

Data preprocessing & apply model of healthcare datasets

1.import dependencies & dataset

```
# pip install gender_guesser
```

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

```
df=pd.read_csv('Healthcare Providers.csv')
df
```



	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	C c Proc
0	8774979	1891106191	UPADHYAYULA	SATYASREE	NaN	M.D.	
1	3354385	1346202256	JONES	WENDY	P	M.D.	
2	3001884	1306820956	DUROCHER	RICHARD	W	DPM	
3	7594822	1770523540	FULLARD	JASPER	NaN	MD	
4	746159	1073627758	PERROTTI	ANTHONY	E	DO	
...	
99995	3837311	1386938868	PAPES	JOAN	NaN	PT	
99996	2079360	1215091327	HAYNER	MARGARET	S	ARNP	
99997	8927965	1902868185	VALENCIA	DANA	NaN	M.D.	
99998	8854571	1891941183	GONZALEZ-LAMOS	RAFAELA	NaN	NaN	
99999	3547535	1356772156	RAMEZANI	ELIIAN	NaN	NaN	

100000 rows × 27 columns

df.info()



```
<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

```

2.check the missing values

```
df.isnull().sum()
```

```

↪ index 0
National Provider Identifier 0
Last Name/Organization Name of the Provider 0
First Name of the Provider 4255
Middle Initial of the Provider 29331
Credentials of the Provider 7209
Gender of the Provider 4254
Entity Type of the Provider 0
Street Address 1 of the Provider 0
Street Address 2 of the Provider 59363
City of the Provider 0
Zip Code of the Provider 0
State Code of the Provider 0
Country Code of the Provider 0
Provider Type 0
Medicare Participation Indicator 0
Place of Service 0
HCPCS Code 0
HCPCS Description 0
HCPCS Drug Indicator 0
Number of Services 0
Number of Medicare Beneficiaries 0
Number of Distinct Medicare Beneficiary/Per Day Services 0
Average Medicare Allowed Amount 0
Average Submitted Charge Amount 0
Average Medicare Payment Amount 0
Average Medicare Standardized Amount 0
dtype: int64

```

3.filling missing value in neccassary features

- here in gender of the provider new category add unknown for gender
- In credentials of the provider new category add unknown_cred

```
# in gender & credentials missing value to unknown

df['Gender of the Provider'].fillna('Unknown', inplace=True)
df['Credentials of the Provider'].fillna('Unknown_cred', inplace=True)
df
```



	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	Pr
0	8774979	1891106191	UPADHYAYULA	SATYASREE	NaN	M.D.	
1	3354385	1346202256	JONES	WENDY	P	M.D.	
2	3001884	1306820956	DUROCHER	RICHARD	W	DPM	
3	7594822	1770523540	FULLARD	JASPER	NaN	MD	
4	746159	1073627758	PERROTTI	ANTHONY	E	DO	
...	
99995	3837311	1386938868	PAPES	JOAN	NaN	PT	
99996	2079360	1215091327	HAYNER	MARGARET	S	ARNP	
99997	8927965	1902868185	VALENCIA	DANA	NaN	M.D.	
99998	8854571	1891941183	GONZALEZ-LAMOS	RAFAELA	NaN	Unknown_cred	
99999	3547535	1356772156	RAMEZANI	ELIAN	NaN	Unknown_cred	

100000 rows × 27 columns

```
df['Gender of the Provider'].value_counts()
```



Gender of the Provider	
M	66641
F	29105
Unknown	4254
Name: count, dtype: int64	

4.remove the unnecassary features & unique identifiers

- 'index', 'National Provider Identifier', 'Last Name/Organization Name of the Provider', 'First Name of the Provider', 'Middle Initial of the Provider','Street Address 1 of the Provider', 'Street Address 2 of the Provider','Zip Code of the Provider' above some colums are unique identifier and not depends results.

```
# drop down unnecesaries columns columns
```

```
df.drop(['index', 'National Provider Identifier',  
        'Last Name/Organization Name of the Provider',  
        'First Name of the Provider', 'Middle Initial of the Provider', 'Street Address 1 of the Provider',  
        'Street Address 2 of the Provider', 'Zip Code of the Provider'], axis=1, inplace=True)  
df
```



	Credentials of the Provider	Gender of the Provider	Entity Type of the Provider	City of the Provider	State Code of the Provider	Country Code of the Provider	Provider Type
0	M.D.	F	I	SAINT LOUIS	MO	US	Internal Medicine
1	M.D.	F	I	FAYETTEVILLE	NC	US	Obstetrics & Gynecology
2	DPM	M	I	NORTH HAVEN	CT	US	Podiatry
3	MD	M	I	KANSAS CITY	MO	US	Internal Medicine
4	DO	M	I	JUPITER	FL	US	Internal Medicine
...
99995	PT	F	I	WILMINGTON	IL	US	Physical Therapist in Private Practice
99996	ARNP	F	I	REDMOND	OR	US	Nurse Practitioner
99997	M.D.	M	I	SAINT LOUIS	MO	US	Cardiology
99998	Unknown_cred	F	I	LARCHMONT	NY	US	Internal Medicine
99999	Unknown_cred	F	I	GREAT NECK	NY	US	Physical Therapist in Private Practice

100000 rows × 19 columns

Next steps:

Generate code with df

View recommended plots

```
df.isnull().sum()
```



Credentials of the Provider	0
Gender of the Provider	0
Entity Type of the Provider	0
City of the Provider	0
State Code of the Provider	0
Country Code of the Provider	0
Provider Type	0
Medicare Participation Indicator	0
Place of Service	0

```

HCPCS Code                                0
HCPCS Description                         0
HCPCS Drug Indicator                      0
Number of Services                       0
Number of Medicare Beneficiaries         0
Number of Distinct Medicare Beneficiary/Per Day Services 0
Average Medicare Allowed Amount          0
Average Submitted Charge Amount          0
Average Medicare Payment Amount          0
Average Medicare Standardized Amount     0
dtype: int64

```

```
df['Gender of the Provider'].value_counts()
```

```

Gender of the Provider
M          66641
F          29105
Unknown    4254
Name: count, dtype: int64

```

6. Data type for numerical columns check & convert int & float for training propose

```
df.info()
```

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 19 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Credentials of the Provider           100000 non-null object
1   Gender of the Provider                100000 non-null object
2   Entity Type of the Provider          100000 non-null object
3   City of the Provider                 100000 non-null object
4   State Code of the Provider           100000 non-null object
5   Country Code of the Provider         100000 non-null object
6   Provider Type                        100000 non-null object
7   Medicare Participation Indicator      100000 non-null object
8   Place of Service                     100000 non-null object
9   HCPCS Code                           100000 non-null object
10  HCPCS Description                     100000 non-null object
11  HCPCS Drug Indicator                 100000 non-null object
12  Number of Services                   100000 non-null object
13  Number of Medicare Beneficiaries     100000 non-null object
14  Number of Distinct Medicare Beneficiary/Per Day Services 100000 non-null object
15  Average Medicare Allowed Amount       100000 non-null object
16  Average Submitted Charge Amount       100000 non-null object
17  Average Medicare Payment Amount       100000 non-null object
18  Average Medicare Standardized Amount 100000 non-null object
dtypes: object(19)
memory usage: 14.5+ MB

```

```
def RemoveComma(x):
    return x.replace(",", "")
df["Number of Services"] = pd.to_numeric(df["Number of Services"].apply(lambda x: RemoveComma(x)),
                                         errors= "ignore")
df["Number of Medicare Beneficiaries"] = pd.to_numeric(df["Number of Medicare Beneficiaries"].apply(lambda x: RemoveComma(x)),
                                                         errors= "ignore")
df["Number of Distinct Medicare Beneficiary/Per Day Services"] = pd.to_numeric(df["Number of Distinct Medicare Beneficiary/Per Day Services"].apply(lambda x: RemoveComma(x)),
                                                                                  errors= "ignore")
df["Average Medicare Allowed Amount"] = pd.to_numeric(df["Average Medicare Allowed Amount"].apply(lambda x: RemoveComma(x)),
                                                         errors= "ignore")
df["Average Submitted Charge Amount"] = pd.to_numeric(df["Average Submitted Charge Amount"].apply(lambda x: RemoveComma(x)),
                                                         errors= "ignore")
df["Average Medicare Payment Amount"] = pd.to_numeric(df["Average Medicare Payment Amount"].apply(lambda x: RemoveComma(x)),
                                                         errors= "ignore")
df["Average Medicare Standardized Amount"] = pd.to_numeric(df["Average Medicare Standardized Amount"].apply(lambda x: RemoveComma(x)),
                                                             errors= "ignore")
df.info()
```

```
>>> <class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 19 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   Credentials of the Provider               100000 non-null object
1   Gender of the Provider                   100000 non-null object
2   Entity Type of the Provider              100000 non-null object
3   City of the Provider                    100000 non-null object
4   State Code of the Provider              100000 non-null object
5   Country Code of the Provider            100000 non-null object
6   Provider Type                           100000 non-null object
7   Medicare Participation Indicator         100000 non-null object
8   Place of Service                        100000 non-null object
9   HCPCS Code                              100000 non-null object
10  HCPCS Description                       100000 non-null object
11  HCPCS Drug Indicator                   100000 non-null object
12  Number of Services                     100000 non-null float64
13  Number of Medicare Beneficiaries        100000 non-null int64
14  Number of Distinct Medicare Beneficiary/Per Day Services 100000 non-null int64
15  Average Medicare Allowed Amount         100000 non-null float64
16  Average Submitted Charge Amount         100000 non-null float64
17  Average Medicare Payment Amount         100000 non-null float64
18  Average Medicare Standardized Amount    100000 non-null float64
dtypes: float64(5), int64(2), object(12)
memory usage: 14.5+ MB
```

```
df['Number of Services']= df['Number of Services'].astype(np.int64)
df['Number of Medicare Beneficiaries']= df['Number of Medicare Beneficiaries'].astype(np.int64)
df['Number of Distinct Medicare Beneficiary/Per Day Services']= df['Number of Distinct Medicare Beneficiary/Per Day Services'].astype(np.int64)
df['Average Medicare Allowed Amount']=df['Average Medicare Allowed Amount'].astype(np.float64)
df['Average Submitted Charge Amount']=df['Average Submitted Charge Amount'].astype(np.float64)
df['Average Medicare Payment Amount']=df['Average Medicare Payment Amount'].astype(np.float64)
df['Average Medicare Standardized Amount'] = df['Average Medicare Standardized Amount'].astype(np.float64)
df.info()
```

```
>>> <class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 19 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   Credentials of the Provider               100000 non-null object
1   Gender of the Provider                   100000 non-null object
2   Entity Type of the Provider              100000 non-null object
3   City of the Provider                    100000 non-null object
4   State Code of the Provider              100000 non-null object
5   Country Code of the Provider            100000 non-null object
6   Provider Type                           100000 non-null object
7   Medicare Participation Indicator         100000 non-null object
```



```

8 Place of Service 100000 non-null object
9 HCPCS Code 100000 non-null object
10 HCPCS Description 100000 non-null object
11 HCPCS Drug Indicator 100000 non-null object
12 Number of Services 100000 non-null int64
13 Number of Medicare Beneficiaries 100000 non-null int64
14 Number of Distinct Medicare Beneficiary/Per Day Services 100000 non-null int64
15 Average Medicare Allowed Amount 100000 non-null float64
16 Average Submitted Charge Amount 100000 non-null float64
17 Average Medicare Payment Amount 100000 non-null float64
18 Average Medicare Standardized Amount 100000 non-null float64
dtypes: float64(4), int64(3), object(12)
memory usage: 14.5+ MB

```

7. creating new feature from existing features

```

# 1. Service Intensity
df['medicare Service Intensity'] = df['Number of Medicare Beneficiaries']/df['Number of Services']

# 2. Price Differential
df['Price Differential'] = df['Average Submitted Charge Amount'] - df['Average Medicare Allowed Amount']

# 3. Standardization Factor
df['Standardization Factor'] = df['Average Medicare Standardized Amount'] / df['Average Medicare Payment Amount']

```

```
df.head()
```



	Credentials of the Provider	Gender of the Provider	Entity Type of the Provider	City of the Provider	State Code of the Provider	Country Code of the Provider	Provider Type	Par
0	M.D.	F	I	SAINT LOUIS	MO	US	Internal Medicine	
1	M.D.	F	I	FAYETTEVILLE	NC	US	Obstetrics & Gynecology	
2	DPM	M	I	NORTH HAVEN	CT	US	Podiatry	
3	MD	M	I	KANSAS CITY	MO	US	Internal Medicine	
4	DO	M	I	JUPITER	FL	US	Internal Medicine	

5 rows × 22 columns

8. checking the numbers of uniques value for encoding

- unique values>50 then used label encoding
- unique values>50 then used one hot encoding

```

print(len(df['Credentials of the Provider'].unique()))
print(len(df['Gender of the Provider'].unique()))
print(len(df['Entity Type of the Provider'].unique()))
print(len(df['City of the Provider'].unique()))
print(len(df['State Code of the Provider'].unique()))
print(len(df['Country Code of the Provider'].unique()))
print(len(df['Provider Type'].unique()))
print(len(df['Medicare Participation Indicator'].unique()))
print(len(df['Place of Service'].unique()))
print(len(df['HCPCS Code'].unique()))
print(len(df['HCPCS Description'].unique()))
print(len(df['HCPCS Drug Indicator'].unique()))

```

```

1855
3
2
5846
58
4
90
2
2
2631
2455
2

```

9. Apply the label encoding for categorical columns

prompt: i have label encoding for Credentials of the Provider City of the Provider State Code of the Provider Pr

```

from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()

```

```

df['Credentials of the Provider'] = le.fit_transform(df['Credentials of the Provider'])
df['City of the Provider'] = le.fit_transform(df['City of the Provider'])
df['State Code of the Provider'] = le.fit_transform(df['State Code of the Provider'])
df['Provider Type'] = le.fit_transform(df['Provider Type'])
df['HCPCS Code'] = le.fit_transform(df['HCPCS Code'])
df['HCPCS Description'] = le.fit_transform(df['HCPCS Description'])

```

```
df.head()
```



	Credentials of the Provider	Gender of the Provider	Entity Type of the Provider	City of the Provider	State Code of the Provider	Country Code of the Provider	Provider Type	Medi Participa Indic
0	667	F	I	4541	28	US	39	
1	667	F	I	1624	31	US	54	
2	438	M	I	3666	9	US	71	
3	956	M	I	2522	28	US	39	
4	410	M	I	2508	12	US	39	

5 rows × 22 columns

10. Apply the one hot encoding for categorical columns

```
# prompt: i have one hot encoding for Gender of the Provider Entity Type of the Provider Country Code of the Provi
```

```
df = pd.get_dummies(df, columns=['Gender of the Provider', 'Entity Type of the Provider', 'Country Code of the Pro
df.head()
```



	Credentials of the Provider	City of the Provider	State Code of the Provider	Provider Type	HCPCS Code	HCPCS Description	Number of Services	Numbe Medi Beneficia
0	667	4541	28	39	2251	967	27	
1	667	1624	31	54	2374	2054	175	
2	438	3666	9	71	2295	665	32	
3	956	2522	28	39	1329	2330	20	
4	410	2508	12	39	2163	973	33	

5 rows × 31 columns

```
df.info()
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 31 columns):
#   Column                                                                 Non-Null Count  Dtype
---  -
0   Credentials of the Provider                                           100000 non-null  int64
1   City of the Provider                                                  100000 non-null  int64
2   State Code of the Provider                                           100000 non-null  int64
3   Provider Type                                                         100000 non-null  int64
4   HCPCS Code                                                            100000 non-null  int64
5   HCPCS Description                                                     100000 non-null  int64
6   Number of Services                                                    100000 non-null  int64
7   Number of Medicare Beneficiaries                                     100000 non-null  int64
8   Number of Distinct Medicare Beneficiary/Per Day Services            100000 non-null  int64
9   Average Medicare Allowed Amount                                       100000 non-null  float64
10  Average Submitted Charge Amount                                       100000 non-null  float64
11  Average Medicare Payment Amount                                       100000 non-null  float64
12  Average Medicare Standardized Amount                                 100000 non-null  float64
13  medicare Service Intensity                                             100000 non-null  float64
14  Price Differential                                                    100000 non-null  float64
15  Standardization Factor                                                100000 non-null  float64
16  Gender of the Provider_F                                              100000 non-null  int64
17  Gender of the Provider_M                                              100000 non-null  int64
18  Gender of the Provider_Unknown                                         100000 non-null  int64
19  Entity Type of the Provider_I                                          100000 non-null  int64
20  Entity Type of the Provider_O                                          100000 non-null  int64
21  Country Code of the Provider_DE                                        100000 non-null  int64
22  Country Code of the Provider_JP                                        100000 non-null  int64
23  Country Code of the Provider_TR                                        100000 non-null  int64
24  Country Code of the Provider_US                                        100000 non-null  int64
25  Medicare Participation Indicator_N                                     100000 non-null  int64
26  Medicare Participation Indicator_Y                                     100000 non-null  int64
27  Place of Service_F                                                    100000 non-null  int64
28  Place of Service_O                                                    100000 non-null  int64
29  HCPCS Drug Indicator_N                                                100000 non-null  int64
30  HCPCS Drug Indicator_Y                                                100000 non-null  int64
dtypes: float64(7), int64(24)
memory usage: 23.7 MB
```

11. Apply the standard scaler after succesfull encoding for all categorical columns as well as numerical columns


```
# prompt: all colums apply standard scaler

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
df_scaled = scaler.fit_transform(df)

# Create a new DataFrame with the scaled data
df_scaled = pd.DataFrame(df_scaled, columns=df.columns)

# Print the first few rows of the scaled data
df_scaled.head()
```



	Credentials of the Provider	City of the Provider	State Code of the Provider	Provider Type	HCPCS Code	HCPCS Description	Number of Services	N Benef
0	-0.525655	0.998857	-0.021772	-0.037953	0.782766	-0.336239	-0.085301	.
1	-0.525655	-0.727201	0.175119	0.660735	0.958212	1.105866	-0.025939	
2	-1.092238	0.481099	-1.268747	1.452582	0.845527	-0.736897	-0.083296	.
3	0.189377	-0.195833	-0.021772	-0.037953	-0.532364	1.472031	-0.088109	.
4	-1.161514	-0.204118	-1.071857	-0.037953	0.657244	-0.328279	-0.082895	.

5 rows × 31 columns

12.apply the Principal Component Analysis

- 20 components converted


```
# prompt: pca 20 dimension

from sklearn.decomposition import PCA

pca = PCA(n_components=20)
pca_components = pca.fit_transform(df_scaled)

# Create a new DataFrame with the PCA components
pca_df = pd.DataFrame(pca_components, columns=[f"PC{i+1}" for i in range(20)])

# Print the first few rows of the PCA DataFrame
pca_df.head()
```



	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	
0	0.500139	-0.730306	-0.082374	-0.428770	-0.360424	-0.109642	1.783237	-1.761939	(
1	-0.114045	0.137931	-0.250051	1.272301	-0.312189	-0.037351	1.972540	-0.129650	(
2	-0.844076	-0.504543	0.104084	1.447918	0.105589	-0.011592	-0.360755	1.352774	(
3	-0.856746	-0.215941	-0.082006	0.012066	0.183853	0.018137	-0.759451	0.923332	(
4	-0.995749	-0.508510	0.063426	0.591117	0.128362	-0.009369	-0.596144	1.089525	-(

Next steps:

Generate code with `pca_df`

 View recommended plots

APPLY MACHINE LEARNING TECHNIQUES

1.Isolation forest

```
#apply isolation forest

from sklearn.ensemble import IsolationForest

# Define the model
model = IsolationForest(n_estimators=100, contamination=0.007)

# Fit the model
model.fit(df_scaled)


# Predict the outlier scores
outlier_scores = model.decision_function(df_scaled)

# Get the outlier predictions
outlier_predictions = model.predict(df_scaled)

# Create a new DataFrame with the outlier scores and predictions
df_outliers = pd.DataFrame({
    'Outlier Score': outlier_scores,
    'Outlier Prediction': outlier_predictions
})

# Merge the outlier information with the original data
df_with_outliers = pd.merge(df, df_outliers, left_index=True, right_index=True)

# Print the first few rows of the data with outlier information
df_with_outliers.head()
```

 /usr/local/lib/python3.10/dist-packages/sklearn/base.py:439: UserWarning: X does not
warnings.warn(

	Credentials of the Provider	City of the Provider	State Code of the Provider	Provider Type	HCPCS Code	HCPCS Description	Number of Services	Numbe Medi Beneficia
0	667	4541	28	39	2251	967	27	
1	667	1624	31	54	2374	2054	175	
2	438	3666	9	71	2295	665	32	
3	956	2522	28	39	1329	2330	20	
4	410	2508	12	39	2163	973	33	

5 rows × 33 columns

```
list(outlier_predictions).count(-1)
```

 700

```
# diff graph of feature1 vs feature2 with outlier prediction
```

```
import matplotlib.pyplot as plt
```

```
# Separate the data by outlier prediction
```

```
outliers = df_with_outliers[df_with_outliers['Outlier Prediction'] == -1]
```

```
regular = df_with_outliers[df_with_outliers['Outlier Prediction'] == 1]
```

```
# Plot the data points
```

```
plt.scatter(outliers['Number of Services'], outliers['Number of Medicare Beneficiaries'], color='red', label='Outl
```

```
plt.scatter(regular['Number of Services'], regular['Number of Medicare Beneficiaries'], color='blue', label='Regul
```

```
# Add labels and title
```

```
plt.xlabel('Number of Services')
```

```
plt.ylabel('Number of Medicare Beneficiaries')
```

```
plt.title('Number of Services vs Number of Medicare Beneficiaries with Outlier Predictions')
```

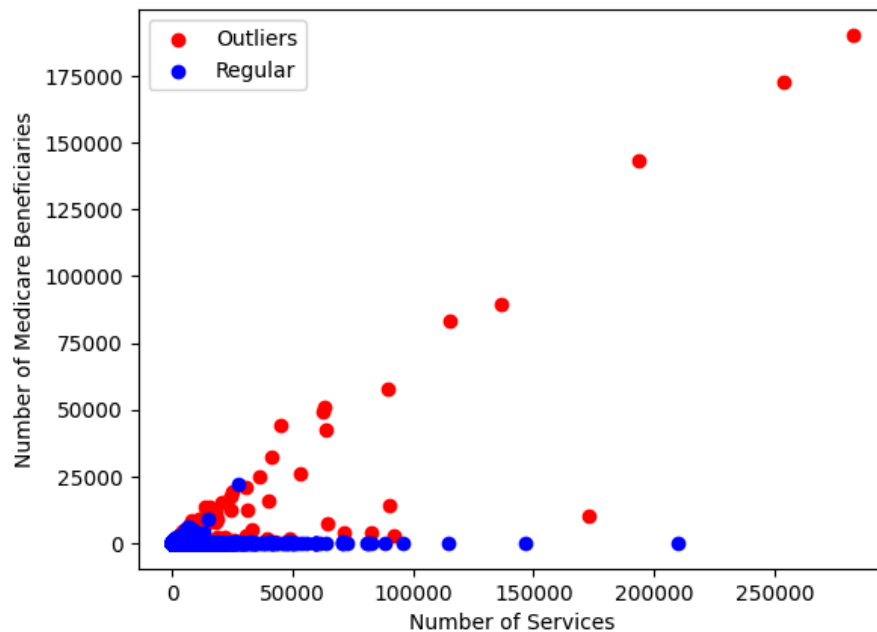
```
# Add legend and show plot
```

```
plt.legend()
```

```
plt.show()
```



Number of Services vs Number of Medicare Beneficiaries with Outlier Predictions



```
plt.scatter(outliers['Number of Services'], outliers['Number of Distinct Medicare Beneficiary/Per Day Services'],
plt.scatter(regular['Number of Services'], regular['Number of Distinct Medicare Beneficiary/Per Day Services'], cc
```

```
# Add labels and title
```

```
plt.xlabel('Number of Services')
```

```
plt.ylabel('Number of Distinct Medicare Beneficiary/Per Day Services')
```

```
plt.title('Number of Services vs Number of Distinct Medicare Beneficiary/Per Day Servicess with Outlier Predictor
```

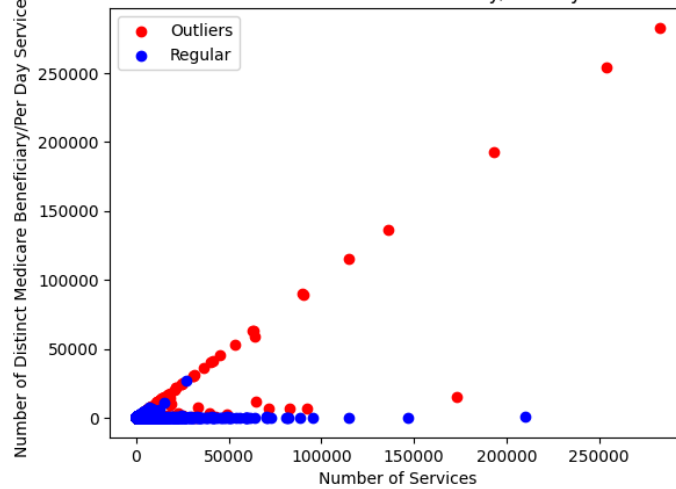
```
# Add legend and show plot
```

```
plt.legend()
```

```
plt.show()
```



Number of Services vs Number of Distinct Medicare Beneficiary/Per Day Servicess with Outlier Predictions



```
plt.scatter(outliers['Average Submitted Charge Amount'], outliers['Average Medicare Allowed Amount'], color='red',
plt.scatter(regular['Average Submitted Charge Amount'], regular['Average Medicare Allowed Amount'], color='blue',
```

```
# Add labels and title
```

```
plt.xlabel('Average Submitted Charge Amount')
```

```
plt.ylabel('Average Medicare Allowed Amount')
```

```
plt.title('Average Submitted Charge Amount vs Average Medicare Allowed Amount with Outlier Predictions')
```

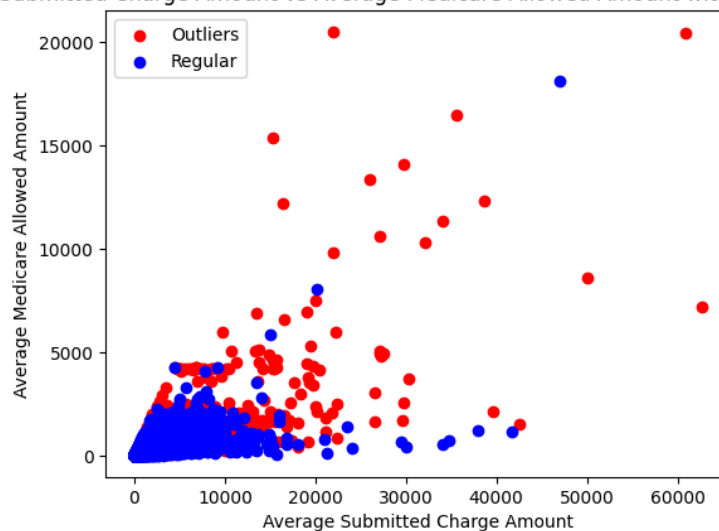
```
# Add legend and show plot
```

```
plt.legend()
```

```
plt.show()
```



Average Submitted Charge Amount vs Average Medicare Allowed Amount with Outlier Predictions



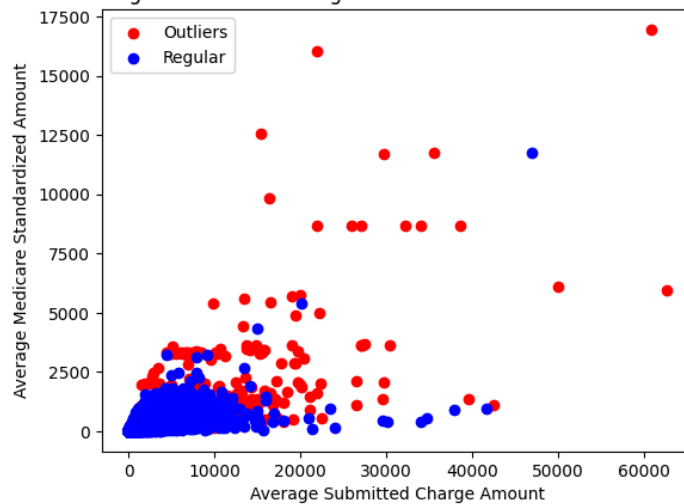
```
plt.scatter(outliers['Average Submitted Charge Amount'], outliers['Average Medicare Standardized Amount'], color='red')
plt.scatter(regular['Average Submitted Charge Amount'], regular['Average Medicare Standardized Amount'], color='blue')

# Add labels and title
plt.xlabel('Average Submitted Charge Amount')
plt.ylabel('Average Medicare Standardized Amount')
plt.title('Average Submitted Charge Amount vs Average Medicare Standardized Amount with Outlier Predictions')

# Add legend and show plot
plt.legend()
plt.show()
```



Average Submitted Charge Amount vs Average Medicare Standardized Amount with Outlier Predictions




```
# prompt: ANOTHER COLUMNS SUGGEST ME TYPE GRAPH WITH DETECTION MODEL OUTLIER PREDICTION
```

```
# Create a scatter plot of the 'Number of Services' and 'Average Medicare Payment Amount' columns, colored by outl
plt.scatter(outliers['Number of Services'], outliers['Average Medicare Payment Amount'], color='red', label='Outli
plt.scatter(regular['Number of Services'], regular['Average Medicare Payment Amount'], color='blue', label='Regula
```

```
# Add labels and title
```

```
plt.xlabel('Number of Services')
```

```
plt.ylabel('Average Medicare Payment Amount')
```

```
plt.title('Number of Services vs Average Medicare Payment Amount with Outlier Predictions')
```

```
# Add legend and show plot
```

```
plt.legend()
```

```
plt.show()
```

```
# Create a scatter plot of the 'Number of Medicare Beneficiaries' and 'Average Medicare Payment Amount' columns, c
plt.scatter(outliers['Number of Medicare Beneficiaries'], outliers['Average Medicare Payment Amount'], color='red'
plt.scatter(regular['Number of Medicare Beneficiaries'], regular['Average Medicare Payment Amount'], color='blue',
```

```
# Add labels and title
```

```
plt.xlabel('Number of Medicare Beneficiaries')
```

```
plt.ylabel('Average Medicare Payment Amount')
```

```
plt.title('Number of Medicare Beneficiaries vs Average Medicare Payment Amount with Outlier Predictions')
```

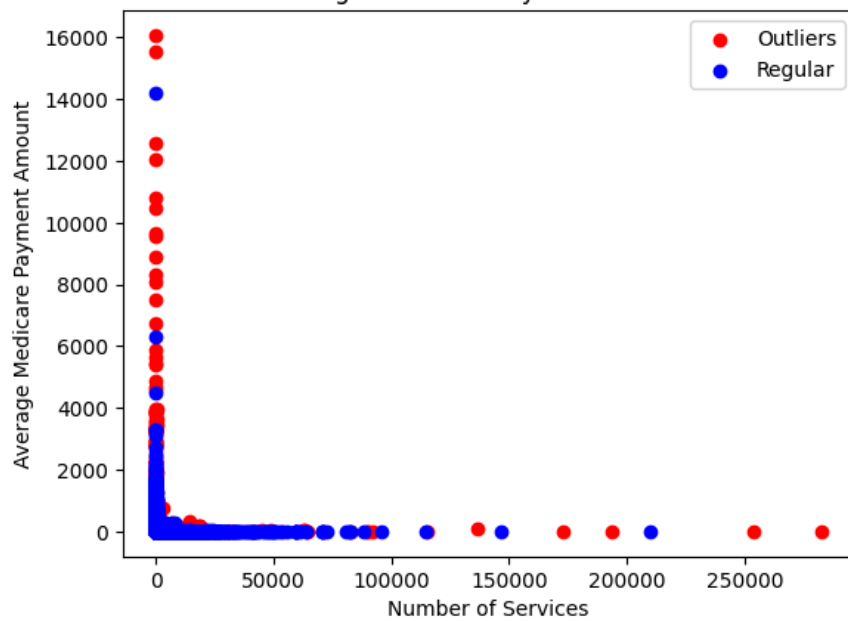
```
# Add legend and show plot
```

```
plt.legend()
```

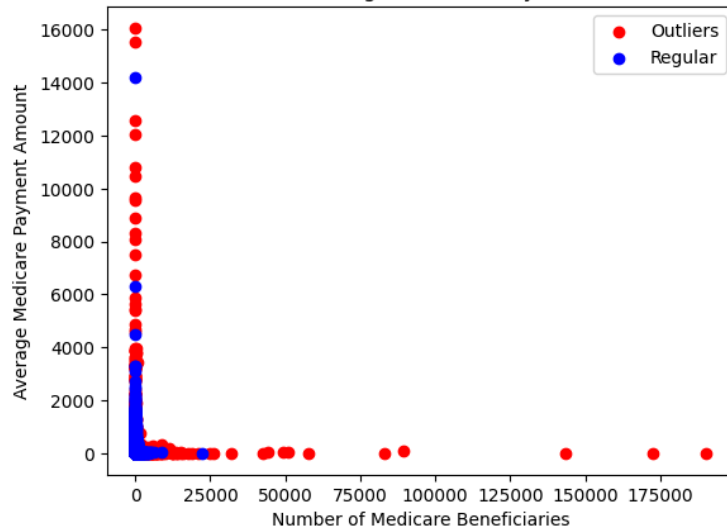
```
plt.show()
```



Number of Services vs Average Medicare Payment Amount with Outlier Predictions



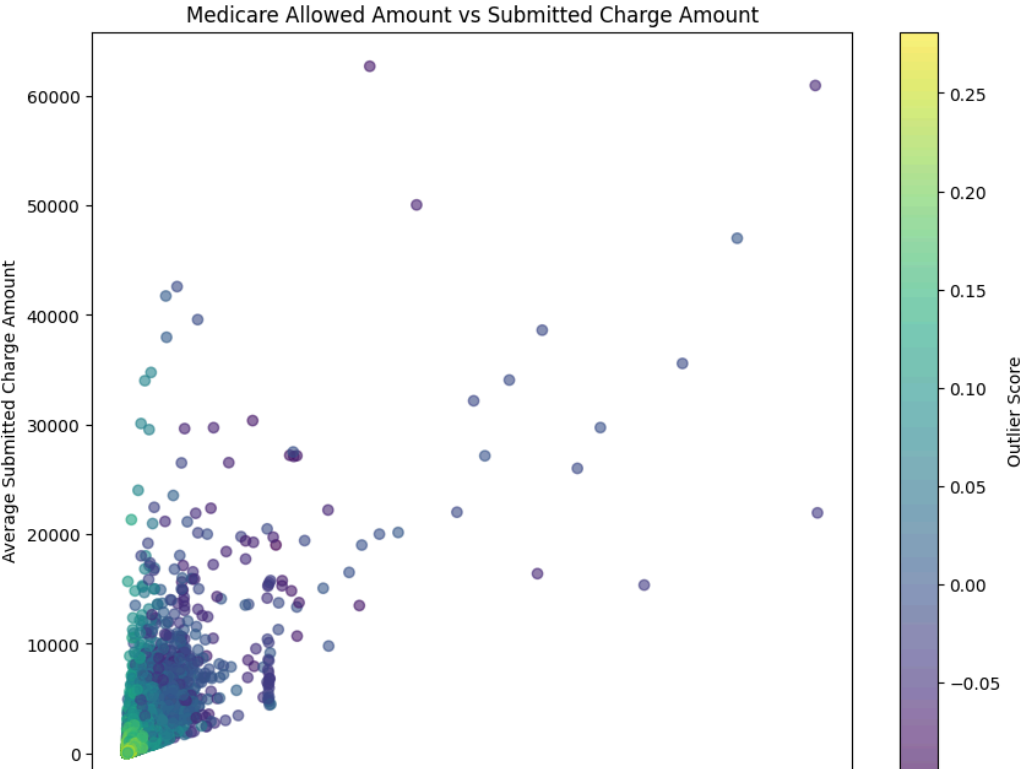
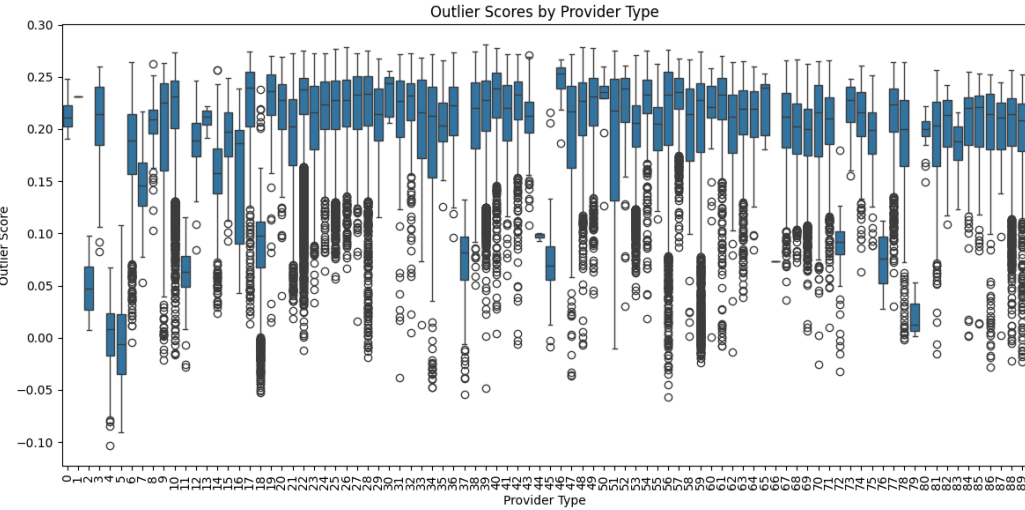
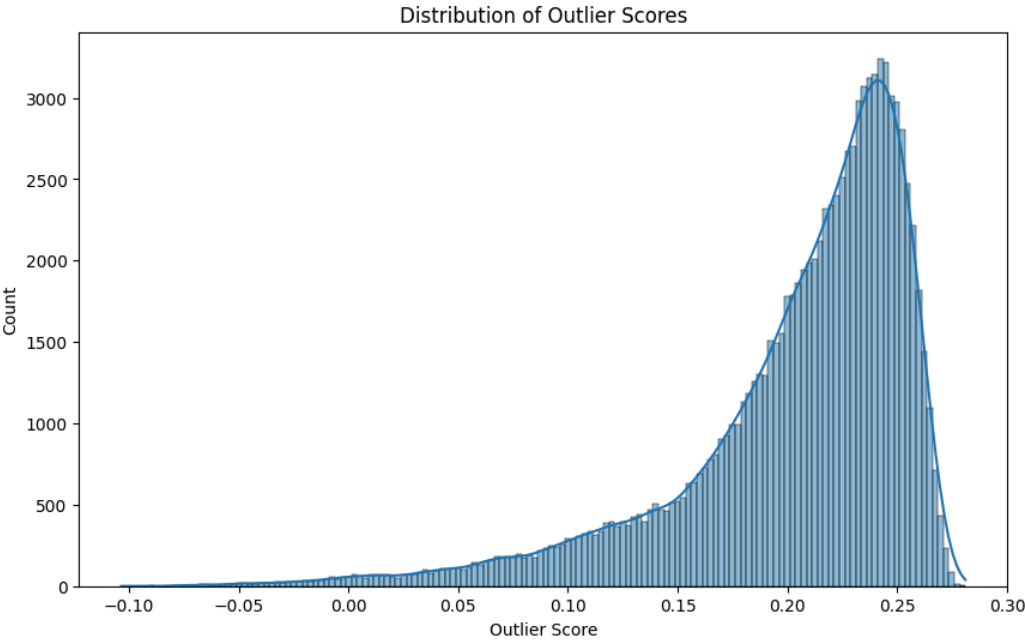
Number of Medicare Beneficiaries vs Average Medicare Payment Amount with Outlier Predictions



```
# 2. Histogram of Outlier Scores
plt.figure(figsize=(10, 6))
sns.histplot(df_with_outliers['Outlier Score'], kde=True)
plt.title('Distribution of Outlier Scores')
plt.xlabel('Outlier Score')
plt.ylabel('Count')
plt.show()

# 3. Box plot of Outlier Scores by Provider Type
plt.figure(figsize=(12, 6))
sns.boxplot(x='Provider Type', y='Outlier Score', data=df_with_outliers)
plt.xticks(rotation=90)
plt.title('Outlier Scores by Provider Type')
plt.tight_layout()
plt.show()

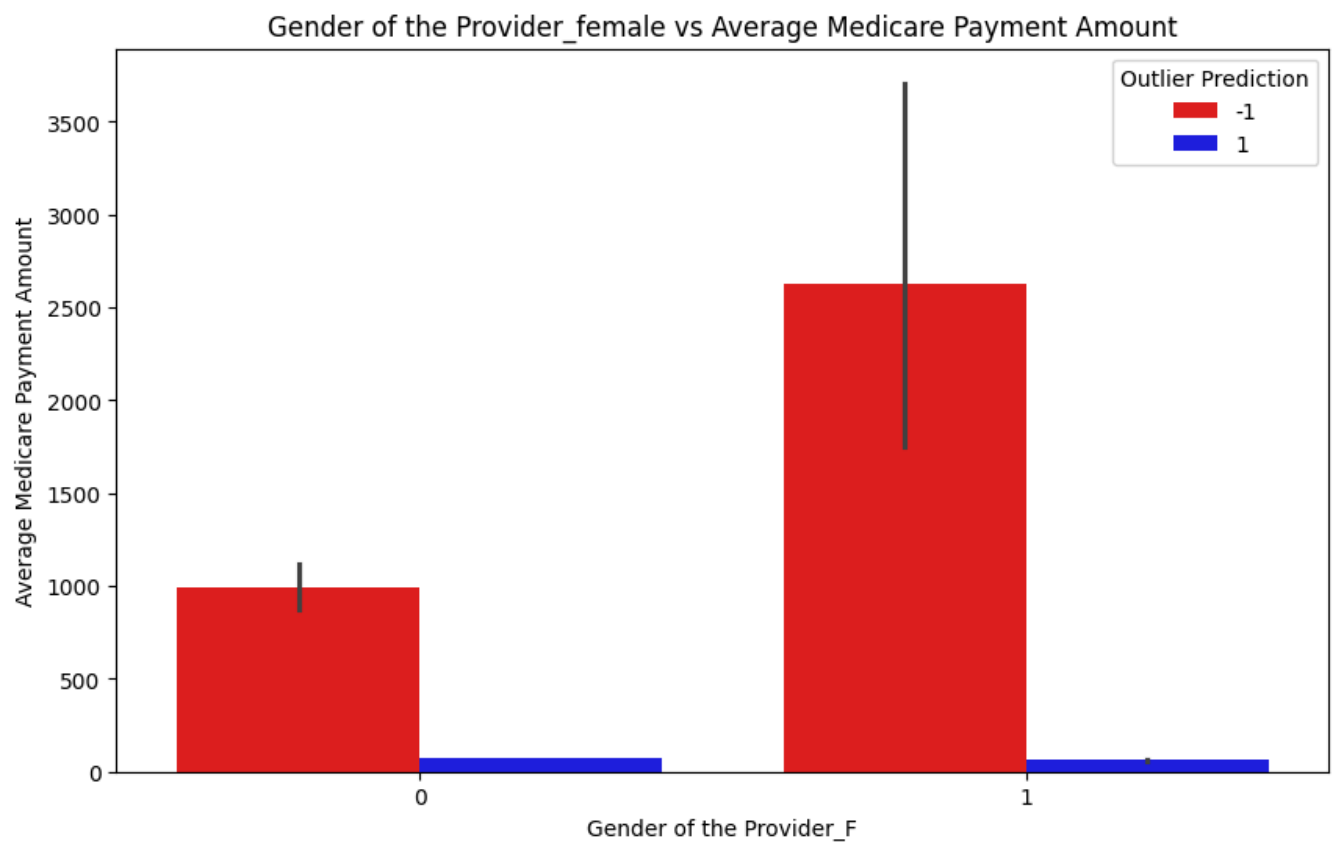
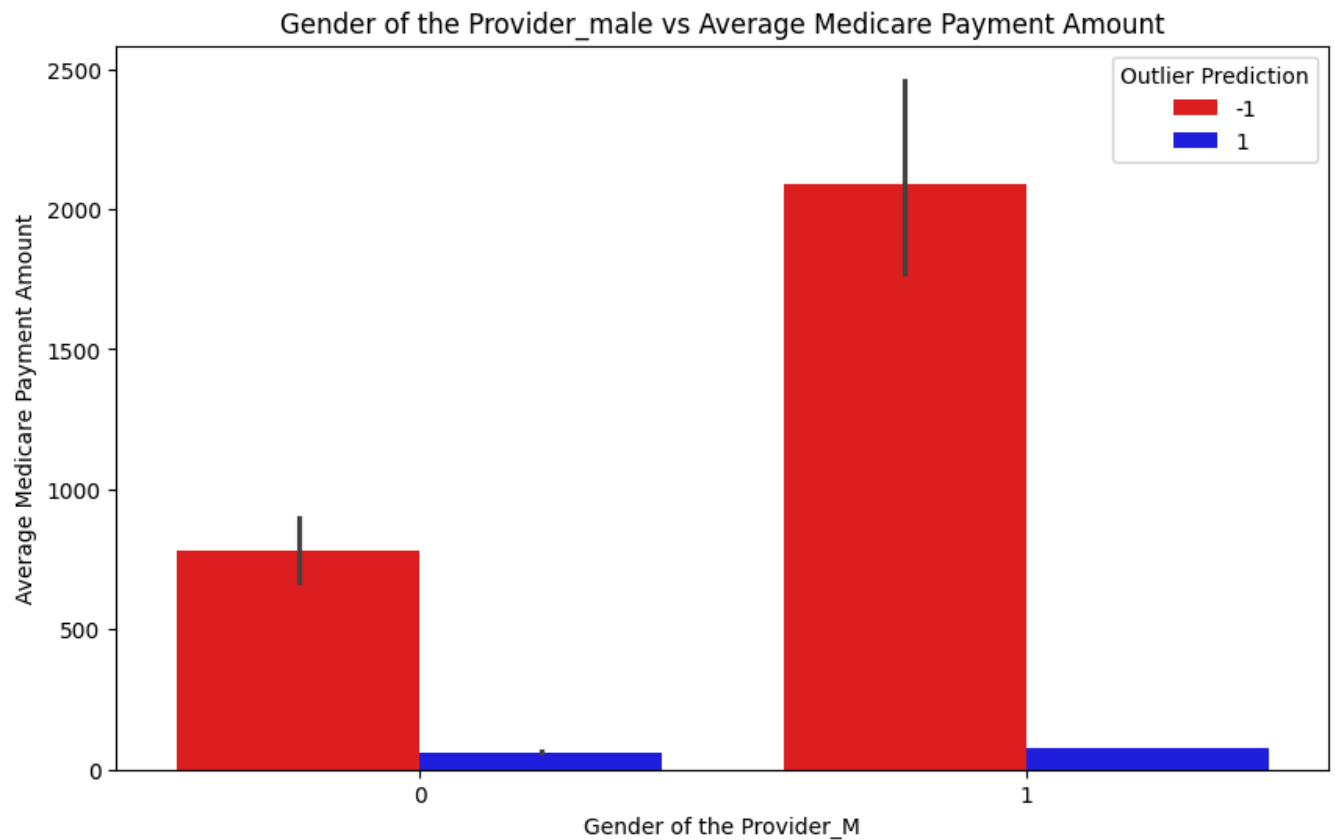
# 4. Scatter plot of Average Medicare Allowed Amount vs Average Submitted Charge Amount
plt.figure(figsize=(10, 8))
plt.scatter(df_with_outliers['Average Medicare Allowed Amount'].astype(float),
            df_with_outliers['Average Submitted Charge Amount'].astype(float),
            c=df_with_outliers['Outlier Score'],
            cmap='viridis',
            alpha=0.6)
plt.colorbar(label='Outlier Score')
plt.title('Medicare Allowed Amount vs Submitted Charge Amount')
plt.xlabel('Average Medicare Allowed Amount')
plt.ylabel('Average Submitted Charge Amount')
plt.show()
```





```
plt.figure(figsize=(10, 6))
sns.barplot(x='Gender of the Provider_M', y='Average Medicare Payment Amount',
            hue='Outlier Prediction', data=df_with_outliers,
            palette={-1: 'red', 1: 'blue'})
plt.title('Gender of the Provider_male vs Average Medicare Payment Amount')
plt.show()
```

```
plt.figure(figsize=(10, 6))
sns.barplot(x='Gender of the Provider_F', y='Average Medicare Payment Amount',
            hue='Outlier Prediction', data=df_with_outliers,
            palette={-1: 'red', 1: 'blue'})
plt.title('Gender of the Provider_female vs Average Medicare Payment Amount')
plt.show()
```



2. Elliptical envelope

```
# elliptical envelope technique apply
```

```
from sklearn.covariance import EllipticEnvelope
```

```

# Define the model
model = EllipticEnvelope(contamination=0.005)

# Fit the model
model.fit(df_scaled)

# Predict the outlier scores
outlier_scores = model.decision_function(df_scaled)

# Get the outlier predictions
outlier_predictions = model.predict(df_scaled)

# Create a new DataFrame with the outlier scores and predictions
df_outliers = pd.DataFrame({
    'Outlier Score': outlier_scores,
    'Outlier Prediction': outlier_predictions
})

# Merge the outlier information with the original data
df_with_outliers = pd.merge(df, df_outliers, left_index=True, right_index=True)

# Print the first few rows of the data with outlier information
df_with_outliers.head()

# Separate the data by outlier prediction
outliers = df_with_outliers[df_with_outliers['Outlier Prediction'] == -1]
regular = df_with_outliers[df_with_outliers['Outlier Prediction'] == 1]

# Plot the data points
plt.scatter(outliers['Number of Services'], outliers['Number of Medicare Beneficiaries'], color='red', label='Outli
plt.scatter(regular['Number of Services'], regular['Number of Medicare Beneficiaries'], color='blue', label='Regula

# Add labels and title
plt.xlabel('Number of Services')
plt.ylabel('Number of Medicare Beneficiaries')
plt.title('Number of Services vs Number of Medicare Beneficiaries with Outlier Predictions')

# Add legend and show plot
plt.legend()
plt.show()

plt.scatter(outliers['Number of Services'], outliers['Number of Distinct Medicare Beneficiary/Per Day Services'], c
plt.scatter(regular['Number of Services'], regular['Number of Distinct Medicare Beneficiary/Per Day Services'], col

# Add labels and title
plt.xlabel('Number of Services')
plt.ylabel('Number of Distinct Medicare Beneficiary/Per Day Services')
plt.title('Number of Services vs Number of Distinct Medicare Beneficiary/Per Day Servicess with Outlier Predictions

# Add legend and show plot
plt.legend()
plt.show()

plt.scatter(outliers['Average Submitted Charge Amount'], outliers['Average Medicare Allowed Amount'], color='red',
plt.scatter(regular['Average Submitted Charge Amount'], regular['Average Medicare Allowed Amount'], color='blue', 1

# Add labels and title
plt.xlabel('Average Submitted Charge Amount')
plt.ylabel('Average Medicare Allowed Amount')
plt.title('Average Submitted Charge Amount vs Average Medicare Allowed Amount with Outlier Predictions')

# Add legend and show plot
plt.legend()
plt.show()

```

```
plt.show()
```

```
plt.scatter(outliers['Average Submitted Charge Amount'], outliers['Average Medicare Standardized Amount'], color='r')  
plt.scatter(regular['Average Submitted Charge Amount'], regular['Average Medicare Standardized Amount'], color='blue')
```

```
# Add labels and title
```

```
plt.xlabel('Average Submitted Charge Amount')
```

```
plt.ylabel('Average Medicare Standardized Amount')
```

```
plt.title('Average Submitted Charge Amount vs Average Medicare Standardized Amount with Outlier Predictions')
```

```
# Add legend and show plot
```

```
plt.legend()
```

```
plt.show()
```


https://colab.research.google.com/drive/1cFRg-tKsaRZ3c43PfBG71o82SIPIB8QF#scrollTo=8Mau_ojbDsqz&printMode=true

[illegible]

h research google.com/drive/1cEBq-tKsaB73c43PfBG71o82SIPB8OE#scrollTo=8Mau_oibDsgz&printMode=true

https://colab.research.google.com/drive/1cFRg-tKsaRZ3c43PfBG71o82SIPIB8QF#scrollTo=8Mau_oibDsqr&printMode=true

https://colab.research.google.com/drive/1cFRg-tKsaRZ3c43PfBG71o82SIPIB8QF#scrollTo=8Mau_ojbDsqz&printMode=true

[illegible]

https://colab.research.google.com/drive/1cFRg-tKsaRZ3c43PfBG71o82SIPIB8QF#scrollTo=8Mau_ojbDsqz&printMode=true

https://colab.research.google.com/drive/1cFRq-tKsaRZ3c43PfBG71o82SIPIB8QF#scrollTo=8Mau_oibDsqq&printMode=true

[illegible]

b.research.google.com/drive/1cFRg-tKsaRZ3c43PfBG71o82SIPIB8QF#scrollTo=8Mau_ojbDsQz&printMode=true

https://colab.research.google.com/drive/1cFRq-tKsaRZ3c43PfBG71o82SIPIB8QF#scrollTo=8Mau_ojbDsQz&printMode=true

```
/usr/local/lib/python3.10/dist-packages/sklearn/covariance/_robust_covariance.py:184
```

$\chi^2 = 0.97$, d.f. = 1, $p = 0.62$

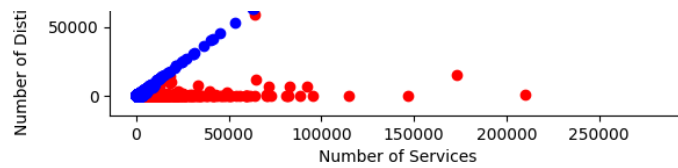
https://colab.research.google.com/drive/1cFRq-tKsaRZ3c43PfBG71o82SIPIB8QF#scrollTo=8Mau_oibDsqr&printMode=true

[illegible]

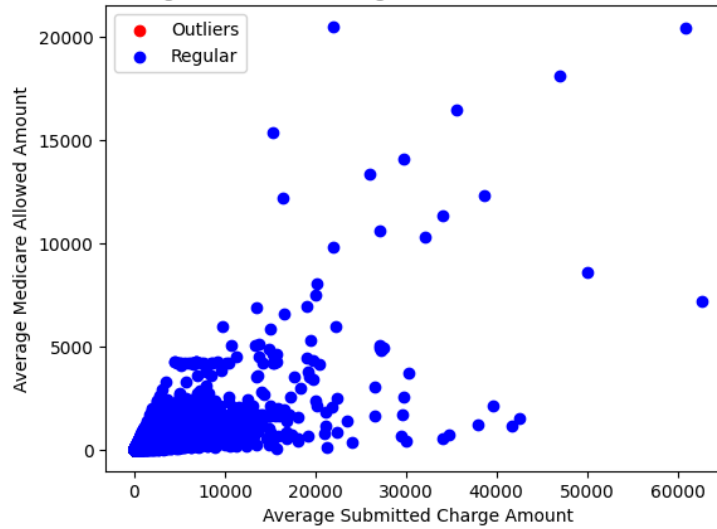
b.research.google.com/drive/1cFRg-tKsaRZ3c43PfbG71o82SIpB8QF#scrollTo=8Mau_oibDsqz&printMode=true

Number of Services vs Number of Medicare Beneficiaries with Outlier Predictions

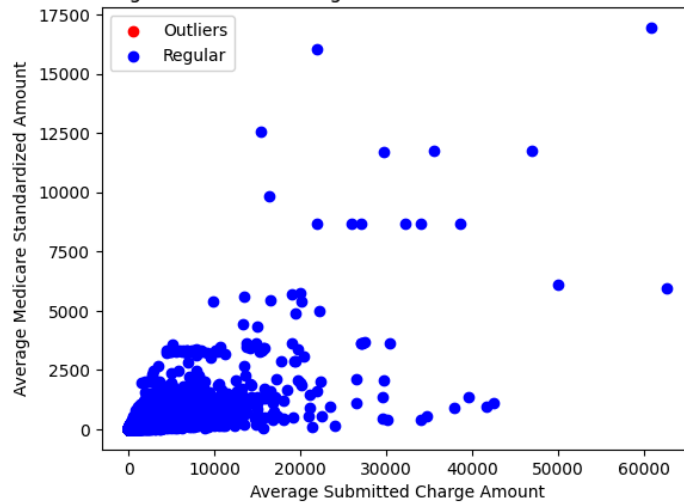




Average Submitted Charge Amount vs Average Medicare Allowed Amount with Outlier Predictions



Average Submitted Charge Amount vs Average Medicare Standardized Amount with Outlier Predictions



```
# Create a scatter plot of the 'Number of Services' and 'Average Medicare Payment Amount' columns, colored by outlier status
plt.scatter(outliers['Number of Services'], outliers['Average Medicare Payment Amount'], color='red', label='Outliers')
plt.scatter(regular['Number of Services'], regular['Average Medicare Payment Amount'], color='blue', label='Regular')

# Add labels and title
plt.xlabel('Number of Services')
plt.ylabel('Average Medicare Payment Amount')
plt.title('Number of Services vs Average Medicare Payment Amount with Outlier Predictions')

# Add legend and show plot
plt.legend()
plt.show()

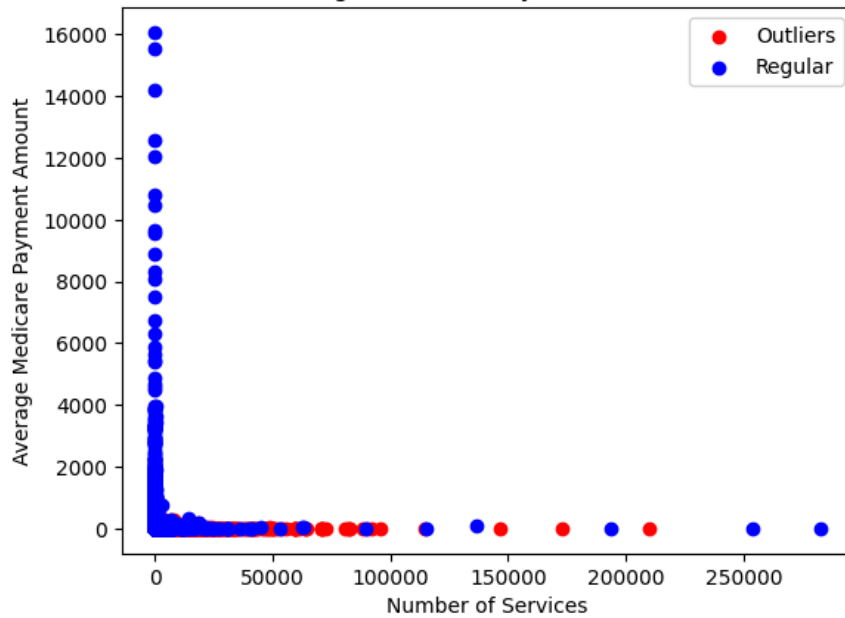
# Create a scatter plot of the 'Number of Medicare Beneficiaries' and 'Average Medicare Payment Amount' columns, colored by outlier status
plt.scatter(outliers['Number of Medicare Beneficiaries'], outliers['Average Medicare Payment Amount'], color='red', label='Outliers')
plt.scatter(regular['Number of Medicare Beneficiaries'], regular['Average Medicare Payment Amount'], color='blue', label='Regular')

# Add labels and title
plt.xlabel('Number of Medicare Beneficiaries')
plt.ylabel('Average Medicare Payment Amount')
plt.title('Number of Medicare Beneficiaries vs Average Medicare Payment Amount with Outlier Predictions')

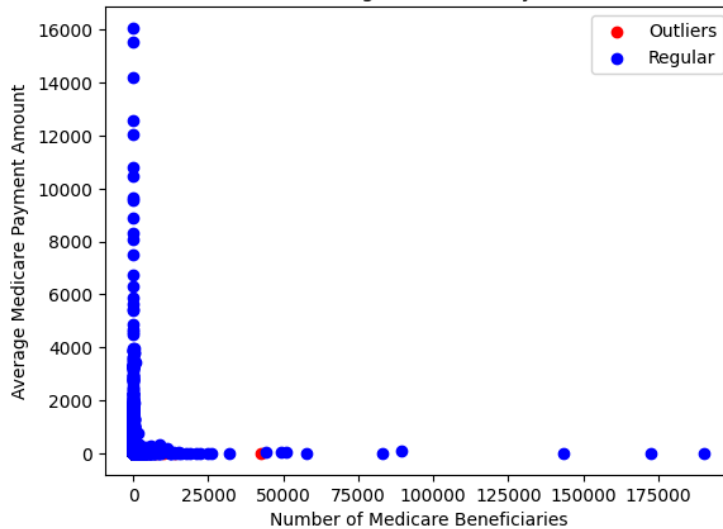
# Add legend and show plot
plt.legend()
plt.show()
```



Number of Services vs Average Medicare Payment Amount with Outlier Predictions



Number of Medicare Beneficiaries vs Average Medicare Payment Amount with Outlier Predictions

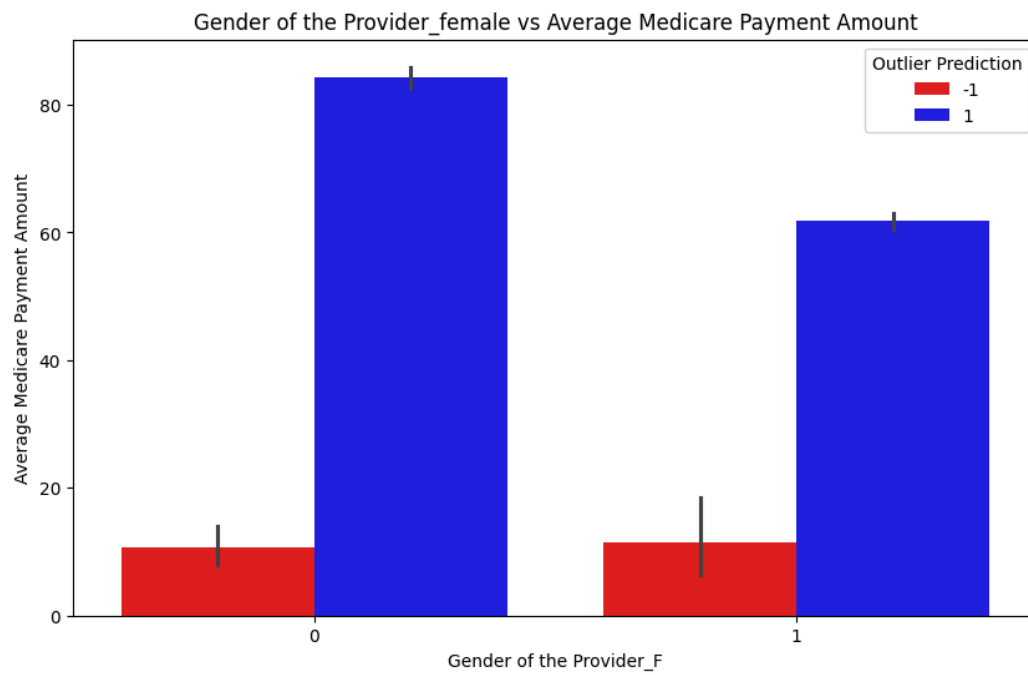
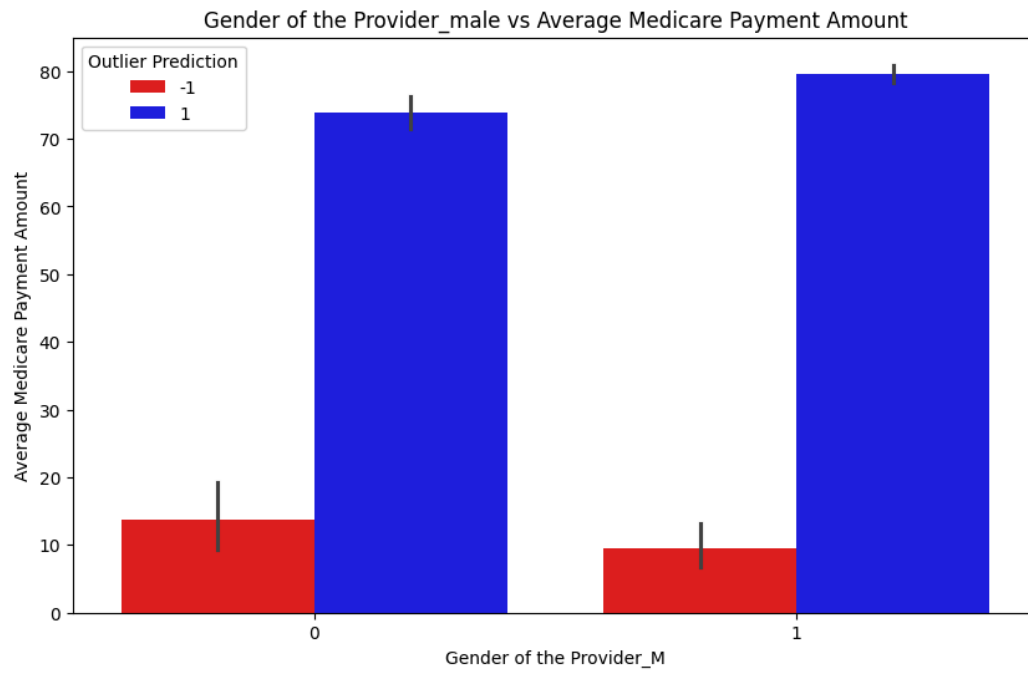


```
## 2. Histogram of Outlier Scores
# plt.figure(figsize=(10, 6))
# sns.histplot(df_with_outliers['Outlier Score'], kde=True)
# plt.title('Distribution of Outlier Scores')
# plt.xlabel('Outlier Score')
# plt.ylabel('Count')
# plt.show()

## 3. Box plot of Outlier Scores by Provider Type
# plt.figure(figsize=(12, 6))
# sns.boxplot(x='Provider Type', y='Outlier Score', data=df_with_outliers)
# plt.xticks(rotation=90)
# plt.title('Outlier Scores by Provider Type')
# plt.tight_layout()
# plt.show()
```

```
plt.figure(figsize=(10, 6))
sns.barplot(x='Gender of the Provider_M', y='Average Medicare Payment Amount',
            hue='Outlier Prediction', data=df_with_outliers,
            palette={-1: 'red', 1: 'blue'})
plt.title('Gender of the Provider_male vs Average Medicare Payment Amount')
plt.show()

plt.figure(figsize=(10, 6))
sns.barplot(x='Gender of the Provider_F', y='Average Medicare Payment Amount',
            hue='Outlier Prediction', data=df_with_outliers,
            palette={-1: 'red', 1: 'blue'})
plt.title('Gender of the Provider_female vs Average Medicare Payment Amount')
plt.show()
```



```
list(outlier_predictions).count(-1)
```



500

3.One class svm

```
# one class svm apply

from sklearn.svm import OneClassSVM

# Define the model
model = OneClassSVM(nu=0.007)

# Fit the model
```