notebook para limpieza de GEIH 2022

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import statsmodels.api as sm
import statsmodels.formula.api as smf
```

In [2]:
```
enero_personas = pd.read_csv("Enero/Características generales, seguridad social en sal
enero_ocupados = pd.read_csv("Enero/Ocupados.csv", sep=",", encoding = "latin-1")
```

```
C:\Users\anama\AppData\Local\Temp\ipykernel_23768\1339535389.py:2: DtypeWarning: Colu
mns (17,68,88,102,113) have mixed types. Specify dtype option on import or set low_me
mory=False.
  enero_ocupados = pd.read_csv("Enero/Ocupados.csv", sep=",", encoding = "latin-1")
```

In [3]:
```
enero_ocupados
```

Out[3]:

	DIRECTORIO	SECUENCIA_P	ORDEN	HOGAR	P3044	P6440	P6450	P6460	P6460S1
0	5000000	1	1	1	AGUA EMBOTELLADA	1	1.0	NaN	NaN
1	5000001	1	1	1	SERVICIO DE REPARTO DE VOLANTES PUBLICITARIO A...	1	1.0	NaN	NaN
2	5000001	1	2	1	DESODORANTE COLONIAS	2	NaN	NaN	NaN
3	5000001	1	3	1	SERVICIO DE DOMICILIOS A PARTICULARES	2	NaN	NaN	NaN
4	5000002	1	1	1	LICOR	1	1.0	NaN	NaN
...	...	...	...	...	...	...	...	...	...
31814	5032617	1	1	1	SERVICIO DE LAVAR PLANCHAR Y ASEO	1	1.0	NaN	NaN
31815	5032618	1	1	1	ASEO Y LIMPIEZA	1	1.0	NaN	NaN
31816	5032619	1	2	1	ASEO Y LIMPIEZA	1	1.0	NaN	NaN
31817	5032620	1	1	1	SERVICIO DE CONTROL DE MIGRACION	1	2.0	1.0	NaN
31818	5032620	1	2	1	SERVICIO DE EDUCACION	1	2.0	1.0	NaN

31819 rows × 199 columns

In [4]:
```
enero=pd.merge(enero_personas,enero_ocupados, on=["DIRECTORIO",'SECUENCIA_P', 'ORDEN']
enero.tail()
```

Out[4]:

	DIRECTORIO	SECUENCIA_P	ORDEN	P6016	P3271	P6030S1	P6030S3	P6040	P6050	P6083
31814	5032617	1	1	1	2	4.0	1960.0	61	1	3
31815	5032618	1	1	1	2	NaN	NaN	45	1	2
31816	5032619	1	2	2	2	NaN	NaN	41	2	2
31817	5032620	1	1	1	1	11.0	1971.0	50	1	2
31818	5032620	1	2	2	2	3.0	1973.0	48	2	2

5 rows × 270 columns

In [5]:
```
febrero_personas = pd.read_csv("Febrero/Características generales, seguridad social er
febrero_ocupados = pd.read_csv("Febrero/Ocupados.csv", sep=";", encoding = "latin-1")
```

```
C:\Users\anama\AppData\Local\Temp\ipykernel_23768\856714140.py:2: DtypeWarning: Colum
ns (17,68,88,94,97,102,113) have mixed types. Specify dtype option on import or set l
ow_memory=False.
  febrero_ocupados = pd.read_csv("Febrero/Ocupados.csv", sep=";", encoding = "latin-
1")
```

In [6]:
```
febrero_ocupados
```

Out[6]:

	DIRECTORIO	SECUENCIA_P	ORDEN	HOGAR	P3044	P6440	P6450	P6460	P64
0	5032621	1	1	1	PAPA RELLENAS EMPENADAS GASEOSAS RELLENA	2	NaN	NaN	
1	5032621	1	2	1	PAPAS RELLENAS EMPENADAS RELLENA GASEOSA	1	1.0	NaN	
2	5032622	1	1	1	CUIDADO DE PACIENTES	1	2.0	2.0	
3	5032627	1	1	1	SERVICIO DE SUMINISTRO DE AGUA POTABLE EN EL A...	1	2.0	2.0	
4	5032628	1	2	1	SERVICIO DE EDUCACION PREESCOLAR BASICA PRIMAR...	1	2.0	1.0	
...	...	...	...	...	...	...	...	...	
32974	5063808	1	1	1	CANCHAS DE TENNIS NATACION ATLETISMO Y TRIATLO...	1	1.0	NaN	
32975	5063809	1	1	1	SERVICIOS DE APOYO Y ACOMPAÑAMIENTO A PERSONA ...	1	1.0	NaN	
32976	5063810	1	2	1	MECANICA	1	2.0	2.0	
32977	5063810	1	3	1	MECANICA	1	1.0	NaN	
32978	5063810	1	4	1	EDUCACION	1	2.0	1.0	

32979 rows × 199 columns

In [7]:
```
febrero =pd.merge(febrero_personas,febrero_ocupados, on=["DIRECTORIO",'SECUENCIA_P', '
febrero.tail()
```

Out[7]:

	DIRECTORIO	SECUENCIA_P	ORDEN	P6016	P3271	P6030S1	P6030S3	P6040	P6050	P6083
32974	5063808	1	1	1	1	3.0	1988.0	33	1	2
32975	5063809	1	1	1	2	3.0	1950.0	71	1	3
32976	5063810	1	2	2	1	5.0	1969.0	52	6	3
32977	5063810	1	3	3	1	7.0	1959.0	62	6	3
32978	5063810	1	4	4	2	7.0	1980.0	41	3	1

5 rows × 270 columns

In [8]:
```
marzo_personas = pd.read_csv("Marzo/Características generales, seguridad social en sal
marzo_ocupados = pd.read_csv("Marzo/Ocupados.csv", sep=";", encoding = "latin-1")
```

```
C:\Users\anama\AppData\Local\Temp\ipykernel_23768\858966042.py:2: DtypeWarning: Colum
ns (17,68,88,97,113) have mixed types. Specify dtype option on import or set low_memo
ry=False.
  marzo_ocupados = pd.read_csv("Marzo/Ocupados.csv", sep=";", encoding = "latin-1")
```

In [9]:
```
marzo =pd.merge(marzo_personas,marzo_ocupados, on=["DIRECTORIO",'SECUENCIA_P', 'ORDEN'
marzo.tail()
```

Out[9]:

	DIRECTORIO	SECUENCIA_P	ORDEN	P6016	P3271	P6030S1	P6030S3	P6040	P6050	P6083
32870	7030785	1	1	1	1	3.0	1990.0	32	1	3
32871	7030786	1	1	1	2	2.0	1996.0	26	1	2
32872	7030787	1	1	1	1	8.0	1985.0	36	1	2
32873	7030788	1	1	1	2	3.0	1997.0	25	1	2
32874	7030789	1	1	1	2	2.0	1996.0	26	1	2

5 rows × 270 columns

In [10]:
```
abril_personas = pd.read_csv("Abril/Características generales, seguridad social en sal
abril_ocupados = pd.read_csv("Abril/Ocupados.csv", sep=";", encoding = "latin-1")
```

```
C:\Users\anama\AppData\Local\Temp\ipykernel_23768\499192029.py:2: DtypeWarning: Colum
ns (28,89,135,150,161) have mixed types. Specify dtype option on import or set low_me
mory=False.
  abril_ocupados = pd.read_csv("Abril/Ocupados.csv", sep=";", encoding = "latin-1")
```

In [11]:
```
abril =pd.merge(abril_personas,abril_ocupados, on=["DIRECTORIO",'SECUENCIA_P', 'ORDEN'
abril.tail()
```

Out[11]:

	PERIODO_x	MES_x	PER_x	DIRECTORIO	SECUENCIA_P	ORDEN	HOGAR_x	REGIS_x	AREA_x
31909	20220417	4	2022	7061802	1	1	1	10	47.0
31910	20220417	4	2022	7061802	1	3	1	10	47.0
31911	20220417	4	2022	7061802	1	4	1	10	47.0
31912	20220416	4	2022	7061803	1	1	1	10	47.0
31913	20220416	4	2022	7061803	1	3	1	10	47.0

5 rows × 271 columns

In [12]:
```
mayo_personas = pd.read_csv("Mayo/Características generales, seguridad social en saluc
mayo_ocupados = pd.read_csv("Mayo/Ocupados.csv", sep=";", encoding = "latin-1")
```

```
C:\Users\anama\AppData\Local\Temp\ipykernel_23768\344780937.py:2: DtypeWarning: Colum
ns (28,89,135,161) have mixed types. Specify dtype option on import or set low_memory
=False.
  mayo_ocupados = pd.read_csv("Mayo/Ocupados.csv", sep=";", encoding = "latin-1")
```

In [13]:
```
mayo =pd.merge(mayo_personas,mayo_ocupados, on=["DIRECTORIO",'SECUENCIA_P', 'ORDEN'],
mayo.tail()
```

Out[13]:

	PERIODO_x	MES_x	PER_x	DIRECTORIO	SECUENCIA_P	ORDEN	HOGAR_x	REGIS_x	AREA_x
32533	20220521	5	2022	7092467	1	1	1	10	63.0
32534	20220521	5	2022	7092468	1	1	1	10	63.0
32535	20220521	5	2022	7092470	1	1	1	10	63.0
32536	20220521	5	2022	7092472	1	1	1	10	63.0
32537	20220521	5	2022	7092472	1	5	1	10	63.0

5 rows × 271 columns

In [14]:
```
junio_personas = pd.read_csv("Junio/Características generales, seguridad social en sal
junio_ocupados = pd.read_csv("Junio/Ocupados.csv", sep=";", encoding = "latin-1")
```

```
C:\Users\anama\AppData\Local\Temp\ipykernel_23768\2184435766.py:2: DtypeWarning: Colu
mns (28,89,135,150,161) have mixed types. Specify dtype option on import or set low_m
emory=False.
  junio_ocupados = pd.read_csv("Junio/Ocupados.csv", sep=";", encoding = "latin-1")
```

In [15]:
```
junio =pd.merge(junio_personas,junio_ocupados, on=["DIRECTORIO",'SECUENCIA_P', 'ORDEN'
junio.tail()
```

Out[15]:

	PERIODO_x	MES_x	PER_x	DIRECTORIO	SECUENCIA_P	ORDEN	HOGAR_x	REGIS_x	AREA_x
32517	20220624	6	2022	7126190	1	1	1	10	88.0
32518	20220626	6	2022	7126191	1	1	1	10	NaN
32519	20220626	6	2022	7126192	1	1	1	10	NaN
32520	20220626	6	2022	7126195	1	1	1	10	NaN
32521	20220626	6	2022	7126197	1	1	1	10	NaN

5 rows × 271 columns

In [16]:
```
julio_personas = pd.read_csv("Julio/Características generales, seguridad social en sal
julio_ocupados = pd.read_csv("Julio/Ocupados.csv", sep=";", encoding = "latin-1")
```

```
C:\Users\anama\AppData\Local\Temp\ipykernel_23768\367737641.py:2: DtypeWarning: Colum
ns (89,135,161) have mixed types. Specify dtype option on import or set low_memory=Fa
lse.
  julio_ocupados = pd.read_csv("Julio/Ocupados.csv", sep=";", encoding = "latin-1")
```

In [17]:
```
julio =pd.merge(julio_personas,julio_ocupados, on=["DIRECTORIO",'SECUENCIA_P', 'ORDEN'
julio.tail()
```

Out[17]:

	PERIODO_x	MES_x	PER_x	DIRECTORIO	SECUENCIA_P	ORDEN	HOGAR_x	REGIS_x	AREA_x
31526	20220730	7	2022	7156449	1	1	1	10	NaN
31527	20220730	7	2022	7156449	1	2	1	10	NaN
31528	20220728	7	2022	7156451	1	1	1	10	47.0
31529	20220730	7	2022	7156452	1	1	1	10	47.0
31530	20220730	7	2022	7156452	1	7	1	10	47.0

5 rows × 271 columns

In [18]:
```
agosto_personas = pd.read_csv("Agosto/Características generales, seguridad social en s
agosto_ocupados = pd.read_csv("Agosto/Ocupados.csv", sep=";", encoding = "latin-1")
```

```
C:\Users\anama\AppData\Local\Temp\ipykernel_23768\4143199729.py:2: DtypeWarning: Colu
mns (28,89,135,161) have mixed types. Specify dtype option on import or set low_memor
y=False.
  agosto_ocupados = pd.read_csv("Agosto/Ocupados.csv", sep=";", encoding = "latin-1")
```

In [19]:
```
agosto =pd.merge(agosto_personas,agosto_ocupados, on=["DIRECTORIO",'SECUENCIA_P', 'ORD
agosto.tail()
```

Out[19]:

	PERIODO_x	MES_x	PER_x	DIRECTORIO	SECUENCIA_P	ORDEN	HOGAR_x	REGIS_x	AREA_x
31815	20220834	8	2022	7185903	1	1	1	10	18.0
31816	20220834	8	2022	7185903	1	2	1	10	18.0
31817	20220834	8	2022	7185908	1	1	1	10	47.0
31818	20220834	8	2022	7185908	1	2	1	10	47.0
31819	20220834	8	2022	7185912	1	2	1	10	68.0

5 rows × 271 columns

In [20]:
```
septiembre_personas = pd.read_csv("Septiembre/Características generales, seguridad soc
septiembre_ocupados = pd.read_csv("Septiembre/Ocupados.csv", sep=";", encoding = "lati
```

```
C:\Users\anama\AppData\Local\Temp\ipykernel_23768\2184485926.py:2: DtypeWarning: Colu
mns (28,89,135,148,161) have mixed types. Specify dtype option on import or set low_m
emory=False.
  septiembre_ocupados = pd.read_csv("Septiembre/Ocupados.csv", sep=";", encoding = "l
atin-1")
```

In [21]:
```
septiembre =pd.merge(septiembre_personas,septiembre_ocupados, on=["DIRECTORIO",'SECUEN
septiembre.tail()
```

Out[21]:

	PERIODO_x	MES_x	PER_x	DIRECTORIO	SECUENCIA_P	ORDEN	HOGAR_x	REGIS_x	AREA_x
31862	20220939	9	2022	7215413	1	1	1	10	NaN
31863	20220939	9	2022	7215414	1	1	1	10	NaN
31864	20220939	9	2022	7215414	1	3	1	10	NaN
31865	20220939	9	2022	7215414	1	4	1	10	NaN
31866	20220939	9	2022	7215415	1	1	1	10	NaN

5 rows × 271 columns

In [22]:
```
octubre_personas = pd.read_csv("Octubre/Características generales, seguridad social er
octubre_ocupados = pd.read_csv("Octubre/Ocupados.csv", sep=";", encoding = "latin-1")
```

```
C:\Users\anama\AppData\Local\Temp\ipykernel_23768\2065219422.py:2: DtypeWarning: Colu
mns (28,89,135,161) have mixed types. Specify dtype option on import or set low_memor
y=False.
  octubre_ocupados = pd.read_csv("Octubre/Ocupados.csv", sep=";", encoding = "latin-
1")
```

In [23]:
```
octubre =pd.merge(octubre_personas,octubre_ocupados, on=["DIRECTORIO",'SECUENCIA_P', '
octubre.tail()
```

Out[23]:

	PERIODO_x	MES_x	PER_x	DIRECTORIO	SECUENCIA_P	ORDEN	HOGAR_x	REGIS_x	AREA_x
30953	20221043	10	2022	7246612	1	2	1	10	27.0
30954	20221043	10	2022	7246613	1	1	1	10	27.0
30955	20221042	10	2022	7246614	1	2	1	10	27.0
30956	20221042	10	2022	7246615	1	1	1	10	27.0
30957	20221043	10	2022	7246616	1	1	1	10	27.0

5 rows × 271 columns

In [24]:
```
noviembre_personas = pd.read_csv("Noviembre/Características generales, seguridad socia
noviembre_ocupados = pd.read_csv("Noviembre/Ocupados.csv", sep=";", encoding = "latin-
```

```
C:\Users\anama\AppData\Local\Temp\ipykernel_23768\4061901441.py:2: DtypeWarning: Colu
mns (28,135,161) have mixed types. Specify dtype option on import or set low_memory=F
alse.
  noviembre_ocupados = pd.read_csv("Noviembre/Ocupados.csv", sep=";", encoding = "lat
in-1")
```

In [25]:
```
noviembre =pd.merge(noviembre_personas,noviembre_ocupados, on=["DIRECTORIO",'SECUENCIA
noviembre.tail()
```

Out[25]:

	PERIODO_x	MES_x	PER_x	DIRECTORIO	SECUENCIA_P	ORDEN	HOGAR_x	REGIS_x	AREA_x
31128	20221147	11	2022	7277068	1	2	1	10	NaN
31129	20221147	11	2022	7277068	1	3	1	10	NaN
31130	20221145	11	2022	7277087	1	1	1	10	NaN
31131	20221145	11	2022	7277087	1	2	1	10	NaN
31132	20221145	11	2022	7277087	1	3	1	10	NaN

5 rows × 271 columns

In [26]:
```
diciembre_personas = pd.read_csv("Diciembre/Características generales, seguridad socia
diciembre_ocupados = pd.read_csv("Diciembre/Ocupados.csv", sep=";", encoding = "latin-
```

```
C:\Users\anama\AppData\Local\Temp\ipykernel_23768\961976400.py:2: DtypeWarning: Colum
ns (28,89,135,161) have mixed types. Specify dtype option on import or set low_memory
=False.
  diciembre_ocupados = pd.read_csv("Diciembre/Ocupados.csv", sep=";", encoding = "lat
in-1")
```

In [27]:
```
diciembre =pd.merge(diciembre_personas,diciembre_ocupados, on=["DIRECTORIO",'SECUENCIA
diciembre.tail()
```

Out[27]:

	PERIODO_x	MES_x	PER_x	DIRECTORIO	SECUENCIA_P	ORDEN	HOGAR_x	REGIS_x	AREA_x
30500	20221251	12	2022	7309048	1	2	1	10	66.0
30501	20221251	12	2022	7309048	1	3	1	10	66.0
30502	20221251	12	2022	7309049	1	1	1	10	NaN
30503	20221251	12	2022	7309050	1	1	1	10	NaN
30504	20221251	12	2022	7309050	1	2	1	10	NaN

5 rows × 271 columns

In [28]:
```
septiembre.columns.tolist()
```

Out[28]:
```
['PERIODO_x',
 'MES_x',
 'PER_x',
 'DIRECTORIO',
 'SECUENCIA_P',
 'ORDEN',
 'HOGAR_x',
 'REGIS_x',
 'AREA_x',
 'CLASE_x',
 'FEX_C18_x',
 'DPTO_x',
 'PT',
 'P6016',
 'P3271',
 'P6040',
 'P6030S1',
 'P6030S3',
 'P6050',
 'P6083',
 'P6083S1',
 'P6081',
 'P6081S1',
 'P2057',
 'P2059',
 'P2061',
 'P6080',
 'P6080S1',
 'P6080S1A1',
 'P6070',
 'P6071',
 'P6071S1',
 'P3147S1',
 'P3147S2',
 'P3147S3',
 'P3147S4',
 'P3147S5',
 'P3147S6',
 'P3147S7',
 'P3147S8',
 'P3147S9',
 'P3147S10',
 'P3147S11',
 'P3147S10A1',
 'P6090',
 'P6100',
 'P6110',
 'P6120',
 'P1906S1',
 'P1906S2',
 'P1906S3',
 'P1906S4',
 'P1906S5',
 'P1906S6',
 'P1906S7',
 'P1906S8',
 'P6160',
 'P6170',
 'P3041',
 'P3042',
```

```
'P3042S1',
'P3042S2',
'P3043',
'P3043S1',
'P3038',
'P3039',
'POB_MAY18',
'LGB_Numerica',
'LGB_sectores',
'LGBT_Numerica',
'Trans_numerica',
'Discapacidad',
'Dificultad',
'Campesina',
'PERIODO_y',
'MES_y',
'PER_y',
'HOGAR_y',
'REGIS_y',
'AREA_y',
'CLASE_y',
'FEX_C18_y',
'DPTO_y',
'FT',
'P3044S2',
'P6440',
'P6450',
'P6460',
'P6460S1',
'P6400',
'P6410',
'P6422',
'P6420S2',
'P6424S1',
'P6424S2',
'P6424S3',
'P6424S5',
'P6426',
'P6430',
'P6430S1',
'P3045S1',
'P3045S2',
'P3045S3',
'P3046',
'P3363',
'P9440',
'P6500',
'P3364',
'P3364S1',
'P6510',
'P6510S1',
'P6510S2',
'P6590',
'P6590S1',
'P6600',
'P6600S1',
'P6610',
'P6610S1',
'P6620',
'P6620S1',
```

```
                'P6585S1',
                'P6585S1A1',
                'P6585S1A2',
                'P6585S2',
                'P6585S2A1',
                'P6585S2A2',
                'P6585S3',
                'P6585S3A1',
                'P6585S3A2',
                'P6585S4',
                'P6585S4A1',
                'P6585S4A2',
                'P6545',
                'P6545S1',
                'P6545S2',
                'P6580',
                'P6580S1',
                'P6580S2',
                'P6630S1',
                'P6630S1A1',
                'P6630S2',
                'P6630S2A1',
                'P6630S3',
                'P6630S3A1',
                'P6630S4',
                'P6630S4A1',
                'P6630S6',
                'P6630S6A1',
                'P6640',
                'P6640S1',
                'P1800',
                'P1800S1',
                'P1801S1',
                'P1801S2',
                'P1801S3',
                'P1802',
                'P3047',
                'P3048',
                'P3049',
                'P6765',
                'P6765S1',
                'P3051',
                'P3051S1',
                'P3052',
                'P3052S1',
                'P3053',
                'P3365',
                'P3365S1',
                'P3054',
                'P3054S1',
                'P3055',
                'P3055S1',
                'P3056',
                'P3057',
                'P6760',
                'P3058S1',
                'P3058S2',
                'P3058S3',
                'P3058S4',
                'P3058S5',
```

```
                'P3059',
                'P3061',
                'P3062S1',
                'P3062S2',
                'P3062S3',
                'P3062S4',
                'P3062S5',
                'P3062S6',
                'P3062S7',
                'P3062S8',
                'P3062S9',
                'P3063',
                'P3063S1',
                'P3064',
                'P3064S1',
                'P3065',
                'P3066',
                'P3067',
                'P3067S1',
                'P3067S2',
                'P6775',
                'P3068',
                'P6750',
                'P3073',
                'P550',
                'P6780',
                'P6780S1',
                'P1879',
                'P1805',
                'P6790',
                'P6800',
                'P6810',
                'P6810S1',
                'P6850',
                'P6830',
                'P6830S1',
                'P3366',
                'P3069',
                'P6880',
                'P6880S1',
                'P6915',
                'P6915S1',
                'P6920',
                'P6930',
                'P6940',
                'P6960',
                'P6990',
                'P9450',
                'P7020',
                'P760',
                'P7026',
                'P7028',
                'P7028S1',
                'P1880',
                'P1880S1',
                'P7040',
                'P7045',
                'P7050',
                'P7070',
                'P7075',
```

```
 'P7077',
 'P7090',
 'P7100',
 'P7110',
 'P7120',
 'P7130',
 'P7140S1',
 'P7140S2',
 'P7140S3',
 'P7140S4',
 'P7140S5',
 'P7140S6',
 'P7140S7',
 'P7140S8',
 'P7140S9',
 'P7150',
 'P7160',
 'P7170S1',
 'P7170S5',
 'P7170S6',
 'P7180',
 'P514',
 'P515',
 'P1881',
 'P1882',
 'P7240',
 'OCI',
 'INGLABO',
 'RAMA2D_R4',
 'RAMA4D_R4',
 'OFICIO_C8']
```

In [29]:
```
GEIH2022 = pd.concat([enero,febrero,marzo,abril, mayo, junio, julio, agosto, septiembr
```

In [30]:
```
GEIH2022= GEIH2022.rename(columns = {"P3271":"Sexo", "P6080":"Etnia", "P6090":"CotizaE
                                     "P3042":"NivelEducativo", "P6430":"TipoEmpleo", '
                                     "P1800":"Empleados", "P3067":"CCCyRegistro_emplea
                                     "P3068":"Contabilidad_gastos", "P6920":"Pension",
                                     "P6050":"JefeHogar", "P6110": "PagoEPS", "P3069":
```

In [31]:
```
print(GEIH2022.columns.tolist())
```

```
['DIRECTORIO', 'SECUENCIA_P', 'ORDEN', 'P6016', 'Sexo', 'P6030S1', 'P6030S3', 'P604
0', 'JefeHogar', 'P6083', 'P6083S1', 'P6081', 'P6081S1', 'P2057', 'P2059', 'P2061',
'Etnia', 'P6080S1', 'P6080S1A1', 'P6070', 'P6071', 'P6071S1', 'P3147S1', 'P3147S2',
'P3147S3', 'P3147S4', 'P3147S5', 'P3147S6', 'P3147S7', 'P3147S8', 'P3147S9', 'P3147S1
0', 'P3147S11', 'P3147S10A1', 'CotizaEPS', 'P6100', 'PagoEPS', 'P6120', 'P1906S1', 'P
1906S2', 'P1906S3', 'P1906S4', 'P1906S5', 'P1906S6', 'P1906S7', 'P1906S8', 'P6160',
'P6170', 'P3041', 'NivelEducativo', 'P3042S1', 'P3042S2', 'P3043', 'P3043S1', 'HOGAR_
x', 'CLASE_x', 'P3038', 'P3039', 'AREA_x', 'MES_x', 'PERIODO_x', 'DPTO_x', 'pt', 'Fac
torExpansion', 'LGB_Numerica', 'LGB_sectores', 'Trans_numerica', 'LGBT_Numerica', 'Di
scapacidad', 'Dificultad', 'Campesina', 'POB_MAY18', 'PER_x', 'REGIS_x', 'HOGAR_y',
'P3044', 'P6440', 'P6450', 'P6460', 'P6460S1', 'P6400', 'P6410', 'P6422', 'P6424S1',
'P6424S2', 'P6424S3', 'P6424S5', 'P6426', 'P6430S1', 'P3045S1', 'P3045S2', 'P3045S3',
'P3046', 'P9440', 'P6500', 'P6510', 'P6510S1', 'P6510S2', 'P6590', 'P6590S1', 'P660
0', 'P6600S1', 'P6610', 'P6610S1', 'P6620', 'P6620S1', 'P6585S1', 'P6585S1A1', 'P6585
S1A2', 'P6585S2', 'P6585S2A1', 'P6585S2A2', 'P6585S3', 'P6585S3A1', 'P6585S3A2', 'P65
85S4', 'P6585S4A1', 'P6585S4A2', 'P6545', 'P6545S1', 'P6545S2', 'P6580', 'P6580S1',
'P6580S2', 'P6630S1', 'P6630S1A1', 'P6630S2', 'P6630S2A1', 'P6630S3', 'P6630S3A1', 'P
6630S4', 'P6630S4A1', 'P6630S6', 'P6630S6A1', 'P6640', 'P3047', 'P3048', 'P3049', 'P6
765', 'P6765S1', 'P3053', 'P3054', 'P3055', 'P3056', 'P6760', 'P3061', 'P3063', 'P306
4', 'CCCyRegistro', 'P3066', 'CCCyRegistro_empleador', 'RenovacionRegistro', 'P3067S
2', 'P6775', 'Contabilidad_gastos', 'P6750', 'P3073', 'P550', 'P6780', 'P6780S1', 'P1
879', 'P1805', 'P6790', 'P6800', 'P6810', 'P6810S1', 'P6850', 'P6830', 'P6830S1', 'To
tal_Empleados', 'P6880', 'P6880S1', 'P6915', 'P6915S1', 'Pension', 'P6930', 'P6940',
'P6960', 'P6990', 'P9450', 'P7020', 'P760', 'P7026', 'P7028', 'P7028S1', 'P1880', 'P1
880S1', 'P7040', 'P7045', 'P7050', 'P7070', 'P7075', 'P7077', 'P7090', 'P7100', 'P711
0', 'P7120', 'P7130', 'P7140S1', 'P7140S2', 'P7140S3', 'P7140S4', 'P7140S5', 'P7140S
6', 'P7140S7', 'P7140S8', 'P7140S9', 'P7150', 'P7160', 'P7170S1', 'P7170S5', 'P7170S
6', 'P7180', 'P514', 'P515', 'P1881', 'P1882', 'P7240', 'CLASE_y', 'OCI', 'AREA_y',
'TipoEmpleo', 'RAMA4D_R4', 'OFICIO_C8', 'RAMA2D_R4', 'MES_y', 'PERIODO_y', 'Departame
nto', 'FT', 'FEX_C18_y', 'INGLABO', 'P3363', 'P3364', 'P3364S1', 'P6640S1', 'Empleado
s', 'P1800S1', 'P1801S1', 'P1801S2', 'P1801S3', 'P1802', 'P3051', 'P3051S1', 'P3052',
'P3052S1', 'P3365', 'P3365S1', 'P3054S1', 'P3055S1', 'P3057', 'P3058S1', 'P3058S2',
'P3058S3', 'P3058S4', 'P3058S5', 'P3059', 'P3062S1', 'P3062S2', 'P3062S3', 'P3062S4',
'P3062S5', 'P3062S6', 'P3062S7', 'P3062S8', 'P3062S9', 'P3063S1', 'P3064S1', 'P3366',
'PER_y', 'REGIS_y', 'PT', 'P3044S2', 'P6420S2']
```

In [32]:
```
GEIH_limpio=GEIH2022.drop(columns=['P6016', 'P6030S1', 'P6030S3', 'P6040', 'P6083', 'P
```

In [33]:
```
GEIH_2022= GEIH_limpio.copy()
```

Estas son las condiciones que da el DANE para 2022 de informalidad, creando una nueva columna dummy que toma el valor 1 (True) si cumple alguno de los criterios

In [34]:
```
informalidad = [
    (GEIH_2022["TipoEmpleo"] == 1) & (GEIH_2022["CCCyRegistro"]== 2),
    (GEIH_2022["TipoEmpleo"] == 1) & (GEIH_2022["Contabilidad_gastos"] == 2),
    (GEIH_2022["TipoEmpleo"] == 4) & (GEIH_2022["CCCyRegistro_empleador"]== 2),
    (GEIH_2022["TipoEmpleo"] == 5) & (GEIH_2022["CCCyRegistro_empleador"]== 2),
    (GEIH_2022["TipoEmpleo"] == 4) & (GEIH_2022["Contabilidad_gastos"] == 2),
    (GEIH_2022["TipoEmpleo"] == 5) & (GEIH_2022["Contabilidad_gastos"] == 2),
    (GEIH_2022["TipoEmpleo"] == 4) & (GEIH_2022["CCCyRegistro_empleador"]== 1) & (GEIH
    (GEIH_2022["TipoEmpleo"] == 1) & ((GEIH_2022["Total_Empleados"] == 1) |(GEIH_2022[
    (GEIH_2022["TipoEmpleo"] == 4) & ((GEIH_2022["Total_Empleados"] == 1) |(GEIH_2022[
    (GEIH_2022["TipoEmpleo"] == 5) & ((GEIH_2022["Total_Empleados"] == 1) |(GEIH_2022[
    (GEIH_2022["TipoEmpleo"] == 6), (GEIH_2022["TipoEmpleo"] == 9),
    (GEIH_2022["TipoEmpleo"] == 1) & (GEIH_2022["CotizaEPS"]== 2) & (GEIH_2022["Pensio

Return = [1,1,1,1,1,1,1,1,1,1,1,1]
```

```
GEIH_2022["Ocupacion_informal"] = np.select(informalidad, Return)

GEIH_2022
```

Out[34]:

	DIRECTORIO	SECUENCIA_P	ORDEN	Sexo	JefeHogar	Etnia	CotizaEPS	PagoEPS	NivelEduca
0	5000000	1	1	1	1	6	2	NaN	
1	5000001	1	1	1	1	6	1	NaN	
2	5000001	1	2	2	2	6	1	NaN	
3	5000001	1	3	1	3	6	1	NaN	
4	5000002	1	1	2	1	6	2	NaN	
...	...	...	...	...	...	...	...	...	
30500	7309048	1	2	1	2	6	1	3.0	
30501	7309048	1	3	1	3	6	1	3.0	
30502	7309049	1	1	2	1	6	1	1.0	
30503	7309050	1	1	1	1	6	1	NaN	
30504	7309050	1	2	2	2	6	2	NaN	

382461 rows × 19 columns

In [35]:
```
GEIH_2022.FactorExpansion
```

Out[35]:
```
0        1432.463323
1        1088.796266
2        1088.796266
3        1088.796266
4        2066.712422
            ...
30500     365.671281
30501     365.671281
30502      25.636473
30503      25.167948
30504      25.167948
Name: FactorExpansion, Length: 382461, dtype: float64
```

In [36]:
```
GEIH_2022["Ponderacion"] = GEIH_2022["Ocupacion_informal"] * GEIH_2022["FactorExpansic
```

In [37]:
```
Total_Expansion = GEIH_2022["FactorExpansion"].sum()
Total_Ponderacion = GEIH_2022["Ponderacion"].sum()
```

In [38]:
```
OI_TASA= (Total_Ponderacion/Total_Expansion)*100
OI_TASA
```

Out[38]:
```
53.47689153052556
```

In [39]:
```
OInformal= (Total_Ponderacion/Total_Expansion)
OInformal
```

Out[39]:
```
0.5347689153052556
```

In [40]:
```
Total_Ponderacion
```

Out[40]:
```
141385499.83136684
```

In [41]:
```
GEIH_dummy= GEIH_2022.copy()
GEIH_dummy
```

Out[41]:

	DIRECTORIO	SECUENCIA_P	ORDEN	Sexo	JefeHogar	Etnia	CotizaEPS	PagoEPS	NivelEduca
0	5000000	1	1	1	1	6	2	NaN	
1	5000001	1	1	1	1	6	1	NaN	
2	5000001	1	2	2	2	6	1	NaN	
3	5000001	1	3	1	3	6	1	NaN	
4	5000002	1	1	2	1	6	2	NaN	
...	...	...	...	...	...	...	...	...	
30500	7309048	1	2	1	2	6	1	3.0	
30501	7309048	1	3	1	3	6	1	3.0	
30502	7309049	1	1	2	1	6	1	1.0	
30503	7309050	1	1	1	1	6	1	NaN	
30504	7309050	1	2	2	2	6	2	NaN	

382461 rows × 20 columns

In [42]:
```
GEIH_dummy["JefeHogar"]=GEIH_dummy["JefeHogar"].replace([1], "JEFEHO")
GEIH_dummy["Etnia"] = GEIH_dummy["Etnia"].replace([1,2,3,4,5,6],["Indígena", "Gitano",
GEIH_dummy["Sexo"]=GEIH_dummy["Sexo"].replace([1,2], ["Hombre", "Mujer"])
GEIH_dummy["NivelEducativo"]=GEIH_dummy["NivelEducativo"].replace([1,2,3,4,5,6,7,8,9,1
GEIH_dummy["Departamento"]=GEIH_dummy["Departamento"].replace([91,5,81,8,11,13,15,17,1
GEIH_dummy
```

Out[42]:

	DIRECTORIO	SECUENCIA_P	ORDEN	Sexo	JefeHogar	Etnia	CotizaEPS	PagoEPS	Nivel
0	5000000	1	1	Hombre	JEFEHO	Ninguno	2	NaN	B
1	5000001	1	1	Hombre	JEFEHO	Ninguno	1	NaN	
2	5000001	1	2	Mujer	2	Ninguno	1	NaN	
3	5000001	1	3	Hombre	3	Ninguno	1	NaN	
4	5000002	1	1	Mujer	JEFEHO	Ninguno	2	NaN	B
...	...	...	...	...	...	...	...	...	
30500	7309048	1	2	Hombre	2	Ninguno	1	3.0	B
30501	7309048	1	3	Hombre	3	Ninguno	1	3.0	U
30502	7309049	1	1	Mujer	JEFEHO	Ninguno	1	1.0	T
30503	7309050	1	1	Hombre	JEFEHO	Ninguno	1	NaN	
30504	7309050	1	2	Mujer	2	Ninguno	2	NaN	B

382461 rows × 20 columns

In [43]:
```
GEIH_dummy.columns
```

Out[43]:
```
Index(['DIRECTORIO', 'SECUENCIA_P', 'ORDEN', 'Sexo', 'JefeHogar', 'Etnia',
       'CotizaEPS', 'PagoEPS', 'NivelEducativo', 'FactorExpansion',
       'CCCyRegistro', 'CCCyRegistro_empleador', 'RenovacionRegistro',
       'Contabilidad_gastos', 'Total_Empleados', 'Pension', 'TipoEmpleo',
       'Departamento', 'Ocupacion_informal', 'Ponderacion'],
      dtype='object')
```

In [44]:
```
GEIH_dummy.to_csv("GEIH_2022.csv")
```

In [45]:
```
GEIH_limpionan= GEIH_dummy.drop(columns = [
       'CotizaEPS', 'PagoEPS',
       'CCCyRegistro', 'CCCyRegistro_empleador', 'RenovacionRegistro',
       'Contabilidad_gastos', 'Total_Empleados', 'Pension', 'TipoEmpleo'])
GEIH_limpionan
```

Out[45]:

	DIRECTORIO	SECUENCIA_P	ORDEN	Sexo	JefeHogar	Etnia	NivelEducativo	FactorExpa
0	5000000	1	1	Hombre	JEFEHO	Ninguno	Bachillerato	1432.4
1	5000001	1	1	Hombre	JEFEHO	Ninguno	Primaria	1088.7
2	5000001	1	2	Mujer	2	Ninguno	Primaria	1088.7
3	5000001	1	3	Hombre	3	Ninguno	Primaria	1088.7
4	5000002	1	1	Mujer	JEFEHO	Ninguno	Bachillerato	2066.7
...	...	...	...	...	...	...	...	
30500	7309048	1	2	Hombre	2	Ninguno	Bachillerato	365.6
30501	7309048	1	3	Hombre	3	Ninguno	Universitaria	365.6
30502	7309049	1	1	Mujer	JEFEHO	Ninguno	Tecnologica	25.6
30503	7309050	1	1	Hombre	JEFEHO	Ninguno	Primaria	25.1
30504	7309050	1	2	Mujer	2	Ninguno	Bachillerato	25.1

382461 rows × 11 columns

In [46]:
```
GEIH_limpionan = GEIH_limpionan.dropna(axis=0)
GEIH_limpionan
```

Out[46]:

	DIRECTORIO	SECUENCIA_P	ORDEN	Sexo	JefeHogar	Etnia	NivelEducativo	FactorExpa
0	5000000	1	1	Hombre	JEFEHO	Ninguno	Bachillerato	1432.4
1	5000001	1	1	Hombre	JEFEHO	Ninguno	Primaria	1088.7
2	5000001	1	2	Mujer	2	Ninguno	Primaria	1088.7
3	5000001	1	3	Hombre	3	Ninguno	Primaria	1088.7
4	5000002	1	1	Mujer	JEFEHO	Ninguno	Bachillerato	2066.7
...	...	...	...	...	...	...	...	
30500	7309048	1	2	Hombre	2	Ninguno	Bachillerato	365.6
30501	7309048	1	3	Hombre	3	Ninguno	Universitaria	365.6
30502	7309049	1	1	Mujer	JEFEHO	Ninguno	Tecnologica	25.6
30503	7309050	1	1	Hombre	JEFEHO	Ninguno	Primaria	25.1
30504	7309050	1	2	Mujer	2	Ninguno	Bachillerato	25.1

382461 rows × 11 columns

In [47]:
```
GEIH_dummy = GEIH_limpionan.astype({'Sexo': 'category', 'JefeHogar': 'category', 'Etni
```

In [48]:
```
dummies = GEIH_dummy.select_dtypes(include = ["category"]).columns
```

In [49]:
```
GEIH_encoding = pd.get_dummies(GEIH_dummy[dummies])
GEIH_encoding
```

Out[49]:

	Sexo_Hombre	Sexo_Mujer	JefeHogar_2	JefeHogar_3	JefeHogar_4	JefeHogar_5	JefeHogar_6
0	1	0	0	0	0	0	0
1	1	0	0	0	0	0	0
2	0	1	1	0	0	0	0
3	1	0	0	1	0	0	0
4	0	1	0	0	0	0	0
...	...	...	...	...	...	...	...
30500	1	0	1	0	0	0	0
30501	1	0	0	1	0	0	0
30502	0	1	0	0	0	0	0
30503	1	0	0	0	0	0	0
30504	0	1	1	0	0	0	0

382461 rows × 68 columns

In [50]:
```
GEIH_encoding.columns
```

Out[50]:
```
Index(['Sexo_Hombre', 'Sexo_Mujer', 'JefeHogar_2', 'JefeHogar_3',
       'JefeHogar_4', 'JefeHogar_5', 'JefeHogar_6', 'JefeHogar_7',
       'JefeHogar_8', 'JefeHogar_9', 'JefeHogar_10', 'JefeHogar_11',
       'JefeHogar_12', 'JefeHogar_13', 'JefeHogar_JEFEHO', 'Etnia_Gitano',
       'Etnia_Indígena', 'Etnia_Negro_mulato_afrod_afroc', 'Etnia_Ninguno',
       'Etnia_Palenquero', 'Etnia_Raizal_SAI', 'NivelEducativo_99.0',
       'NivelEducativo_B_tecnico', 'NivelEducativo_Bachillerato',
       'NivelEducativo_Doctorado', 'NivelEducativo_Especializacion',
       'NivelEducativo_Maestria', 'NivelEducativo_Ninguno',
       'NivelEducativo_Normalista', 'NivelEducativo_Preescolar',
       'NivelEducativo_Primaria', 'NivelEducativo_Secundaria',
       'NivelEducativo_Tecnica', 'NivelEducativo_Tecnologica',
       'NivelEducativo_Universitaria', 'Departamento_Amazonas',
       'Departamento_Antioquia', 'Departamento_Arauca',
       'Departamento_Atlantico', 'Departamento_Bogota', 'Departamento_Bolivar',
       'Departamento_Boyaca', 'Departamento_Caldas', 'Departamento_Caqueta',
       'Departamento_Casanare', 'Departamento_Cauca', 'Departamento_Cesar',
       'Departamento_Choco', 'Departamento_Cordoba',
       'Departamento_Cundinamarca', 'Departamento_Guainia',
       'Departamento_Guajira', 'Departamento_Guaviare', 'Departamento_Huila',
       'Departamento_Magdalena', 'Departamento_Meta', 'Departamento_Narino',
       'Departamento_NorteSantander', 'Departamento_Putumayo',
       'Departamento_Quindio', 'Departamento_Risaralda', 'Departamento_SAI',
       'Departamento_Santander', 'Departamento_Sucre', 'Departamento_Tolima',
       'Departamento_Valle', 'Departamento_Vaupes', 'Departamento_Vichada'],
      dtype='object')
```

In [51]:
```
GEIH_encoding= GEIH_encoding.drop(columns =['JefeHogar_3', 'JefeHogar_4', 'JefeHogar_5
       'JefeHogar_6', 'JefeHogar_7', 'JefeHogar_8', 'JefeHogar_9',
```

```
       'JefeHogar_10', 'JefeHogar_11', 'JefeHogar_12', 'JefeHogar_13', 'JefeHogar_2'])
GEIH_encoding
```

Out[51]:

	Sexo_Hombre	Sexo_Mujer	JefeHogar_JEFEHO	Etnia_Gitano	Etnia_Indígena	Etnia_Negro_mulat
0	1	0	1	0	0	
1	1	0	1	0	0	
2	0	1	0	0	0	
3	1	0	0	0	0	
4	0	1	1	0	0	
...	...	...	...	...	...	
30500	1	0	0	0	0	
30501	1	0	0	0	0	
30502	0	1	1	0	0	
30503	1	0	1	0	0	
30504	0	1	0	0	0	

382461 rows × 56 columns

In [52]:
```
GEIH_logit = pd.concat([GEIH_limpionan, GEIH_encoding], axis=1)
GEIH_logit.columns
```

Out[52]:
```
Index(['DIRECTORIO', 'SECUENCIA_P', 'ORDEN', 'Sexo', 'JefeHogar', 'Etnia',
       'NivelEducativo', 'FactorExpansion', 'Departamento',
       'Ocupacion_informal', 'Ponderacion', 'Sexo_Hombre', 'Sexo_Mujer',
       'JefeHogar_JEFEHO', 'Etnia_Gitano', 'Etnia_Indígena',
       'Etnia_Negro_mulato_afrod_afroc', 'Etnia_Ninguno', 'Etnia_Palenquero',
       'Etnia_Raizal_SAI', 'NivelEducativo_99.0', 'NivelEducativo_B_tecnico',
       'NivelEducativo_Bachillerato', 'NivelEducativo_Doctorado',
       'NivelEducativo_Especializacion', 'NivelEducativo_Maestria',
       'NivelEducativo_Ninguno', 'NivelEducativo_Normalista',
       'NivelEducativo_Preescolar', 'NivelEducativo_Primaria',
       'NivelEducativo_Secundaria', 'NivelEducativo_Tecnica',
       'NivelEducativo_Tecnologica', 'NivelEducativo_Universitaria',
       'Departamento_Amazonas', 'Departamento_Antioquia',
       'Departamento_Arauca', 'Departamento_Atlantico', 'Departamento_Bogota',
       'Departamento_Bolivar', 'Departamento_Boyaca', 'Departamento_Caldas',
       'Departamento_Caqueta', 'Departamento_Casanare', 'Departamento_Cauca',
       'Departamento_Cesar', 'Departamento_Choco', 'Departamento_Cordoba',
       'Departamento_Cundinamarca', 'Departamento_Guainia',
       'Departamento_Guajira', 'Departamento_Guaviare', 'Departamento_Huila',
       'Departamento_Magdalena', 'Departamento_Meta', 'Departamento_Narino',
       'Departamento_NorteSantander', 'Departamento_Putumayo',
       'Departamento_Quindio', 'Departamento_Risaralda', 'Departamento_SAI',
       'Departamento_Santander', 'Departamento_Sucre', 'Departamento_Tolima',
       'Departamento_Valle', 'Departamento_Vaupes', 'Departamento_Vichada'],
      dtype='object')
```

In [53]:
```
GEIH_logit.to_csv("Logit.csv")
```

In [54]:
```
GEIH_2022.to_csv("GEIH2022.csv")
```

In [ ]:
```

```