- Diana Zaray Corado #191025
- Pablo Alejandro Méndez #19195
- Orlando Osberto Cabrera #19943
- Javier Alejandro Mejía Alecio #20304
- Erick Raúl Alvarez Melgar #20900 # Proyecto - Análisis Exploratorio

Guatemala es conocido como el país de la eterna primavera, debido a su diversidad de flora y fauna, además de contar con un clima conocido como templado, es decir no existen climas extremos a lo largo del año. Sin embargo, así como es conocido por su belleza natural, también es fuertemente reconocible por la corrupción y los bajos índices de calidad de vida, y uno de ellos es el de mortalidad, la cual expresa la frecuencia con la cual ocurren las defunciones en una población dada, en el caso de Guatemala, para el año 2019 fue de 4.72% (Datosmacro, 2021), es decir en promedio 5 muertes por cada 1000 habitantes. A continuación se realizará un análisis exploratorio, sobre los datos de defunciones reportados por el Instituto Nacional de Estadística de Guatemala -INE- de los años 2011 a 2020, y dentro de los cuales se pretende encontrar una situación problemática la cual a su vez pueda ser resuelta con los datos analizados.

In [ ]:
```
import pandas as pd
import numpy as np
import scipy.stats as sp
import seaborn as sns
import matplotlib.pyplot as plt

# Estilos
plt.style.use('ggplot')
```

In [ ]:
```
# General functions
def calculate_frequency(data, column, index='index', head = False, use = False):
    data_f = pd.DataFrame({
        'frequency': data[column].value_counts(),
        'relative_frequency (%)': data[column].value_counts(normalize=True)*100,
        'relative_acc_frequency': data[column].value_counts(normalize=True).cumsum()
    })
    data_f.reset_index(level=[0], inplace=True)
    data_f.rename(columns={index:column}, inplace=True)
    if head:
        left_aligned_df = data_f.head(20).style.set_properties(**{'text-align': 'cente
    else:
        left_aligned_df = data_f.style.set_properties(**{'text-align': 'center'})
    display(left_aligned_df)

    if use:
        return data_f

    return None

def is_normal(column, tolerancia=0.05):
    return sp.normaltest(column).pvalue > tolerancia
```

In [ ]:
```
# Cargar los datos
deaths = pd.read_csv('final.csv')
# Como ya se cuenta con la causa de muerte (categorizada) la descripción sale sobrando
# De igual forma, el periodo de edad se puede obtener mediante la edad así que es info
deaths = deaths.loc[:, deaths.columns != 'Perdif']
deaths = deaths.loc[:, deaths.columns != 'caudef.descrip']
```

```
C:\Users\Orlando\AppData\Local\Temp\ipykernel_17676\3782891449.py:2: DtypeWarning: Co
lumns (7,12,18,22,26) have mixed types. Specify dtype option on import or set low_mem
ory=False.
  deaths = pd.read_csv('final.csv')
```

Exploración de los datos

// Rúbrica

- Estudia las variables cuantitativas mediante técnicas de estadística descriptiva
- Hace gráficos exploratorios como histogramas, diagramas de cajas y bigotes, gráficos de dispersión, que ayudan a explicar los datos.
- Analiza las correlaciones entre las variables, trata de explicar los outliers (puntos atípicos) y toma decisiones acertadas ante la presencia de valores faltantes.
- Estudia las variables categóricas.
- Elabora gráficos de barra, tablas de frecuencia y de proporciones
- Explica muy bien todos los procedimientos y los hallazgos que va haciendo.
- Determina el mejor número de clusters a utilizar.
- Hace el agrupamiento con cualquiera de los algoritmos estudiados.
- Verifica la calidad del agrupamiento usando el método de la silueta.
- Interpreta los grupos, usando para eso las variables numéricas y categóricas dentro de cada grupo.

¿Cuáles son los datos?

Comience describiendo cuantas variables y observaciones tiene disponibles, el tipo de cada una de las variables.

In [ ]:
```
deaths.shape
```

Out[ ]:
```
(809296, 28)
```

Para la elaboración del análisis se cuenta con los datos encontrados en la página del Instituto Nacional de Estadística de Guatemala -INE- para las defunciones del rango de años desde 2011 a 2020. Al unificar los datos provistos en cada uno de los años, en total, se cuenta con 809296 observaciones y con un total de 28 variables.

Clasificación de variables

Cuantitativas discretas

- EDADIF → Edad del difunto

Cualitativas Ordinal

- ESCODIF → Escolaridad del difunto

Cualitativa Nominal

- DEPREG → Departamento de registro
- MUPREG → Municipio de registro
- MESREG → Mes de registro
- AÑOREG → Año de registro
- DEPOCU → Departamento de ocurrencia
- MUPOCU → Municipio de ocurrencia
- AREGAG → Área geográfica de ocurrencia
- SEXO
- DIAOCU → Día de ocurrencia
- MESOCU → Mes de ocurrencia
- AÑOOCU → Año de ocurrencia
- PUEDIF → Pueblo de pertenencia del difunto
- ECIDIF → Estado civil del difunto
- CIUODIF → Ocupación del difunto
- PNADIF → País de nacimiento
- DNADIF → Departamento de nacimiento
- MNADIF → Municipio de nacimiento
- NACDIF → Nacionalidad del difunto
- PREDIF → País de residencia
- DREDIF → Departamento de residencia
- MREDIF → Municipio de residencia
- CAUDEF → Causa de defunción
- ASIST → Asistencia recibida
- OCUR → Sitio de ocurrencia
- CERDEF → Quién certifica

Preprocesamiento

Análisis de valores atípicos y tratamientos de valores faltantes

In [ ]:
```
# Debido a que edad se tomará como variable cuantitativa será necesario reeplazar el v
# los datos estén en el mismo formato
deaths['Edadif'] = deaths['Edadif'].replace(['Ignorado'], -1)
deaths['Edadif'] = deaths['Edadif'].astype('int')
```

Variables cuantitativas y su distribución

Haga un resumen de las variables numéricas e investigue si siguen una distribución normal.

In [ ]:
```
cuantitative = ["Edadif"]
# Cambiando el valor de "Ignorado" por -1, para evitar errores en el futuro
clean_cuantitative = deaths[cuantitative].copy()
clean_cuantitative['Edadif'] = [-1 if year == "Ignorado" else int (year) for year in c

clean_cuantitative.describe()
```

Out[ ]:

	Edadif
count	809296.000000
mean	53.701852
std	28.428669
min	-1.000000
25%	31.000000
50%	60.000000
75%	78.000000
max	122.000000

In [ ]:
```
for col in cuantitative:
    if is_normal(clean_cuantitative[col]):
        print(f'{col} tiene una distribución normal.')
    else:
        print(f'{col} no tiene una distribución normal.')
del clean_cuantitative
```

```
Edadif no tiene una distribución normal.
```

Variables cualitativas y su frecuencia

Elabore tablas de frecuencia para las variables categóricas, escriba lo que vaya encontrando. // Acá también pueden ir preguntas

Departamento de registro

In [ ]:
```
calculate_frequency(deaths, 'Depreg')
```

	Depreg	frequency	relative_frequency (%)	relative_acc_frequency
0	Guatemala	238866	29.515282	0.295153
1	Quetzaltenango	49558	6.123594	0.356389
2	Alta Verapaz	49071	6.063418	0.417023
3	San Marcos	47997	5.930710	0.476330
4	Escuintla	46155	5.703105	0.533361
5	Huehuetenango	45374	5.606601	0.589427
6	Quiché	32499	4.015712	0.629584
7	Suchitepequez	30460	3.763765	0.667222
8	Chimaltenango	28316	3.498843	0.702210
9	Jutiapa	24779	3.061797	0.732828
10	Santa Rosa	22106	2.731510	0.760143
11	Chiquimula	20496	2.532572	0.785469
12	Izabal	19278	2.382070	0.809290
13	Totonicapán	19217	2.374533	0.833035
14	Retalhuleu	17160	2.120361	0.854239
15	Petén	16743	2.068835	0.874927
16	Jalapa	16041	1.982093	0.894748
17	Sololá	15280	1.888061	0.913629
18	Zacapa	14402	1.779571	0.931424
19	Sacatepéquez	13716	1.694806	0.948372
20	Baja Verapaz	12056	1.489690	0.963269
21	El Progreso	8385	1.036086	0.973630
22	Quiche	7402	0.914622	0.982776
23	Totonicapan	4419	0.546030	0.988237
24	Peten	3415	0.421972	0.992456
25	Sacatepequez	3055	0.377489	0.996231
26	Solola	3050	0.376871	1.000000

Municipio de registro

In [ ]:
```
# Para no mostrar muchos datos, solo poner un head de 20
calculate_frequency(deaths, 'Mupreg', head=True)
```

	Mupreg	frequency	relative_frequency (%)	relative_acc_frequency
0	Guatemala	200377	24.759421	0.247594
1	Quetzaltenango	18111	2.237871	0.269973
2	Cobán	12975	1.603245	0.286005
3	Escuintla	12595	1.556291	0.301568
4	Mazatenango	8470	1.046589	0.312034
5	Chiquimula	8411	1.039298	0.322427
6	Huehuetenango	8278	1.022864	0.332656
7	Coatepeque	8222	1.015945	0.342815
8	Cuilapa	8165	1.008902	0.352904
9	Jalapa	8088	0.999387	0.362898
10	San Pedro Carchá	7955	0.982953	0.372728
11	Chimaltenango	7920	0.978628	0.382514
12	Retalhuleu	7778	0.961082	0.392125
13	Totonicapán	7659	0.946378	0.401589
14	Puerto Barrios	7369	0.910544	0.410694
15	Jutiapa	7052	0.871375	0.419408
16	Antigua Guatemala	6032	0.745339	0.426861
17	Santa Cruz del Quiché	6008	0.742374	0.434285
18	Amatitlán	5912	0.730511	0.441590
19	Chichicastenango	5896	0.728534	0.448875

Mes de registro

In [ ]:
```
calculate_frequency(deaths, 'Mesreg')
```

	Mesreg	frequency	relative_frequency (%)	relative_acc_frequency
0	Julio	74719	9.232592	0.092326
1	Enero	73762	9.114341	0.183469
2	Agosto	70181	8.671858	0.270188
3	Octubre	69355	8.569794	0.355886
4	Mayo	67973	8.399028	0.439876
5	Noviembre	66530	8.220725	0.522083
6	Septiembre	66261	8.187486	0.603958
7	Junio	65637	8.110382	0.685062
8	Abril	64861	8.014497	0.765207
9	Diciembre	64323	7.948019	0.844687
10	Marzo	63145	7.802460	0.922712
11	Febrero	62549	7.728816	1.000000

Año de registro

In [ ]:
```
calculate_frequency(deaths, 'Añoreg')
```

	Añoreg	frequency	relative_frequency (%)	relative_acc_frequency
0	2020	95100	11.750954	0.117510
1	2019	85476	10.561772	0.223127
2	2018	82755	10.225554	0.325383
3	2016	82420	10.184160	0.427224
4	2017	81475	10.067392	0.527898
5	2015	81040	10.013641	0.628035
6	2014	77582	9.586357	0.723898
7	2013	76618	9.467241	0.818571
8	2012	72115	8.910831	0.907679
9	2011	71144	8.790850	0.995588
10	2021	3571	0.441248	1.000000

Departamento de ocurrencia

In [ ]:
```
calculate_frequency(deaths, 'Depocu')
```

	Depocu	frequency	relative_frequency (%)	relative_acc_frequency
0	Guatemala	237388	29.332655	0.293327
1	Quetzaltenango	50797	6.276690	0.356093
2	Alta Verapaz	48846	6.035616	0.416450
3	Escuintla	47028	5.810976	0.474559
4	San Marcos	47003	5.807887	0.532638
5	Huehuetenango	45282	5.595233	0.588591
6	Quiché	32167	3.974689	0.628337
7	Suchitepequez	30514	3.770438	0.666042
8	Chimaltenango	28161	3.479691	0.700839
9	Jutiapa	23582	2.913891	0.729978
10	Santa Rosa	22661	2.800088	0.757979
11	Chiquimula	20476	2.530100	0.783280
12	Izabal	19333	2.388866	0.807168
13	Totonicapán	19217	2.374533	0.830914
14	Petén	17043	2.105904	0.851973
15	Retalhuleu	16726	2.066735	0.872640
16	Jalapa	15791	1.951202	0.892152
17	Sololá	15722	1.942676	0.911579
18	Sacatepéquez	15024	1.856428	0.930143
19	Zacapa	14611	1.805396	0.948197
20	Baja Verapaz	11849	1.464112	0.962838
21	El Progreso	8684	1.073031	0.973568
22	Quiche	7352	0.908444	0.982653
23	Totonicapan	4353	0.537875	0.988032
24	Peten	3439	0.424937	0.992281
25	Sacatepequez	3192	0.394417	0.996225
26	Solola	3055	0.377489	1.000000

Municipio de ocurrencia

In [ ]:
```
calculate_frequency(deaths, 'Mupocu', head=True)
```

	Mupocu	frequency	relative_frequency (%)	relative_acc_frequency
0	Guatemala	158214	19.549584	0.195496
1	Mixco	20192	2.495008	0.220446
2	Escuintla	19321	2.387384	0.244320
3	Quetzaltenango	18274	2.258012	0.266900
4	Villa Nueva	13046	1.612018	0.283020
5	Cobán	12886	1.592248	0.298943
6	Mazatenango	10429	1.288651	0.311829
7	Coatepeque	9326	1.152360	0.323353
8	Huehuetenango	9296	1.148653	0.334839
9	Amatitlán	9182	1.134566	0.346185
10	Cuilapa	8692	1.074020	0.356925
11	Jalapa	8249	1.019281	0.367118
12	Chiquimula	8198	1.012979	0.377248
13	Puerto Barrios	8157	1.007913	0.387327
14	Chimaltenango	8083	0.998769	0.397314
15	San Pedro Carchá	7998	0.988266	0.407197
16	Jutiapa	7966	0.984312	0.417040
17	Totonicapán	7889	0.974798	0.426788
18	San Juan Sacatepéquez	7835	0.968125	0.436469
19	Antigua Guatemala	7340	0.906961	0.445539

Área geográfica de ocurrencia

In [ ]:
```
calculate_frequency(deaths, 'Areag')
```

	Areag	frequency	relative_frequency (%)	relative_acc_frequency
0	Urbano	295064	54.177561	0.541776
1	Rural	239389	43.954912	0.981325
2	Ignorado	10171	1.867527	1.000000

Sexo

In [ ]:
```
calculate_frequency(deaths, 'Sexo')
```

	Sexo	frequency	relative_frequency (%)	relative_acc_frequency
0	Hombre	454900	56.209347	0.562093
1	Mujer	354396	43.790653	1.000000

Día de ocurrencia

In [ ]:
```
calculate_frequency(deaths, 'Diaocu')
```

	Diaocu	frequency	relative_frequency (%)	relative_acc_frequency
0	1	27544	3.403452	0.034035
1	5	27062	3.343894	0.067473
2	4	27045	3.341793	0.100891
3	2	27043	3.341546	0.134307
4	25	26899	3.323753	0.167544
5	10	26857	3.318563	0.200730
6	15	26822	3.314239	0.233872
7	6	26769	3.307690	0.266949
8	3	26736	3.303612	0.299985
9	16	26682	3.296940	0.332955
10	20	26644	3.292244	0.365877
11	17	26607	3.287672	0.398754
12	27	26599	3.286684	0.431621
13	24	26566	3.282606	0.464447
14	28	26555	3.281247	0.497259
15	7	26525	3.277540	0.530035
16	14	26520	3.276922	0.562804
17	11	26456	3.269014	0.595494
18	12	26440	3.267037	0.628164
19	23	26428	3.265554	0.660820
20	9	26389	3.260735	0.693427
21	18	26369	3.258264	0.726010
22	8	26319	3.252086	0.758531
23	26	26319	3.252086	0.791052
24	22	26311	3.251097	0.823563
25	13	26290	3.248502	0.856048
26	19	26196	3.236887	0.888417
27	21	26050	3.218847	0.920605
28	29	24736	3.056484	0.951170
29	30	24316	3.004587	0.981216
30	31	15202	1.878423	1.000000

Mes de ocurrencia

In [ ]:
```
calculate_frequency(deaths, 'Mesocu')
```

	Mesocu	frequency	relative_frequency (%)	relative_acc_frequency
0	Julio	73553	9.088516	0.090885
1	Agosto	69865	8.632812	0.177213
2	Diciembre	69332	8.566952	0.262883
3	Octubre	68468	8.460193	0.347485
4	Enero	68213	8.428684	0.431772
5	Junio	67813	8.379258	0.515564
6	Marzo	66918	8.268668	0.598251
7	Septiembre	66696	8.241237	0.680663
8	Mayo	66551	8.223320	0.762896
9	Noviembre	66100	8.167593	0.844572
10	Abril	65316	8.070718	0.925280
11	Febrero	60471	7.472050	1.000000

Año de ocurrencia

In [ ]:
```
calculate_frequency(deaths, 'Añoocu')
```

	Añoocu	frequency	relative_frequency (%)	relative_acc_frequency
0	2020	96001	11.862285	0.118623
1	2019	85600	10.577094	0.224394
2	2018	83071	10.264600	0.327040
3	2016	82565	10.202077	0.429061
4	2017	81726	10.098407	0.530045
5	2015	80876	9.993377	0.629978
6	2014	77807	9.614158	0.726120
7	2013	76639	9.469836	0.820818
8	2012	72657	8.977803	0.910596
9	2011	72354	8.940363	1.000000

Pueblo de pertenencia

In [ ]:
```
calculate_frequency(deaths, 'Puedif')
```

	Puedif	frequency	relative_frequency (%)	relative_acc_frequency
0	No indigena	404695	50.005808	0.500058
1	Indigena	220157	27.203520	0.772093
2	Ignorado	177070	21.879510	0.990888
3	Otro	7374	0.911162	1.000000

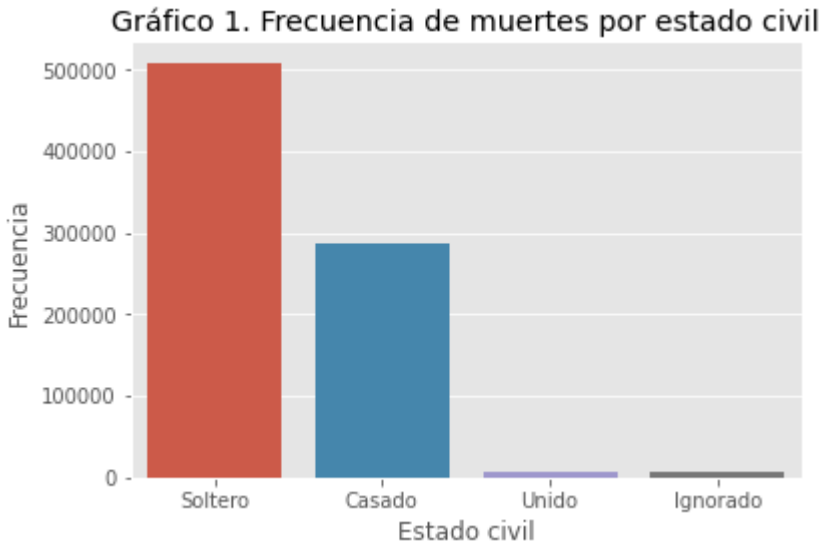
Estado civil del difunto

¿Cuál es el estado civil predominante en las muertes?

Como se puede notar en el gráfico inferior, el estado civil predominante en las defunciones es el soltero. Este es un resultado esperado, ya que dentro de los datos se encuentran personas de todas las edades, por lo que muchos de los fallecidos son aún niños cuando mueren, por lo tal, su estado civil es soltero, y también están todos aquellos que mueren a una edad relativamente joven por lo que aún no han contraído matrimonio.

In [ ]:
```
estado = calculate_frequency(deaths, 'Ecidif', use=True)

sns.barplot(x='Ecidif',y='frequency',data=estado);
plt.title('Gráfico 1. Frecuencia de muertes por estado civil')
plt.ylabel('Frecuencia')
plt.xlabel('Estado civil')

plt.show()

del estado
```

	Ecidif	frequency	relative_frequency (%)	relative_acc_frequency
0	Soltero	508861	62.876994	0.628770
1	Casado	286024	35.342322	0.982193
2	Unido	7923	0.978999	0.991983
3	Ignorado	6488	0.801684	1.000000



Escolaridad del difunto

In [ ]:
```
calculate_frequency(deaths, 'Escodif')
```

	Escodif	frequency	relative_frequency (%)	relative_acc_frequency
0	Ninguna	438866	54.228119	0.542281
1	Primaria	207679	25.661686	0.798898
2	Ignorado	64824	8.009925	0.878997
3	Diversificado	49919	6.168201	0.940679
4	Básico	35846	4.429282	0.984972
5	Universitario	11944	1.475851	0.999731
6	Post grado	218	0.026937	1.000000

Ocupación del difunto

In [ ]:
```
calculate_frequency(deaths, 'Ciuodif', head=True)
```

	Ciuodif	frequency	relative_frequency (%)	relative_acc_frequency
0	No especificado en otro grupo	404982	50.041270	0.500413
1	Peones agropecuarios, pesqueros y forestales	127064	15.700559	0.657418
2	Agricultores y trabajadores calificados de explotaciones agropecuarias con destino al mercado	75575	9.338363	0.750802
3	Ignorado	61801	7.636390	0.827166
4	Vendedores	27354	3.379975	0.860966
5	Oficiales y operarios de la construcción excluyendo electricistas	17402	2.150264	0.882468
6	Conductores de vehículos y operadores de equipos pesados móviles	13684	1.690852	0.899377
7	Operarios y oficiales de procesamiento de alimentos, de la confección, ebanistas, otros artesanos y afines	9466	1.169659	0.911073
8	Profesionales de la enseñanza	8543	1.055609	0.921629
9	Oficiales y operarios de la metalurgia, la construcción mecánica y afines	8338	1.030278	0.931932
10	Operadores de instalaciones fijas y máquinas	6058	0.748552	0.939418
11	Empleados en trato directo con el público	4582	0.566171	0.945079
12	Empleados contables y encargados del registro de materiales	4393	0.542817	0.950508
13	Profesionales en derecho, en ciencias sociales y culturales	3773	0.466208	0.955170
14	Artesanos y operarios de las artes gráficas	3330	0.411469	0.959284
15	Oficinistas	3131	0.386879	0.963153
16	Trabajadores de los servicios personales	3128	0.386509	0.967018
17	Profesionales de las ciencias y la ingeniería de nivel medio	2690	0.332388	0.970342
18	Personal de los servicios de protección	2670	0.329916	0.973641
19	Peones de la minería, la construcción, la industria manufacturera y el transporte	2589	0.319908	0.976840

País de nacimiento

In [ ]:
```
calculate_frequency(deaths, 'Pnadif', head=True)
```

	Pnadif	frequency	relative_frequency (%)	relative_acc_frequency
0	Guatemala	725117	98.395396	0.983954
1	Ignorado	4891	0.663689	0.990591
2	El Salvador	2672	0.362579	0.994217
3	Honduras	963	0.130675	0.995523
4	Nicaragua	604	0.081960	0.996343
5	México	567	0.076940	0.997112
6	Estados Unidos De América	364	0.049393	0.997606
7	España	239	0.032431	0.997931
8	Estados Unidos de América	180	0.024425	0.998175
9	Colombia	149	0.020219	0.998377
10	Belice	120	0.016284	0.998540
11	Alemania	115	0.015605	0.998696
12	Italia	89	0.012077	0.998817
13	Costa Rica	84	0.011398	0.998931
14	Cuba	82	0.011127	0.999042
15	China	64	0.008685	0.999129
16	Canadá	49	0.006649	0.999195
17	Argentina	47	0.006378	0.999259
18	Francia	46	0.006242	0.999322
19	Perú	37	0.005021	0.999372

Departamento de nacimiento

In [ ]:
```
calculate_frequency(deaths, 'Dnadif')
```

	Dnadif	frequency	relative_frequency (%)	relative_acc_frequency
0	Guatemala	135407	16.731455	0.167315
1	San Marcos	58711	7.254577	0.239860
2	Alta Verapaz	55594	6.869427	0.308555
3	Quetzaltenango	50619	6.254695	0.371102
4	Huehuetenango	50190	6.201686	0.433118
5	Escuintla	41228	5.094304	0.484061
6	Jutiapa	40376	4.989028	0.533952
7	Quiché	39436	4.872877	0.582681
8	Suchitepequez	34727	4.291013	0.625591
9	Chimaltenango	34292	4.237263	0.667963
10	Santa Rosa	32785	4.051052	0.708474
11	Chiquimula	27616	3.412349	0.742597
12	Totonicapán	23423	2.894244	0.771540
13	Jalapa	23029	2.845560	0.799995
14	Zacapa	19440	2.402088	0.824016
15	Retalhuleu	18802	2.323254	0.847249
16	Baja Verapaz	17978	2.221437	0.869463
17	Sololá	17024	2.103557	0.890499
18	Izabal	15183	1.876075	0.909259
19	Sacatepéquez	14428	1.782784	0.927087
20	El Progreso	14376	1.776359	0.944851
21	Quiche	8784	1.085388	0.955705
22	Petén	8590	1.061416	0.966319
23	Extranjero	7700	0.951444	0.975833
24	Ignorado	6043	0.746698	0.983300
25	Totonicapan	5240	0.647476	0.989775
26	Solola	3406	0.420860	0.993984
27	Sacatepequez	3124	0.386015	0.997844
28	Peten	1741	0.215125	0.999995
29	9999	4	0.000494	1.000000

Municipio de nacimiento

In [ ]:
```
calculate_frequency(deaths, 'Mnadif', head=True)
```

	Mnadif	frequency	relative_frequency (%)	relative_acc_frequency
0	Guatemala	85342	10.545215	0.105452
1	San Pedro Carchá	12515	1.546406	0.120916
2	Escuintla	11920	1.472885	0.135645
3	Quetzaltenango	11846	1.463741	0.150282
4	Cobán	10947	1.352657	0.163809
5	Jutiapa	10425	1.288157	0.176691
6	Jalapa	9933	1.227363	0.188964
7	Totonicapán	9501	1.173983	0.200704
8	San Juan Sacatepéquez	8772	1.083905	0.211543
9	Chiquimula	7752	0.957870	0.221122
10	Extranjero	7700	0.951444	0.230636
11	Mazatenango	7675	0.948355	0.240120
12	Coatepeque	7555	0.933527	0.249455
13	Chichicastenango	7296	0.901524	0.258470
14	Zacapa	7165	0.885337	0.267324
15	Huehuetenango	7160	0.884720	0.276171
16	Momostenango	7064	0.872857	0.284899
17	Retalhuleu	7016	0.866926	0.293569
18	Tiquisate	6622	0.818242	0.301751
19	Santa Cruz del Quiché	6616	0.817501	0.309926

Nacionalidad del difunto

In [ ]:
```
calculate_frequency(deaths, 'Nacdif')
```

	Nacdif	frequency	relative_frequency (%)	relative_acc_frequency
0	Guatemala	795884	98.342757	0.983428
1	Ignorado	5711	0.705675	0.990484
2	El Salvador	2944	0.363773	0.994122
3	Honduras	1083	0.133820	0.995460
4	Nicaragua	666	0.082294	0.996283
5	México	625	0.077228	0.997055
6	Estados Unidos De América	364	0.044977	0.997505
7	España	263	0.032497	0.997830
8	Estados Unidos de América	239	0.029532	0.998126
9	Colombia	175	0.021624	0.998342
10	Belice	142	0.017546	0.998517
11	Alemania	122	0.015075	0.998668
12	Italia	102	0.012604	0.998794
13	Costa Rica	92	0.011368	0.998908
14	Cuba	91	0.011244	0.999020
15	China	71	0.008773	0.999108
16	Canadá	58	0.007167	0.999180
17	Argentina	55	0.006796	0.999247
18	Francia	51	0.006302	0.999311
19	Perú	40	0.004943	0.999360
20	Rep. Boliv. De Venezuela	31	0.003830	0.999398
21	Chile	30	0.003707	0.999435
22	República De Corea	30	0.003707	0.999472
23	Corea Del Sur	28	0.003460	0.999507
24	Panamá	28	0.003460	0.999542
25	G. Bretaña E Irl. Del N.	23	0.002842	0.999570
26	Suiza	22	0.002718	0.999597
27	Ecuador	20	0.002471	0.999622
28	Polonia	20	0.002471	0.999647
29	República Dominicana	16	0.001977	0.999666
30	Brasil	16	0.001977	0.999686
31	Puerto Rico	13	0.001606	0.999702
32	Bolivia	13	0.001606	0.999718

	Nacdif	frequency	relative_frequency (%)	relative_acc_frequency
33	Bélgica	12	0.001483	0.999733
34	Israel	11	0.001359	0.999747
35	Austria	11	0.001359	0.999760
36	Jordania	10	0.001236	0.999773
37	Holanda (Paises Bajos)	10	0.001236	0.999785
38	Holanda	9	0.001112	0.999796
39	Gran Bretaña	9	0.001112	0.999807
40	Filipinas	8	0.000989	0.999817
41	Rusia	8	0.000989	0.999827
42	Japón	7	0.000865	0.999836
43	Uruguay	7	0.000865	0.999844
44	Palestina	6	0.000741	0.999852
45	Turquia	5	0.000618	0.999858
46	India	5	0.000618	0.999864
47	Hungría	5	0.000618	0.999870
48	Paraguay	5	0.000618	0.999876
49	Ucrania	5	0.000618	0.999883
50	Suecia	5	0.000618	0.999889
51	Reino Unido	4	0.000494	0.999894
52	Venezuela	4	0.000494	0.999899
53	Australia	4	0.000494	0.999904
54	Siria	4	0.000494	0.999909
55	Líbano	4	0.000494	0.999914
56	Grecia	4	0.000494	0.999918
57	Inglaterra	3	0.000371	0.999922
58	Armenia	3	0.000371	0.999926
59	Trinidad Y Tobago	3	0.000371	0.999930
60	Portugal	3	0.000371	0.999933
61	República Checa	3	0.000371	0.999937
62	Noruega	3	0.000371	0.999941
63	Marruecos	3	0.000371	0.999944
64	Estonia	3	0.000371	0.999948
65	Egipto	3	0.000371	0.999952
66	Dinamarca	3	0.000371	0.999956

	Nacdif	frequency	relative_frequency (%)	relative_acc_frequency
67	Hong Kong, China	2	0.000247	0.999958
68	Corea	2	0.000247	0.999960
69	Hong Kong	2	0.000247	0.999963
70	Guam	2	0.000247	0.999965
71	Jamaica	2	0.000247	0.999968
72	Angola	2	0.000247	0.999970
73	Irlanda	2	0.000247	0.999973
74	Macao, China	1	0.000124	0.999974
75	República Unida de Tanzania	1	0.000124	0.999975
76	Guinea	1	0.000124	0.999977
77	Checoslovaquia	1	0.000124	0.999978
78	Bahamas	1	0.000124	0.999979
79	Chad	1	0.000124	0.999980
80	Nueva Zelanda	1	0.000124	0.999981
81	Tailandia	1	0.000124	0.999983
82	Barbados	1	0.000124	0.999984
83	Bulgaria	1	0.000124	0.999985
84	Sudáfrica	1	0.000124	0.999986
85	República de China (Taiwan)	1	0.000124	0.999988
86	Rumania	1	0.000124	0.999989
87	Malasia	1	0.000124	0.999990
88	Bermudas	1	0.000124	0.999991
89	Serbia	1	0.000124	0.999993
90	Finlandia	1	0.000124	0.999994
91	Eslovenia	1	0.000124	0.999995
92	Camerún	1	0.000124	0.999996
93	Indonesia	1	0.000124	0.999998
94	Bangladesh	1	0.000124	0.999999
95	Lituania	1	0.000124	1.000000

País de residencia

In [ ]:
```
calculate_frequency(deaths, 'Predif')
```

	Predif	frequency	relative_frequency (%)	relative_acc_frequency
0	Guatemala	679028	92.141308	0.921413
1	Ignorado	56788	7.705898	0.998472
2	El Salvador	340	0.046137	0.998933
3	Estados Unidos De América	229	0.031074	0.999244
4	Honduras	131	0.017776	0.999422
5	México	125	0.016962	0.999592
6	721	92	0.012484	0.999716
7	Nicaragua	70	0.009499	0.999811
8	Belice	48	0.006513	0.999877
9	Estados Unidos de América	33	0.004478	0.999921
10	484	8	0.001086	0.999932
11	Colombia	7	0.000950	0.999942
12	Canadá	4	0.000543	0.999947
13	Francia	4	0.000543	0.999953
14	Alemania	3	0.000407	0.999957
15	Chile	2	0.000271	0.999959
16	India	2	0.000271	0.999962
17	Costa Rica	2	0.000271	0.999965
18	G. Bretaña E Irl. Del N.	2	0.000271	0.999967
19	Italia	2	0.000271	0.999970
20	España	2	0.000271	0.999973
21	República Dominicana	1	0.000136	0.999974
22	Ecuador	1	0.000136	0.999976
23	Cuba	1	0.000136	0.999977
24	Jordania	1	0.000136	0.999978
25	Gran Bretaña	1	0.000136	0.999980
26	Ucrania	1	0.000136	0.999981
27	Venezuela	1	0.000136	0.999982
28	Rusia	1	0.000136	0.999984
29	República de Corea	1	0.000136	0.999985
30	Puerto Rico	1	0.000136	0.999986
31	Panamá	1	0.000136	0.999988
32	1045	1	0.000136	0.999989

	Predif	frequency	relative_frequency (%)	relative_acc_frequency
33	458	1	0.000136	0.999991
34	Austria	1	0.000136	0.999992
35	Polonia	1	0.000136	0.999993
36	Israel	1	0.000136	0.999995
37	Guadalupe	1	0.000136	0.999996
38	Rep. Boliv. De Venezuela	1	0.000136	0.999997
39	República Unida de Tanzania	1	0.000136	0.999999
40	Holanda	1	0.000136	1.000000

Departamento de residencia

In [ ]:
```
calculate_frequency(deaths, 'Dredif')
```

	Dredif	frequency	relative_frequency (%)	relative_acc_frequency
0	Guatemala	186794	23.081048	0.230810
1	Ignorado	62239	7.690511	0.307716
2	San Marcos	47246	5.837913	0.366095
3	Alta Verapaz	45726	5.650096	0.422596
4	Escuintla	43787	5.410505	0.476701
5	Quetzaltenango	43312	5.351812	0.530219
6	Huehuetenango	42618	5.266058	0.582879
7	Quiché	30555	3.775504	0.620634
8	Suchitepequez	29625	3.660589	0.657240
9	Chimaltenango	27865	3.443116	0.691672
10	Jutiapa	26835	3.315845	0.724830
11	Santa Rosa	21434	2.648475	0.751315
12	Izabal	20786	2.568405	0.776999
13	Chiquimula	20474	2.529853	0.802297
14	Totonicapán	19056	2.354639	0.825844
15	Retalhuleu	18256	2.255788	0.848402
16	Jalapa	17047	2.106399	0.869466
17	Petén	16607	2.052030	0.889986
18	Zacapa	14712	1.817876	0.908165
19	Sololá	14584	1.802060	0.926185
20	Sacatepéquez	13599	1.680349	0.942989
21	Baja Verapaz	13164	1.626599	0.959255
22	El Progreso	10411	1.286427	0.972119
23	Quiche	7306	0.902760	0.981147
24	Totonicapan	4354	0.537998	0.986527
25	Peten	3441	0.425184	0.990778
26	Solola	3107	0.383914	0.994618
27	Sacatepequez	3086	0.381319	0.998431
28	Extranjero	1270	0.156927	1.000000

Municipio de residencia

In [ ]:
```
calculate_frequency(deaths, 'Mredif', head=True)
```

	Mredif	frequency	relative_frequency (%)	relative_acc_frequency
0	Guatemala	90500	11.182559	0.111826
1	Ignorado	62239	7.690511	0.188731
2	Mixco	23527	2.907095	0.217802
3	Villa Nueva	17847	2.205250	0.239854
4	Quetzaltenango	11145	1.377123	0.253625
5	Escuintla	10471	1.293841	0.266564
6	San Juan Sacatepéquez	9344	1.154584	0.278110
7	San Pedro Carchá	9293	1.148282	0.289592
8	Cobán	8461	1.045477	0.300047
9	Jalapa	7855	0.970597	0.309753
10	Totonicapán	7463	0.922160	0.318975
11	Villa Canales	7392	0.913386	0.328109
12	Jutiapa	6930	0.856300	0.336672
13	Santa Lucía Cotzumalguapa	6859	0.847527	0.345147
14	Puerto Barrios	6800	0.840236	0.353549
15	Amatitlán	6537	0.807739	0.361627
16	Coatepeque	6441	0.795877	0.369585
17	Chichicastenango	6132	0.757696	0.377162
18	Retalhuleu	6000	0.741385	0.384576
19	Chiquimula	5932	0.732983	0.391906

Causa de defunción

In [ ]:
```
calculate_frequency(deaths, 'Caudef', head=True)
```

	Caudef	frequency	relative_frequency (%)	relative_acc_frequency
0	I219	55564	6.865720	0.068657
1	J189	47766	5.902167	0.127679
2	R98X	32185	3.976913	0.167448
3	E149	31541	3.897338	0.206421
4	X599	25770	3.184249	0.238264
5	K746	22998	2.841729	0.266681
6	R54X	18621	2.300889	0.289690
7	X959	17982	2.221931	0.311909
8	I64X	17212	2.126787	0.333177
9	R99X	15682	1.937734	0.352555
10	J180	15219	1.880523	0.371360
11	N189	14007	1.730764	0.388667
12	V899	13278	1.640685	0.405074
13	C169	12984	1.604357	0.421118
14	C229	11115	1.373416	0.434852
15	E46X	9954	1.229958	0.447152
16	E119	9909	1.224398	0.459396
17	N19X	9800	1.210929	0.471505
18	A099	9672	1.195113	0.483456
19	A419	9598	1.185969	0.495316

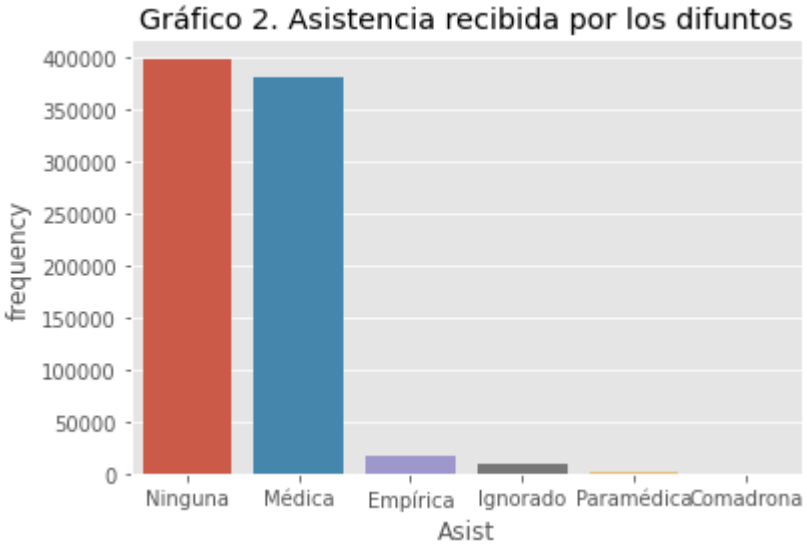
Asistencia recibida

¿Cuál es el tipo de asistencia recibida predominante en las defunciones?

Como se puede observar en el gráfico 2, en un 48% de las defunciones no se recibió ningún tipo de asistencia, sin embargo, algo interesante a notar es que un 47% de los difuntos recibieron asistencia médica. Dejando así solo aproximadamente un 1% de diferencia entre la asistencia predominante y la segunda más utilizada.

In [ ]:
```
assist = calculate_frequency(deaths, 'Asist', use=True)

sns.barplot(x='Asist',y='frequency',data=assist);
plt.title('Gráfico 2. Asistencia recibida por los difuntos')
plt.show()

del assist
```

	Asist	frequency	relative_frequency (%)	relative_acc_frequency
0	Ninguna	397811	49.155192	0.491552
1	Médica	380022	46.957108	0.961123
2	Empírica	18318	2.263449	0.983757
3	Ignorado	9482	1.171636	0.995474
4	Paramédica	3028	0.374152	0.999215
5	Comadrona	635	0.078463	1.000000



Sitio de ocurrencia

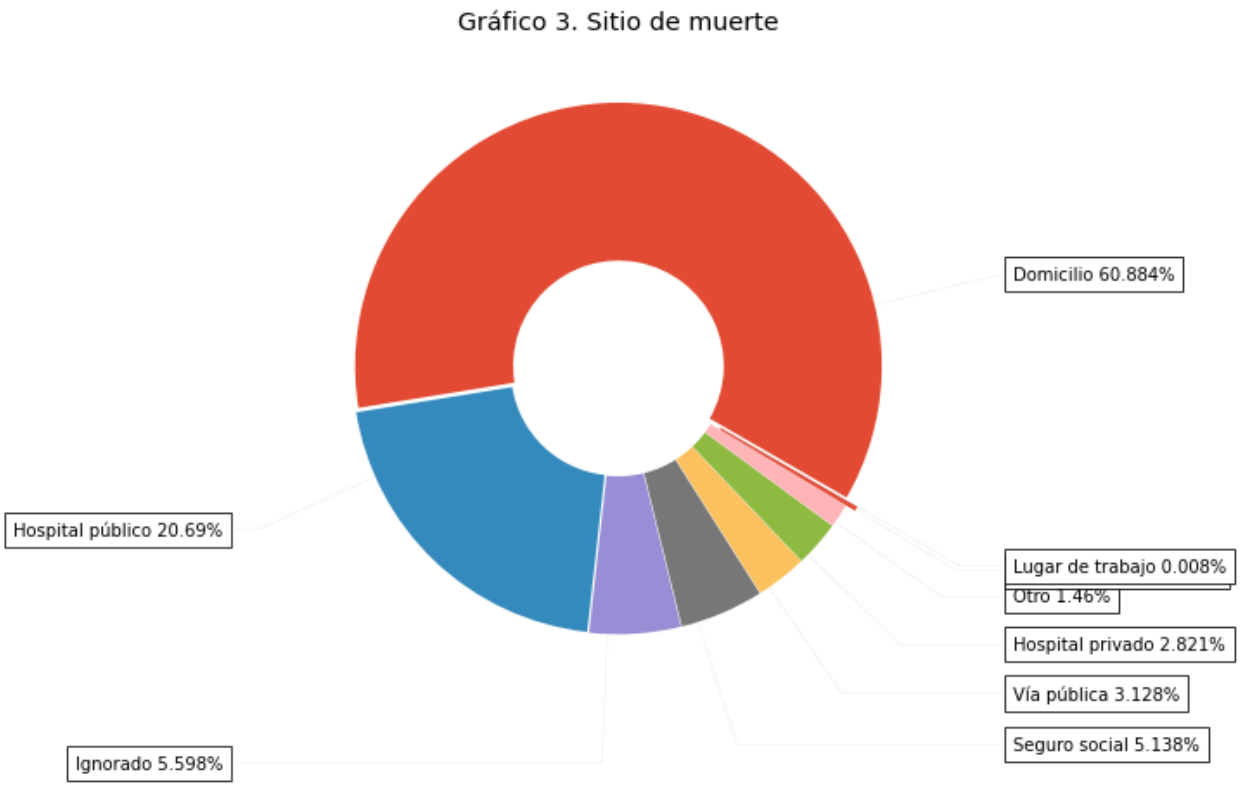
¿Cuál es la proporción de personas que mueren en sus casas con respecto a las que mueren en hospitales?

Como se puede observar en el **gráfico 3** el principal sitio de ocurrencia de las muertes en en los domicilios de los difuntos, estos representando un 69% de las muertes reportadas, y es destacable notar que únicamente un 21% de los difuntos fallecieron en el hospital.

In [ ]:
```
ocur = calculate_frequency(deaths, 'Ocur', use=True)

fig, ax = plt.subplots(figsize=(7, 7), subplot_kw=dict(aspect="equal"))
explode = (0.01, 0.01, 0.01, 0.01, 0.01, 0.01, 0.01, 0.05, 0.05)
wedges, texts = ax.pie(ocur['relative_frequency (%)'], wedgeprops=dict(width=0.6), sta

bbox_props = dict(boxstyle="square,pad=0.5", fc="w", ec="k", lw=0.77)
kw = dict(arrowprops=dict(arrowstyle="-"),
          bbox=bbox_props, zorder=0, va="center")

for i, p in enumerate(wedges):
    ang = (p.theta2 - p.theta1)/5. + p.theta1
    y = np.sin(np.deg2rad(ang))
    x = np.cos(np.deg2rad(ang))
    horizontalalignment = {-1: "right", 1: "left"}[int(np.sign(x))]
    connectionstyle = "angle,angleA=0,angleB={}".format(ang)
    kw["arrowprops"].update({"connectionstyle": connectionstyle})
```

```
    ax.annotate(ocur['Ocur'][i] + ' ' + str(round(ocur['relative_frequency (%)'][i], 3
                horizontalalignment=horizontalalignment, **kw)
plt.title('Gráfico 3. Sitio de muerte')
plt.show()

del ocur, fig, ax, explode, wedges, texts, bbox_props, kw,
```

	Ocur	frequency	relative_frequency (%)	relative_acc_frequency
0	Domicilio	492731	60.883904	0.608839
1	Hospital público	167440	20.689587	0.815735
2	Ignorado	45301	5.597581	0.871711
3	Seguro social	41581	5.137922	0.923090
4	Vía pública	25316	3.128151	0.954371
5	Hospital privado	22830	2.820970	0.982581
6	Otro	11815	1.459911	0.997180
7	Centro de salud	2221	0.274436	0.999925
8	Lugar de trabajo	61	0.007537	1.000000



Quién certifica

In [ ]:
```
calculate_frequency(deaths, 'Cerdef')
```

	Cerdef	frequency	relative_frequency (%)	relative_acc_frequency
0	Médico	452385	55.898583	0.558986
1	Ignorado	295654	36.532245	0.924308
2	Medico	49374	6.100858	0.985317
3	Paramédico	5217	0.644634	0.991763
4	Autoridad	5054	0.624493	0.998008
5	Paramedico	1612	0.199185	1.000000

Correlación entre los datos

In [ ]:
```
useless = ['Unnamed: 0', 'Edadif', 'Escodif']
otros = deaths.loc[:, ~deaths.columns.isin(useless)]
```

Definiendo elementos clave

Cruce las variables que considere que son las más importantes para hallar los elementos clave que lo pueden llevar a comprender lo que está causando el problema encontrado.

Durante el análisis exploratorio de los datos y mediante las diversas preguntas que se plantearon para conocer dichos datos se pudo encontrar que existen una gran parte de las defunciones las cuales NO reciben asistencia médica, y esto es más notorio cuando se habla de asistencia médica por pueblo de pertenencia. Por lo tal, y basado en la correlación de los datos, las variables que se consideran clave para conocer por qué una gran parte de la población no recibe asistencia económica son las siguientes:

- Asistencia
- Sexo
- Escolaridad
- Año de ocurrencia
- Población de pertenencia
- Área geográfica
- Departamento de Ocurrencia
- Sitio de Ocurrencia

Estado de los datos

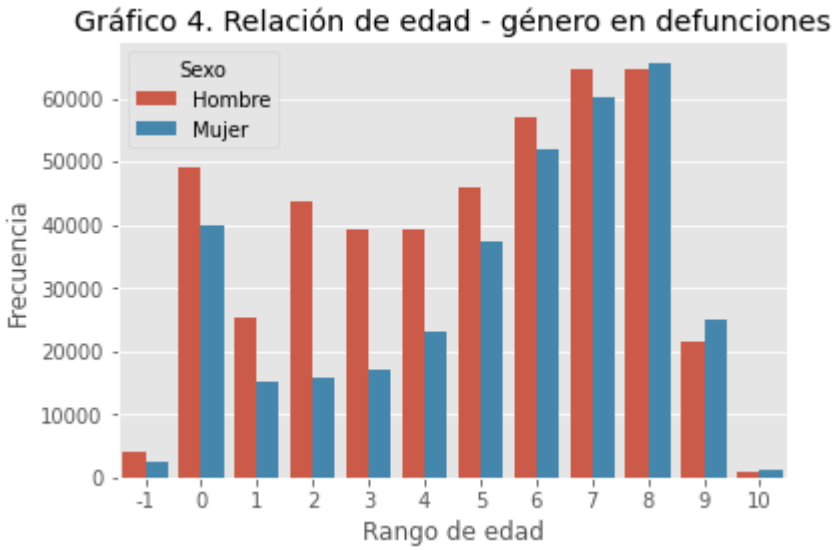
Haga gráficos exploratorios que le dé ideas del estado de los datos. // Acá van las preguntas

¿Cuál es la relación entre el género y la edad en cuanto a las defunciones en Guatemala? ¿Las mujeres tienden a morir más jóvenes que los hombres?

En el gráfico 4, con relación al género y edad, se puede observar que hasta los 80 años, se han reportado más muertes masculinas que femeninas en cada uno de los rangos de edad. Sin

embargo, luego de los 80 años la relación cambia, y en este caso, se reportaron más muertes de mujeres que de hombres. Por lo tal, no se puede inferir que las mujeres tienden a morir más jóvenes que los hombres ya que por lo contrario, se han reportado más muertes de hombres jóvenes que de mujeres.

In [ ]:
```
# Debido a que es una diversa cantidad de edades, se decide hacer grupos
new_deaths = deaths.copy()

for x in range(11):
    if x < 10:
        new_deaths.loc[(new_deaths.Edadif >= x * 10) & (new_deaths.Edadif < (x+1) * 10
    else:
       new_deaths.loc[(new_deaths.Edadif >= 100), 'Edadif'] = 10

age_gender = new_deaths.groupby(by=['Sexo', 'Edadif']).count()
age_gender.reset_index(level=[1], inplace=True)

ax = sns.barplot(x="Edadif", y="Depreg", hue=age_gender.index, data=age_gender)
plt.ylabel('Frecuencia')
plt.xlabel('Rango de edad')
plt.title('Gráfico 4. Relación de edad - género en defunciones')
plt.show()

del new_deaths, age_gender
```

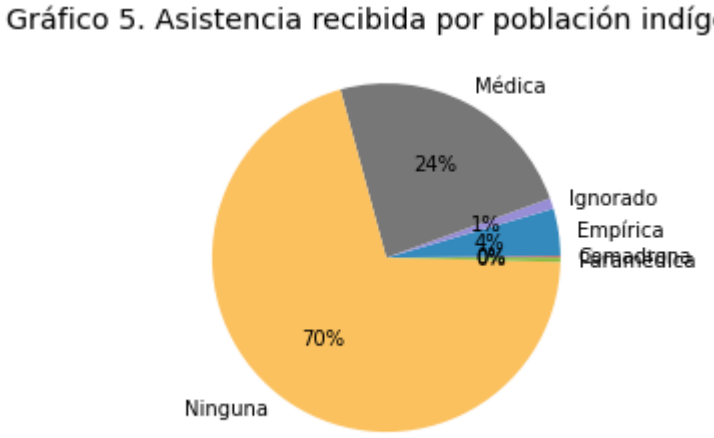


¿Cuál es el porcentaje de población indígena que recibe asistencia hospitalaria?

Como se muestra en el **gráfico 5** solamente un 24% de las defunciones reportadas del pueblo indígena han recibido asistencia médica a diferencia de un 70% de los cuales no ha recibido ningún tipo de asistencia.

In [ ]:
```
etnic_asist = deaths.groupby(by=['Puedif', 'Asist']).count()
etnic_asist.reset_index(level=[1], inplace=True)
etnic_asist = etnic_asist.loc[etnic_asist.index == 'Indigena']
etnic_asist = etnic_asist.loc[ :, ['Asist', 'Depreg']]
etnic_asist['frequency (%)'] = (etnic_asist['Depreg'] / etnic_asist['Depreg'].sum()) *
```

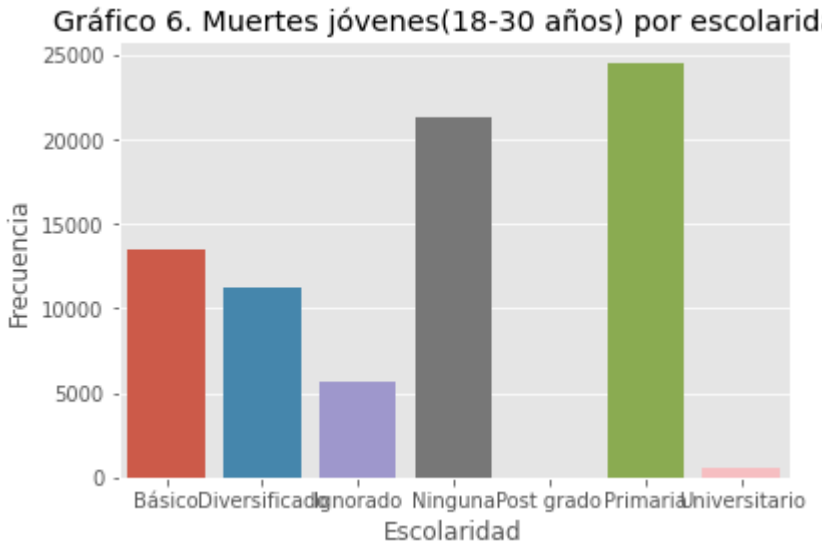
```
plt.pie(etnic_asist.Depreg, labels = etnic_asist.Asist, autopct='%.0f%%')
plt.title('Gráfico 5. Asistencia recibida por población indígena')
plt.show()

del etnic_asist
```



¿Cómo es la distribución de las muertes de los jóvenes entre 18-30 que tienen una escolaridad por encima de diversificado?

De forma sencilla de interpretar el gráfico 6 representa la distribución de las muertes de personas entre 18-30 años agrupados por escolaridad. Como se puede notar, con personas con grado académico igual y por encima de diversificado, la mayor cantidad de muertes se da justamente con grado académico de diversificado. Este resultado es bastante interesante ya que si bien esto no presenta una relación entre nivel académico y muertes sino más bien, representa la desigualdad en acceso a educación que se vive dentro del país, ya que como se puede notar, la cantidad de jóvenes, que fallecen y son universitarios es muy poca en comparación con aquellos que mueren y están estudiando o tienen grado académico de primaria.

In [ ]:
```
young = deaths.loc[(deaths['Edadif'] >= 18) & (deaths['Edadif'] <= 30), ]
young = young.groupby(by=['Escodif']).count()

sns.barplot(x=young.index, y="Depreg", data=young)
plt.ylabel('Frecuencia')
plt.xlabel('Escolaridad')
plt.title('Gráfico 6. Muertes jóvenes(18-30 años) por escolaridad')
plt.show()

del young
```



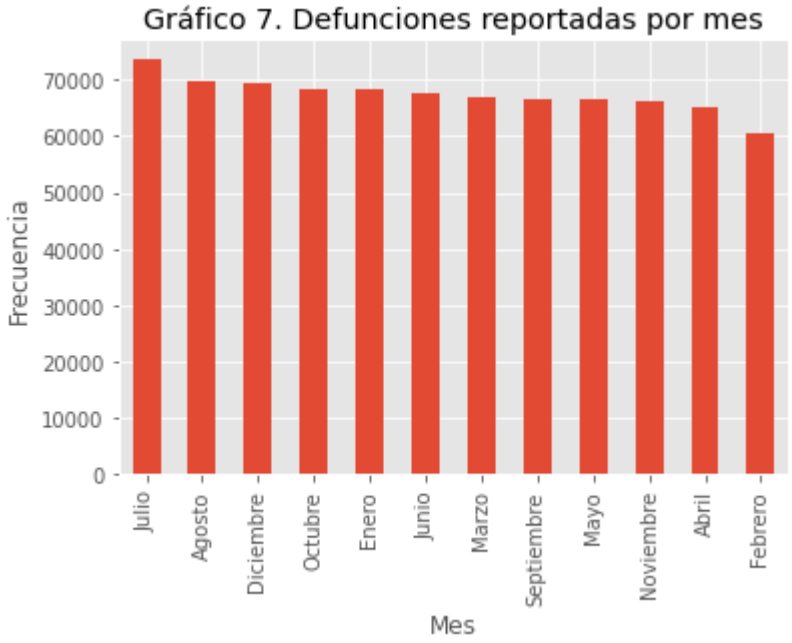
¿En qué meses se dieron la mayor cantidad de defunciones?

Como se puede observar en la **gráfica 7**, los meses 3 con mayores defunciones son:

- Julio con 73553 muertes reportas,
- Agosto con 69865
- Diciembre con 69332

In [ ]:
```
month = deaths.groupby(["Mesocu"])["Mesocu"].count().sort_values(ascending=False)
display(month)

month.plot.bar()
plt.title('Gráfico 7. Defunciones reportadas por mes')
plt.xlabel('Mes')
plt.ylabel('Frecuencia')
del month
```

```
Mesocu
Julio         73553
Agosto        69865
Diciembre     69332
Octubre       68468
Enero         68213
Junio         67813
Marzo         66918
Septiembre    66696
Mayo          66551
Noviembre     66100
Abril         65316
Febrero       60471
Name: Mesocu, dtype: int64
```



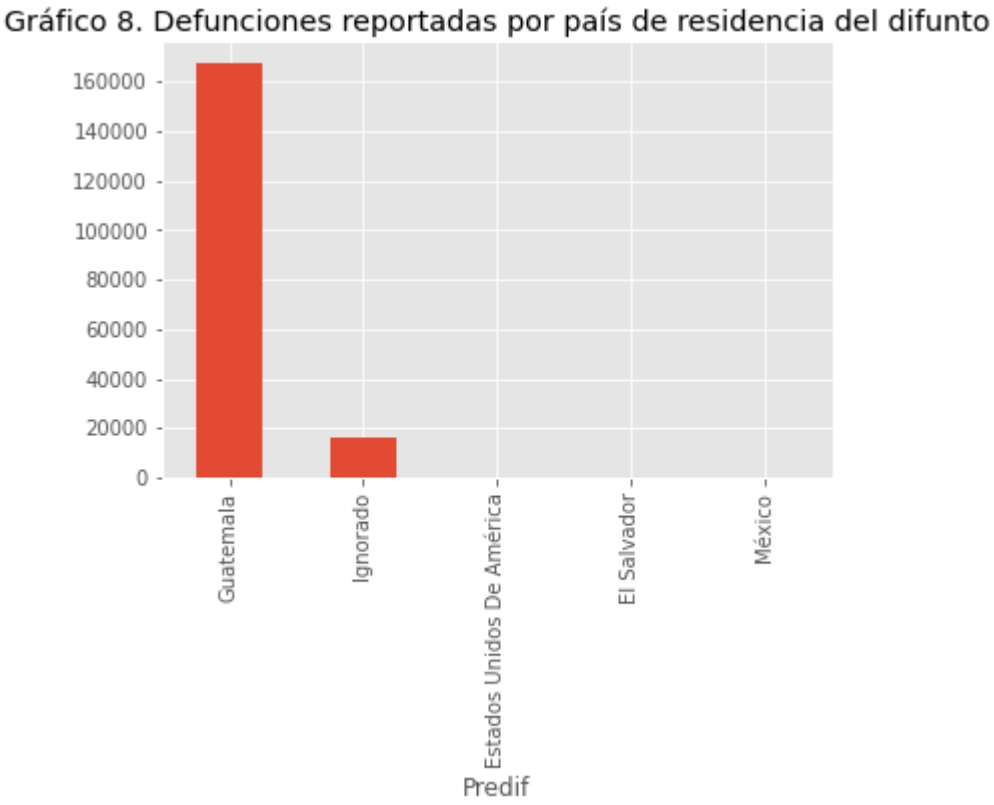
¿En qué países ocurrieron la mayor cantidad de defunciones en los años 2019 y 2020 (pandemia)?

El país con mayor defunciones es Guatemala. Cabe mencionar que este resultado se debe, principalmente, a que la mayoría de datos que se recolectaron fueron dentro del país, obviando así muchos casos de defunciones de guatemaltecos que residen en el extranjero.

In [ ]:
```
pandemia = pd.DataFrame(deaths.query("Añoreg>=2019"))
pandemia['Edadif'] = [-1 if year == "Ignorado" else year for year in pandemia["Edadif"
pais = pandemia.groupby(["Predif"])["Predif"].count().sort_values(ascending=False).hea
display(pais)

pais.plot.bar()
plt.title('Gráfico 8. Defunciones reportadas por país de residencia del difunto')
```

```
Predif
Guatemala                   167596
Ignorado                     16302
Estados Unidos De América      148
El Salvador                     37
México                          26
Name: Predif, dtype: int64
```

Out[ ]:
```
Text(0.5, 1.0, 'Gráfico 8. Defunciones reportadas por país de residencia del difunt
o')
```



Durante estos años, las personas que fallecieron tuvieron tratamiento médico?

Como se puede observar, la cantidad de defunciones que hubieron en estos años de pandemia y que recibieron asistencia médica fueron de 93212 muertes, luego siguen las personas que no recibieron ningún tipo de asistencia médica. Esto se puede deber a que en esos tiempo, muchas personas se enfermaban y cuando necesitaban ir a un hospital, estos se encontraban llenos y no lograban encontrar cupo.

In [ ]:
```
pandemia.groupby("Asist")["Asist"].count().sort_values(ascending=False)
```

Out[ ]:
```
Asist
Médica        93212
Ninguna       80592
Ignorado       5860
Empírica       3683
Paramédica      671
Comadrona       129
Name: Asist, dtype: int64
```

¿Existe alguna relación entre el sexo y la atención recibida que tuvieron los fallecidos en los años 2011 a 2020?

In [ ]:
```
atencionGenero = deaths.groupby(["Asist","Sexo"])["Asist"].count().sort_values(ascendi
display(atencionGenero)

atencionGenero.plot.bar()
plt.title('Gráfico 9. Atención recibida por sexo')
plt.xlabel('Asitencia, Sexo')
```

```
plt.ylabel('Frecuencia')
del atencionGenero
```

```
Asist       Sexo
Ninguna     Hombre    228798
Médica      Hombre    207723
            Mujer     172299
Ninguna     Mujer     169013
Empírica    Hombre      9465
            Mujer       8853
Ignorado    Hombre      6777
            Mujer       2705
Paramédica  Hombre      1893
            Mujer       1135
Comadrona   Mujer        391
            Hombre       244
Name: Asist, dtype: int64
```

Como se puede observar en la **gráfica 9**, no hay alguna relación entre el género y la atención recibida, por lo que no se puede concluir que el género afecte en algo hacia la atención médica que se puede recibir. Sin embargo, se puede notar que sin importar el género, existe una gran cantidad de personas que no recibe atención médica antes de morir.

¿En Guatemala, en qué lugares predominan las defunciones? ¿El servicio que ofrecen los hospitales privados es mejor que el de los hospitales públicos?

In [ ]:
```
guatemala = pd.DataFrame(deaths.query("Predif=='Guatemala'"))
```

In [ ]:
```
municipios = guatemala.groupby("Mredif")["Mredif"].count().sort_values(ascending=False
display(municipios)

municipios.plot.pie()
```

```
plt.title('Gráfica 10. Defunciones por departamento')
plt.ylabel('')
plt.xlabel('')
```

```
Mredif
Guatemala               80791
Mixco                   21792
Villa Nueva             16477
Quetzaltenango          10186
Escuintla                9532
San Pedro Carchá         8691
San Juan Sacatepéquez    8587
Cobán                    7677
Jalapa                   7093
Totonicapán              6781
Name: Mredif, dtype: int64
Text(0.5, 0, '')
```

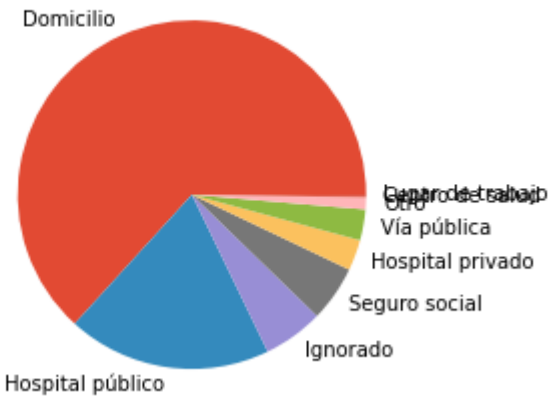
Out[ ]:

In [ ]:

```
lugar = guatemala.groupby("Ocur")["Ocur"].count().sort_values(ascending=False).head(10
display(lugar)

lugar.plot.pie()
plt.title('Gráfico 11. Lugar de ocurrencia de las defunciones')
plt.ylabel('')
plt.xlabel('')

del guatemala, lugar, municipios
```

```
Ocur
Domicilio           429523
Hospital público    128415
Ignorado             37938
Seguro social        35107
Hospital privado     19461
Vía pública          19401
Otro                  7572
Centro de salud       1566
Lugar de trabajo        45
Name: Ocur, dtype: int64
```

Gráfico 11. Lugar de ocurrencia de las defunciones



Cómo se puede observar en las dos gráficas de pie, el municipio con mayor cantidad de defunciones es Guatemala, luego le sigue Mixco, Villa Nueva y Quetzaltenango. Por otro lado, se puede ver en el **gráfico 11** en los hospitales públicos existió una mayor cantidad de defunciones que en hospitales privados, lo que podría indicar que el servicio de los hospitales privados es mejor, sin embargo, como se vio en la pregunta anterior y en la siguiente pregunta, el número de personas que tiene acceso a tratamientos médicos privados es muy baja, por lo tal no es posible concluir que los hospitales privados son mejores que los públicos debido a que la cantidad de personas que cuentan con los recursos económicos para asistir a un hospital privado es menos de la mitad de la población guatemalteca.

¿Las personas fallecidas de 50 años o más recibieron buen tratamiento médico?

In [ ]:
```
datos2 = deaths.copy()
datos2['Edadif'] = [-1 if year == "Ignorado" else int (year) for year in datos2["Edadi
mayores = pd.DataFrame(datos2.query("Edadif>49"))

display(mayores.groupby("Asist")["Asist"].count().sort_values(ascending=False))

del mayores, datos2
```

```
Asist
Ninguna       244804
Médica        233154
Empírica       11730
Ignorado        3915
Paramédica      1433
Comadrona        138
Name: Asist, dtype: int64
```

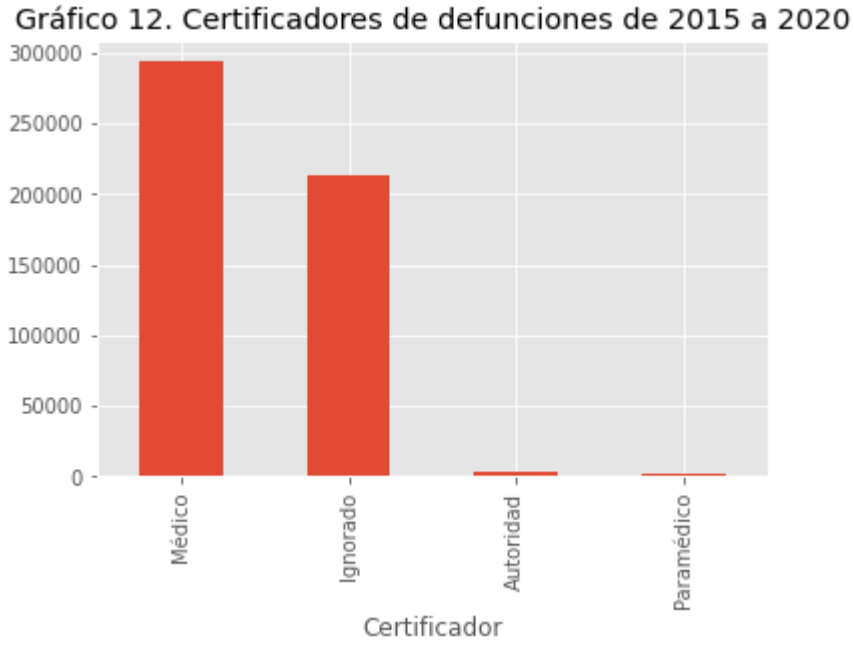
Cómo se mencionó anteriormente, aquí se puede observar que las personas mayores a los 50 años no recibieron asistencia médica a la hora de su muerte, con una cantidad de 200701 defunciones registradas para ese caso. Luego siguen las defunciones que sí recibieron asistencia médica, con un valor de 196730.

¿Quienes fueron las personas que más certificaron durante los

años 2015 a 2020?

In [ ]:
```
datos2 = pd.DataFrame(deaths.query("Añoreg>2014"))

cert = datos2.groupby("Cerdef")["Cerdef"].count().sort_values(ascending=False)
display(cert)

cert.plot.bar()
plt.title('Gráfico 12. Certificadores de defunciones de 2015 a 2020')
plt.xlabel('Certificador')
del datos2
```

```
Cerdef
Médico        293734
Ignorado      213060
Autoridad       2915
Paramédico      2128
Name: Cerdef, dtype: int64
```



Como se puede observar en la gráfica de arriba, las personas que mayor certificaron son los médicos, seguidos por los que se registraron como "ignorados".

In [ ]:
```
deaths.shape
```

Out[ ]:
```
(809296, 28)
```

In [ ]:
```
deaths.columns
```

Out[ ]:
```
Index(['Unnamed: 0', 'Depreg', 'Mupreg', 'Mesreg', 'Añoreg', 'Depocu',
       'Mupocu', 'Areag', 'Sexo', 'Diaocu', 'Mesocu', 'Añoocu', 'Edadif',
       'Puedif', 'Ecidif', 'Escodif', 'Ciuodif', 'Pnadif', 'Dnadif', 'Mnadif',
       'Nacdif', 'Predif', 'Dredif', 'Mredif', 'Caudef', 'Asist', 'Ocur',
       'Cerdef'],
      dtype='object')
```

¿En qué años se han dado la mayor cantidad de

defunciones?

In [ ]:
```
años = deaths["Añoreg"].value_counts().to_frame()
años
```

Out[ ]:

	Añoreg
2020	95100
2019	85476
2018	82755
2016	82420
2017	81475
2015	81040
2014	77582
2013	76618
2012	72115
2011	71144
2021	3571

La mayor cantidad de muertes se dió en el 2020. Podríamos relacionarlo al inicio de la Pandemia, puesto que hubo un aumento a comparación del 2019 de 9,624 muertes

¿En qué departamento se han dado la mayor cantidad de defunciones?

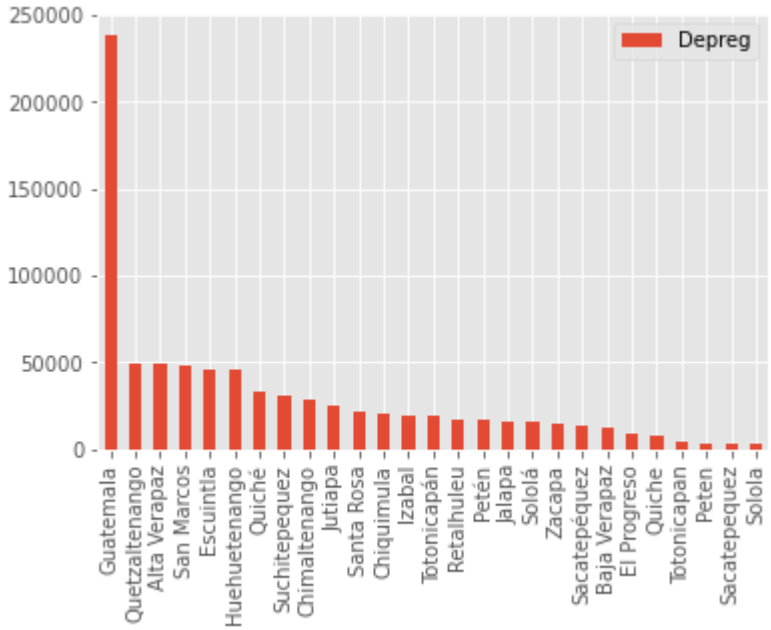
In [ ]:
```
dpto = deaths["Depreg"].value_counts().to_frame()
dpto
```

Out[ ]:

	Depreg
Guatemala	238866
Quetzaltenango	49558
Alta Verapaz	49071
San Marcos	47997
Escuintla	46155
Huehuetenango	45374
Quiché	32499
Suchitepequez	30460
Chimaltenango	28316
Jutiapa	24779
Santa Rosa	22106
Chiquimula	20496
Izabal	19278
Totonicapán	19217
Retalhuleu	17160
Petén	16743
Jalapa	16041
Sololá	15280
Zacapa	14402
Sacatepéquez	13716
Baja Verapaz	12056
El Progreso	8385
Quiche	7402
Totonicapan	4419
Peten	3415
Sacatepequez	3055
Solola	3050

In [ ]:
```
dpto.plot(kind = "bar")
```

Out[ ]:
```
<AxesSubplot:>
```

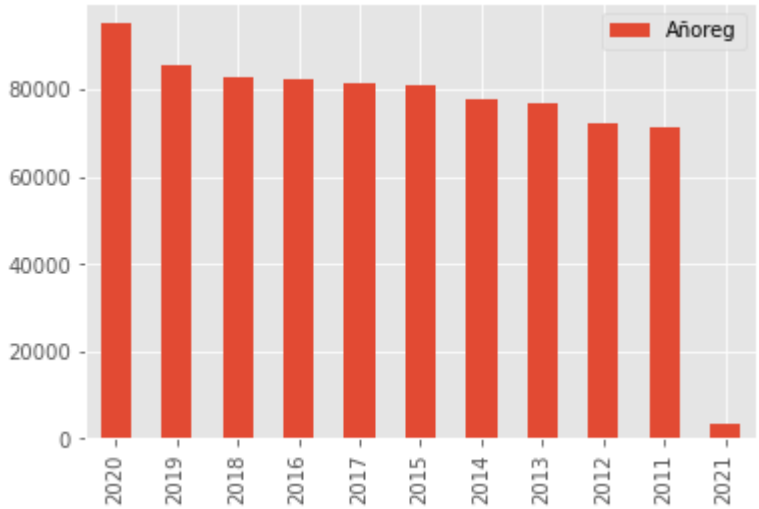


Guatemala presenta la mayor cantidad de muertes registradas. Esto era de esperarse puesto que la densidad poblacional es mucho mayor en comparación con los demás departamentos. Luego le sigue Quetzaltenango, departamento con una alta densidad poblacional.

¿Cuál fue el número de defunciones por año ?

In [ ]:
```
años = deaths["Añoreg"].value_counts().to_frame()
años.plot(kind = "bar")
```

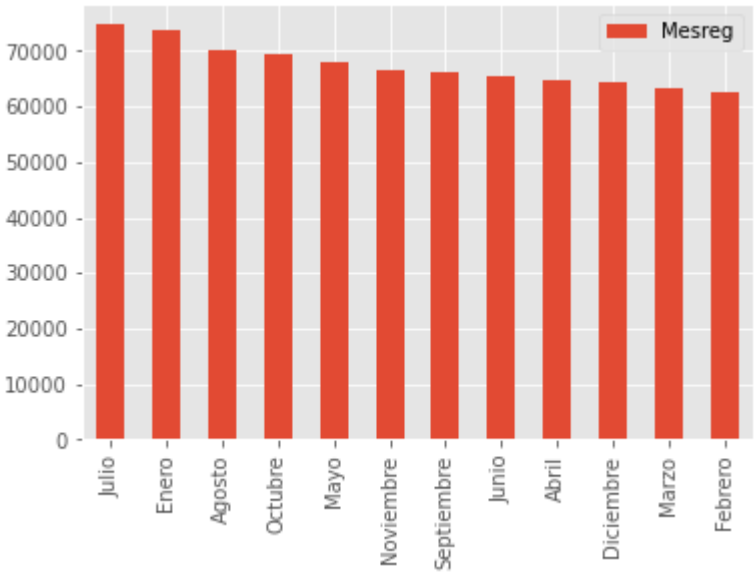
Out[ ]:
```
<AxesSubplot:>
```



Realice un gráfico de barras por el mes de ocurrencia de la defunción

In [ ]:
```
mes = deaths["Mesreg"].value_counts().to_frame()
mes.plot(kind = "bar")
```

Out[ ]: <AxesSubplot:>

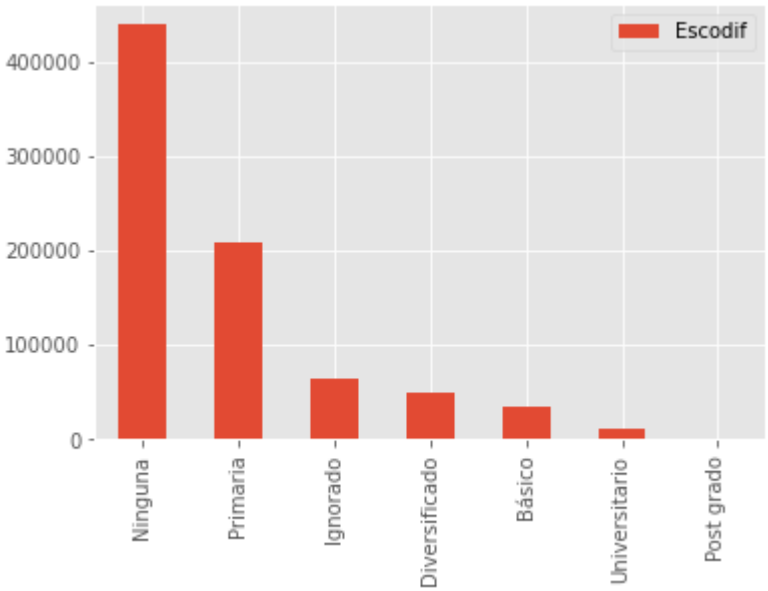


Es curioso ver como en el mes de julio es donde se dan la mayor cantidad de muertes. Podemos relacionarlo con el pago de la bonificación del bono 14, ya que en este mes las extorsiones aumentan. Esto nos indica que el índice de violencia puede ser superior por dicho bono, por lo que las muertes aumentan. Para Julio de 2020 además, se tuvieron picos en la pandemia Covid19, lo que quizá pudo darle peso a esta variable.

¿Qué escolaridad presenta mayores defunciones? ¿Influye el tener una educación superior a reducir el número de defunciones?

In [ ]:
```
esc = deaths["Escodif"].value_counts().to_frame()
esc.plot(kind = "bar")
```

Out[ ]: <AxesSubplot:>



La mayoría de muertes se dan en personas con escolaridad nula. Es interesante observar como estudiantes universitarios y de posgrado son los que presentan la menor cantidad de muertes en los últimos 10 años.

Agrupamiento

Haga un agrupamiento (clustering) e interprete los resultados. Para la elaboración del agrupamiento, debido a que se cuenta con una gran cantidad de datos, sin embargo no se tiene a disposición un ordenador capaz de procesar dicha cantidad de datos, entonces se decidió utilizar una muestra de la población.

Debido a que se busca que la muestra sea lo más significativa posible, se decidió tomar aproximadamente un 1% de los datos de cada año y así garantizar que en la muestra existirá una proporción de defunciones de todos los años descritos en la población original.

In [ ]:
```
# Obtención de la muestra
deaths_sample = deaths.groupby('Añoocu', group_keys=False).apply(lambda x: x.sample(fr

usable = ['Asist', 'Ocur', 'Edadif', 'Escodif', 'Puedif', 'Areag']
deaths_sample = deaths_sample[usable]

deaths_sample = pd.get_dummies(deaths_sample)
deaths_sample.to_csv('project_sample.csv')
```

In [ ]:
```
# Obtención de la muestra
deaths_sample = deaths.groupby('Añoocu', group_keys=False).apply(lambda x: x.sample(fr

usable = ['Asist', 'Ocur', 'Edadif', 'Escodif', 'Puedif', 'Areag']
deaths_sample = deaths_sample[usable]
# deaths_sample.shape
# Pasando a número asistencia recibida
deaths_sample.loc[(deaths_sample.Asist == 'Médica'), 'Asist'] = 1
deaths_sample.loc[(deaths_sample.Asist == 'Comadrona'), 'Asist'] = 2
deaths_sample.loc[(deaths_sample.Asist == 'Empírica'), 'Asist'] = 3
deaths_sample.loc[(deaths_sample.Asist == 'Ignorado'), 'Asist'] = 4
deaths_sample.loc[(deaths_sample.Asist == 'Ninguna'), 'Asist'] = 5
deaths_sample.loc[(deaths_sample.Asist == 'Paramédica'), 'Asist'] = 6

# Pasando a número la ocurrencia
deaths_sample.loc[(deaths_sample.Ocur == 'Centro de salud'), 'Ocur'] = 1
deaths_sample.loc[(deaths_sample.Ocur == 'Domicilio'), 'Ocur'] = 2
deaths_sample.loc[(deaths_sample.Ocur == 'Hospital privado'), 'Ocur'] = 3
deaths_sample.loc[(deaths_sample.Ocur == 'Hospital público'), 'Ocur'] = 4
deaths_sample.loc[(deaths_sample.Ocur == 'Ignorado'), 'Ocur'] = 5
deaths_sample.loc[(deaths_sample.Ocur == 'Otro'), 'Ocur'] = 6
deaths_sample.loc[(deaths_sample.Ocur == 'Seguro social'), 'Ocur'] = 7
deaths_sample.loc[(deaths_sample.Ocur == 'Vía pública'), 'Ocur'] = 8
deaths_sample.loc[(deaths_sample.Ocur == 'Lugar de trabajo'), 'Ocur'] = 9

# Paso a rangos la edad
for x in range(1, 12):
    if x < 11:
        deaths_sample.loc[(deaths_sample.Edadif >= x * 10) & (deaths_sample.Edadif < (
    else:
       deaths_sample.loc[(deaths_sample.Edadif >= 100), 'Edadif'] = 10
```

```
# Pasando a número la escolaridad
deaths_sample.loc[(deaths_sample.Escodif == 'Básico'), 'Escodif'] = 1
deaths_sample.loc[(deaths_sample.Escodif == 'Diversificado'), 'Escodif'] = 2
deaths_sample.loc[(deaths_sample.Escodif == 'Ignorado'), 'Escodif'] = 3
deaths_sample.loc[(deaths_sample.Escodif == 'Ninguna'), 'Escodif'] = 4
deaths_sample.loc[(deaths_sample.Escodif == 'Post grado'), 'Escodif'] = 5
deaths_sample.loc[(deaths_sample.Escodif == 'Primaria'), 'Escodif'] = 6
deaths_sample.loc[(deaths_sample.Escodif == 'Universitario'), 'Escodif'] = 7

# Pasando a número el pueblo de pertenencia
deaths_sample.loc[(deaths_sample.Puedif == 'Indigena'), 'Puedif'] = 1
deaths_sample.loc[(deaths_sample.Puedif == 'No indigena'), 'Puedif'] = 2
deaths_sample.loc[(deaths_sample.Puedif == 'Ignorado'), 'Puedif'] = 3
deaths_sample.loc[(deaths_sample.Puedif == 'Otro'), 'Puedif'] = 4

# Pasando a número el area geográfica
deaths_sample.loc[(deaths_sample.Areag == 'Rural'), 'Areag'] = 1
deaths_sample.loc[(deaths_sample.Areag == 'Urbano'), 'Areag'] = 2
deaths_sample.loc[(deaths_sample.Areag == 'Ignorado'), 'Areag'] = 3

deaths_sample.to_csv('project_sample.csv')

del deaths_sample
```

Número óptimo de clusters a utilizar.

In [ ]:

```
numeroClusters = range(1,11)
wcss = []
for i in numeroClusters:
    kmeans = cluster.KMeans(n_clusters=i)
    kmeans.fit(X_scale)
    wcss.append(kmeans.inertia_)

plt.plot(numeroClusters, wcss)
plt.xlabel("Número de clusters")
plt.ylabel("Score")
plt.title("Gráfico de Codo")
plt.show()

# get it from https://towardsdatascience.com/elbows-and-silhouettes-hands-on-customer-
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.cluster import KMeans, MeanShift, estimate_bandwidth
from sklearn.preprocessing import StandardScaler, PowerTransformer
from sklearn.metrics import silhouette_score
from sklearn.model_selection import train_test_split

from kmodes.kprototypes import KPrototypes

from yellowbrick.cluster import KElbowVisualizer, SilhouetteVisualizer, InterclusterDi
from kneed import KneeLocator

from sklearn.decomposition import PCA
```

```
from tqdm import tqdm
import sys
import warnings
warnings.filterwarnings("ignore")
# elbow score plot with Yellowbrick
nK = 12
RNDN = 42
def elbowplot(df, elbowmetric, model):
    print("Elbow Score Plot (" + str(elbowmetric) + " metric):")
    vis = KElbowVisualizer(
        model,
        k=(2,nK),
        metric=elbowmetric,
        locate_elbow=True,
        timings=False)
    vis.fit(df)
    print("elbow value = optimal k:", f'{vis.elbow_value_:.0f}', \
            " | elbow score:", f'{vis.elbow_score_:,.3f}')
    vis.show()



# call elbow plot for each of 3 alternative metrics
    # distortion = mean sum of squared distances to center
    # silhouette = mean ratio of intra-cluster and nearest-cluster distance
    # calinski = ratio of within to between cluster dispersion

model = KMeans(random_state=RNDN)
_ = [elbowplot(X, m, model) for m in tqdm(["distortion", "silhouette", "calinski_harab
```

Algoritmo de agrupamiento.

Calidad del agrupamiento usando el método de la silueta.

Interpretación de la agrupación

Usando para eso las variables numéricas y categóricas dentro de cada grupo.

Análisis de resultados

Situación problemática

Describa la situación problemática que lo lleva a acotar un problema a resolver.

Durante el análisis exploratorio realizada en el apartado anterior se pudo encontrar que aproximadamente el 50% de los datos reportados para las defunciones de los años de 2011 a 2020 no recibe ningún tipo de asistencia al morir. De igual forma,

Problema científico

Enuncie un problema científico y unos objetivos preliminares.

¿Qué se tiene para responder el problema?

Describa los datos que tiene para responder el problema planteado. Esto incluye el estado en que encontró el o los conjuntos de datos y las operaciones de limpieza que le realizó, en caso de que hayan sido necesarias.

Conclusiones

Escriba unas conclusiones con los hallazgos encontrados durante el análisis exploratorio