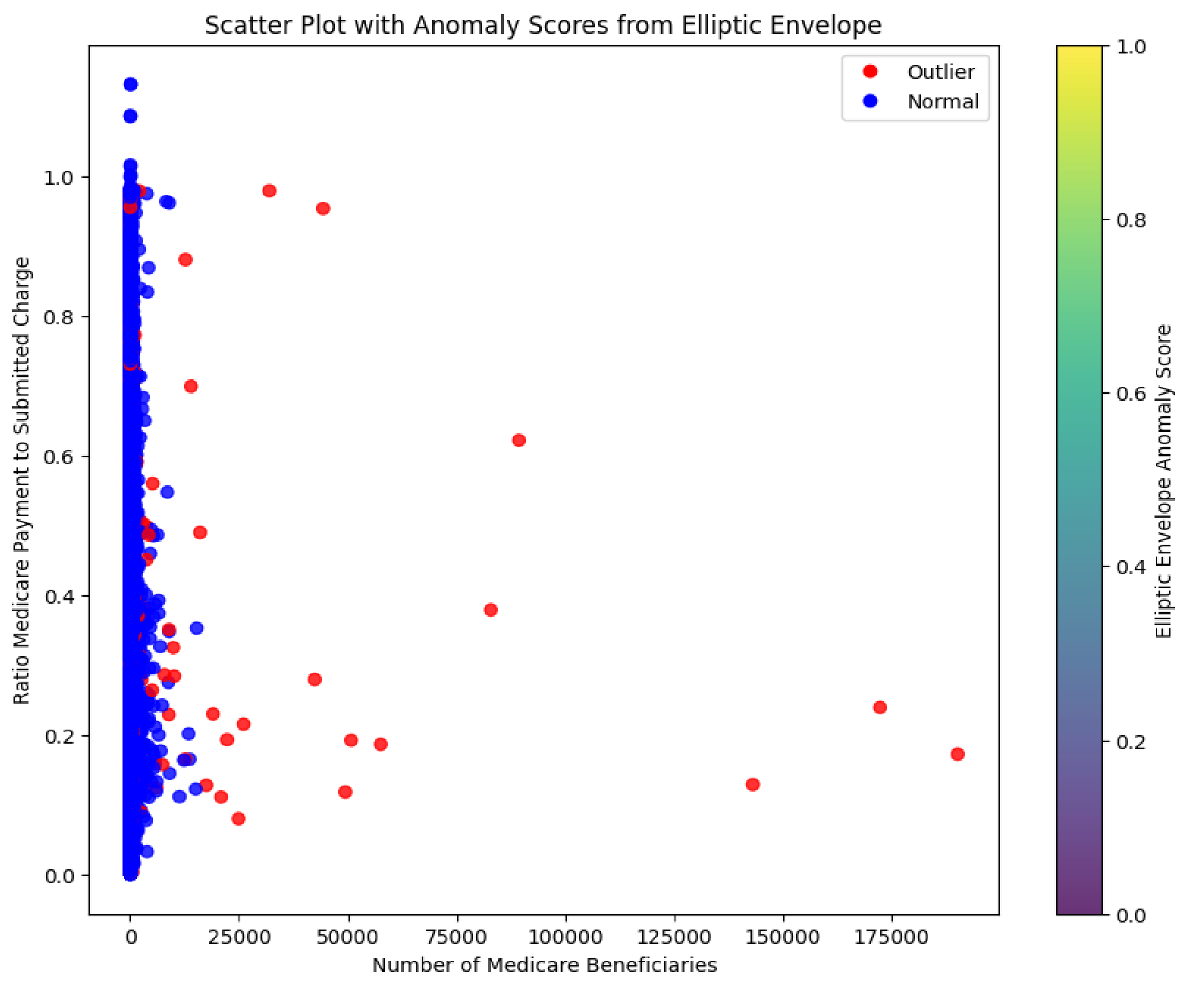
Anomalies Detected = 1000

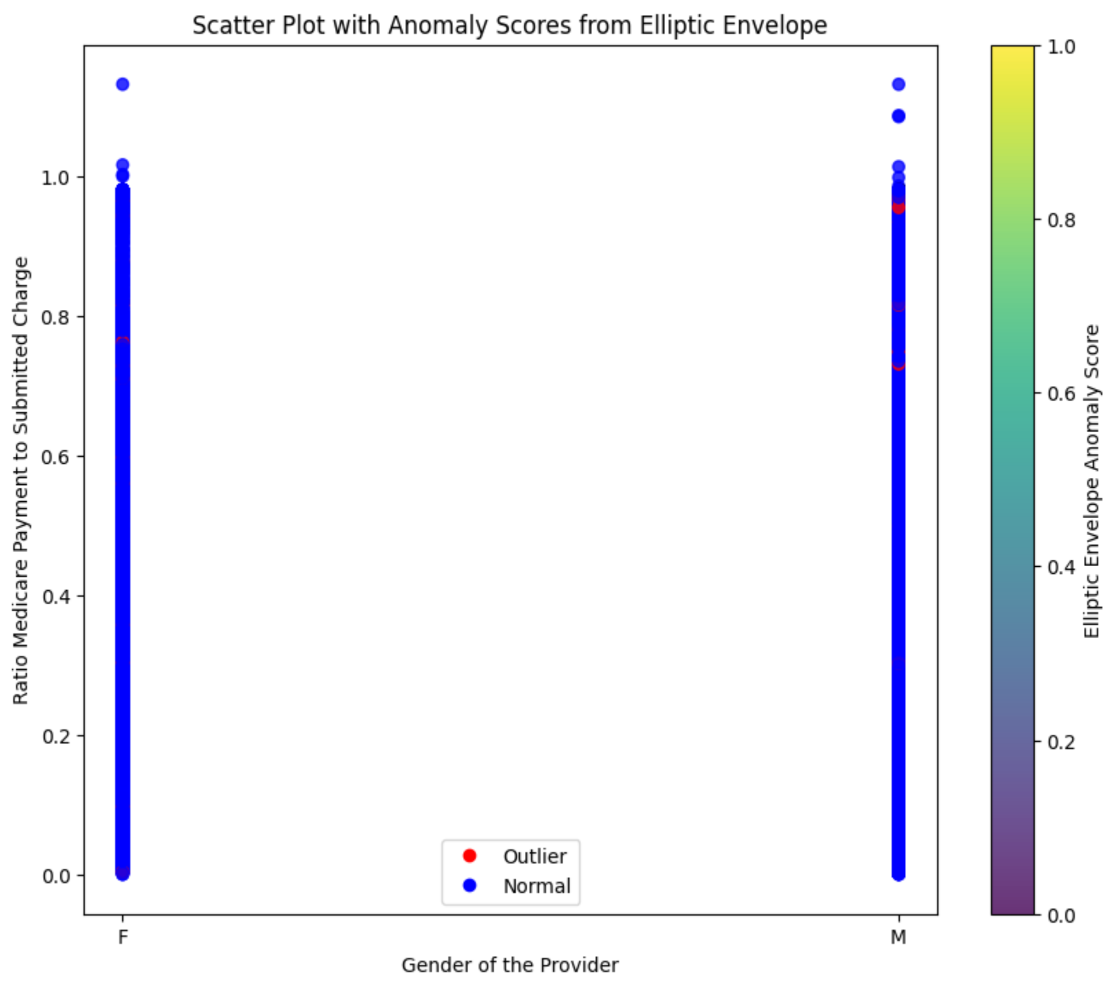




ML ALGORITHM (ELLIPTIC ENVELOPE)

Anomalies Detected = 1000



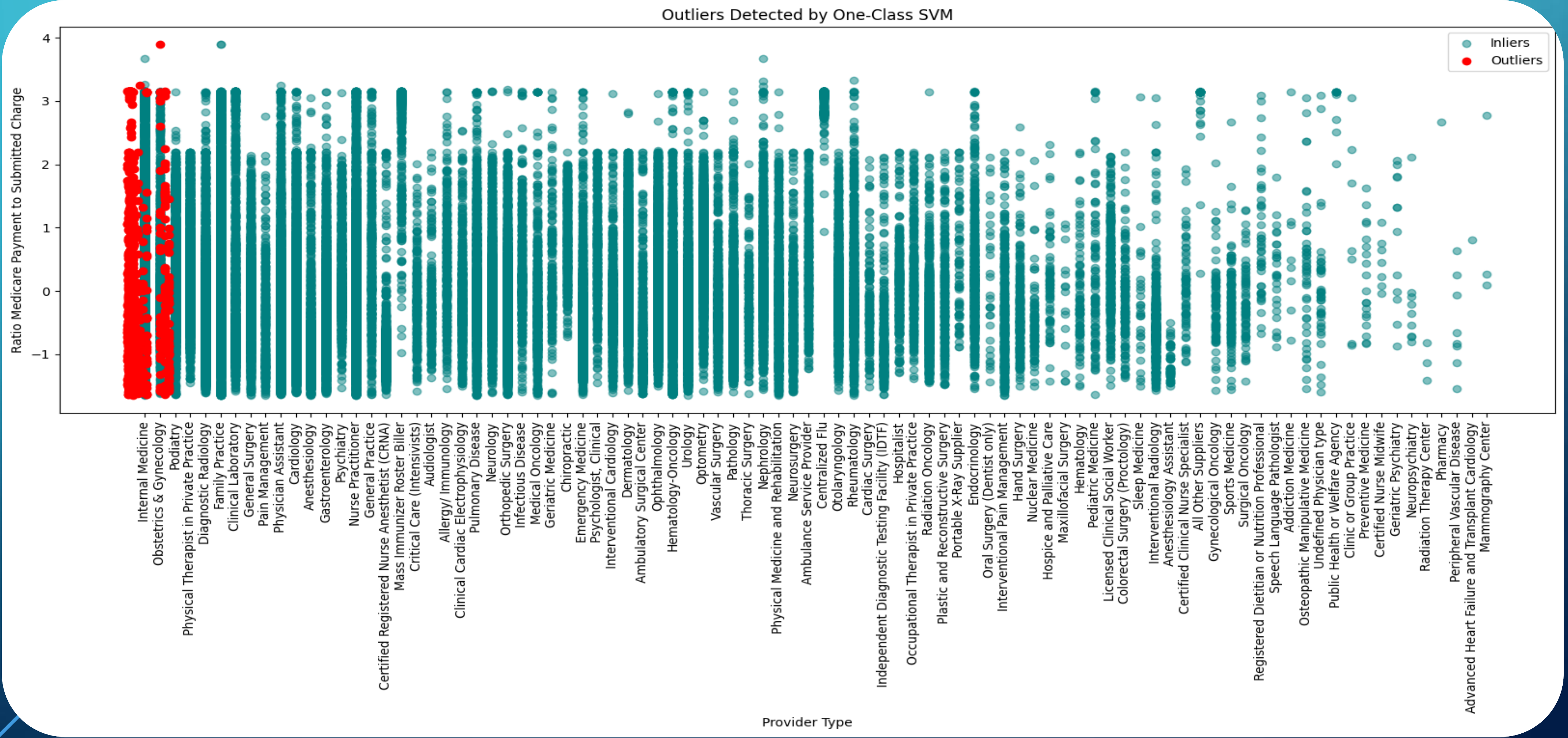


ML ALGORITHM (ONE CLASS SVM)

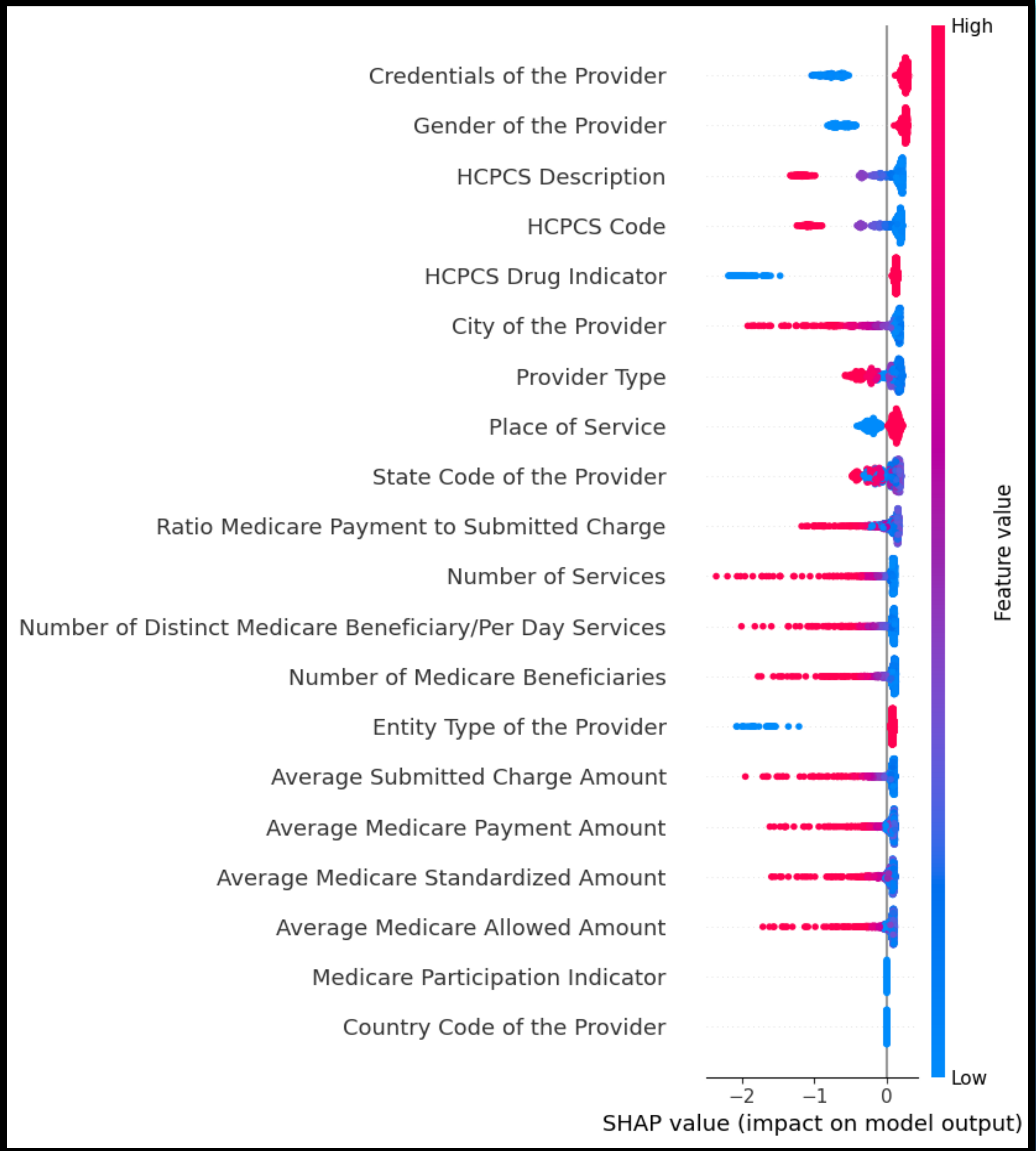
Anomalies Detected = 1002

Here, in One Class SVM Model, the total number of anomalies detected by setting the 'nu' value as '0.01' are **1002.**

ML ALGORITHM (ONE CLASS SVM) (CONTD.)



SHAP ANALYSIS(ISOLATION FOREST)



The following attributes have the highest positive impact on the output:

- Credentials of the Provider
- Gender of the Provider
- HCPS Drug Indicator
- Place of Service
- Entity Type of the Provider

The following attributes have the highest negative impact on the output:

- City of the Provider
- Number of Services
- Number of Distinct Medicare Beneficiary/Per Day Services
- Number of Medicare Beneficiaries
- Average Medicare Allowed Amount
- Average Submitted Charge Amount
- Average Medicare Payment Amount
- Average Medicare Standardized Amount

AUTO ENCODERS

Anomalies Detected = 1000

Model: "functional_1"

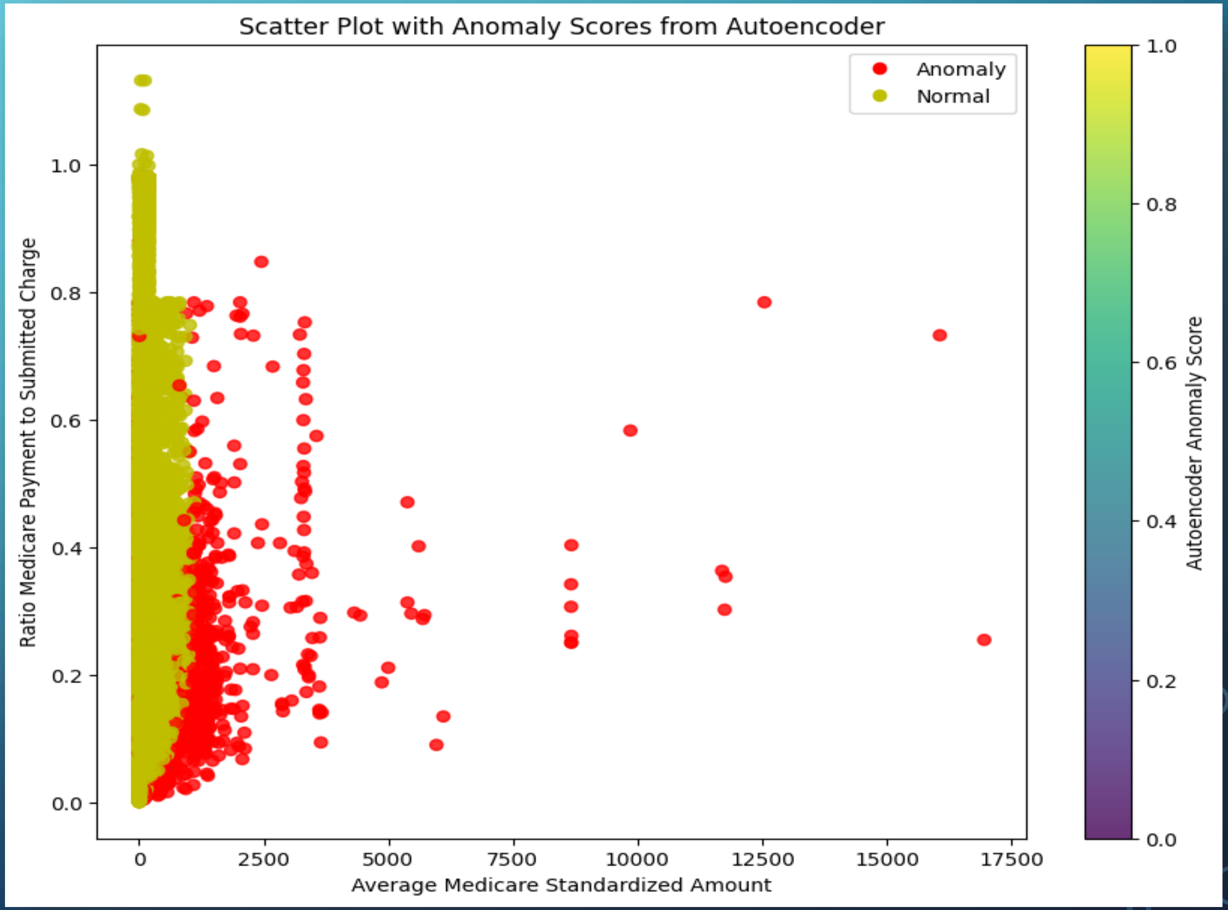
Layer (type)	Output Shape	Param #
input_layer (InputLayer)	(None, 20)	0
dense (Dense)	(None, 64)	1,344
dense_1 (Dense)	(None, 20)	1,300

Total params: 2,644 (10.33 KB)

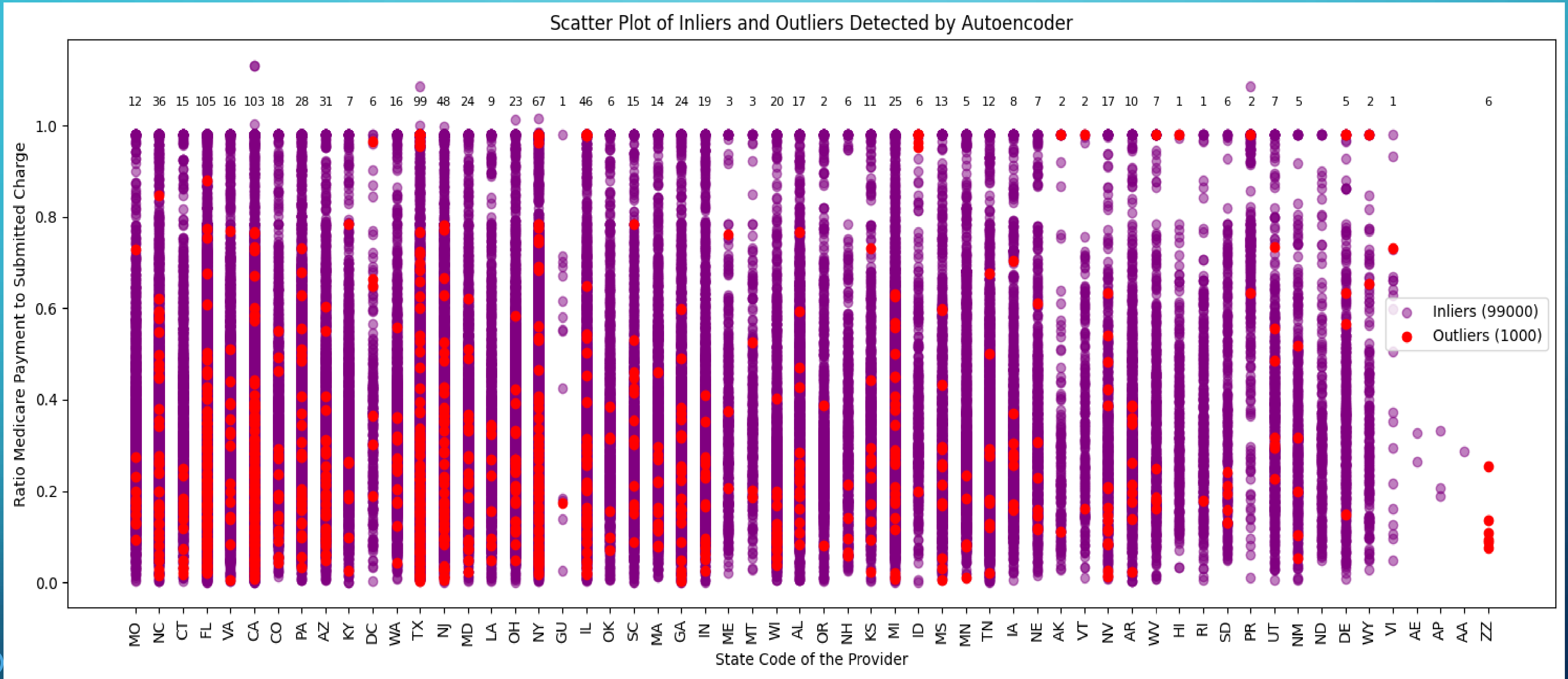
Trainable params: 2,644 (10.33 KB)

Non-trainable params: 0 (0.00 B)

- The input layer is having shape as (None, 20) and 0 parameters.
- There are 2 dense layers having shape (None, 64) and (None, 20).
- Total number of epochs = 30.



AUTO ENCODERS (CONTD.)



Most number of anomalies: **Florida, California, and Texas.**

Least number of anomalies: **Armed Forces Europe, Armed Forces Central/South America, and Armed Forces Pacific.**


THANK YOU