DATA ANALYSIS PROJECT ON Air_Traffic_Passenger DATA-SET

Importing Libraries

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

Loading the dataset

In [2]:
```
df = pd.read_csv(r'Air_Traffic_Passenger_Statistics.csv')
pd.set_option('display.max_columns', None)
df
```

Out[2]:

	Activity Period	Operating Airline	Operating Airline IATA Code	Published Airline	Published Airline IATA Code	GEO Summary	GEO Region	Activity Type Code	F Cate C
0	200507	ATA Airlines	TZ	ATA Airlines	TZ	Domestic	US	Deplaned	Low
1	200507	ATA Airlines	TZ	ATA Airlines	TZ	Domestic	US	Enplaned	Low
2	200507	ATA Airlines	TZ	ATA Airlines	TZ	Domestic	US	Thru / Transit	Low
3	200507	Air Canada	AC	Air Canada	AC	International	Canada	Deplaned	C
4	200507	Air Canada	AC	Air Canada	AC	International	Canada	Enplaned	C
...	...	...	...	...	...	...	...	...	
15002	201603	Virgin America	VX	Virgin America	VX	Domestic	US	Enplaned	Low
15003	201603	Virgin America	VX	Virgin America	VX	International	Mexico	Deplaned	Low
15004	201603	Virgin America	VX	Virgin America	VX	International	Mexico	Enplaned	Low
15005	201603	Virgin Atlantic	VS	Virgin Atlantic	VS	International	Europe	Deplaned	C
15006	201603	Virgin Atlantic	VS	Virgin Atlantic	VS	International	Europe	Enplaned	C

15007 rows × 16 columns

INTRODUCTION OF PROJECT

- TARGET COLUMN == GEO Summary
- GEO Summary "International" :- International Flights.
- GEO Summary "Domastic" :-Domastic Flights.

PROBLEM STATEMENT

This project aims to uncover the factors contributing to GEO Summary. Through analysis of various dataset features, we seek to understand their impact International and Domastic Flights on Airline.

Exploratory Data Analysis and Data Cleaning

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

Out[3]:

	Activity Period	Operating Airline	Operating Airline IATA Code	Published Airline	Published Airline IATA Code	GEO Summary	GEO Region	Activity Type Code	Price Category Code
0	200507	ATA Airlines	TZ	ATA Airlines	TZ	Domestic	US	Deplaned	Low Fare
1	200507	ATA Airlines	TZ	ATA Airlines	TZ	Domestic	US	Enplaned	Low Fare
2	200507	ATA Airlines	TZ	ATA Airlines	TZ	Domestic	US	Thru / Transit	Low Fare
3	200507	Air Canada	AC	Air Canada	AC	International	Canada	Deplaned	Other
4	200507	Air Canada	AC	Air Canada	AC	International	Canada	Enplaned	Other

In [4]:
```
df.tail()
```

Out[4]:

	Activity Period	Operating Airline	Operating Airline IATA Code	Published Airline	Published Airline IATA Code	GEO Summary	GEO Region	Activity Type Code	P Categ C
15002	201603	Virgin America	VX	Virgin America	VX	Domestic	US	Enplaned	Low
15003	201603	Virgin America	VX	Virgin America	VX	International	Mexico	Deplaned	Low
15004	201603	Virgin America	VX	Virgin America	VX	International	Mexico	Enplaned	Low
15005	201603	Virgin Atlantic	VS	Virgin Atlantic	VS	International	Europe	Deplaned	C
15006	201603	Virgin Atlantic	VS	Virgin Atlantic	VS	International	Europe	Enplaned	C

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

Out[5]:
```
(15007, 16)
```

In [6]:
```
df.columns
```

Out[6]:
```
Index(['Activity Period', 'Operating Airline', 'Operating Airline IATA Code',
       'Published Airline', 'Published Airline IATA Code', 'GEO Summary',
       'GEO Region', 'Activity Type Code', 'Price Category Code', 'Terminal',
       'Boarding Area', 'Passenger Count', 'Adjusted Activity Type Code',
       'Adjusted Passenger Count', 'Year', 'Month'],
      dtype='object')
```

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15007 entries, 0 to 15006
Data columns (total 16 columns):
 #   Column                        Non-Null Count  Dtype 
---  ------                        --------------  ----- 
 0   Activity Period               15007 non-null  int64 
 1   Operating Airline             15007 non-null  object
 2   Operating Airline IATA Code   14953 non-null  object
 3   Published Airline             15007 non-null  object
 4   Published Airline IATA Code   14953 non-null  object
 5   GEO Summary                   15007 non-null  object
 6   GEO Region                    15007 non-null  object
 7   Activity Type Code            15007 non-null  object
 8   Price Category Code           15007 non-null  object
 9   Terminal                      15007 non-null  object
 10  Boarding Area                 15007 non-null  object
 11  Passenger Count               15007 non-null  int64 
 12  Adjusted Activity Type Code   15007 non-null  object
 13  Adjusted Passenger Count      15007 non-null  int64 
 14  Year                          15007 non-null  int64 
 15  Month                         15007 non-null  object
dtypes: int64(4), object(12)
memory usage: 1.8+ MB
```

In [8]:
```
for col in df.describe(include='object').columns: # check all categorical co
 print(col)
 print(df[col].unique())
 print('-'*50)
```

```
Operating Airline
['ATA Airlines' 'Air Canada ' 'Air China' 'Air France' 'Air New Zealand'
 'AirTran Airways' 'Alaska Airlines' 'All Nippon Airways'
 'American Airlines' 'American Eagle Airlines' 'Asiana Airlines'
 'Atlantic Southeast Airlines' 'BelAir Airlines' 'British Airways'
 'Cathay Pacific' 'China Airlines' 'Delta Air Lines' 'EVA Airways'
 'Frontier Airlines' 'Hawaiian Airlines' 'Horizon Air ' 'Icelandair'
 'Independence Air' 'Japan Airlines' 'KLM Royal Dutch Airlines'
 'Korean Air Lines' 'Lufthansa German Airlines' 'Mesa Airlines'
 'Mexicana Airlines' 'Midwest Airlines' 'Northwest Airlines'
 'Philippine Airlines' 'Singapore Airlines' 'SkyWest Airlines'
 'Sun Country Airlines' 'TACA' 'US Airways' 'United Airlines'
 'United Airlines - Pre 07/01/2013' 'Virgin Atlantic' 'WestJet Airlines'
 'Boeing Company' 'Miami Air International' 'Air Canada Jazz'
 'Qantas Airways' 'Ameriflight' 'Spirit Airlines' 'Xtra Airways'
 'Evergreen International Airlines' 'Aeromexico' 'JetBlue Airways '
 'ExpressJet Airlines' 'Southwest Airlines' 'Virgin America' 'Aer Lingus'
 'Allegiant Air' 'Jet Airways' 'Emirates ' 'Mesaba Airlines'
 'World Airways' 'Air Berlin' 'Republic Airlines' 'Servisair'
 'Pacific Aviation' 'Swiss International' 'LAN Peru' 'Swissport USA'
 'XL Airways France' 'China Eastern' 'SAS Airlines' 'Atlas Air, Inc'
 'Compass Airlines' 'Etihad Airways' 'China Southern' 'Turkish Airlines'
 'COPA Airlines, Inc.' 'Air India Limited']
--------------------------------------------------
Operating Airline IATA Code
['TZ' 'AC' 'CA' 'AF' 'NZ' 'FL' 'AS' 'NH' 'AA' 'MQ' 'OZ' 'EV' '4T' 'BA'
 'CX' 'CI' 'DL' 'BR' 'F9' 'HA' 'QX' 'FI' 'DH' 'JL' 'KL' 'KE' 'LH' 'YV'
 'MX' 'YX' 'NW' 'PR' 'SQ' 'OO' 'SY' 'TA' 'US ' 'UA' 'VS' 'WS' nan 'GL'
 'QK' 'QF' 'A8' 'NK' 'XP' 'EZ' 'AM' 'B6' 'XE' 'WN' 'VX' 'EI' 'G4' '9W'
 'BBB' 'EK' 'XJ' 'WO' 'AB' 'RW' 'LX' 'LP' 'SE' 'MU' 'SK' '5Y' 'CP' 'EY'
 'CZ' 'TK' 'CM' 'AI']
--------------------------------------------------
Published Airline
['ATA Airlines' 'Air Canada ' 'Air China' 'Air France' 'Air New Zealand'
 'AirTran Airways' 'Alaska Airlines' 'All Nippon Airways'
 'American Airlines' 'Asiana Airlines' 'Delta Air Lines' 'BelAir Airlines'
 'British Airways' 'Cathay Pacific' 'China Airlines' 'EVA Airways'
 'Frontier Airlines' 'Hawaiian Airlines' 'Icelandair' 'Independence Air'
 'Japan Airlines' 'KLM Royal Dutch Airlines' 'Korean Air Lines'
 'Lufthansa German Airlines' 'US Airways' 'Mexicana Airlines'
 'Midwest Airlines' 'Northwest Airlines' 'Philippine Airlines'
 'Singapore Airlines' 'United Airlines - Pre 07/01/2013'
 'Sun Country Airlines' 'TACA' 'United Airlines' 'Virgin Atlantic'
 'WestJet Airlines' 'Boeing Company' 'Miami Air International'
 'Qantas Airways' 'Ameriflight' 'Spirit Airlines' 'Xtra Airways'
 'Evergreen International Airlines' 'Aeromexico' 'JetBlue Airways '
 'Southwest Airlines' 'Virgin America' 'Aer Lingus' 'Allegiant Air'
 'Jet Airways' 'Emirates ' 'World Airways' 'Air Berlin'
 'Republic Airlines' 'Servisair' 'Pacific Aviation' 'Swiss International'
 'LAN Peru' 'Swissport USA' 'XL Airways France' 'China Eastern'
 'SAS Airlines' 'Atlas Air, Inc' 'Etihad Airways' 'China Southern'
 'Turkish Airlines' 'COPA Airlines, Inc.' 'Air India Limited']
--------------------------------------------------
Published Airline IATA Code
['TZ' 'AC' 'CA' 'AF' 'NZ' 'FL' 'AS' 'NH' 'AA' 'OZ' 'DL' '4T' 'BA' 'CX'
 'CI' 'BR' 'F9' 'HA' 'FI' 'DH' 'JL' 'KL' 'KE' 'LH' 'US ' 'MX' 'YX' 'NW'
 'PR' 'SQ' 'UA' 'SY' 'TA' 'VS' 'WS' nan 'GL' 'QF' 'A8' 'NK' 'XP' 'EZ' 'AM'
```

```
 'B6' 'WN' 'VX' 'EI' 'G4' '9W' 'BBB' 'EK' 'WO' 'AB' 'RW' 'LX' 'LP' 'SE'
 'MU' 'SK' '5Y' 'EY' 'CZ' 'TK' 'CM' 'AI']
--------------------------------------------------
GEO Summary
['Domestic' 'International']
--------------------------------------------------
GEO Region
['US' 'Canada' 'Asia' 'Europe' 'Australia / Oceania' 'Mexico'
 'Central America' 'Middle East' 'South America']
--------------------------------------------------
Activity Type Code
['Deplaned' 'Enplaned' 'Thru / Transit']
--------------------------------------------------
Price Category Code
['Low Fare' 'Other']
--------------------------------------------------
Terminal
['Terminal 1' 'International' 'Terminal 3' 'Other' 'Terminal 2']
--------------------------------------------------
Boarding Area
['B' 'G' 'A' 'E' 'C' 'F' 'Other' 'D']
--------------------------------------------------
Adjusted Activity Type Code
['Deplaned' 'Enplaned' 'Thru / Transit * 2']
--------------------------------------------------
Month
['July' 'August' 'September' 'October' 'November' 'April' 'December'
 'January' 'February' 'March' 'May' 'June']
--------------------------------------------------
```

In [9]:
```
pd.isnull(df)
```

Out[9]:

	Activity Period	Operating Airline	Operating Airline IATA Code	Published Airline	Published Airline IATA Code	GEO Summary	GEO Region	Activity Type Code	Price Category Code
0	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	..
15002	False	False	False	False	False	False	False	False	False
15003	False	False	False	False	False	False	False	False	False
15004	False	False	False	False	False	False	False	False	False
15005	False	False	False	False	False	False	False	False	False
15006	False	False	False	False	False	False	False	False	False

15007 rows × 16 columns

In [10]:
```
pd.isnull(df).sum()
```

Out[10]:
```
Activity Period                 0
Operating Airline               0
Operating Airline IATA Code    54
Published Airline               0
Published Airline IATA Code    54
GEO Summary                     0
GEO Region                      0
Activity Type Code              0
Price Category Code             0
Terminal                        0
Boarding Area                   0
Passenger Count                 0
Adjusted Activity Type Code     0
Adjusted Passenger Count        0
Year                            0
Month                           0
dtype: int64
```

In [11]:
```
df.describe()
```

Out[11]:

	Activity Period	Passenger Count	Adjusted Passenger Count	Year
count	15007.000000	15007.000000	15007.000000	15007.000000
mean	201045.073366	29240.521090	29331.917105	2010.385220
std	313.336196	58319.509284	58284.182219	3.137589
min	200507.000000	1.000000	1.000000	2005.000000
25%	200803.000000	5373.500000	5495.500000	2008.000000
50%	201011.000000	9210.000000	9354.000000	2010.000000
75%	201308.000000	21158.500000	21182.000000	2013.000000
max	201603.000000	659837.000000	659837.000000	2016.000000

In [12]:
```
df.describe(include='object')
```

Out[12]:

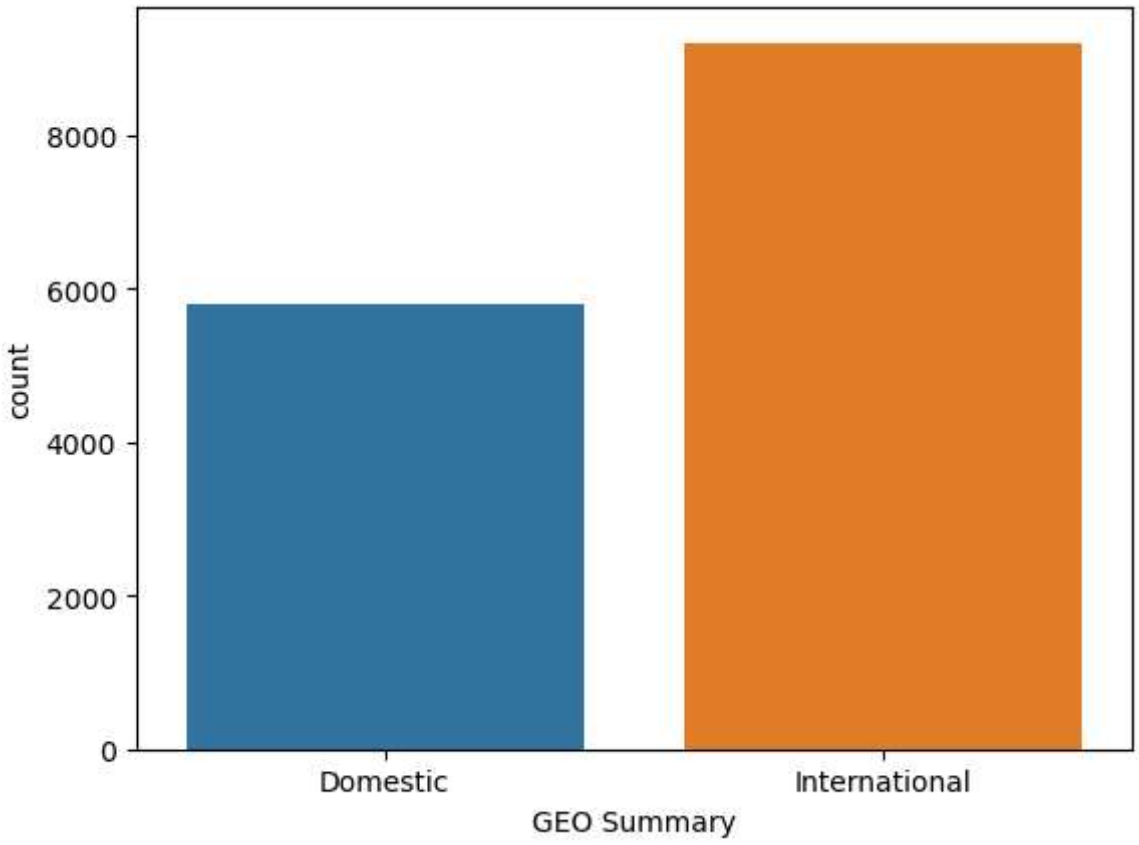
	Operating Airline	Operating Airline IATA Code	Published Airline	Published Airline IATA Code	GEO Summary	GEO Region	Activity Type Code	Price Category Code
count	15007	14953	15007	14953	15007	15007	15007	15007
unique	77	73	68	64	2	9	3	2
top	United Airlines - Pre 07/01/2013	UA	United Airlines - Pre 07/01/2013	UA	International	US	Deplaned	Other
freq	2154	3046	2645	3752	9210	5797	7071	13087

In [ ]:

Data Analysis and Visulaization

ANALYSIS ON CATEGORICAL COLUMN WRT TARGET COLUMN(GEO Summary)

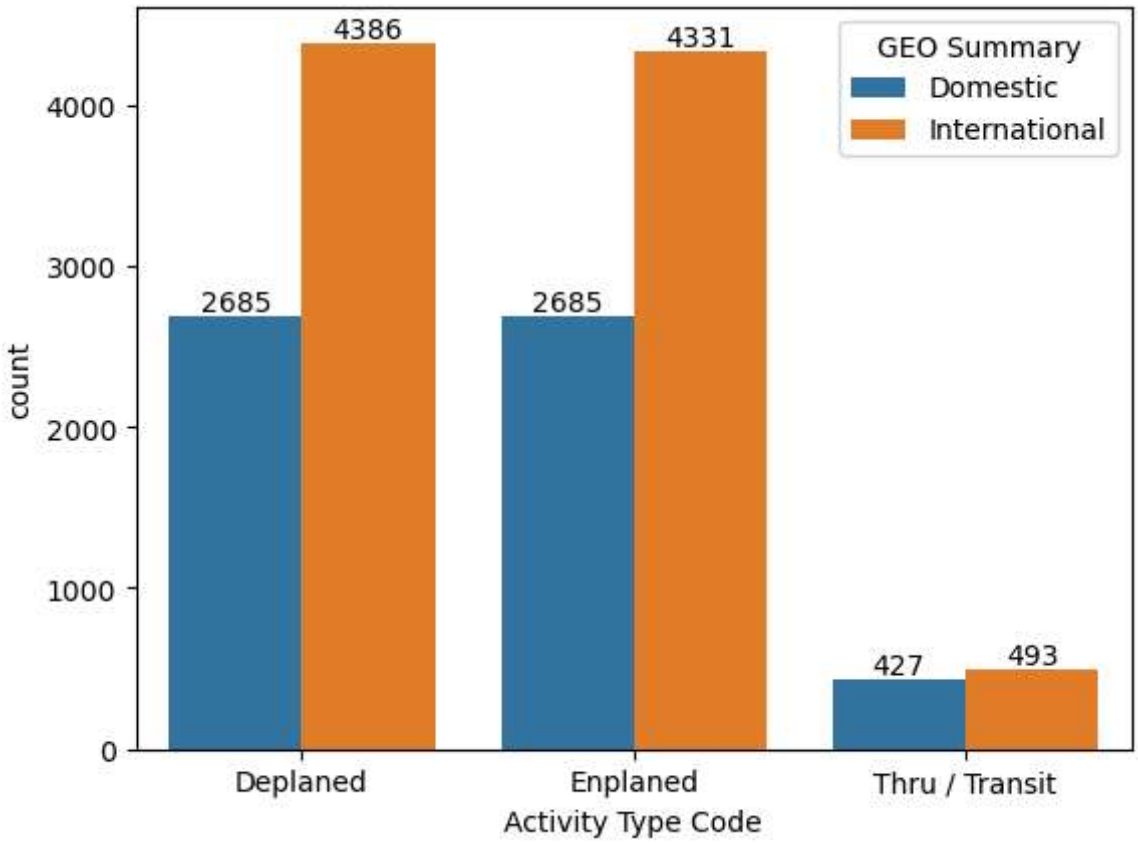
In [13]:
```
ax=sns.countplot(data=df, x='GEO Summary')
plt.show()
```



observation:- Data of GEO Summary say that it has more number of "Interntional" flights and less number of "Domestic" Flights

IMPACT OF BUSINESS TRAVEL ON ATTRITION

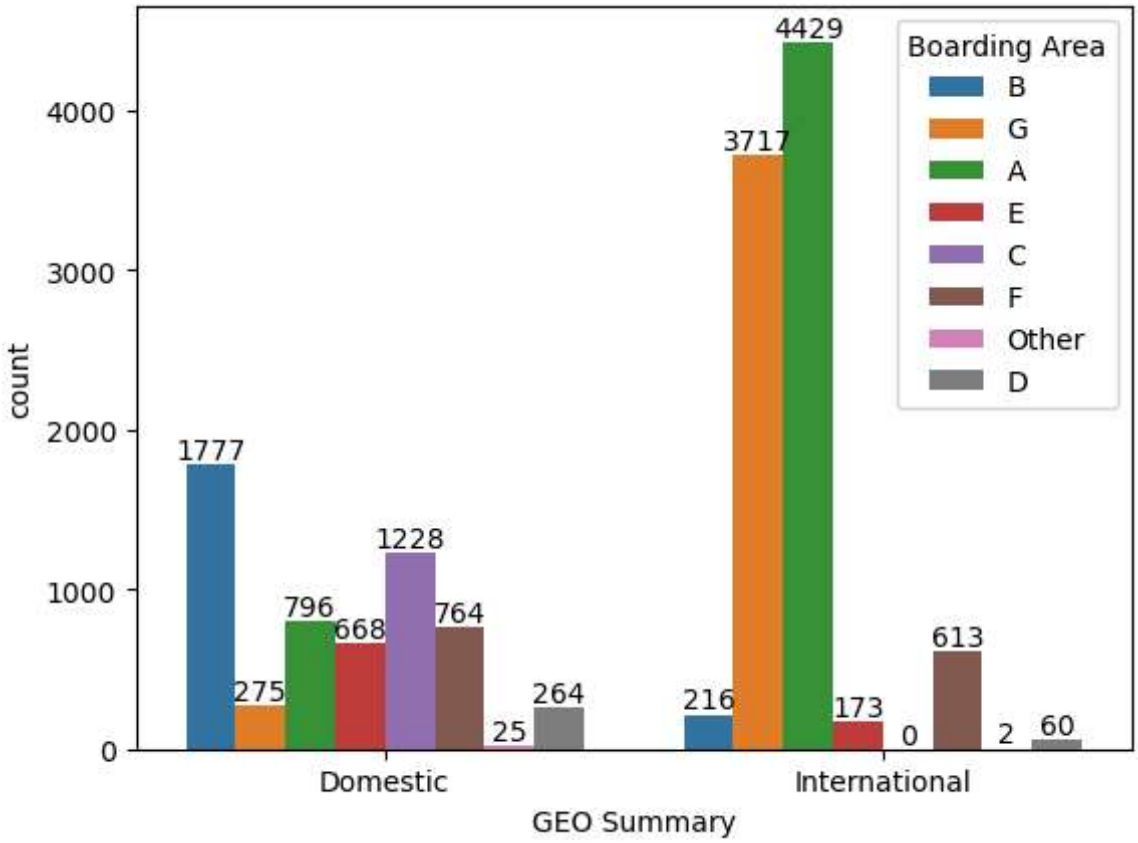
In [14]:
```
ax=sns.countplot(data=df, x='Activity Type Code', hue='GEO Summary')
for bars in ax.containers:
 ax.bar_label(bars)
plt.show()
```



Observation:- -Graph tells us that Airline has more count or more number of Deplaned passenger arriving at the airport on a International flight -There are more Enplaned passenger who depart from airport on a international flight -Transit passenger have least count as well as least GEO Summary

IMPACT OF EDUCATION FIELD ON ATTRITION

In [23]:
```
ax2=sns.countplot(data=df, x='GEO Summary', hue='Boarding Area')
for bars in ax2.containers:
 ax2.bar_label(bars)
plt.show()
```

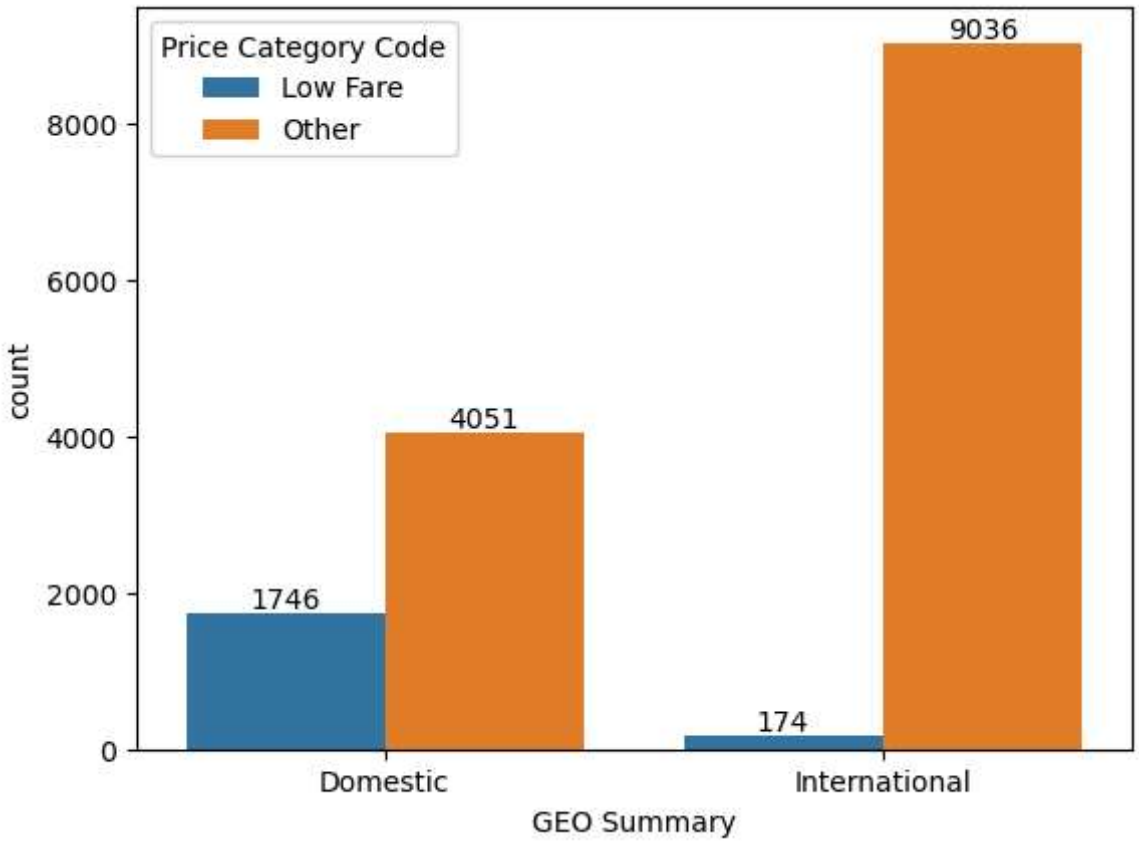


Observation:-

- In this graph shows that more passengers are Boarding from area A and area G have taken international flights
- Nearly 3025 number of passengers are there who are from B Boarding Area has take Domastic Flights
- As we conclude from analysis of Boarding Area and GEO summary, here also D boarding area having least count pf passengers geo summary

IMPACT OF GENDER ON ATTRITION

In [24]:
```
ax3=sns.countplot(data=df, x='GEO Summary', hue='Price Category Code')
for bars in ax3.containers:
 ax3.bar_label(bars)
plt.show()
```



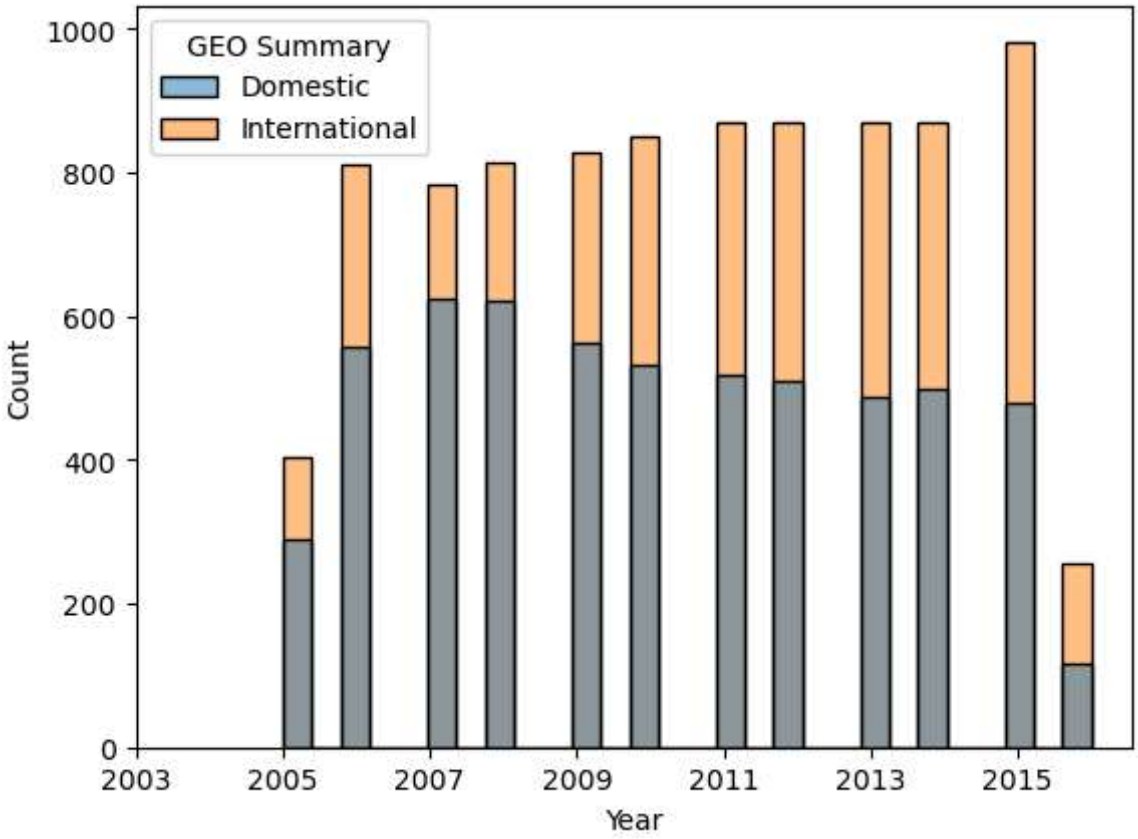
observation:-

- other fare paying passenger are more as compared to low fare paying passenger
- other fare paying passenger are more likely to get domastic flights rather than low fare paying passenger

ANALYSIS ON CONTINUOS DATA W.R.T TARGET COLUMN

IMPACT OF ON ATTRITION

In [16]:
```
sns.histplot(data=df, x='Year', hue='GEO Summary')
plt.xticks(range(2003,2017,2))
plt.show()
```

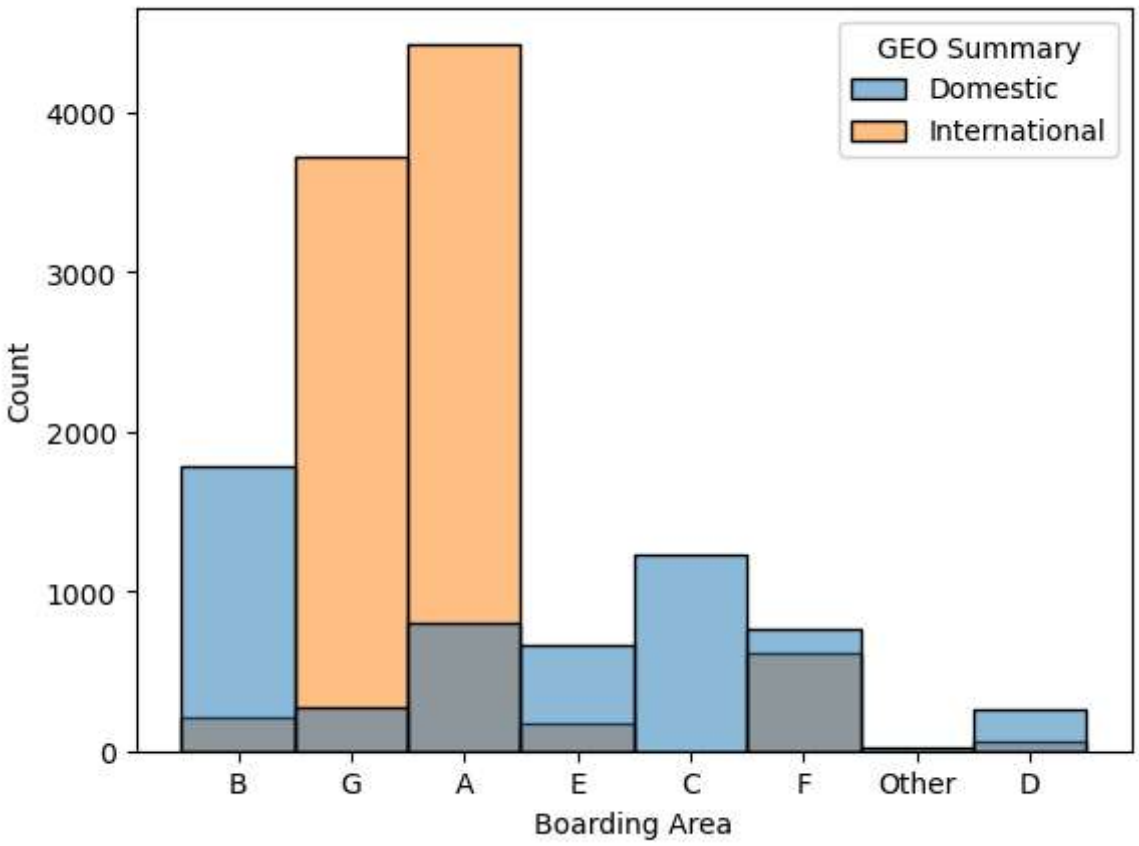


observation:-

- In the histplot shows Passengers from year 2005 to 2013 are more likely to take Domastic flights.
- After year 2013, the distribution tells us that as the year increasing counts of passenger taking intenatioanl flights are also increasing.

IMPACT OF Boarding Area AND GEO Summary

In [17]:
```
sns.histplot(data=df, x='Boarding Area', hue='GEO Summary')
plt.show()
```

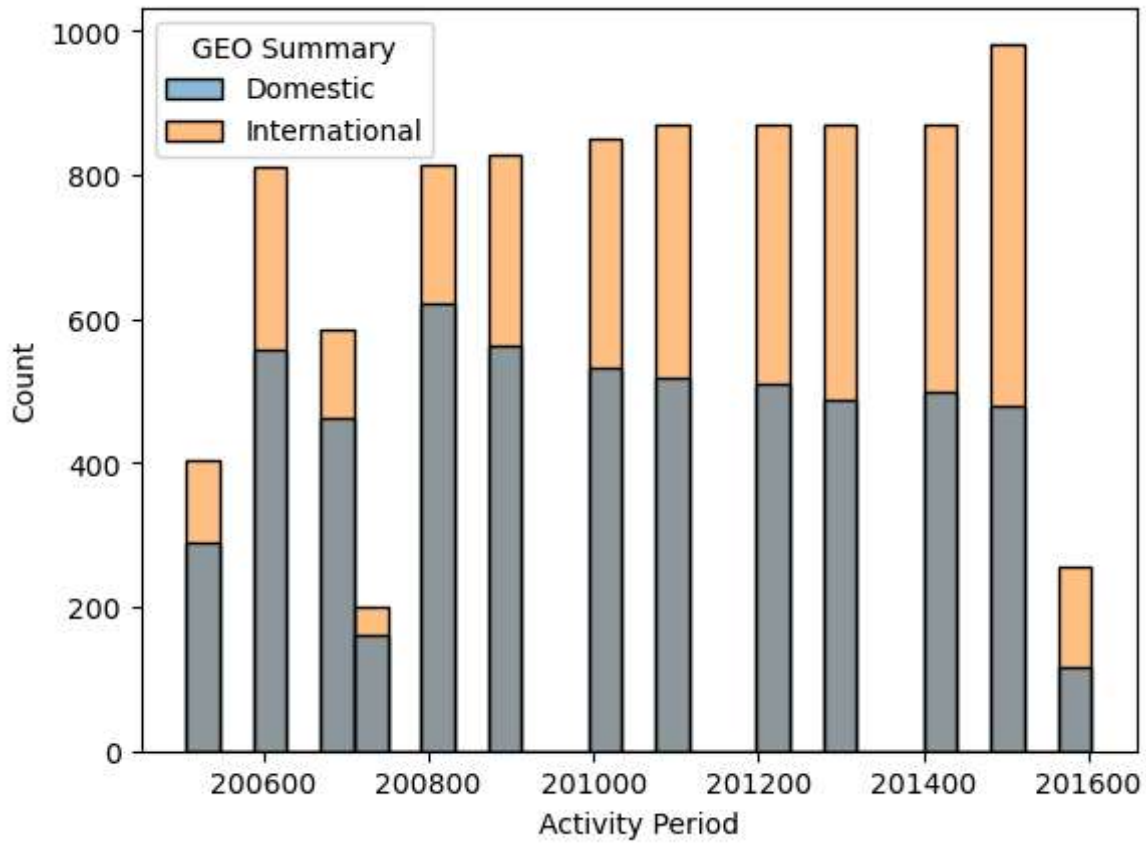


Observation:

- In this graph shows that more passengers are Boarding from area A and area G have taken international flights
- From boarding area E to D passengers are more likely to take Domastic flight

HOW Activity Period GIVES TRENDS W.R.T GEO Summary

In [18]:
```
sns.histplot(data=df, x='Activity Period', hue='GEO Summary')
plt.show()
```



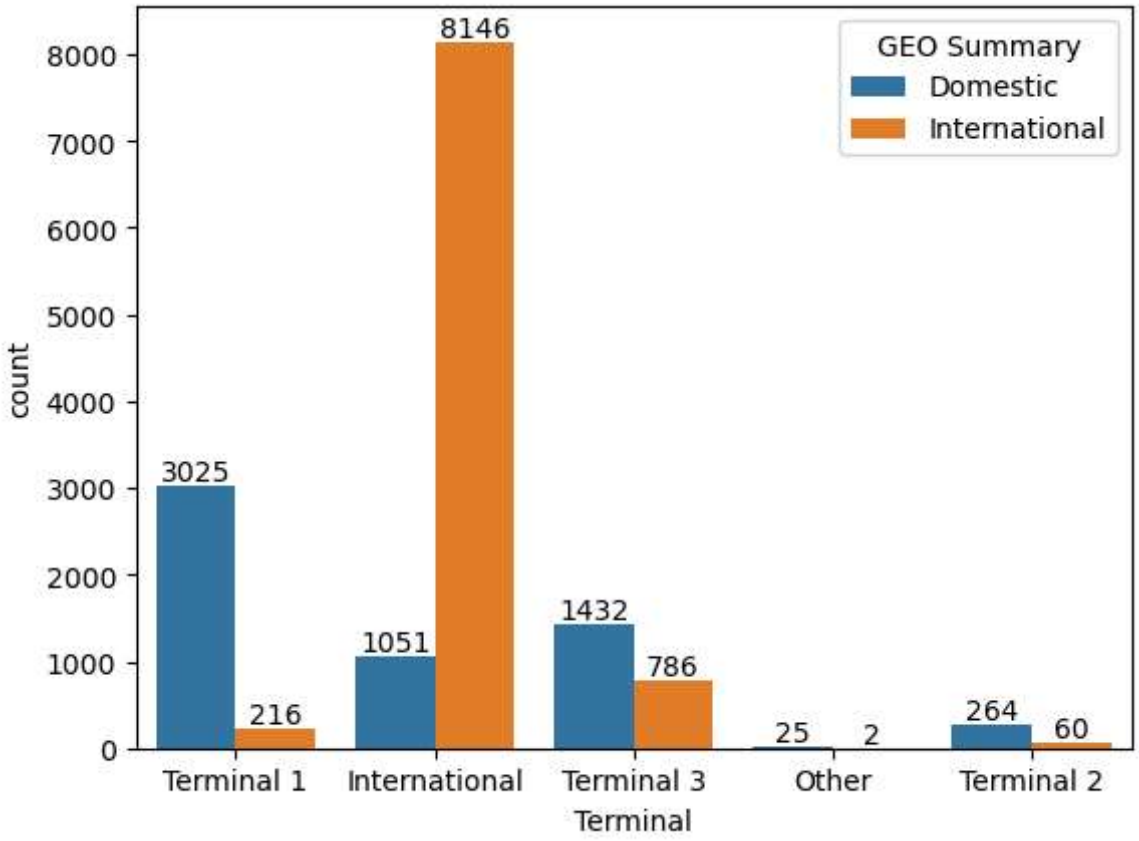
Observation:-

- In this histplot shows the maximum acitivity period of passenger is 201500 are more likely to take internatinal flight
- Activity period of passenger taking domastic flight is more as compare to passenger taking international flight

ANALYSIS ON DISCRETE DATA W.R.TTARGET COLUMN

IMPACT OF Terminal ON GEO Summary

In [26]:
```
ax5=sns.countplot(data=df, x='Terminal', hue='GEO Summary')
for bars in ax5.containers:
 ax5.bar_label(bars)
plt.show()
```

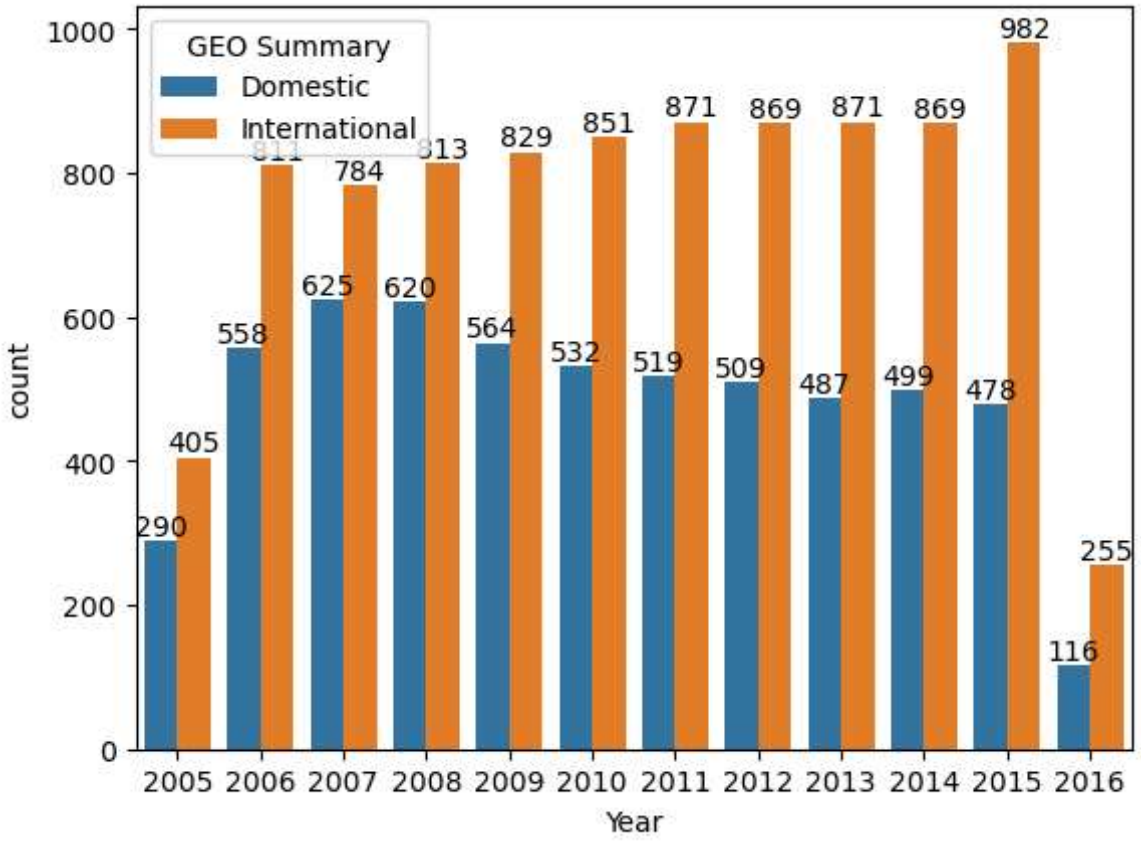


Observation:-

- There are more number of Passengers are takinginternational flights terminated internationally
- passengers taking domastic flights are very less atv other teminals except terminal 1.

IMPACT OF Year ON GEO Summary

In [30]:
```
ax6=sns.countplot(data=df, x='Year', hue='GEO Summary')
for bars in ax6.containers:
 ax6.bar_label(bars)
plt.show()
```



Observation:-

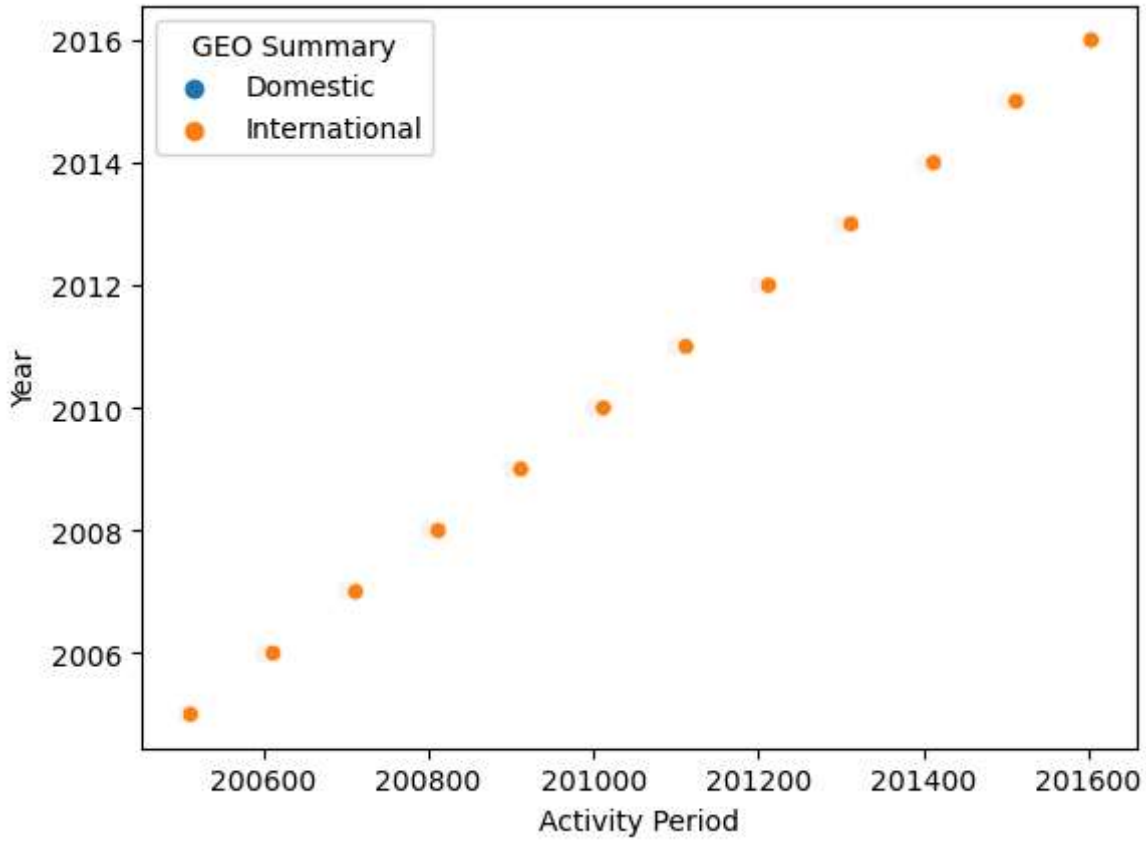
- the highest rate of passernger taking internatrional flight isin 2015
- from year 2006 as the year incresses count of passenger taking donmastic flight decreases

In [ ]:

Explore Relationships

Relationship 2 major factors of GEO Summary activity period of passenger and years

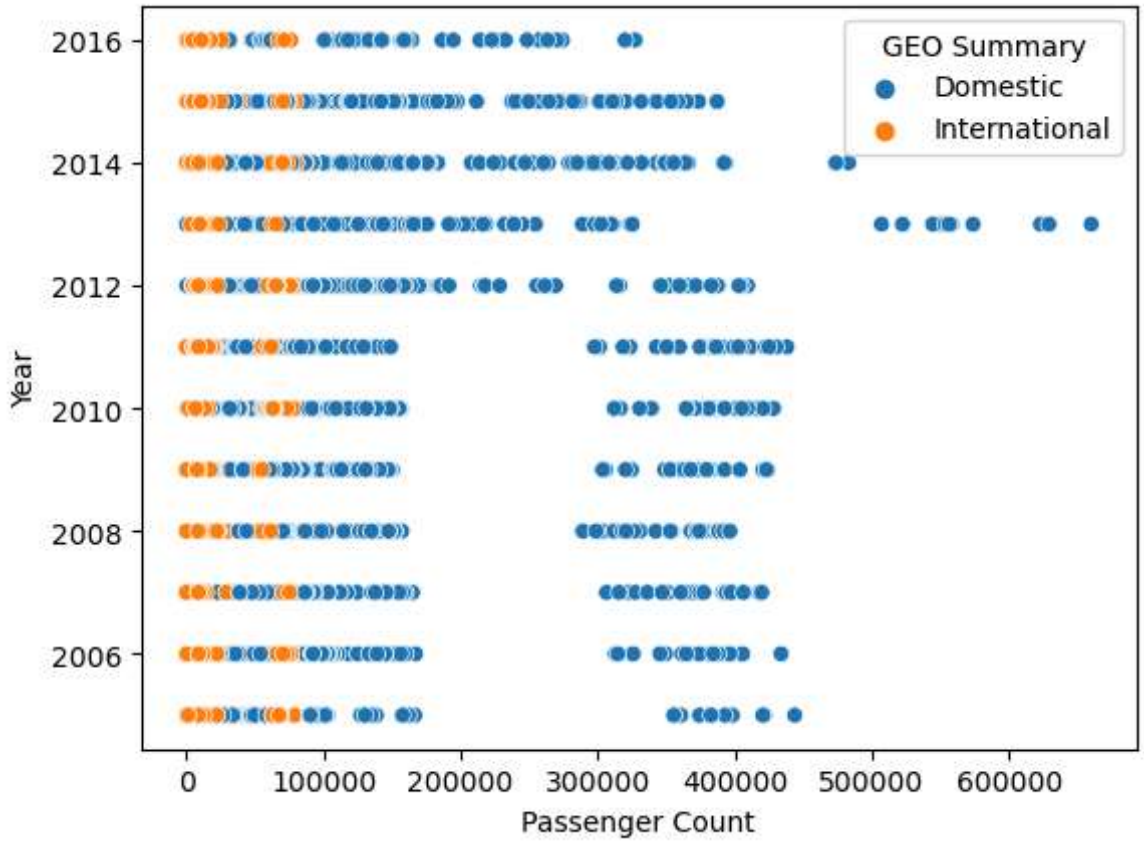
In [19]:
```
sns.scatterplot(data=df, x='Activity Period', y='Year', hue='GEO Summary')
plt.show()
```



observation:- In this relationship in the scatterplot shows that as the yeae increases activity period of passengers who are taking international flights are also increases

Relationship 2 major factors of GEO Summary Passenger Count and years of airlines

In [20]:
```
sns.scatterplot(data=df, x='Passenger Count', y='Year', hue='GEO Summary')
plt.show()
```

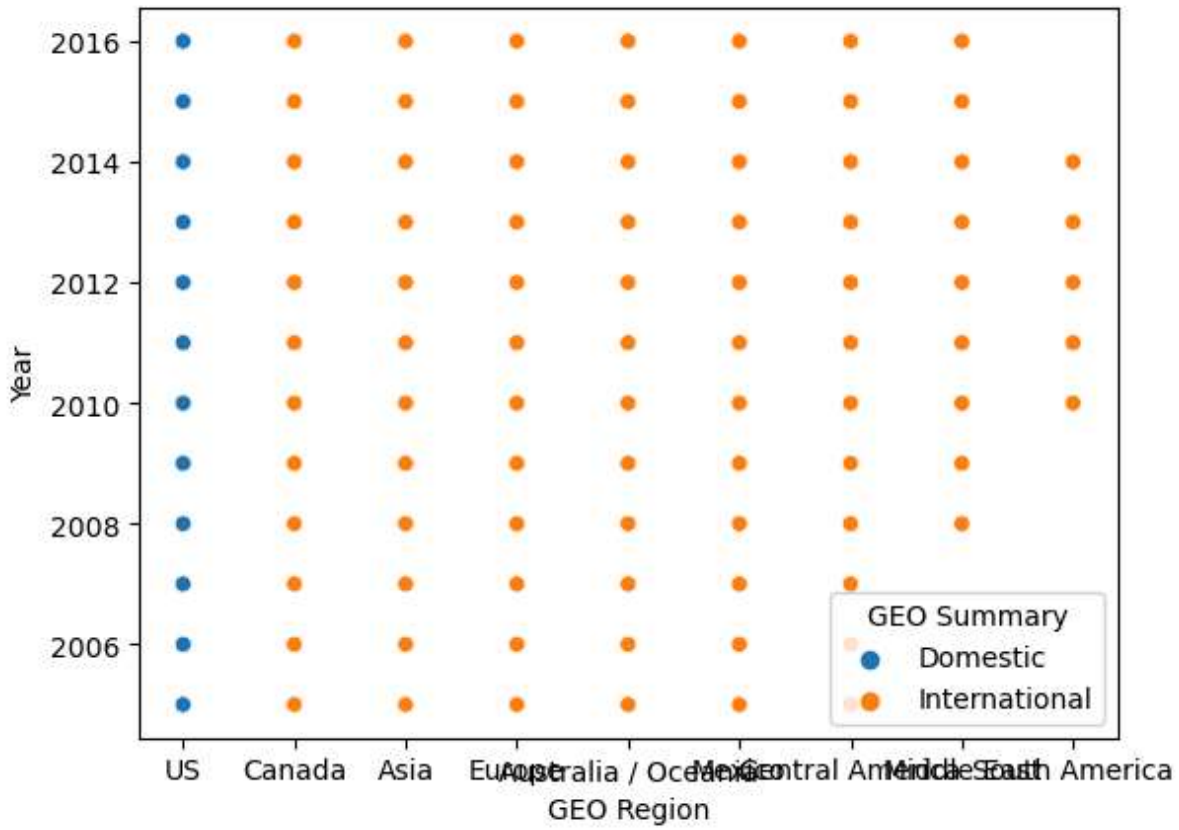


In [ ]:
```
observation:- In this relationship in the scatterplot shows that the count of r
    passengers taking domastic flights are more as compare to number of
    passengers taking international flights
```

In [ ]:
```

```

In [32]:
```
sns.scatterplot(data=df, x='GEO Region', y='Year', hue='GEO Summary')
plt.show()
```



Observation:-

- Only passenger from US Region are taking domastic flights other Region passenger are taking international flights
- count of passenger taking international flight are more as compare to passenger taking Domastic flight

In [ ]:
```
```