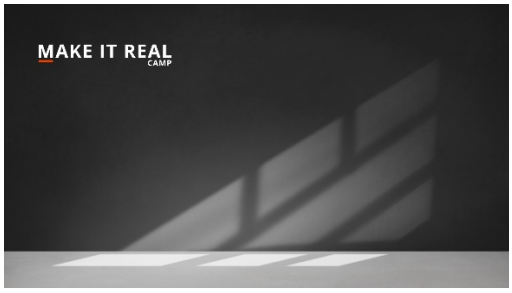
Análisis de Datos - MITIC 2024/2025

Make it Real



Profesor: Juan Sebastián Paniagua

Alumno: Fabrizio Eduardo Aquino Rojas.

Cedula Identidad: 4206219

Proyecto Final: Análisis de Datos - MITIC

Archivo de Datos: Patients

Objetivo y alcance del Proyecto.

El objetivo es demostrar que he aprendido a aplicar las herramientas y técnicas de análisis de datos para resolver problemas.

Introducción

A través del análisis de datos, se busca identificar patrones y relaciones entre distintos factores y la tasa de fallecimientos, utilizando enfoques estadísticos y exploratorios.

El estudio se centra en analizar la relación entre los fallecimientos en la población y diversos factores sociodemográficos. Algunas de las preguntas clave que se abordarán incluyen:

- **¿Influye el género en la tasa de mortalidad?**
- **¿La ciudad tiene algún impacto en los fallecimientos?**
- **¿Existe una relación entre la edad y la probabilidad de fallecer?**
- **¿La raza puede influir en el fallecimiento?**

Hipótesis de Género

Se plantea la hipótesis de que **existen diferencias significativas en la edad entre hombres y mujeres**, lo que podría impactar en las tasas de mortalidad. Para validar esta hipótesis, se realizarán pruebas estadísticas que permitan determinar si dichas diferencias son estadísticamente significativas.

Alcance

Este análisis permitirá no solo explorar correlaciones entre variables, sino también identificar factores clave que pueden afectar la mortalidad. Los hallazgos pueden ser utilizados para orientar futuras investigaciones, mejorar estrategias de salud pública y optimizar la asignación de recursos médicos.

Desarrollo

1-Exploracion de los datos.

```
# Carga de archivo csv.
from google.colab import files
uploaded = files.upload()
```

Primeramente, aplique este método que me permite subir archivos CSV desde mi computadora.

```
# Libreria pandas
import pandas as pd
df = pd.read_csv('patients.csv')
```

Una vez subido el archivo al colab use el Código df (variable que almacena el dataframe creado a a partir de un archivo CSV) pd (abreviatura usada para referirse a la librería pandas) , read_csv(funcion de pandas que lee un archivo CSV y lo convierte en dataframe) que se usa en Python con la librería pandas para leer un archivo CSV y cargarlo en el dataframe.

```
# Información general del DataFrame
df.info()
df.head()
```

Aplique el codigo df.info para que muestre la informacion general del dataframe. Y el codigo df.head muestra Las primeras 5 filas del dataframe.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 974 entries, 0 to 973
Data columns (total 20 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Id           974 non-null    object
 1   BIRTHDATE    974 non-null    object
 2   DEATHDATE    154 non-null    object
 3   PREFIX       974 non-null    object
 4   FIRST        974 non-null    object
 5   LAST         974 non-null    object
 6   SUFFIX       21 non-null     object
 7   MAIDEN       386 non-null    object
 8   MARITAL      973 non-null    object
 9   RACE         974 non-null    object
 10  ETHNICITY    974 non-null    object
 11  GENDER       974 non-null    object
 12  BIRTHPLACE   974 non-null    object
 13  ADDRESS      974 non-null    object
 14  CITY         974 non-null    object
 15  STATE        974 non-null    object
 16  COUNTY       974 non-null    object
 17  ZIP          832 non-null    float64
 18  LAT          974 non-null    float64
 19  LON          974 non-null    float64
dtypes: float64(3), object(17)
memory usage: 152.3+ KB
```

	Id	BIRTHDATE	DEATHDATE	PREFIX	FIRST	LAST	SUFFIX	MAIDEN	MARITAL	RACE	ETHNICITY	GENDER	BIRTHPLACE	ADDRESS	CITY	STATE	COUNTY	ZIP	LAT	LON
0	5605b66b-e92d-c16c-1b83-b8bf7040d51f	1977-03-19	NaN	Mrs.	Nikita578	Erdman779	NaN	Leannon79	M	white	nonhispanic	F	Wakefield Massachusetts US	510 Little Station Unit 69	Quincy	Massachusetts	Norfolk County	2186.0	42.290937	-70.975503
1	6e5ae27c-8038-7988-e2c0-25a103f01bfa	1940-02-19	NaN	Mr.	Zane918	Hodkiewicz467	NaN	NaN	M	white	nonhispanic	M	Brookline Massachusetts US	747 Conn Throughway	Boston	Massachusetts	Suffolk County	2135.0	42.308831	-71.063162
2	8123d076-0886-9007-e956-d5864aa121a7	1958-06-04	NaN	Mr.	Quinn173	Marquardt819	NaN	NaN	M	white	nonhispanic	M	Gardner Massachusetts US	816 Okuneva Extension Apt 91	Quincy	Massachusetts	Norfolk County	2170.0	42.265177	-70.967085
3	770518e4-6133-648e-60c9-071eb2f0e2ce	1928-12-25	2017-09-29	Mr.	Abel832	Smitham825	NaN	NaN	M	white	hispanic	M	Randolph Massachusetts US	127 Cole Way Unit 95	Boston	Massachusetts	Suffolk County	2118.0	42.334304	-71.066801
4	f96addf5-81b9-0aab-7855-d208d3d352c5	1928-12-25	2014-02-23	Mr.	Edwin773	Labadie908	NaN	NaN	M	white	hispanic	M	Stow Massachusetts US	976 Ziemann Gateway	Boston	Massachusetts	Suffolk County	2125.0	42.346771	-71.058813

El dataframe tiene 974 filas, 20columnas. El deathdate tiene solo 154 valores, lo que significa que 820 registros estan vacios(posiblemente personas vivas) Suffix tiene solo 21 valores lo que indica que la mayoria no tiene un sufijo (como JR, Sr,etc) Zip (codigo postal) tiene 832 valores, por lo que 142 estan vacios.

Interpretación del Dataframe:

Datos Demográficos:

El dataframe contiene información demográfica detallada de 974 individuos, incluyendo edad, género, raza, etnicidad y lugar de nacimiento.

Fechas de Nacimiento y Muerte:

La columna BIRTHDATE está completa para todos los individuos.

La columna DEATHDATE tiene 154 valores no nulos, lo que indica que 154 individuos han fallecido. lo que significa que 820 registros están vacíos(posiblemente personas vivas)

Datos Geográficos:

Las columnas CITY, STATE, COUNTY, LAT y LON están completas, proporcionando información geográfica precisa para cada individuo.

La columna ZIP tiene algunos valores faltantes (832 no nulos de 974) por lo que 142 están vacíos.

.

Estado Civil y Otros Detalles:

La mayoría de los individuos tienen información sobre su estado civil (MARITAL), con solo un valor faltante.

Las columnas PREFIX, FIRST, LAST, ADDRESS y GENDER están completas.

Conclusiones:

Distribución de la Edad:

Se puede calcular la edad de los individuos a partir de la columna BIRTHDATE y analizar su distribución.

La edad puede ser un factor importante para analizar la tasa de mortalidad y supervivencia.

Tasa de Mortalidad

Con 154 individuos fallecidos, se puede calcular la tasa de mortalidad general y analizarla en función de diferentes variables como edad, ciudad, raza y género.

Análisis Geográfico:

La información geográfica completa permite realizar análisis espaciales, como la distribución de la mortalidad y supervivencia en diferentes ciudades y estados.

Datos Faltantes:

La columna SUFFIX y MAIDEN tienen muchos valores faltantes, lo que sugiere que estos campos pueden no ser críticos para el análisis principal.

Siguientes Pasos

Calcular la Edad:

Utiliza la columna BIRTHDATE para calcular la edad de cada individuo.

Calcular la Tasa de Mortalidad:

Calcula la tasa de mortalidad general y específica por diferentes categorías.

Visualizaciones:

Crea gráficos y tablas de calor para visualizar la relación entre edad, ciudad, raza.

2- Preprocesamiento de Datos:

```
# Convertir las columnas 'BIRTHDATE' y 'DEATHDATE' a datetime

df['BIRTHDATE'] = pd.to_datetime(df['BIRTHDATE'])
df['DEATHDATE'] = pd.to_datetime(df['DEATHDATE'])
```

Para manipular fechas de manera más eficiente y realizar operaciones de calcular edades, restar entre fechas.

Facilita cálculos: Podemos calcular la edad de los pacientes o el tiempo de supervivencia.

Permite filtrado y análisis de datos por fecha: Por ejemplo, podemos filtrar registros de personas nacidas en cierto año o fallecidas en un rango de tiempo.

```
# 1 si está vivo, 0 si falleció
df ['IS_ALIVE'] = df['DEATHDATE'].isna().astype(int)
```

Ventaja de este método

Fácil interpretación 1= vivo 2=fallecido, Facilita filtros y análisis, segmentación quienes están vivos o fallecidos

```
# Manejo de valores faltantes

df['MAIDEN'].fillna('N/A', inplace=True)
df['SUFFIX'].fillna('N/A', inplace=True)
```

Reemplace los valores faltantes (NaN) en las columnas 'SUFFIX' y 'MAIDEN' por 'N/A'. Evita valores nulos (NaN) en el DataFrame, lo que puede causar problemas en análisis, filtros o modelos de Machine Learning. Mejora la limpieza y consistencia de los datos, permitiendo que no haya celdas vacías o valores NaN que puedan afectar consultas o visualización. Facilita la interpretación de datos, ya que 'N/A' es más claro que un campo vacío o NaN.

```
# Manejo de valores faltantes

df['MARITAL'].fillna('Unknown', inplace=True)
```

Evitar valores nulos en la columna Marital, asegurando que todos los registros tengan información clara y estructurada. Hacer que el análisis de datos sea más fácil y preciso, al asegurarnos de que cada paciente tiene una categoría en la columna MARITAL.

```
# Crear columna 'Edad'

from datetime import datetime

df['AGE'] = df['BIRTHDATE'].apply(lambda x: (datetime.today() - x).days // 365)
```

Obtener la edad de cada persona a partir de su fecha de nacimiento para facilitar análisis estadísticos y visualizaciones. Evitar cálculos manuales y hacer que el DataFrame sea más útil y fácil de analizar.El codigo calcula la edad de la persona en base a su fecha de nacimiento y la fecha actual,para realizar análisis estadísticos más detallados para relacionar edad con la tasa de supervivencia, análisis de tendencias relacionadas con la edad. Y identificar patrones de mortalidad según diferentes grupos de edad.

```
# Crea columna que concatena

df['FULL_NAME'] = df[['PREFIX', 'FIRST', 'LAST', 'SUFFIX']].astype(str).agg(' '.join, axis=1)
df['FULL_NAME'] = df['FULL_NAME'].str.replace(' N/A', '', regex=False)
```

Crea una nueva columna Full name que concatena el prefijo, primer nombre, apellido y sufijo. Si el sufijo es N/A elimnina de la concatenacion final. Esto es útil porque 'N/A' se usó para rellenar valores nulos en SUFFIX, y no queremos que aparezca en el nombre final.

```python
# Información general del DataFrame

print(df.info())
print(df.head())
```

Aplique el Código df.info para que muestre la información general del dataframe. Y el código df.head muestra Las primeras 5 filas del dataframe.

```
Data columns (total 23 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Id          974 non-null    object
 1   BIRTHDATE   974 non-null    datetime64[ns]
 2   DEATHDATE   154 non-null    datetime64[ns]
 3   PREFIX      974 non-null    object
 4   FIRST       974 non-null    object
 5   LAST        974 non-null    object
 6   SUFFIX      974 non-null    object
 7   MAIDEN      974 non-null    object
 8   MARITAL     974 non-null    object
 9   RACE        974 non-null    object
 10  ETHNICITY   974 non-null    object
 11  GENDER      974 non-null    object
 12  BIRTHPLACE  974 non-null    object
 13  ADDRESS     974 non-null    object
 14  CITY        974 non-null    object
 15  STATE       974 non-null    object
 16  COUNTY      974 non-null    object
 17  ZIP         974 non-null    object
 18  LAT         974 non-null    float64
 19  LON         974 non-null    float64
 20  IS_ALIVE    974 non-null    int64
 21  AGE         974 non-null    int64
 22  FULL_NAME   974 non-null    object
dtypes: datetime64[ns](2), float64(2), int64(2), object(17)
memory usage: 175.1+ KB
None
```

```
                                     Id  BIRTHDATE  DEATHDATE PREFIX  \
0  5605b66b-e92d-c16c-1b83-b8bf7040d51f 1977-03-19        NaT   Mrs.
1  6e5ae27c-8038-7988-e2c0-25a103f01bfa 1940-02-19        NaT    Mr.
2  8123d076-0886-9007-e956-d5864aa121a7 1958-06-04        NaT    Mr.
3  770518e4-6133-648e-60c9-071eb2f0e2ce 1928-12-25 2017-09-29    Mr.
4  f96addf5-81b9-0aab-7855-d208d3d352c5 1928-12-25 2014-02-23    Mr.

       FIRST           LAST SUFFIX     MAIDEN MARITAL   RACE  ...  \
0  Nikita578      Erdman779    N/A  Leannon79       M  white  ...
1    Zane918  Hodkiewicz467    N/A        N/A       M  white  ...
2   Quinn173   Marquardt819    N/A        N/A       M  white  ...
3    Abel832     Smitham825    N/A        N/A       M  white  ...
4   Edwin773     Labadie908    N/A        N/A       M  white  ...

                       ADDRESS    CITY          STATE          COUNTY  \
0    510 Little Station Unit 69  Quincy  Massachusetts  Norfolk County
1          747 Conn Throughway  Boston  Massachusetts  Suffolk County
2  816 Okuneva Extension Apt 91  Quincy  Massachusetts  Norfolk County
3         127 Cole Way Unit 95  Boston  Massachusetts  Suffolk County
4          976 Ziemann Gateway  Boston  Massachusetts  Suffolk County

      ZIP        LAT        LON IS_ALIVE  AGE                  FULL_NAME
0  2186.0  42.290937 -70.975503        1   48   Mrs. Nikita578 Erdman779
1  2135.0  42.308831 -71.063162        1   85  Mr. Zane918 Hodkiewicz467
2  2170.0  42.265177 -70.967085        1   66  Mr. Quinn173 Marquardt819
3  2118.0  42.334304 -71.066801        0   96     Mr. Abel832 Smitham825
4  2125.0  42.346771 -71.058813        0   96    Mr. Edwin773 Labadie908

[5 rows x 23 columns]
```

```
#subconjunto del DataFrame con las personas fallecidas.
df[df['DEATHDATE'].notna()]
```

	Id	BIRTHDATE	DEATHDATE	PREFIX	FIRST	LAST	SUFFIX	MAIDEN	MARITAL	RACE	...	ADDRESS	CITY	STATE	COUNTY	ZIP	LAT	LON	IS_ALIVE	AGE	FULL_NAME
3	770518e4-6133-648e-60c9-071eb2f0e2ce	1928-12-25	2017-09-29	Mr.	Abel832	Smitham825	NaN	NaN	M	white	...	127 Cole Way Unit 95	Boston	Massachusetts	Suffolk County	2118.0	42.334304	-71.066801	0	96	Mr. Abel832 Smitham825 nan
4	f96addf5-81b9-0aab-7855-d208d3d352c5	1928-12-25	2014-02-23	Mr.	Edwin773	Labadie908	NaN	NaN	M	white	...	976 Ziemann Gateway	Boston	Massachusetts	Suffolk County	2125.0	42.346771	-71.058813	0	96	Mr. Edwin773 Labadie908 nan
12	a8ceba15-ef47-da92-253d-1858f3ff6c34	1971-10-27	2021-03-26	Mrs.	Helene803	Kilback373	NaN	Walker122	M	white	...	865 McLaughlin Underpass Apt 63	Boston	Massachusetts	Suffolk County	2114.0	42.340495	-71.101117	0	53	Mrs. Helene803 Kilback373 nan
30	d5e6d112-f9ef-173b-fe14-e42b480f82c1	1924-11-19	2014-04-21	Mr.	Napoleon578	Wyman904	NaN	NaN	M	asian	...	921 Bins Course	Melrose	Massachusetts	Middlesex County	NaN	42.439301	-71.084713	0	100	Mr. Napoleon578 Wyman904 nan
31	b6633ebf-174a-d0ed-d917-d8289127b1be	1926-07-14	2016-11-06	Mr.	Quentin28	Schmidt332	NaN	NaN	S	white	...	562 Feil Promenade Suite 27	Chelsea	Massachusetts	Suffolk County	NaN	42.400312	-71.007212	0	98	Mr. Quentin28 Schmidt332 nan
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
934	4fa333c9-2380-ecb9-f346-221c14eeb011	1954-01-14	2020-11-15	Mr.	Gerardo48	Valenzuela371	NaN	NaN	M	black	...	183 Shields Annex	Boston	Massachusetts	Suffolk County	2120.0	42.335731	-71.023055	0	71	Mr. Gerardo48 Valenzuela371 nan
949	c1ddd73b-2af1-19ee-40f3-50048992a999	1937-09-04	2018-05-16	Mr.	Tristan353	Beer512	NaN	NaN	S	asian	...	132 Hyatt Park	Boston	Massachusetts	Suffolk County	2126.0	42.362748	-70.990851	0	87	Mr. Tristan353 Beer512 nan
963	0425b284-e697-937d-a00b-6fda0f1b9710	1948-02-28	2017-01-19	Ms.	Angelo118	Hermiston71	NaN	NaN	S	black	...	712 Hartmann View	Boston	Massachusetts	Suffolk County	2467.0	42.396968	-71.003251	0	77	Ms. Angelo118 Hermiston71 nan
964	c0513da-7f35-b4eb-5bab-28c87ff97a10	1959-03-04	2014-11-16	Mr.	Tomas436	Hermann103	NaN	NaN	M	asian	...	754 Pfannerstill Park	Medford	Massachusetts	Middlesex County	2155.0	42.416156	-71.125391	0	66	Mr. Tomas436 Hermann103 nan
967	7506d350-0f35-7c82-3b8a-d7aa80114352	1942-05-18	2018-03-29	Mrs.	Maren639	Breitenberg711	NaN	Ritchie586	M	white	...	460 Padberg Dale Apt 89	Boston	Massachusetts	Suffolk County	2131.0	42.325310	-71.091714	0	82	Mrs. Maren639 Breitenberg711 nan

154 rows × 23 columns

Filtra las personas que tienen una fecha defunción registrada. Este Código selecciona solo filas del dataframe que no es nan.

```
#Cantidad de personas fallecidas por cada categoría de raza

df[df['IS_ALIVE'] == 0]['RACE'].value_counts()
```

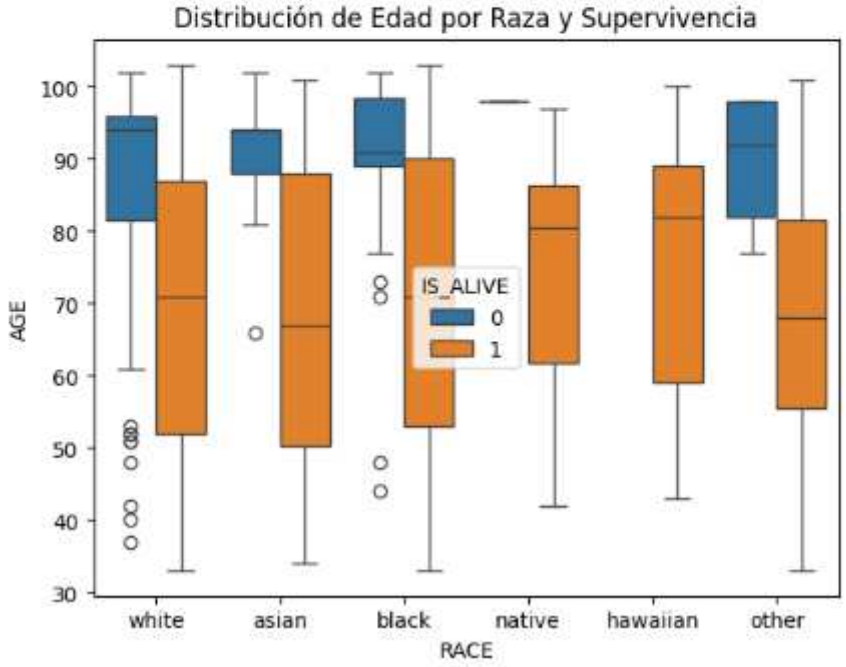
RACE	count
white	103
black	30
asian	13
other	5
native	3

dtype: int64

La raza blanca representa la mayoría de los fallecimientos (alrededor del 69% del total), lo que podría estar relacionado con una mayor proporción de la población blanca en el área estudiada o con otros factores sociodemográficos. Las razas negra y asiática tienen una menor cantidad de fallecimientos en comparación con la raza blanca, con la raza negra alcanzando aproximadamente el 19% y la raza asiática el 8%. Las categorías Other (otros) y Native (nativos) representan una proporción menor de los fallecimientos, con valores del 3.25% y 1.94%, respectivamente.

Nos permite ver si hay diferencias significativas en la cantidad de muertes entre distintos grupos raciales.

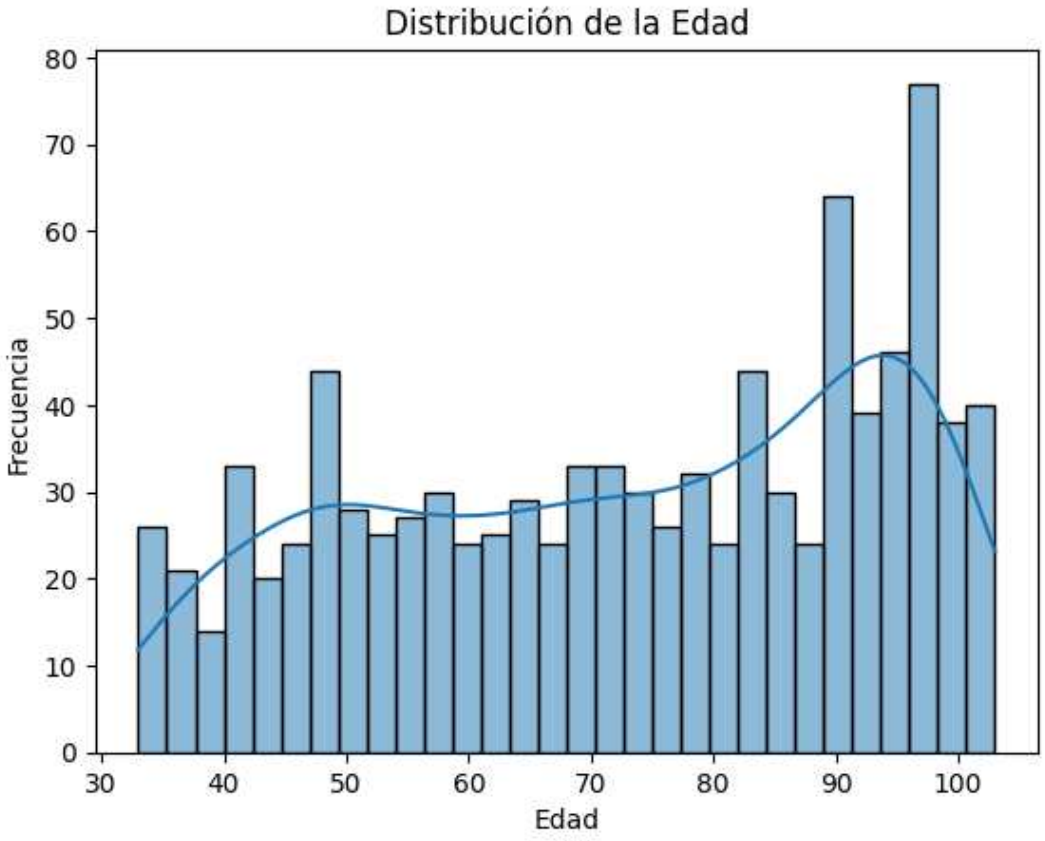
```python
# Boxplot

import seaborn as sns
import matplotlib.pyplot as plt

sns.boxplot(x='RACE', y='AGE', hue='IS_ALIVE', data=df)
plt.title('Distribución de Edad por Raza y Supervivencia')
plt.show()
```



El gráfico muestra que la distribución de la edad varía entre los diferentes grupos raciales.

Busque Comparar la edad de los fallecidos vs. los vivos en cada raza. Identificar posibles outliers (valores extremos). Analizar la variabilidad de edad en cada grupo racial.

```python
#histograma con una curva de densidad KDE

import seaborn as sns
import matplotlib.pyplot as plt

# Histograma de la edad
sns.histplot(df['AGE'], bins=30, kde=True)
plt.title('Distribución de la Edad')
plt.xlabel('Edad')
plt.ylabel('Frecuencia')
plt.show()
```

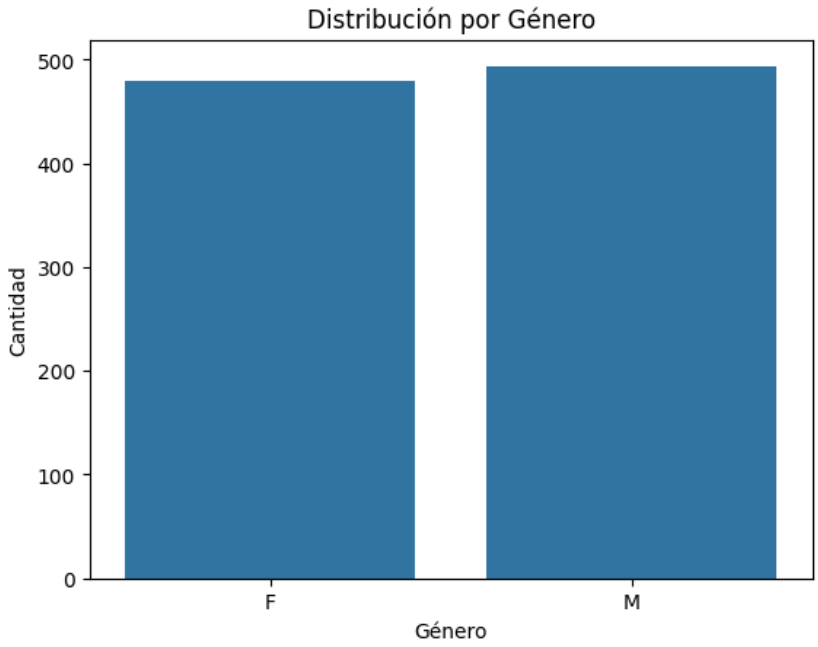


Histograma de la distribución de la edad en el dataframe. Y se visualiza una curva de densisdidal para la distribución suave. El histograma es útil para identificar si la distribución de edades tiene picos, sesgos o una distribución uniforme. Este gráfico es un histograma de la distribución de edades.

Se observa que la distribución de edades no es perfectamente simétrica, parece tener una ligera asimetría hacia la derecha (sesgo positivo). Hay una concentración de personas en el rango de edad de aproximadamente 90 a 100 años. También hay una cantidad considerable de personas en el rango de 60 a 70 años. La curva suavizada muestra una tendencia general

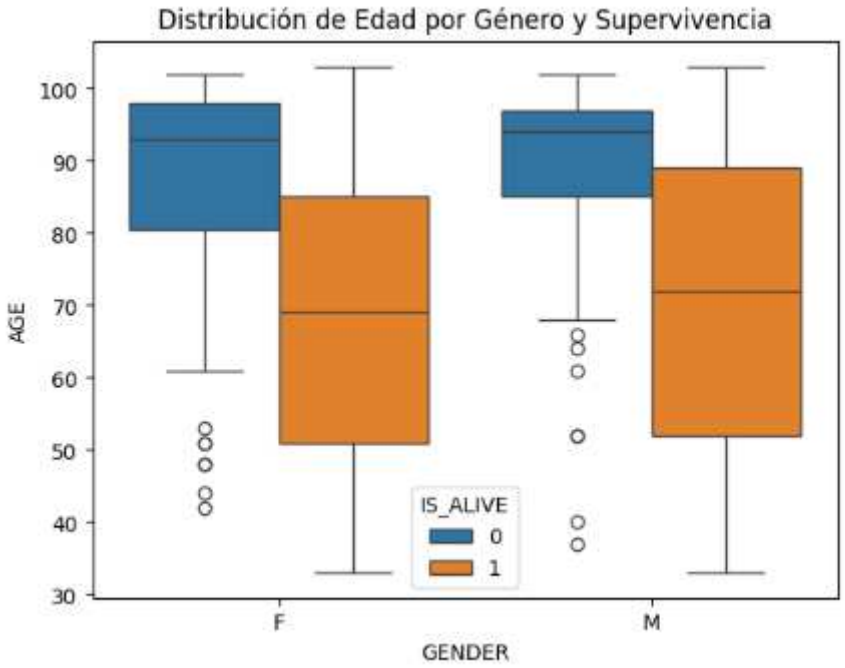
creciente a medida que aumenta la edad, con un pico alrededor de los 90-100 años. En resumen, este gráfico te muestra la distribución de edades en un conjunto de datos, permitiéndote identificar la edad promedio, la dispersión de las edades y la forma general de la distribución.

```python
# --- Análisis Bivariado: Edad y Género ---
import seaborn as sns
import matplotlib.pyplot as plt

# Crear el gráfico de barras
sns.countplot(x='GENDER', data=df)
plt.title('Distribución por Género')
plt.xlabel('Género')
plt.ylabel('Cantidad')
plt.show()
```



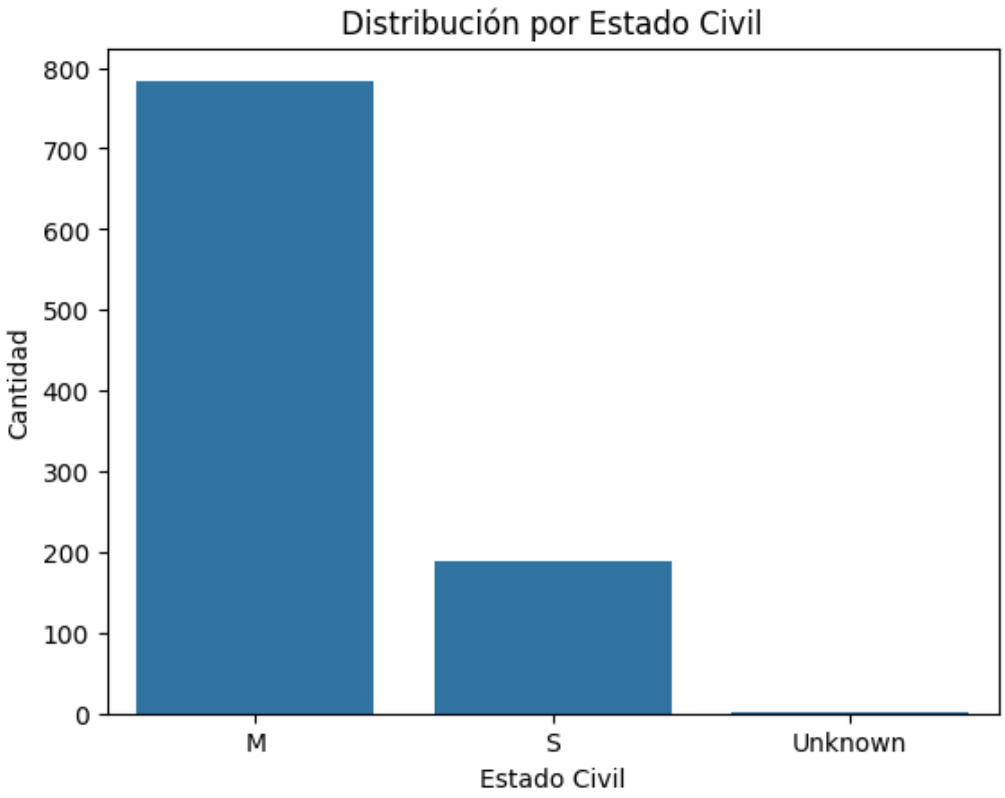
Genere un gráfico de barra que muestra la distribución de los géneros. Este gráfico muestra que la cantidad de hombres y mujeres en el conjunto de datos es muy similar, con una ligera predominancia de hombres. Esto indica una distribución de género bastante equilibrada.

```python
# Boxplot
import seaborn as sns
import matplotlib.pyplot as plt

sns.boxplot(x='GENDER', y='AGE', hue='IS_ALIVE', data=df)
plt.title('Distribución de Edad por Género y Supervivencia')
plt.show()
```



El gráfico de cajas muestra cómo se distribuye la edad según el género y la supervivencia, permitiendo identificar patrones y diferencias entre los grupos. Esutil para visulizar la dispersion, la mediana y los valores atipicos(outliers) en los datos. Mediana(linea central dentro de la caja) muestra la edad central(valor que divide la mitad inferior y superior) Valores atipicos(puntos fuera de los bigotes) si hay valores de edad extremadamente altos o bajos, apareceran como puntos fuera de los bigotes. Conclusiones: Las personas fallecidas tendían a ser mayores y tenían una distribución de edad más amplia que las personas que sobrevivieron. Además, hubo algunos valores atípicos en el grupo de fallecidos, lo que sugiere que hubo algunas personas jóvenes que fallecieron.

```
#Grafico de Barras

sns.countplot(x='MARITAL', data=df)
plt.title('Distribución por Estado Civil')
plt.xlabel('Estado Civil')
plt.ylabel('Cantidad')
plt.show()
```



Este gráfico muestra la cantidad de personas que pertenecen a cada categoría de estado civil, la mayoría de las personas están casadas, seguidas por un número menor de personas solteras, y una cantidad muy pequeña de personas cuyo estado civil se desconoce.

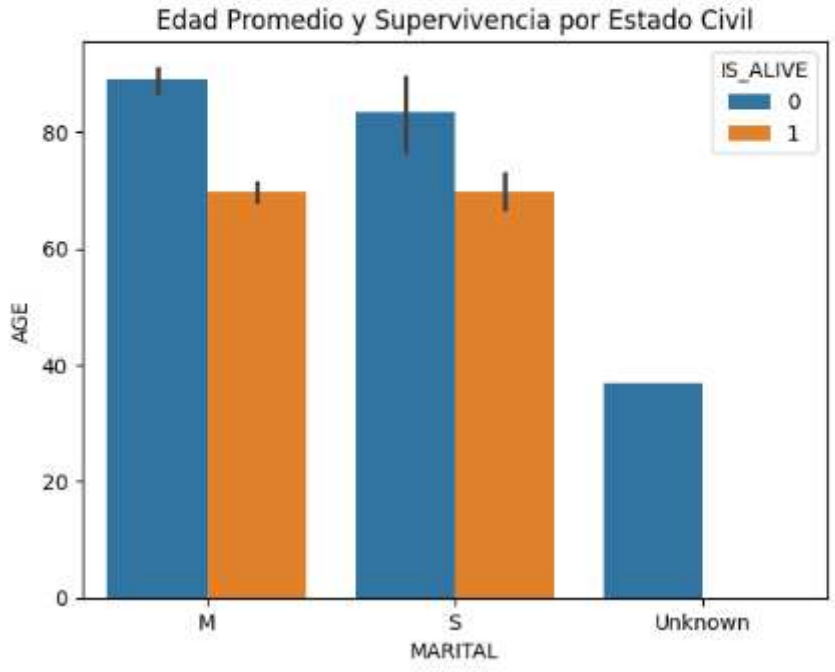
```
#Calcular proporcion de supervivvencia en cada grupo de estado civil

df.groupby('MARITAL')['IS_ALIVE'].mean()
```

MARITAL	IS_ALIVE
M	0.843112
S	0.841270
Unknown	0.000000

dtype: float64

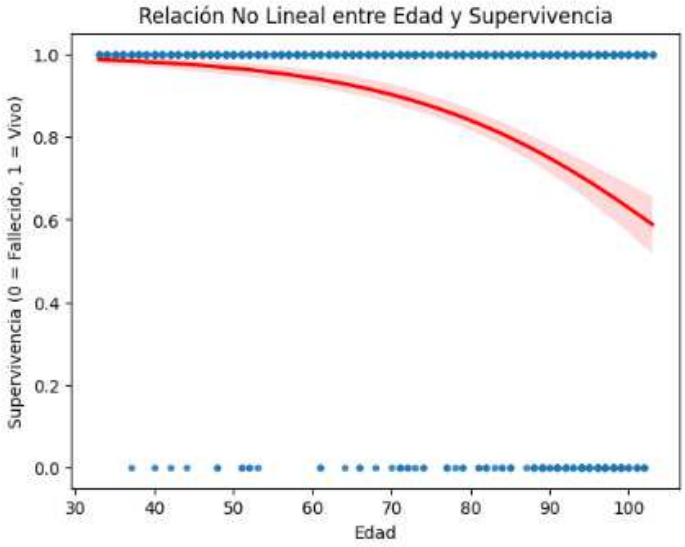
Es un análisis de datos en que se plantea el estado civil puede influir en la probabilidad de supervivencia de las personas. El análisis según los datos obtenidos sugiere que el estado civil no parece ser un factor determinante de supervivencia ya que las tasas son muy similares.

```
# Análisis bivariado utilizando un gráfico de barras

sns.barplot(x='MARITAL', y='AGE', hue='IS_ALIVE', data=df)
plt.title('Edad Promedio y Supervivencia por Estado Civil')
plt.show()
```



Si hay diferencias claras en la edad de fallecidos y vivos según el estado civil

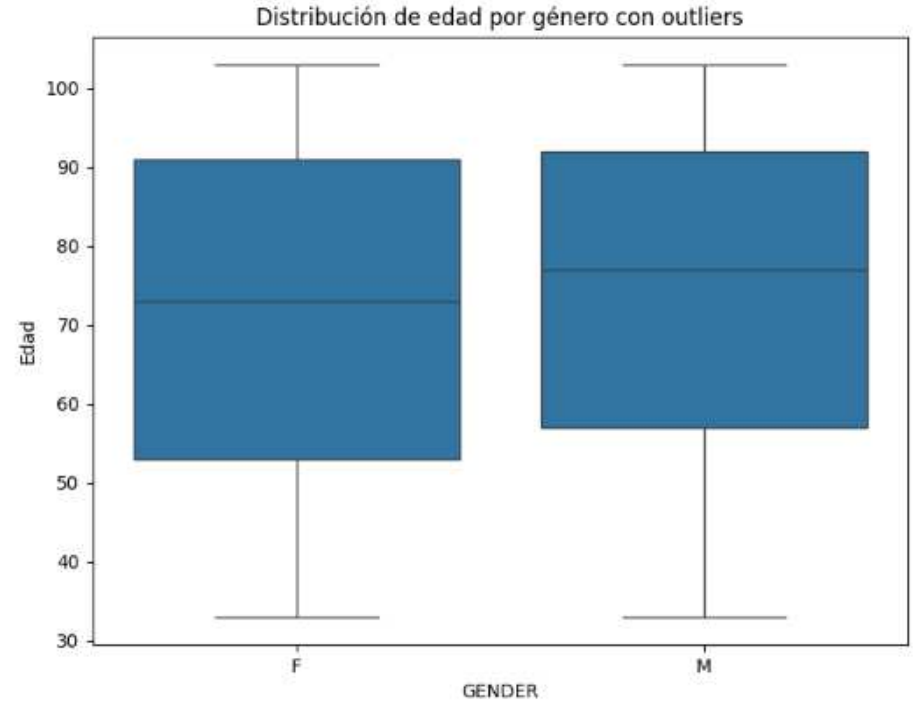
```python
#Análisis bivariado usando regresion logistica
import seaborn as sns
import matplotlib.pyplot as plt

sns.regplot(x='AGE', y='IS_ALIVE', data=df, logistic=True, scatter_kws={'s': 10}, line_kws={'color': 'red'})
plt.title('Relación No Lineal entre Edad y Supervivencia')
plt.xlabel('Edad')
plt.ylabel('Supervivencia (0 = Fallecido, 1 = Vivo)')
plt.show()
```



Relación no lineal: El gráfico muestra claramente que la relación entre la edad y la supervivencia no es lineal. La probabilidad de supervivencia disminuye a medida que aumenta la edad, pero la tasa de disminución no es constante. Disminución de la supervivencia con la edad: La curva roja muestra una tendencia descendente, lo que indica que la probabilidad de supervivencia disminuye significativamente a medida que las personas envejecen. Mayor dispersión en edades avanzadas: Se observa una mayor dispersión de los puntos en las edades más avanzadas, lo que sugiere una mayor variabilidad en la supervivencia en este grupo de edad.
Conclusiones: El gráfico sugiere que la edad es un factor importante que influye en la supervivencia. La probabilidad de supervivencia disminuye a medida que aumenta la edad, y esta disminución no es lineal. En edades avanzadas, la variabilidad en la supervivencia es mayor.

```python
# Análisis Univariado y de Detección de Outliers.

import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Convertir 'BIRTHDATE' a datetime y calcular la edad
df["BIRTHDATE"] = pd.to_datetime(df["BIRTHDATE"], errors="coerce")
df["Edad"] = (pd.to_datetime("today") - df["BIRTHDATE"]).dt.days // 365  # Convertir días a años

# Función para detectar outliers usando el método del Rango Intercuartílico (IQR)
def detectar_outliers(df, columna):
    Q1 = df[columna].quantile(0.25)
    Q3 = df[columna].quantile(0.75)
    IQR = Q3 - Q1
    limite_inferior = Q1 - 1.5 * IQR
    limite_superior = Q3 + 1.5 * IQR
    outliers = df[(df[columna] < limite_inferior) | (df[columna] > limite_superior)]
    return outliers

# Identificar outliers por género
outliers_hombres = detectar_outliers(df[df["GENDER"] == "M"], "Edad")
outliers_mujeres = detectar_outliers(df[df["GENDER"] == "F"], "Edad")

print(f"Número de outliers en hombres: {len(outliers_hombres)}")
print(f"Número de outliers en mujeres: {len(outliers_mujeres)}")

# Visualización con boxplot
plt.figure(figsize=(8, 6))
sns.boxplot(x="GENDER", y="Edad", data=df)
plt.title("Distribución de edad por género con outliers")
plt.show()
```

```
Número de outliers en hombres: 0
Número de outliers en mujeres: 0
```



En este gráfico no se muestran valores atípicos (outliers). Diferencias en la mediana: Hay una ligera diferencia en la mediana de edad entre hombres y mujeres, siendo ligeramente superior para los hombres. No hay valores

atípicos: No hay individuos con edades significativamente diferentes a la mayoría en ninguno de los géneros.

```
#Analisis descriptivo

mean_age = df['AGE'].mean()
median_age = df['AGE'].median()
std_age = df['AGE'].std()

print(f'Media de Edad: {mean_age}')
print(f'Mediana de Edad: {median_age}')
print(f'Desviación Estándar de Edad: {std_age}')
```

```
Media de Edad: 72.52053388090349
Mediana de Edad: 74.0
Desviación Estándar de Edad: 20.504808282832215
```

Calcula 3 estadísticas descriptivas: Media: el valor promedio de las edades. Mediana: es el valor que divide dos mitades. desviación estándar: Mide la dispersión de las edades en relacion con la media, cuanto mayor sea la desviación más dispersa están las edades. La interpretación es que en promedio las personas son mayores a 70,la mediana inidca que la mitad de las personas tienen 74 años o menos y la otra mitad 74 años o mas. La desviación estándar es relativamente alta que las edades estan dispersas en torno a la media.

```
# Prueba t para comparar medias de edad entre géneros
from scipy import stats

# Prueba t entre hombres y mujeres
male_ages = df[df['GENDER'] == 'M']['AGE']
female_ages = df[df['GENDER'] == 'F']['AGE']

t_stat, p_value = stats.ttest_ind(male_ages, female_ages)
print(f'Valor p de la prueba t: {p_value}')

# Interpretación
if p_value < 0.05:
    print('Existen diferencias significativas entre las edades de hombres y mujeres.')
else:
    print('No hay diferencias significativas entre las edades de hombres y mujeres.')
```

```
Valor p de la prueba t: 0.08012849840614014
No hay diferencias significativas entre las edades de hombres y mujeres.
```

Hipótesis nula (H_0): Es la suposición inicial que hacemos antes de realizar la prueba. En este caso, la hipótesis nula es que no hay diferencia

significativa en las edades entre hombres y mujeres.Es un buen análisis estadístico inferencial.

```
# Correlación

# Calcular la correlación entre la edad y la variable IS_ALIVE
correlation_alive = df[['AGE', 'IS_ALIVE']].corr()

# Imprimir el resultado
print(f"Correlación entre la edad y la supervivencia (IS_ALIVE): \n{correlation_alive}")
```

```
Correlación entre la edad y la supervivencia (IS_ALIVE):
               AGE  IS_ALIVE
AGE       1.000000 -0.317986
IS_ALIVE -0.317986  1.000000
```

La correlación entre la edad y supervivencia es negativa moderada, el valor negativo sugiere que las personas mayores tienen más probabilidades de haber fallecido y la relacion moderada significa que la edad no es el único factor que determina si una persona está viva o fallecida. Muy importante para cuantificar la relacion.

```
#Prueba Chi-cuadrado de independencia

import pandas as pd
import scipy.stats as stats

# Crear un DataFrame de ejemplo
data = {
    'GENDER': ['Male', 'Female', 'Male', 'Female', 'Male', 'Female', 'Male', 'Female'],
    'IS_ALIVE': ['Yes', 'No', 'Yes', 'No', 'Yes', 'Yes', 'No', 'Yes'],
    'RACE': ['White', 'Black', 'White', 'Black', 'White', 'Black', 'White', 'Black'],
    'MARITAL_STATUS': ['Single', 'Married', 'Single', 'Married', 'Single', 'Married', 'Single', 'Married']
}

df = pd.DataFrame(data)

# Prueba Chi-cuadrado para Género y Supervivencia
contingency_gender_survival = pd.crosstab(df['GENDER'], df['IS_ALIVE'])
chi2_gender_survival, p_gender_survival, dof_gender_survival, expected_gender_survival = stats.chi2_contingency(contingency_gender_survival)
print(f"Prueba Chi-cuadrado - Género y Supervivencia:")
print(f"Chi-cuadrado: {chi2_gender_survival}")
print(f"Valor p: {p_gender_survival}")
print(f"Grados de libertad: {dof_gender_survival}")
print(f"Tabla esperada: \n{expected_gender_survival}")
print(f"Relación significativa: {'Sí' if p_gender_survival < 0.05 else 'No'}\n")

# Prueba Chi-cuadrado para Raza y Supervivencia
contingency_race_survival = pd.crosstab(df['RACE'], df['IS_ALIVE'])
chi2_race_survival, p_race_survival, dof_race_survival, expected_race_survival = stats.chi2_contingency(contingency_race_survival)
print(f"Prueba Chi-cuadrado - Raza y Supervivencia:")
print(f"Chi-cuadrado: {chi2_race_survival}")
print(f"Valor p: {p_race_survival}")
print(f"Grados de libertad: {dof_race_survival}")
print(f"Tabla esperada: \n{expected_race_survival}")
print(f"Relación significativa: {'Sí' if p_race_survival < 0.05 else 'No'}\n")

# Prueba Chi-cuadrado para Estado Civil y Supervivencia
contingency_marital_survival = pd.crosstab(df['MARITAL_STATUS'], df['IS_ALIVE'])
chi2_marital_survival, p_marital_survival, dof_marital_survival, expected_marital_survival = stats.chi2_contingency(contingency_marital_survival)
print(f"Prueba Chi-cuadrado - Estado Civil y Supervivencia:")
print(f"Chi-cuadrado: {chi2_marital_survival}")
print(f"Valor p: {p_marital_survival}")
print(f"Grados de libertad: {dof_marital_survival}")
print(f"Tabla esperada: \n{expected_marital_survival}")
print(f"Relación significativa: {'Sí' if p_marital_survival < 0.05 else 'No'}\n")
```

```
Prueba Chi-cuadrado - Género y Supervivencia:
Chi-cuadrado: 0.0
Valor p: 1.0
Grados de libertad: 1
Tabla esperada:
[[1.5 2.5]
 [1.5 2.5]]
Relación significativa: No

Prueba Chi-cuadrado - Raza y Supervivencia:
Chi-cuadrado: 0.0
Valor p: 1.0
Grados de libertad: 1
Tabla esperada:
[[1.5 2.5]
 [1.5 2.5]]
Relación significativa: No

Prueba Chi-cuadrado - Estado Civil y Supervivencia:
Chi-cuadrado: 0.0
Valor p: 1.0
Grados de libertad: 1
Tabla esperada:
[[1.5 2.5]
 [1.5 2.5]]
Relación significativa: No
```

Para analizar la relación entre dos variables categóricas (por ejemplo, género y supervivencia, raza y supervivencia, estado civil y supervivencia). Te dirá si la relación es estadísticamente significativa.

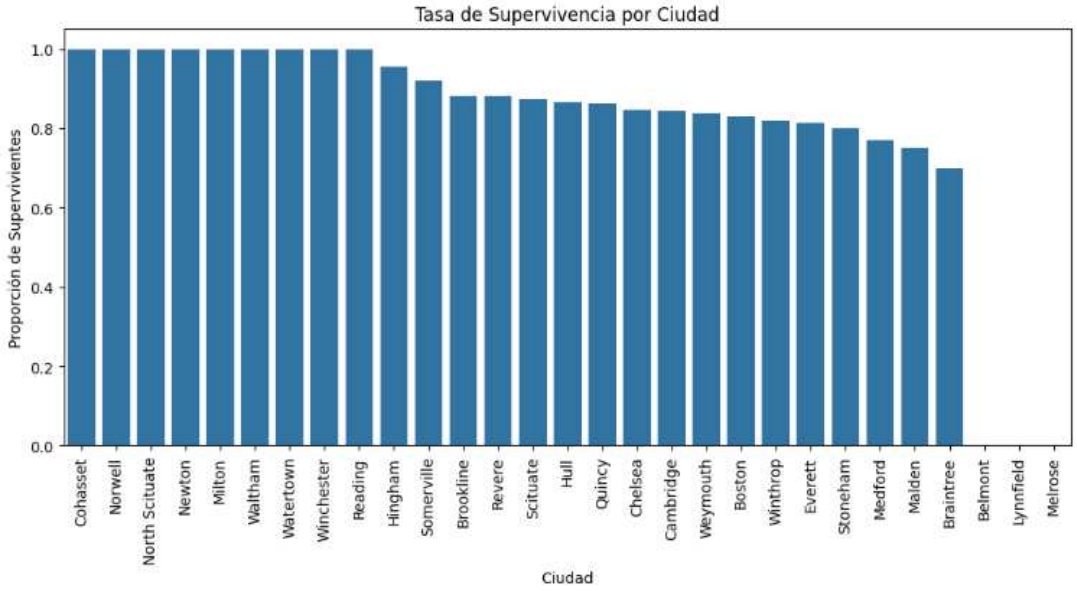
Los valores p para todas las pruebas son 1.0, lo que indica que no hay relación significativa entre las variables y la supervivencia. Esto significa que, según los datos que proporcionas, el género, la raza y el estado civil no parecen estar relacionados de manera significativa con la supervivencia.

```python
#Análisis descriptivo bivariado

# Supervivencia por ciudad
survival_by_city = df.groupby('CITY')['IS_ALIVE'].mean().sort_values(ascending=False)

# Mostrar resultados
print(survival_by_city)
```

```
CITY
Cohasset          1.000000
Norwell           1.000000
North Scituate    1.000000
Newton            1.000000
Milton            1.000000
Waltham           1.000000
Watertown         1.000000
Winchester        1.000000
Reading           1.000000
Hingham           0.954545
Somerville        0.920000
Brookline         0.882353
Revere            0.880952
Scituate          0.875000
Hull              0.866667
Quincy            0.862500
Chelsea           0.846154
Cambridge         0.844444
Weymouth          0.837838
Boston            0.829945
Winthrop          0.818182
Everett           0.812500
Stoneham          0.800000
Medford           0.769231
Malden            0.750000
Braintree         0.700000
Belmont           0.000000
Lynnfield         0.000000
Melrose           0.000000
Name: IS_ALIVE, dtype: float64
```

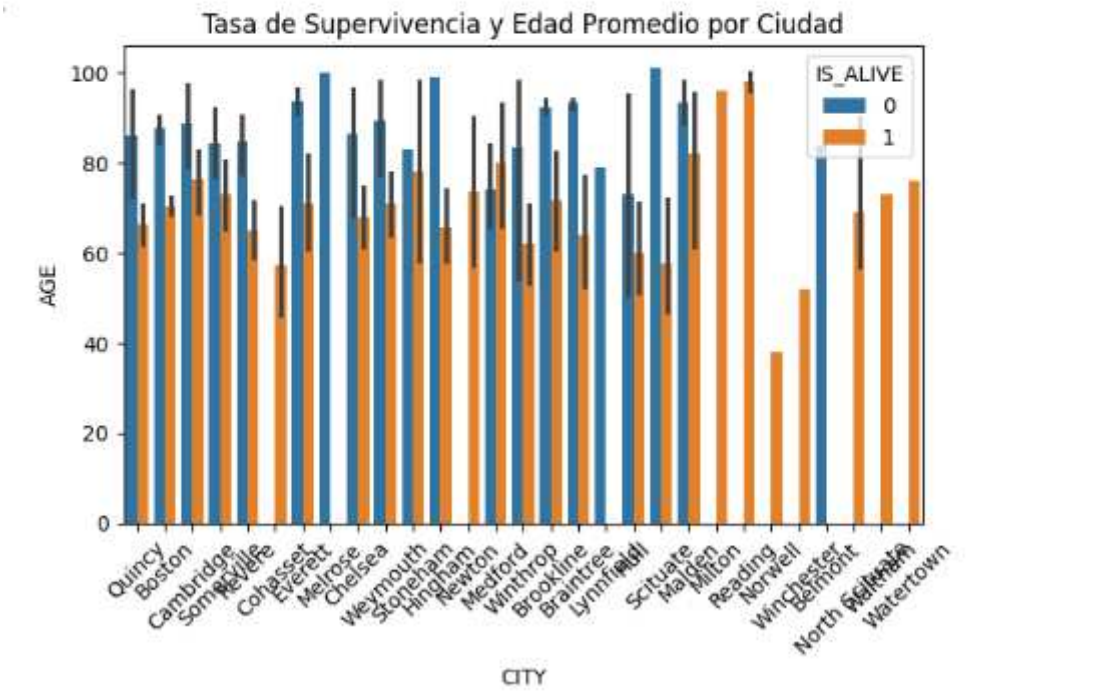
Ayuda a saber si ciertas ciudades tienen tasas de supervivencias mas altas o bajas , se puede interpretar que la ubicacion no tienen un gran impacto. MEelrose,Lynnfield y Belmont podrian ser ciertas ciudades que tienen mayor poblacion de edad avanzada.

```python
#Gráfico de barras utilizando Seaborn y Matplotlib

import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(12, 5))
sns.barplot(x=survival_by_city.index, y=survival_by_city.values)
plt.xticks(rotation=90)
plt.title('Tasa de Supervivencia por Ciudad')
plt.xlabel('Ciudad')
plt.ylabel('Proporción de Supervivientes')
plt.show()
```



Este gráfico proporciona una visión general clara de las diferencias en la tasa de supervivencia entre las ciudades. Permite identificar rápidamente las ciudades con las tasas de supervivencia más altas y más bajas, y sugiere la necesidad de realizar investigaciones adicionales para comprender las causas subyacentes de estas diferencias.

```python
# --- Análisis Bivariado: Edad y Supervivencia Ciudad---

sns.barplot(x='CITY', y='AGE', hue='IS_ALIVE', data=df)
plt.title('Tasa de Supervivencia y Edad Promedio por Ciudad')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



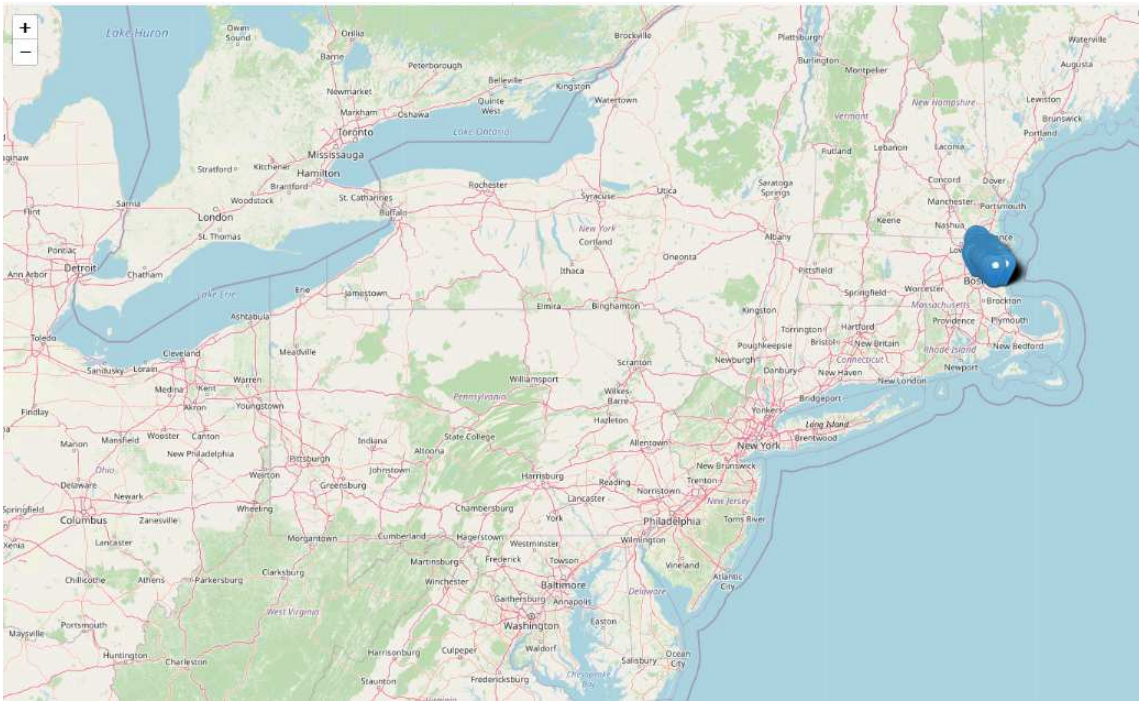
El gráfico proporciona una visión general de cómo la edad promedio varía entre los grupos de supervivencia en diferentes ciudades y muestra que algunas ciudades presentan diferencias más pronunciadas que otras. Este tipo de análisis es solo descriptivo.

```python
# Prueba ANOVA (Análisis de Varianza)

import pandas as pd
import scipy.stats as stats

# Filtramos los datos en base a la columna 'IS_ALIVE'
grupo_vivo = df[df['IS_ALIVE'] == 1]['AGE']
grupo_fallecido = df[df['IS_ALIVE'] == 0]['AGE']

# Realizamos la prueba ANOVA
f_stat, p_value = stats.f_oneway(grupo_vivo, grupo_fallecido)

# Imprimimos los resultados
print("Estadístico F:", f_stat)
print("Valor p:", p_value)

# Interpretación de los resultados
if p_value < 0.05:
    print("La diferencia en la edad promedio entre los grupos es estadísticamente significativ
else:
    print("No hay diferencias significativas en la edad promedio entre los grupos.")
```

```
Estadístico F: 109.33988410083626
Valor p: 2.5227522848836707e-24
La diferencia en la edad promedio entre los grupos es estadísticamente significativa.
```

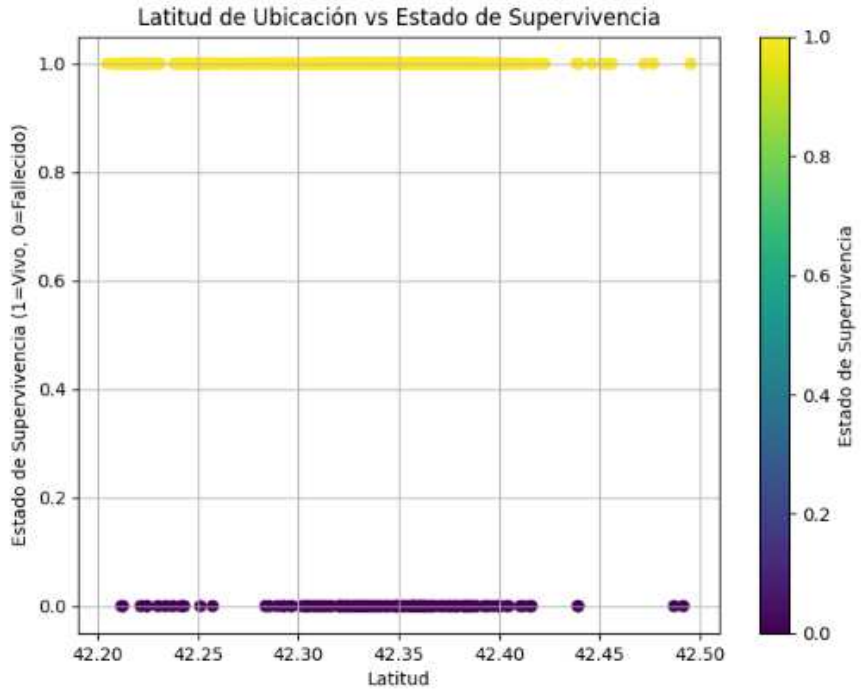
Esto sugiere que la edad juega un papel importante en la supervivencia, y probablemente hay una mayor incidencia de fallecimientos en ciertos rangos de edad. La edad es un factor relevante para la supervivencia, y este análisis confirma que las diferencias en la edad promedio entre los grupos son significativas desde el punto de vista estadístico.

```python
# Análisis Geográfico
import folium
from IPython.display import display

# Crear un mapa centrado en la media de las coordenadas 'LAT' y 'LON'
mapa = folium.Map(location=[df['LAT'].mean(), df['LON'].mean()], zoom_start=10)

# Agregar marcadores al mapa
for _, row in df.iterrows():
    folium.Marker(location=[row['LAT'], row['LON']], popup=row['FULL_NAME']).add_to(mapa)

# Mostrar el mapa en Jupyter Notebook
display(mapa)
```



Este código generará un mapa interactivo donde cada punto representa a una persona, y al hacer clic en un marcador, se mostrará su nombre.

```python
#Análisis Geográfico

# Graficar la latitud de ubicación según si está viva o no
plt.figure(figsize=(8, 6))
plt.scatter(df["LAT"], df['IS_ALIVE'], c=df['IS_ALIVE'], cmap="viridis", label='Alive Status')

# Personalizar el gráfico
plt.title('Latitud de Ubicación vs Estado de Supervivencia')
plt.xlabel('Latitud')
plt.ylabel('Estado de Supervivencia (1=Vivo, 0=Fallecido)')
plt.colorbar(label='Estado de Supervivencia')
plt.grid(True)
```



Este gráfico muestra que no existe una relación aparente entre la latitud de ubicación y el estado de supervivencia. La supervivencia parece ser independiente de la latitud dentro del rango mostrado. Esto sugiere que otros factores, no relacionados con la latitud, son los que influyen en la supervivencia

.

Interpretaciones Finales

1-Distribución de Edad

Análisis descriptivo de la edad

Hemos calculado la media que es aproximadamente de 72,52 años, mediana 74 años y desviación estándar de la edad de 20,50, lo que nos ayuda a entender la distribución general de las edades en el conjunto de datos. Esto es útil para determinar si las edades están distribuidas de manera simétrica (positiva la media y mediana son aproximadamente iguales) o si existen sesgos (la medianada es mayor que la media,sesgo a la izquierda) lo que sugiere que hay mas valores en edades avanzadas.

Histogramas y Boxplots

La visualización de la distribución de la edad muestra una mayor tasa de mortalidad en personas mayores, junto con el análisis de la presencia de outliers, nos ha permitido observar cómo se distribuye la edad entre los individuos, segmentada por género, estado civil y raza.

2-Relación entre Edad y Supervivencia:

Boxplots y análisis bivariado

Se exploró la relación entre la edad y la supervivencia,con el genero, raza , ciudad y estado civil tanto de manera gráfica (con boxplots) como cuantitativa. Esto nos mostró cómo se distribuyen las edades entre los que están vivos y los que han fallecido, y nos dio una idea de si la edad influye significativamente en la probabilidad de supervivencia.

Análisis de regresión logística Se utilizó un gráfico de regresión logística para observar la relación no lineal entre la edad y la supervivencia. Este análisis fue crucial ya que nos indico que la edad es un factor clave en la supervivencia, la relacion no es lineal.

3-Correlacion # Analisis estadistico.

La correlacion estre la edad y supervivencia es negativa moderada, el valor negativo sugiere que las personas mayores tienen mas probabilidades de haber fallecido.

4-Prueba t para comparar medias de edad entre géneros. #Análisis estadístico inferencial.

La hipótesis nula es que no hay diferencia significativa en las edades entre hombres y mujeres.Es un buen análisis estadístico inferencial.

5-Prueba ANOVA (Análisis de Varianza)

El resultado sugiere que la edad es un factor relevante en la supervivencia, ya que hay una diferencia significativa en la edad promedio entre los sobrevivientes y los fallecidos. La probabilidad de supervivencia probablemente varía dependiendo de la edad, con las personas de mayor edad teniendo una mayor probabilidad de fallecer.

6-Análisis de Variables Categóricas:

Chi-cuadrado para variables categóricas: Se realizaron pruebas de Chi-cuadrado para determinar si existen relaciones estadísticamente significativas entre la supervivencia y variables como el género, raza y estado civil. Los resultados de estas pruebas nos ayudaron a entender que estas variables están no estan asociadas con la supervivencia.

7-Mortalidad Y Supervivencia por Ciudad:

Tasa de mortalidad por ciudad: Al analizar la tasa de mortalidad por ciudad, se observó que ciertas ciudades tienen una mayor tasa de mortalidad. Esto podría reflejar desigualdades en el acceso a la salud, condiciones socioeconómicas o factores ambientales que afectan la supervivencia de las personas o hay ciudades con mayor poblacion de edad avanzada.

8-Outliers:

Se identificaron y analizó la presencia de outliers en la variable "Edad", lo cual es importante porque los valores extremos pueden influir significativamente en los resultados de los análisis estadísticos. Algunos jóvenes fallecidos aparecen como outliers sugiere que algunas muertes

ocurrieron en edades inesperadas, lo que podría deberse a condiciones preexistentes, accidentes u otras causas.

9-Análisis Geografico

El gráfico muestra que no existe una relación aparente entre la latitud de ubicación y el estado de supervivencia. La supervivencia parece ser independiente de la latitud dentro del rango mostrado.

Conclusiones:

Los resultados indican que la edad como factor clave en la supervivencia. La edad juega un papel crucial en la probabilidad de supervivencia. Los análisis descriptivos, como la media, la mediana y la desviación estándar, indican que la mayoría de las personas fallecidas eran mayores, con una distribución sesgada a la izquierda, lo que sugiere que las muertes son más comunes en edades avanzadas. Este patrón se confirma mediante el análisis de regresión logística, que muestra una relación no lineal entre la edad y la supervivencia. La correlación negativa moderada entre la edad y la supervivencia también resalta que a medida que aumenta la edad, la probabilidad de fallecimiento se incrementa

Es fundamental recordar que correlación no implica causalidad, pero una fuerte correlación entre dos variables puede ser útil para hacer predicciones. Si bien la edad es un factor asociado a una menor supervivencia, no necesariamente es la causa directa de este fenómeno, ya que pueden intervenir variables adicionales como el estado de salud, el acceso a tratamientos médicos o condiciones preexistentes. Pero uno de los enfoques más utilizados para analizar esta relación es el cálculo de la correlación, que mide la fuerza y dirección del vínculo entre dos variables.

Se observó una mayor tasa de mortalidad en personas mayores, lo cual es consistente con estudios previos sobre la relación entre envejecimiento y riesgo de fallecimiento. Los outliers detectados en la variable de edad, como jóvenes que fallecieron, sugieren que hay factores adicionales, como condiciones preexistentes o accidentes, que podrían explicar muertes en rangos de edad inesperados. A pesar de realizar pruebas de Chi-cuadrado

para explorar la relación entre género, raza y estado civil con la supervivencia, los resultados sugieren que estas variables no tienen una relación estadísticamente significativa con la mortalidad en este conjunto de datos. Esto indica que, al menos en este caso, factores como género, raza y estado civil no influyen significativamente en la supervivencia. La Prueba t y la ANOVA confirmaron que existen diferencias estadísticamente significativas en la edad promedio entre los grupos de sobrevivientes y fallecidos, lo que refuerza la hipótesis de que la edad es un determinante relevante de la supervivencia. Las personas mayores tienen una mayor probabilidad de fallecer, y esta diferencia es estadísticamente significativa.

El análisis de la tasa de mortalidad por ciudad muestra que algunas ciudades tienen una mayor tasa de mortalidad, lo que podría estar relacionado con factores socioeconómicos, acceso a la salud o características demográficas (por ejemplo, ciudades con mayor población de edad avanzada). Sin embargo, no se encontró una relación directa entre la latitud y la supervivencia, lo que sugiere que la ubicación geográfica, en términos de latitud, no es un factor clave en este caso. Desde una perspectiva práctica, estos hallazgos refuerzan la idea de que la edad es un factor de riesgo clave en la mortalidad, lo que puede ser relevante para la formulación de políticas de salud pública, la distribución de recursos médicos y la investigación en estrategias de prevención y tratamiento, subrayando la importancia de considerar la edad como factor de riesgo clave en la mortalidad.