importing libraries, and load dataset file

In [2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from datetime import datetime, date, time

df=pd.read_excel(r"C:\Users\KARTIKI\Downloads\preesia.xlsx")
df
```

Out[2]:

	Client Identifier	Insurance Eligiblity	Intake Category	Pre-visit Status	Visit Month	Amount Collected	Amount Owed	Intake Interview Duration (Minutes)	Visit Count	Client Specialty
0	519	Eligible	No check-in	Completed	2023-07-01	0.00	0.00	NaN	11	Pediatrics
1	519	Eligible	No check-in	Not Completed	2023-07-01	0.00	60.00	NaN	26	Pediatrics
2	519	Eligible	Self-Service	Not Completed	2023-07-01	0.00	0.00	00:11:27.037000	30	Pediatrics
3	519	Eligible	Staff	Not Completed	2023-07-01	70.12	251.07	00:00:00.651000	284	Pediatrics
4	519	Eligible	Self-Service	Completed	2023-07-01	205.00	245.00	00:00:11.313000	368	Pediatrics
...	...	...	...	...	...	...	...	...	...	...
85388	2951	Unsupported Payer	No check-in	Not Completed	2023-08-01	0.00	45.00	NaN	3	Primary Care
85389	2951	Eligible	No check-in	Not Completed	2023-09-01	25.00	9271.21	NaN	614	Primary Care
85390	2951	Error	No check-in	Not Completed	2023-09-01	131.00	756.13	NaN	251	Primary Care
85391	2951	No Coverage/Self Pay	No check-in	Not Completed	2023-09-01	0.00	180.98	NaN	20	Primary Care
85392	2951	Unsupported Payer	No check-in	Not Completed	2023-09-01	0.00	45.00	NaN	2	Primary Care

85393 rows × 10 columns

In [3]:
```
df.head()
```

Out[3]:

	Client Identifier	Insurance Eligiblity	Intake Category	Pre-visit Status	Visit Month	Amount Collected	Amount Owed	Intake Interview Duration (Minutes)	Visit Count	Client Specialty
0	519	Eligible	No check-in	Completed	2023-07-01	0.00	0.00	NaN	11	Pediatrics
1	519	Eligible	No check-in	Not Completed	2023-07-01	0.00	60.00	NaN	26	Pediatrics
2	519	Eligible	Self-Service	Not Completed	2023-07-01	0.00	0.00	00:11:27.037000	30	Pediatrics
3	519	Eligible	Staff	Not Completed	2023-07-01	70.12	251.07	00:00:00.651000	284	Pediatrics
4	519	Eligible	Self-Service	Completed	2023-07-01	205.00	245.00	00:00:11.313000	368	Pediatrics

In [4]:
```
df.shape
```

Out[4]:
```
(85393, 10)
```

In [5]:
```
df.columns
```

Out[5]:
```
Index(['Client Identifier', 'Insurance Eligiblity', 'Intake Category',
       'Pre-visit Status', 'Visit Month', 'Amount Collected', 'Amount Owed',
       'Intake Interview Duration (Minutes)', 'Visit Count',
       'Client Specialty'],
      dtype='object')
```

In [6]:
```
df.isnull().sum()
```

Out[6]:
```
Client Identifier                          0
Insurance Eligiblity                       0
Intake Category                            0
Pre-visit Status                           0
Visit Month                                0
Amount Collected                           0
Amount Owed                                0
Intake Interview Duration (Minutes)    29647
Visit Count                                0
Client Specialty                           0
dtype: int64
```

Data Cleaning process, and fixing the proper data type

In [7]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 85393 entries, 0 to 85392
Data columns (total 10 columns):
 #   Column                               Non-Null Count  Dtype
---  ------                               --------------  -----
 0   Client Identifier                    85393 non-null  int64
 1   Insurance Eligiblity                 85393 non-null  object
 2   Intake Category                      85393 non-null  object
 3   Pre-visit Status                     85393 non-null  object
 4   Visit Month                          85393 non-null  datetime64[ns]
 5   Amount Collected                     85393 non-null  float64
 6   Amount Owed                          85393 non-null  float64
 7   Intake Interview Duration (Minutes)  55746 non-null  object
 8   Visit Count                          85393 non-null  int64
 9   Client Specialty                     85393 non-null  object
dtypes: datetime64[ns](1), float64(2), int64(2), object(5)
memory usage: 6.5+ MB
```

In [8]:
```
df.head()
```

Out[8]:

	Client Identifier	Insurance Eligiblity	Intake Category	Pre-visit Status	Visit Month	Amount Collected	Amount Owed	Intake Interview Duration (Minutes)	Visit Count	Client Specialty
0	519	Eligible	No check-in	Completed	2023-07-01	0.00	0.00	NaN	11	Pediatrics
1	519	Eligible	No check-in	Not Completed	2023-07-01	0.00	60.00	NaN	26	Pediatrics
2	519	Eligible	Self-Service	Not Completed	2023-07-01	0.00	0.00	00:11:27.037000	30	Pediatrics
3	519	Eligible	Staff	Not Completed	2023-07-01	70.12	251.07	00:00:00.651000	284	Pediatrics
4	519	Eligible	Self-Service	Completed	2023-07-01	205.00	245.00	00:00:11.313000	368	Pediatrics

In [9]:
```
df['Intake Interview Duration (Minutes)'] = pd.to_datetime(df['Intake Interview Duration (Minutes)'], format='%H:%M:%S.%f',errors='coerce')
```

In [10]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 85393 entries, 0 to 85392
Data columns (total 10 columns):
 #   Column                               Non-Null Count  Dtype
---  ------                               --------------  -----
 0   Client Identifier                    85393 non-null  int64
 1   Insurance Eligiblity                 85393 non-null  object
 2   Intake Category                      85393 non-null  object
 3   Pre-visit Status                     85393 non-null  object
 4   Visit Month                          85393 non-null  datetime64[ns]
 5   Amount Collected                     85393 non-null  float64
 6   Amount Owed                          85393 non-null  float64
 7   Intake Interview Duration (Minutes)  29406 non-null  datetime64[ns]
 8   Visit Count                          85393 non-null  int64
 9   Client Specialty                     85393 non-null  object
dtypes: datetime64[ns](2), float64(2), int64(2), object(4)
memory usage: 6.5+ MB
```

In [11]:
```
df.isnull().sum()
```

Out[11]:
```
Client Identifier                          0
Insurance Eligiblity                       0
Intake Category                            0
Pre-visit Status                           0
Visit Month                                0
Amount Collected                           0
Amount Owed                                0
Intake Interview Duration (Minutes)    55987
Visit Count                                0
Client Specialty                           0
dtype: int64
```

Null value setting up by fillna with no effect time

In [12]:
```
#df["Intake Interview Duration (Minutes)"]=df.fillna(df["Intake Interview Duration (Minutes)"].mean())

default_datetime = pd.to_datetime('00:00:00')
df["Intake Interview Duration (Minutes)"].fillna(default_datetime, inplace=True)
```

In [13]:
```
df.isnull().sum()
```

Out[13]:
```
Client Identifier                      0
Insurance Eligiblity                   0
Intake Category                        0
Pre-visit Status                       0
Visit Month                            0
Amount Collected                       0
Amount Owed                            0
Intake Interview Duration (Minutes)    0
Visit Count                            0
Client Specialty                       0
dtype: int64
```

In [14]:
```
df.head()
```

Out[14]:

	Client Identifier	Insurance Eligiblity	Intake Category	Pre-visit Status	Visit Month	Amount Collected	Amount Owed	Intake Interview Duration (Minutes)	Visit Count	Client Specialty
0	519	Eligible	No check-in	Completed	2023-07-01	0.00	0.00	2024-01-12 00:00:00.000	11	Pediatrics
1	519	Eligible	No check-in	Not Completed	2023-07-01	0.00	60.00	2024-01-12 00:00:00.000	26	Pediatrics
2	519	Eligible	Self-Service	Not Completed	2023-07-01	0.00	0.00	1900-01-01 00:11:27.037	30	Pediatrics
3	519	Eligible	Staff	Not Completed	2023-07-01	70.12	251.07	1900-01-01 00:00:00.651	284	Pediatrics
4	519	Eligible	Self-Service	Completed	2023-07-01	205.00	245.00	1900-01-01 00:00:11.313	368	Pediatrics

Descriptive Statistics analysis

In [15]:
```
df.describe()
```

Out[15]:

	Client Identifier	Visit Month	Amount Collected	Amount Owed	Intake Interview Duration (Minutes)	Visit Count
count	85393.000000	85393	8.539300e+04	8.539300e+04	85393	85393.000000
mean	1758.099118	2023-08-01 00:01:34.096705792	4.466395e+03	2.146418e+04	1981-04-27 02:43:34.961528832	385.359397
min	519.000000	2023-07-01 00:00:00	0.000000e+00	-6.054956e+04	1900-01-01 00:00:00.001000	1.000000
25%	1109.000000	2023-07-01 00:00:00	0.000000e+00	3.500000e+01	1900-01-01 00:08:23.256999936	5.000000
50%	1781.000000	2023-08-01 00:00:00	6.000000e+01	6.314500e+02	2024-01-12 00:00:00	25.000000
75%	2370.000000	2023-09-01 00:00:00	1.038000e+03	4.784770e+03	2024-01-12 00:00:00	142.000000
max	2951.000000	2023-09-01 00:00:00	2.363843e+06	3.647370e+07	2024-01-12 00:00:00	75121.000000
std	708.066972	NaN	3.210414e+04	2.691850e+05	NaN	2041.493190

In [16]:
```
df.describe(include="object")
#showing the uniques values which contain by columns
```

Out[16]:

	Insurance Eligiblity	Intake Category	Pre-visit Status	Client Specialty
count	85393	85393	85393	85393
unique	4	3	2	10
top	Eligible	Self-Service	Not Completed	Ear, Nose, Throat (ENT)
freq	26335	37461	55340	8652

In [17]:
```
df[["Insurance Eligiblity","Intake Category"]].value_counts()
```

Out[17]:
```
Insurance Eligiblity  Intake Category
Error                 Self-Service       11095
Eligible              Self-Service       10809
No Coverage/Self Pay  Self-Service       10125
Eligible              No check-in        10073
Error                 No check-in         9243
No Coverage/Self Pay  No check-in         7108
Error                 Staff               5564
Eligible              Staff               5453
Unsupported Payer     Self-Service        5432
No Coverage/Self Pay  Staff               4820
Unsupported Payer     No check-in         3092
                      Staff               2579
Name: count, dtype: int64
```

most intake category

In [18]:
```
df.groupby("Intake Category").count()
```

Out[18]:

	Client Identifier	Insurance Eligiblity	Pre-visit Status	Visit Month	Amount Collected	Amount Owed	Intake Interview Duration (Minutes)	Visit Count	Client Specialty
Intake Category									
No check-in	29516	29516	29516	29516	29516	29516	29516	29516	29516
Self-Service	37461	37461	37461	37461	37461	37461	37461	37461	37461
Staff	18416	18416	18416	18416	18416	18416	18416	18416	18416

In [19]:
```
int_cat=df.groupby("Intake Category").agg(
nocheckin=("Intake Category","count")
)
```

In [20]:
```
int_cat=int_cat.reset_index()
```

In [21]:
```
int_cat.sort_values(by="nocheckin")
```

Out[21]:

	Intake Category	nocheckin
2	Staff	18416
0	No check-in	29516
1	Self-Service	37461

In [22]:
```
sns.barplot(data=int_cat,x=int_cat["Intake Category"],y=int_cat["nocheckin"])
plt.grid()
plt.show()
```

In [23]:
```
df.head()
```

Out[23]:

	Client Identifier	Insurance Eligiblity	Intake Category	Pre-visit Status	Visit Month	Amount Collected	Amount Owed	Intake Interview Duration (Minutes)	Visit Count	Client Specialty
0	519	Eligible	No check-in	Completed	2023-07-01	0.00	0.00	2024-01-12 00:00:00.000	11	Pediatrics
1	519	Eligible	No check-in	Not Completed	2023-07-01	0.00	60.00	2024-01-12 00:00:00.000	26	Pediatrics
2	519	Eligible	Self-Service	Not Completed	2023-07-01	0.00	0.00	1900-01-01 00:11:27.037	30	Pediatrics
3	519	Eligible	Staff	Not Completed	2023-07-01	70.12	251.07	1900-01-01 00:00:00.651	284	Pediatrics
4	519	Eligible	Self-Service	Completed	2023-07-01	205.00	245.00	1900-01-01 00:00:11.313	368	Pediatrics

difference between Amount Collected and Amount Owed

In [24]:
```
df[["Amount Collected","Amount Owed"]].describe()
```

Out[24]:

	Amount Collected	Amount Owed
count	8.539300e+04	8.539300e+04
mean	4.466395e+03	2.146418e+04
std	3.210414e+04	2.691850e+05
min	0.000000e+00	-6.054956e+04
25%	0.000000e+00	3.500000e+01
50%	6.000000e+01	6.314500e+02
75%	1.038000e+03	4.784770e+03
max	2.363843e+06	3.647370e+07

In [25]:
```
amt_collected = df["Amount Collected"].sum()
amt_collected
```

Out[25]:
```
381398837.67810005
```

In [26]:
```
amt_owe = df["Amount Owed"].sum()
amt_owe
```

Out[26]:
```
1832890377.7695003
```

In [27]:
```
df["Amount Owed"].sum() - df["Amount Collected"].sum()
```

Out[27]:
```
1451491540.0914001
```

In [28]:
```
df.head()
```

Out[28]:

	Client Identifier	Insurance Eligiblity	Intake Category	Pre-visit Status	Visit Month	Amount Collected	Amount Owed	Intake Interview Duration (Minutes)	Visit Count	Client Specialty
0	519	Eligible	No check-in	Completed	2023-07-01	0.00	0.00	2024-01-12 00:00:00.000	11	Pediatrics
1	519	Eligible	No check-in	Not Completed	2023-07-01	0.00	60.00	2024-01-12 00:00:00.000	26	Pediatrics
2	519	Eligible	Self-Service	Not Completed	2023-07-01	0.00	0.00	1900-01-01 00:11:27.037	30	Pediatrics
3	519	Eligible	Staff	Not Completed	2023-07-01	70.12	251.07	1900-01-01 00:00:00.651	284	Pediatrics
4	519	Eligible	Self-Service	Completed	2023-07-01	205.00	245.00	1900-01-01 00:00:11.313	368	Pediatrics

Most visited client by month

In [29]:
```
df.groupby("Visit Month")["Visit Count"].count()
```

Out[29]:
```
Visit Month
2023-07-01    28453
2023-08-01    28484
2023-09-01    28456
Name: Visit Count, dtype: int64
```

In [30]:
```
visit=df.groupby("Visit Month").agg(
    most_visits=("Visit Count","count")
).reset_index()
visit.sort_values(by="most_visits",ascending=False)
```

Out[30]:

	Visit Month	most_visits
1	2023-08-01	28484
2	2023-09-01	28456
0	2023-07-01	28453

In [31]:
```
df.head(3)
```

Out[31]:

	Client Identifier	Insurance Eligiblity	Intake Category	Pre-visit Status	Visit Month	Amount Collected	Amount Owed	Intake Interview Duration (Minutes)	Visit Count	Client Specialty
0	519	Eligible	No check-in	Completed	2023-07-01	0.0	0.0	2024-01-12 00:00:00.000	11	Pediatrics
1	519	Eligible	No check-in	Not Completed	2023-07-01	0.0	60.0	2024-01-12 00:00:00.000	26	Pediatrics
2	519	Eligible	Self-Service	Not Completed	2023-07-01	0.0	0.0	1900-01-01 00:11:27.037	30	Pediatrics

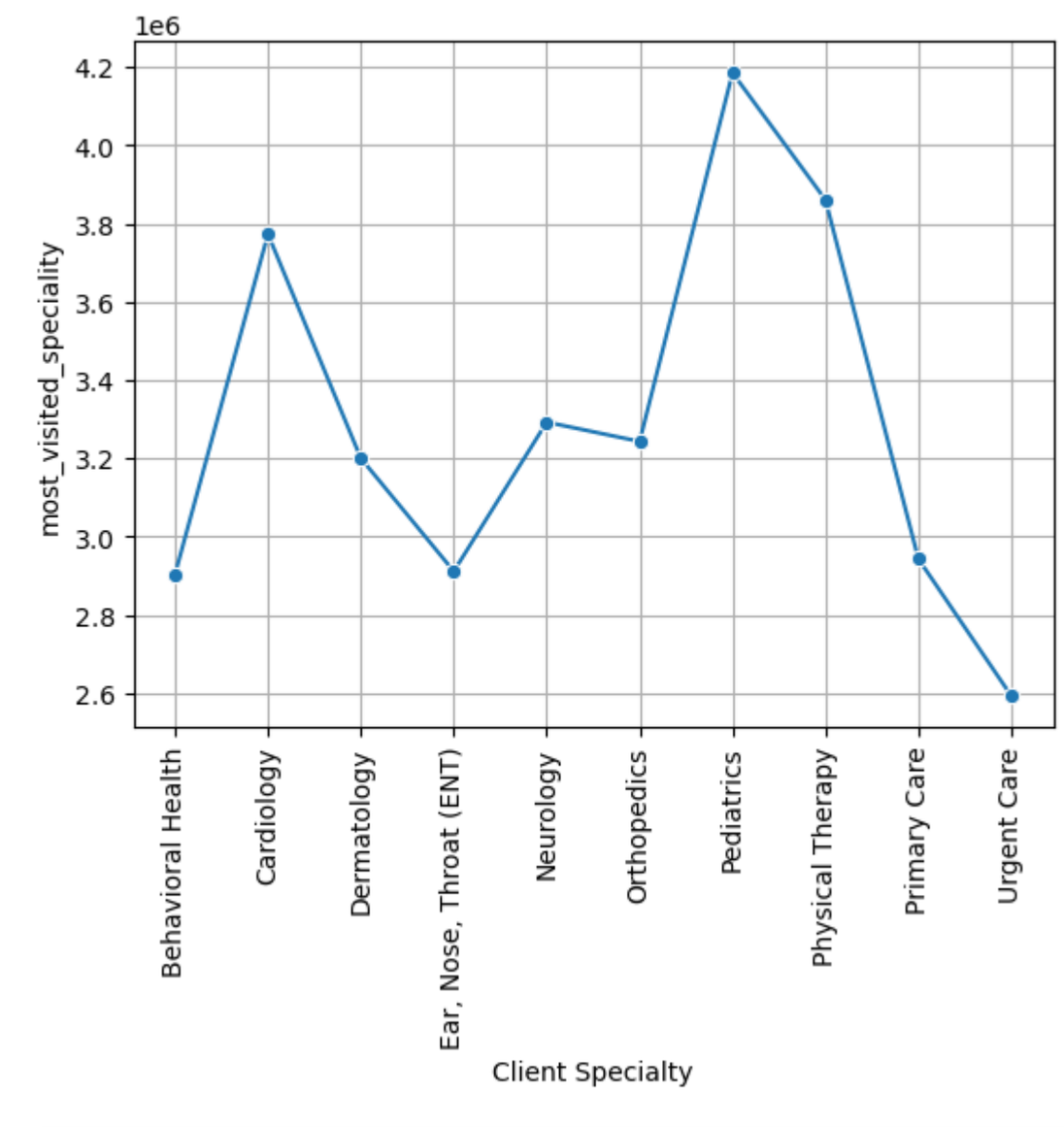
Most Visited speciality

In [32]:
```
splty=df.groupby("Client Specialty").agg(
    most_visited_speciality=("Visit Count","sum")
).reset_index()
splty
splty.sort_values(by="most_visited_speciality",ascending=False)
```

Out[32]:

	Client Specialty	most_visited_speciality
6	Pediatrics	4185319
7	Physical Therapy	3860867
1	Cardiology	3773051
4	Neurology	3292592
5	Orthopedics	3243363
2	Dermatology	3200161
8	Primary Care	2943208
3	Ear, Nose, Throat (ENT)	2911483
0	Behavioral Health	2903920
9	Urgent Care	2593031

In [33]:
```
sns.lineplot(data=splty,x=splty["Client Specialty"],y=splty["most_visited_speciality"],marker="o")
plt.xticks(rotation=90)
plt.grid()
plt.show()
```



In [34]:
```
df.head(2)
```

Out[34]:

	Client Identifier	Insurance Eligiblity	Intake Category	Pre-visit Status	Visit Month	Amount Collected	Amount Owed	Intake Interview Duration (Minutes)	Visit Count	Client Specialty
0	519	Eligible	No check-in	Completed	2023-07-01	0.0	0.0	2024-01-12	11	Pediatrics
1	519	Eligible	No check-in	Not Completed	2023-07-01	0.0	60.0	2024-01-12	26	Pediatrics

pre-visit status

In [35]:
```
df.groupby("Pre-visit Status")["Client Identifier"].count()
```

Out[35]:
```
Pre-visit Status
Completed        30053
Not Completed    55340
Name: Client Identifier, dtype: int64
```

Insurance Eligibility status

In [36]:
```
df.groupby("Insurance Eligiblity")["Client Identifier"].count().reset_index().sort_values(by="Client Identifier",ascending=False)
```

Out[36]:

	Insurance Eligiblity	Client Identifier
0	Eligible	26335
1	Error	25902
2	No Coverage/Self Pay	22053
3	Unsupported Payer	11103

In [37]:
```
df.head()
```

Out[37]:

	Client Identifier	Insurance Eligiblity	Intake Category	Pre-visit Status	Visit Month	Amount Collected	Amount Owed	Intake Interview Duration (Minutes)	Visit Count	Client Specialty
0	519	Eligible	No check-in	Completed	2023-07-01	0.00	0.00	2024-01-12 00:00:00.000	11	Pediatrics
1	519	Eligible	No check-in	Not Completed	2023-07-01	0.00	60.00	2024-01-12 00:00:00.000	26	Pediatrics
2	519	Eligible	Self-Service	Not Completed	2023-07-01	0.00	0.00	1900-01-01 00:11:27.037	30	Pediatrics
3	519	Eligible	Staff	Not Completed	2023-07-01	70.12	251.07	1900-01-01 00:00:00.651	284	Pediatrics
4	519	Eligible	Self-Service	Completed	2023-07-01	205.00	245.00	1900-01-01 00:00:11.313	368	Pediatrics

In [38]:
```
df["Intake Interview Duration (Minutes)"].describe()
```

Out[38]:
```
count                            85393
mean     1981-04-27 02:43:34.961528832
min         1900-01-01 00:00:00.001000
25%      1900-01-01 00:08:23.256999936
50%                2024-01-12 00:00:00
75%                2024-01-12 00:00:00
max                2024-01-12 00:00:00
Name: Intake Interview Duration (Minutes), dtype: object
```

In [39]:
```
df.loc[df["Visit Month"]<="2023-07-31"]
```

Out[39]:

	Client Identifier	Insurance Eligiblity	Intake Category	Pre-visit Status	Visit Month	Amount Collected	Amount Owed	Intake Interview Duration (Minutes)	Visit Count	Client Specialty
0	519	Eligible	No check-in	Completed	2023-07-01	0.00	0.00	2024-01-12 00:00:00.000	11	Pediatrics
1	519	Eligible	No check-in	Not Completed	2023-07-01	0.00	60.00	2024-01-12 00:00:00.000	26	Pediatrics
2	519	Eligible	Self-Service	Not Completed	2023-07-01	0.00	0.00	1900-01-01 00:11:27.037	30	Pediatrics
3	519	Eligible	Staff	Not Completed	2023-07-01	70.12	251.07	1900-01-01 00:00:00.651	284	Pediatrics
4	519	Eligible	Self-Service	Completed	2023-07-01	205.00	245.00	1900-01-01 00:00:11.313	368	Pediatrics
...	...	...	...	...	...	...	...	...	...	...
85380	2951	Error	No check-in	Not Completed	2023-07-01	40.00	6350.26	2024-01-12 00:00:00.000	591	Primary Care
85381	2951	No Coverage/Self Pay	Staff	Not Completed	2023-07-01	217.00	217.00	2024-01-12 00:00:00.000	2	Primary Care
85382	2951	No Coverage/Self Pay	No check-in	Not Completed	2023-07-01	0.00	145.68	2024-01-12 00:00:00.000	16	Primary Care
85383	2951	Unsupported Payer	Staff	Not Completed	2023-07-01	49.42	49.42	2024-01-12 00:00:00.000	2	Primary Care
85384	2951	Unsupported Payer	No check-in	Not Completed	2023-07-01	0.00	627.81	2024-01-12 00:00:00.000	18	Primary Care

28453 rows × 10 columns

In [40]:
```
df.groupby("Visit Month")["Client Identifier"].mean()
```

Out[40]:
```
Visit Month
2023-07-01    1756.569290
2023-08-01    1758.281597
2023-09-01    1759.446127
Name: Client Identifier, dtype: float64
```

outcome:

we can create all three month separate clean files by using to_excefunction

all important analysis i have carried out.

The data represents a time series of monthly visit amounts,

showing a slight increase from 1756.57 in July 2023 to 1759.45 in September 2023.amout owe is more, and most visited speciality is paediatrics with insurance eligible clients

thank you !!!