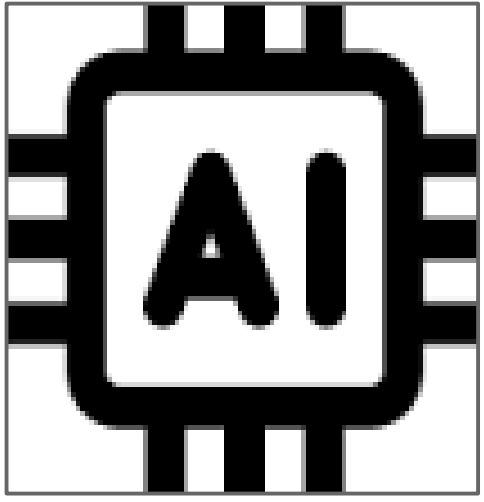
Introducción a la IA



Gráficas y Visualización de Datos

Presentación

Introducción

La visualización de datos es un aspecto fundamental en todo proyecto de Ciencia de Datos.

Mientras que los indicadores y estadísticos permiten presentar valores numéricos que condensan ciertos valores, los gráficos ayudan al fácil entendimiento y comprensión del contenido subyacente los datos.

Tanto indicadores estadísticos como gráficas son de gran utilidad tanto como herramienta para el análisis preliminar como para la evaluación y comunicación efectiva de los resultados.

Las librerías más usadas en Python son:

- Para obtención de indicadores y estadísticos: Pandas
- Para la presentación de gráficas: Matplotlib y Seaborn

Indicadores y Estadísticos con Pandas

Pandas y la Estadística Descriptiva

Pandas permite asistirnos en la obtención de indicadores y gráficas para la Estadística Descriptiva.

Estadísticos e indicadores:

- Media, mediana, cuenta, máximo y mínimo, desvío estándar, cuartiles, varianza, co-varianza, entre otros.

Algunos Gráficos:

- Histogramas, boxplots, líneas y/o barras.

Indicadores y Estadísticos con Pandas

Los principales estadísticos:

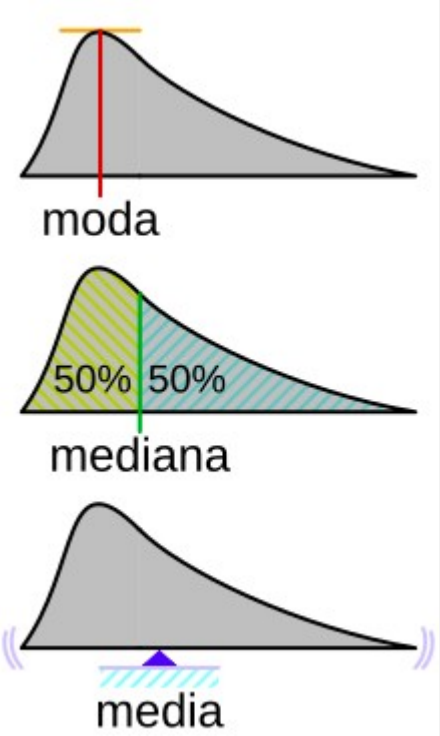
Indicadores de tendencia central

Media: Es el valor promedio.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Mediana: Es el valor no superado por el 50% de los datos.

Moda: es el valor que aparece con mayor frecuencia en un conjunto de datos.



Indicadores y Estadísticos con Pandas

Los principales estadísticos:

Indicadores de dispersión

Varianza: medida de dispersión definida como la esperanza del cuadrado de la desviación de dicha variable respecto a su media.

$$\sigma^2 = \frac{\sum(x-\mu)^2}{n}$$

Varianza Poblacional

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

Varianza Muestral

Desvío estándar: Se calcula como la raíz cuadrada de la varianza. Indica lejanía o cercanía de los datos alrededor del valor medio.

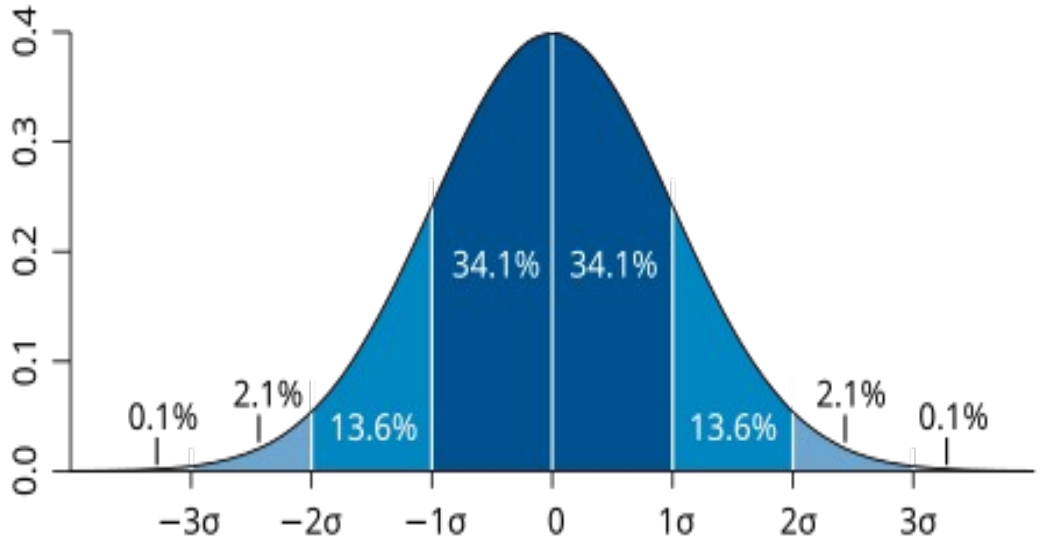
Co-varianza: es un valor que indica el grado de variación conjunta de dos variables aleatorias respecto a sus medias.

Correlación: indica la fuerza y la dirección de una relación lineal y la proporcionalidad entre dos variables estadísticas.

Indicadores y Estadísticos con Pandas

Los principales estadísticos:

Indicadores de dispersión



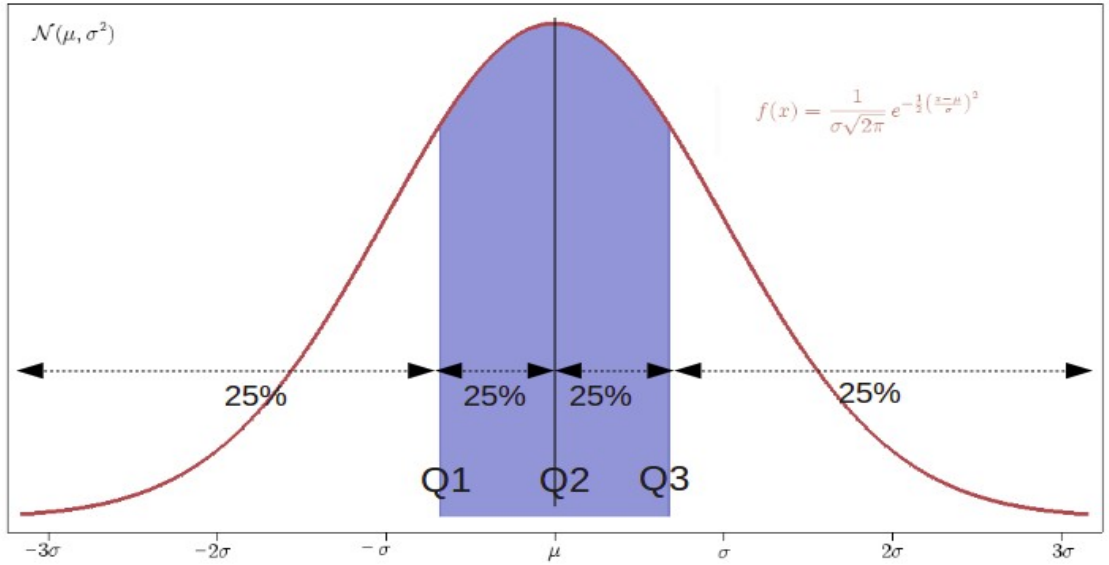
La distribución Normal Estándar y los desvíos estándares (regla 68-95-99.7).

Indicadores y Estadísticos con Pandas

Los principales estadísticos:

Indicadores de posición

Cuartiles: Los cuartiles son medidas de posición. Son tres valores que dividen al conjunto de datos ordenados en cuatro partes iguales.



Indicadores y Estadísticos con Pandas

Los principales estadísticos con Pandas:

Resumen de datos: estadísticos

```
df.describe()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Diabetes
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	428.235091	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	340.485655	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.100000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	205.000000	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	337.000000	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	591.500000	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2329.000000	81.000000	1.000000

Indicadores y Estadísticos con Pandas

Los principales estadísticos con Pandas:

Media y mediana:

```
df.mean()
```

```
Pregnancies                 3.845052
Glucose                   120.894531
BloodPressure              69.105469
SkinThickness              20.536458
Insulin                    79.799479
BMI                        31.992578
DiabetesPedigreeFunction  428.235091
Age                        33.240885
Diabetes                    0.348958
dtype: float64
```

```
df.median()
```

```
Pregnancies                 3.0
Glucose                   117.0
BloodPressure              72.0
SkinThickness              23.0
Insulin                    30.5
BMI                        32.0
DiabetesPedigreeFunction  337.0
Age                        29.0
Diabetes                    0.0
dtype: float64
```

Indicadores y Estadísticos con Pandas

Los principales estadísticos con Pandas:

Desvío estándar:

```
df.std()
```

```
Pregnancies                 3.369578
Glucose                    31.972618
BloodPressure              19.355807
SkinThickness              15.952218
Insulin                   115.244002
BMI                         7.884160
DiabetesPedigreeFunction  340.485655
Age                        11.760232
Diabetes                    0.476951
dtype: float64
```

Indicadores y Estadísticos con Pandas

Los principales estadísticos con Pandas:

Mínimo y máximo:

```
df.min()
```

```
Pregnancies                 0.0
Glucose                     0.0
BloodPressure               0.0
SkinThickness               0.0
Insulin                     0.0
BMI                         0.0
DiabetesPedigreeFunction    0.1
Age                        21.0
Diabetes                    0.0
dtype: float64
```

```
df.max()
```

```
Pregnancies                 17.0
Glucose                    199.0
BloodPressure              122.0
SkinThickness               99.0
Insulin                    846.0
BMI                         67.1
DiabetesPedigreeFunction  2329.0
Age                         81.0
Diabetes                     1.0
dtype: float64
```

Indicadores y Estadísticos con Pandas

Los principales estadísticos con Pandas:

Suma de valores

```
df.sum()
```

```
Pregnancies                 2953.00
Glucose                    92847.00
BloodPressure              53073.00
SkinThickness              15772.00
Insulin                    61286.00
BMI                        24570.30
DiabetesPedigreeFunction  328884.55
Age                        25529.00
Diabetes                     268.00
dtype: float64
```

```
df["Pregnancies"].sum()
```

```
2953
```

Indicadores y Estadísticos con Pandas

Los principales estadísticos con Pandas:

Matriz de correlación

```
df.corr()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Diabetes
Pregnancies	1.000000	0.129459	0.141282	-0.081672	-0.073535	0.017683	-0.026205	0.544341	0.221898
Glucose	0.129459	1.000000	0.152590	0.057328	0.331357	0.221071	0.133163	0.263514	0.466581
BloodPressure	0.141282	0.152590	1.000000	0.207371	0.088933	0.281805	0.051436	0.239528	0.065068
SkinThickness	-0.081672	0.057328	0.207371	1.000000	0.436783	0.392573	0.154274	-0.113970	0.074752
Insulin	-0.073535	0.331357	0.088933	0.436783	1.000000	0.197859	0.185207	-0.042163	0.130548
BMI	0.017683	0.221071	0.281805	0.392573	0.197859	1.000000	0.103575	0.036242	0.292695
DiabetesPedigreeFunction	-0.026205	0.133163	0.051436	0.154274	0.185207	0.103575	1.000000	0.017970	0.176608
Age	0.544341	0.263514	0.239528	-0.113970	-0.042163	0.036242	0.017970	1.000000	0.238356
Diabetes	0.221898	0.466581	0.065068	0.074752	0.130548	0.292695	0.176608	0.238356	1.000000

```
df.corr().iloc[1,0]
```

```
0.12945867149927248
```

Indicadores y Estadísticos con Pandas

Los principales estadísticos con Pandas:

Matriz de covarianza

```
df.cov()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Diabetes
Pregnancies	11.354056	13.947131	9.214538	-4.390041	-28.555231	0.469774	-30.065025	21.570620	0.356618
Glucose	13.947131	1022.248314	94.430956	29.239183	1220.935799	55.726987	1449.640838	99.082805	7.115079
BloodPressure	9.214538	94.430956	374.647271	64.029396	198.378412	43.004695	338.980988	54.523453	0.600697
SkinThickness	-4.390041	29.239183	64.029396	254.473245	802.979941	49.373869	837.938947	-21.381023	0.568747
Insulin	-28.555231	1220.935799	198.378412	802.979941	13281.180078	179.775172	7267.317385	-57.143290	7.175671
BMI	0.469774	55.726987	43.004695	49.373869	179.775172	62.159984	278.041722	3.360330	1.100638
DiabetesPedigreeFunction	-30.065025	1449.640838	338.980988	837.938947	7267.317385	278.041722	115930.481362	71.954039	28.680281
Age	21.570620	99.082805	54.523453	-21.381023	-57.143290	3.360330	71.954039	138.303046	1.336953
Diabetes	0.356618	7.115079	0.600697	0.568747	7.175671	1.100638	28.680281	1.336953	0.227483

```
df.cov().iloc[1,0]
```

```
13.947130663298566
```

Uso de Librerías Gráficas: Matplotlib y Seaborn

Introducción

La visualización de datos es un aspecto fundamental en todo proyecto de Ciencia de Datos.

Los gráficos ayudan al fácil entendimiento y comprensión del contenido subyacente los datos.

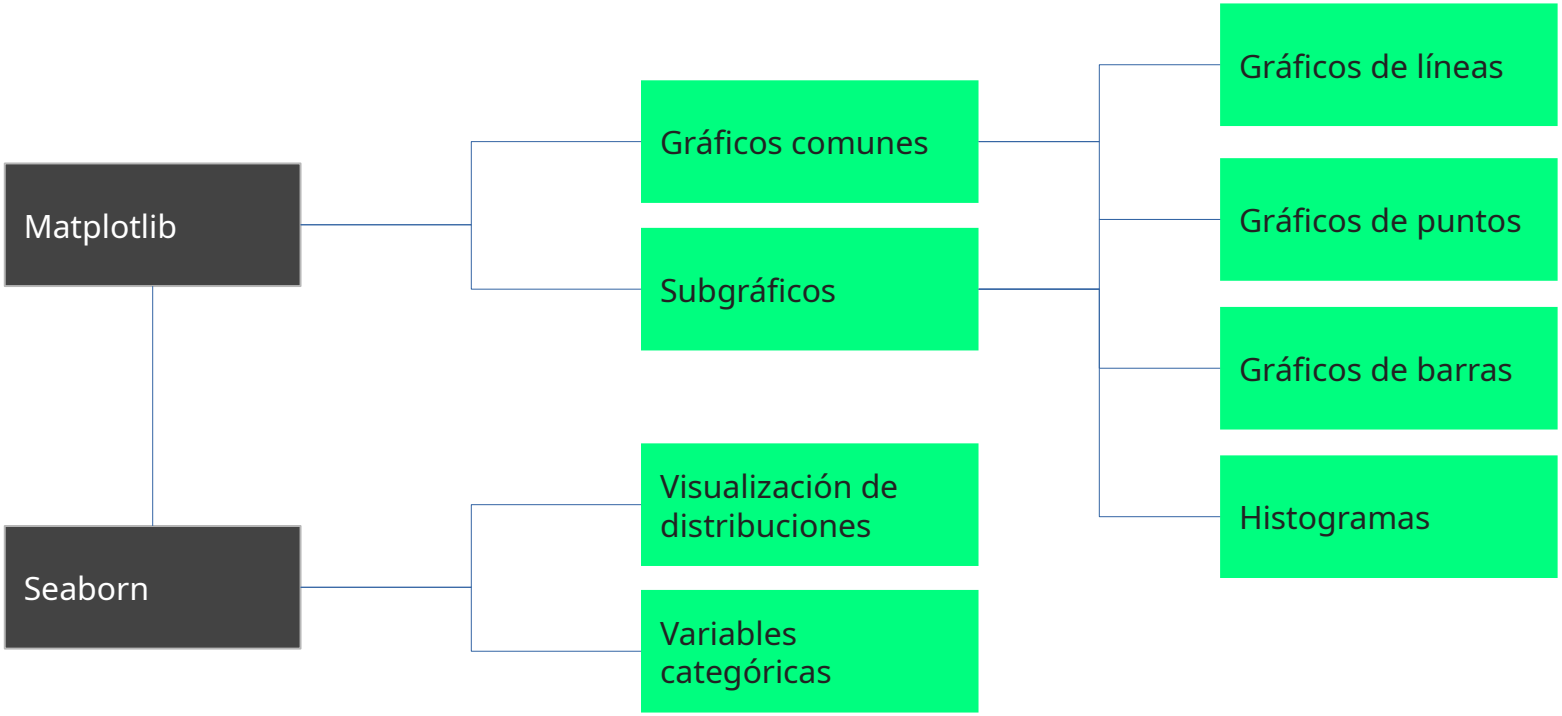
Son de gran utilidad tanto como herramienta para el análisis preliminar como para la comunicación efectiva de los resultados.

Las librerías más usadas en Python para realizar gráficas son:

- Matploptlib
- Seaborn

Uso de Librerías Gráficas: Matplotlib y Seaborn

Mapa de Conceptos



Cuadro extraído del Curso de Data Science de Coderhouse. Link: https://www.coderhouse.com/ar/

Uso de Librerías Gráficas: Matplotlib y Seaborn

Matplotlib:

Matplotlib es una biblioteca para la generación de gráficos en dos dimensiones, a partir de datos contenidos en listas o arrays en el lenguaje de programación Python. Proporciona una API, pylab, diseñada para recordar a la de MATLAB. (Wikipedia)

Autor: John D. Hunter

Lanzamiento inicial: 2003

Licencia: Su propia licencia libre

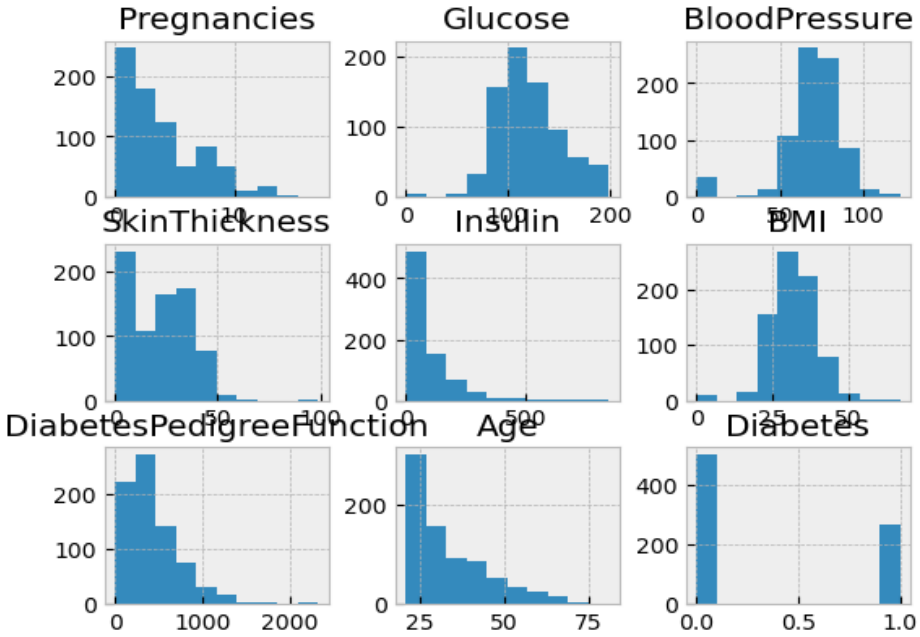
Uso de Librerías Gráficas: Matplotlib y Seaborn

Los principales gráficos con Matplotlib:

Histogramas

Histogramas de todas las variables del df:

df.hist()



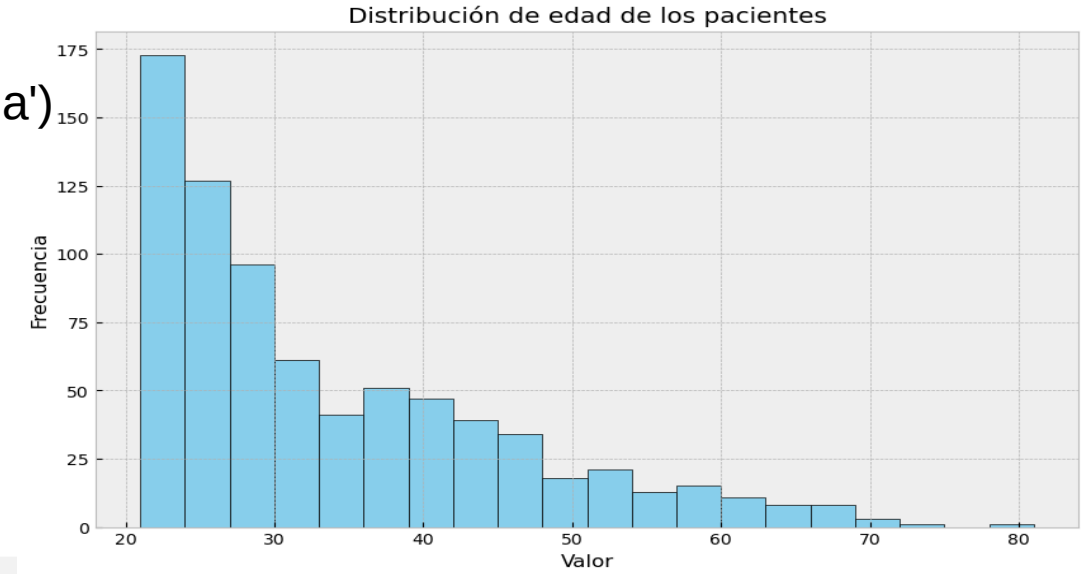
Uso de Librerías Gráficas: Matplotlib y Seaborn

Los principales gráficos con Matplotlib:

Histogramas

Histogramas de alguna de las variables del df: Edad

```
plt.hist(df['Age'], bins=20, color='skyblue', edgecolor='black')
plt.title('Distribución de edad de los pacientes')
plt.xlabel('Valor')
plt.ylabel('Frecuencia')
plt.show()
```



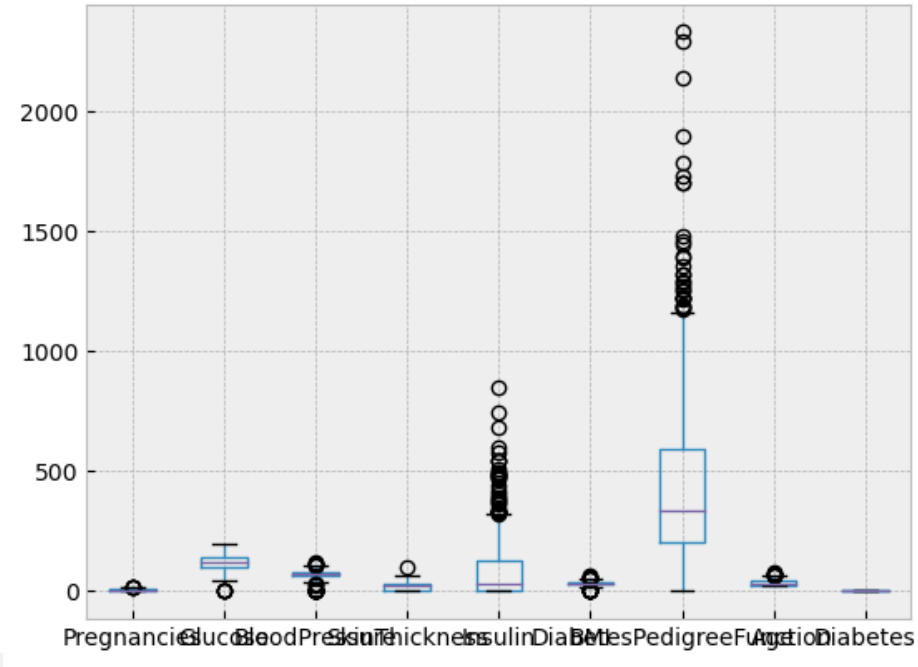
Uso de Librerías Gráficas: Matplotlib y Seaborn

Los principales gráficos con Matplotlib:

Diagrama de cajas

Diagramas de caja de las variables numéricas del Dataframe

df.boxplot()



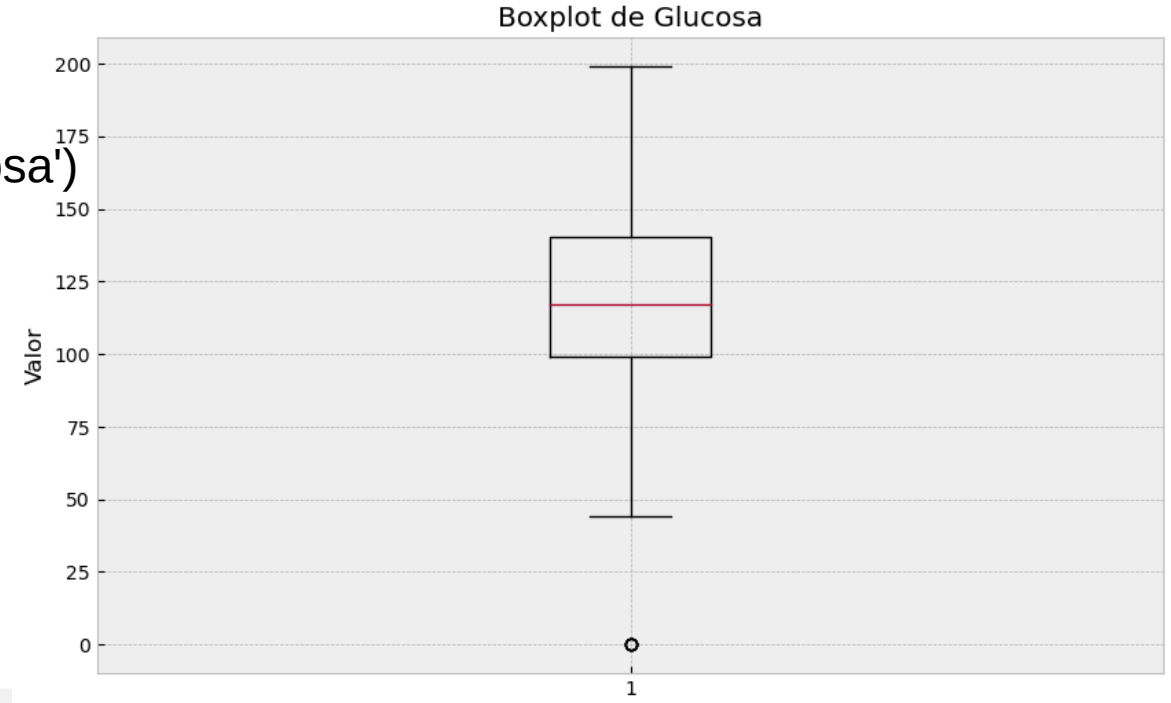
Uso de Librerías Gráficas: Matplotlib y Seaborn

Los principales gráficos con Matplotlib:

Diagrama de cajas

Diagramas de caja de una variable: Glucosa

```
plt.figure(figsize=(10, 6))
plt.boxplot(df['Glucose'])
plt.title('Boxplot de Glucosa')
plt.ylabel('Valor')
plt.show()
```

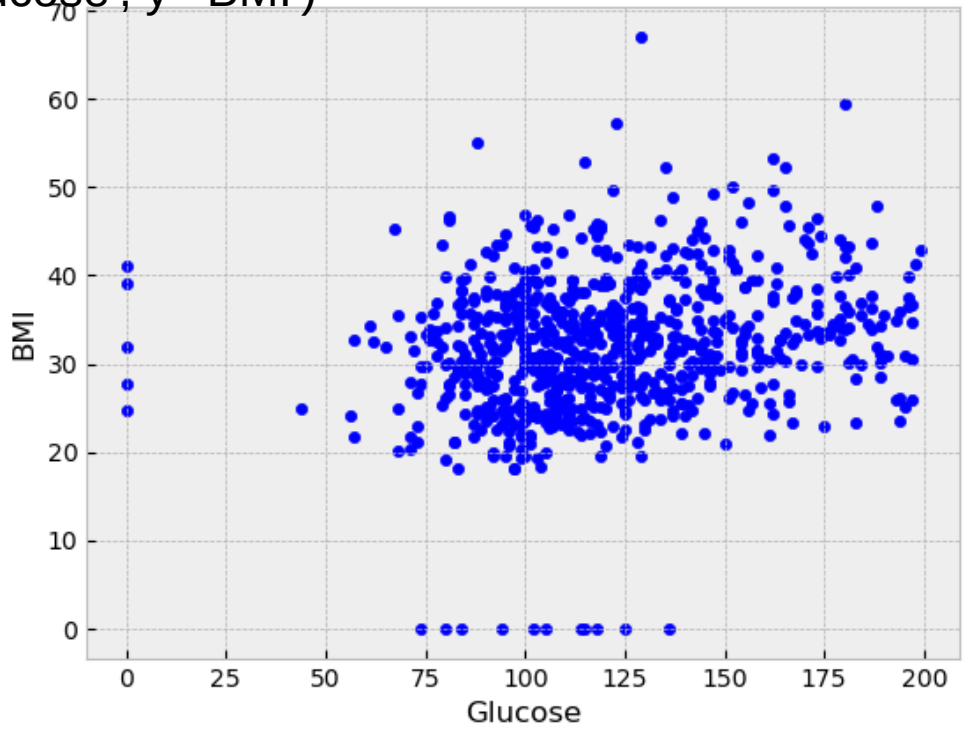


Uso de Librerías Gráficas: Matplotlib y Seaborn

Los principales gráficos con Matplotlib:

Diagrama de dispersión

df.plot.scatter(x='Glucose', y='BMI')



Uso de Librerías Gráficas: Matplotlib y Seaborn

Los principales gráficos con Matplotlib:

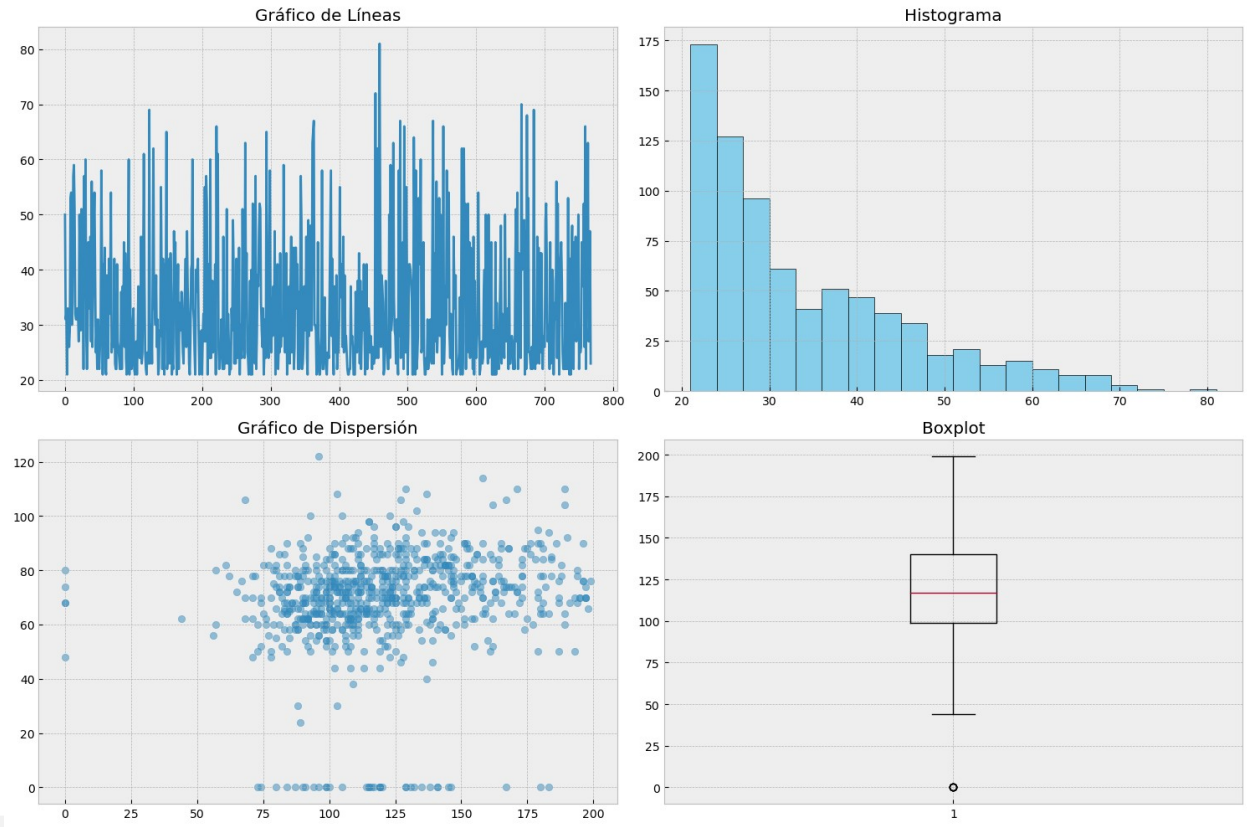
Subgráficos

```
fig, axs = plt.subplots(2, 2, figsize=(15, 10))
axs[0, 0].plot(df['Age'])
axs[0, 0].set_title('Gráfico de Líneas')
axs[0, 1].hist(df['Age'], bins=20, color='skyblue', edgecolor='black')
axs[0, 1].set_title('Histograma')
axs[1, 0].scatter(df['Glucose'], df['BloodPressure'], alpha=0.5)
axs[1, 0].set_title('Gráfico de Dispersión')
axs[1, 1].boxplot(df['Glucose'])
axs[1, 1].set_title('Boxplot')
plt.tight_layout()
plt.show()
```

Uso de Librerías Gráficas: Matplotlib y Seaborn

Los principales gráficos con Matplotlib:

Subgraficos



Uso de Librerías Gráficas: Matplotlib y Seaborn

Seaborn:

Seaborn es una biblioteca de visualización de datos de Python basada en matplotlib. Proporciona una interfaz de alto nivel para dibujar gráficos estadísticos atractivos e informativos[1].

Autor: Michael Waskom.

Creación: 2012

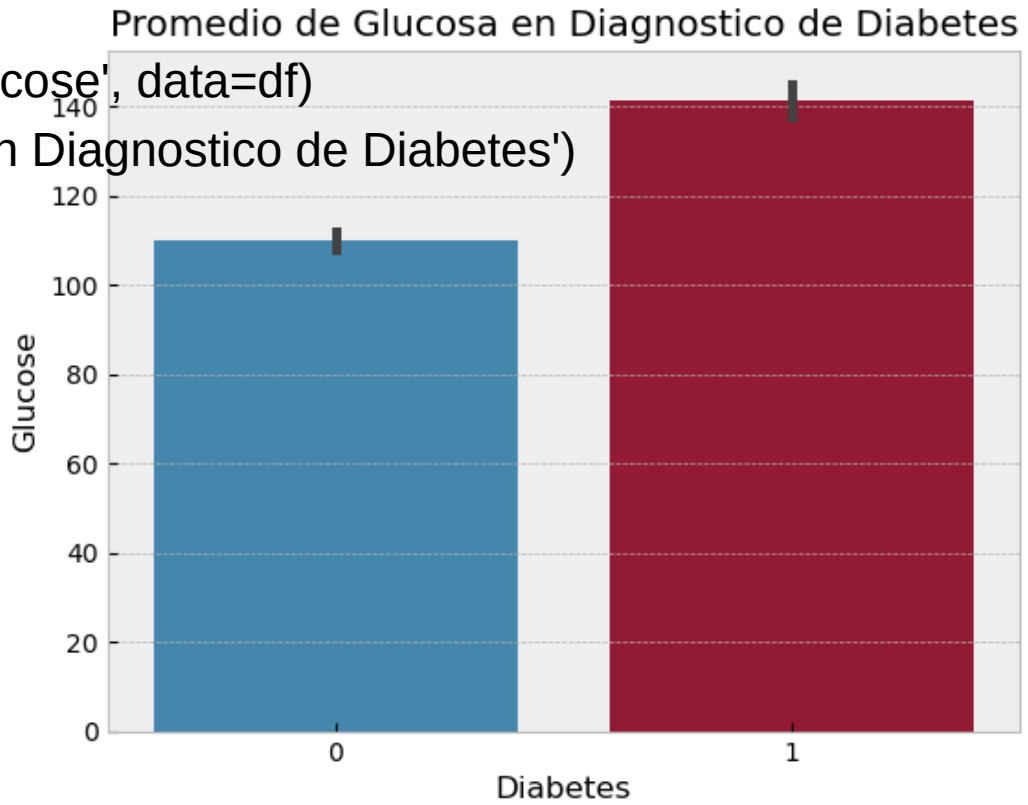
Sitio oficial: https://seaborn.pydata.org/

Uso de Librerías Gráficas: Matplotlib y Seaborn

Los principales gráficos con Seaborn:

Diagramas de Barras

```
sns.barplot(x='Diabetes', y='Glucose', data=df)
plt.title('Promedio de Glucosa en Diagnostico de Diabetes')
plt.show()
```

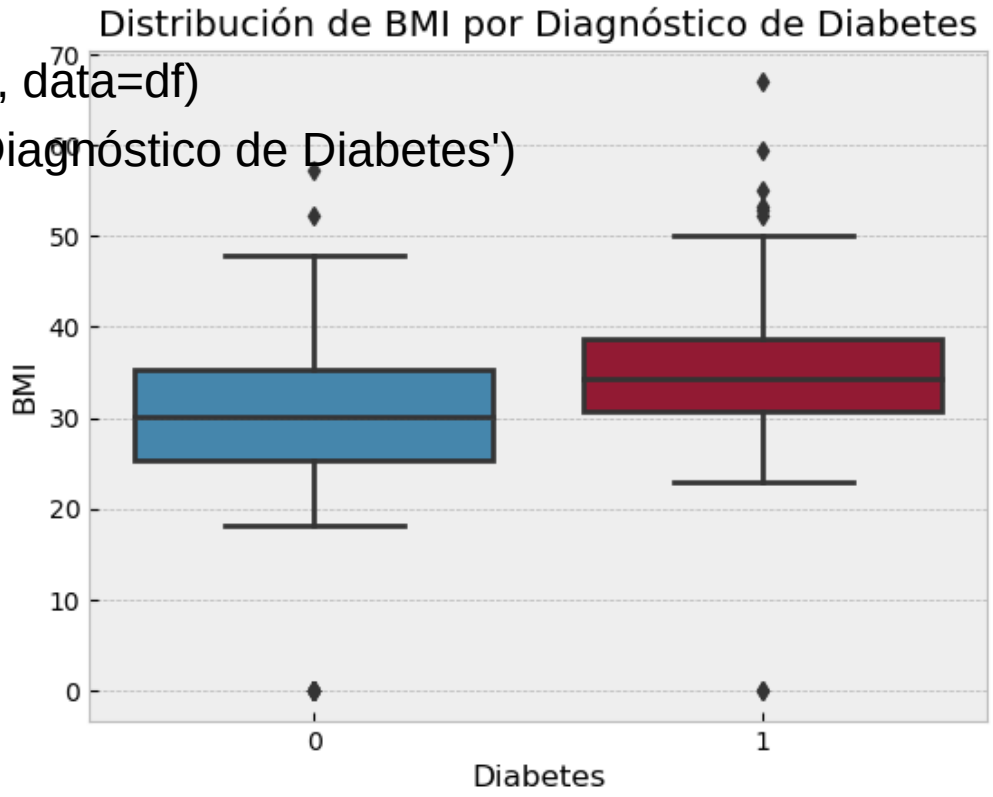


Uso de Librerías Gráficas: Matplotlib y Seaborn

Los principales gráficos con Seaborn:

Diagramas de Cajas

```
sns.boxplot(x='Diabetes', y='BMI', data=df)
plt.title('Distribución de BMI por Diagnóstico de Diabetes')
plt.show()
```

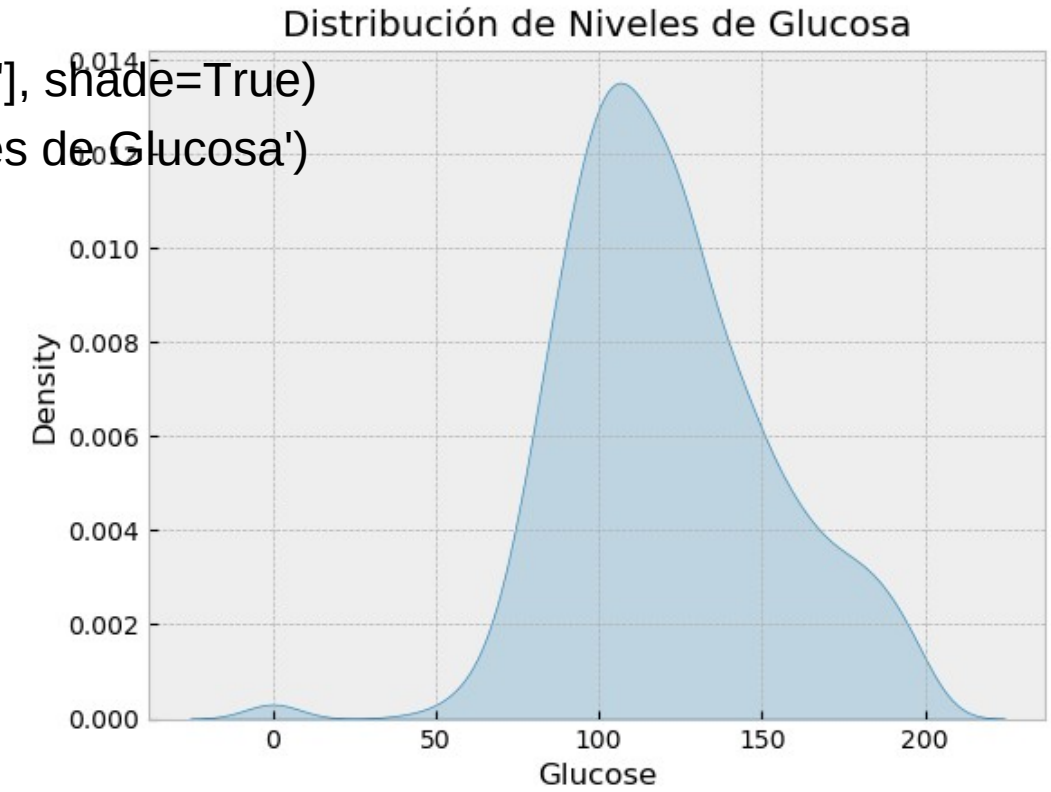


Uso de Librerías Gráficas: Matplotlib y Seaborn

Los principales gráficos con Seaborn:

Diagramas de Densidad

```
sns.kdeplot(data=df['Glucose'], shade=True)
plt.title('Distribución de Niveles de Glucosa')
plt.show()
```

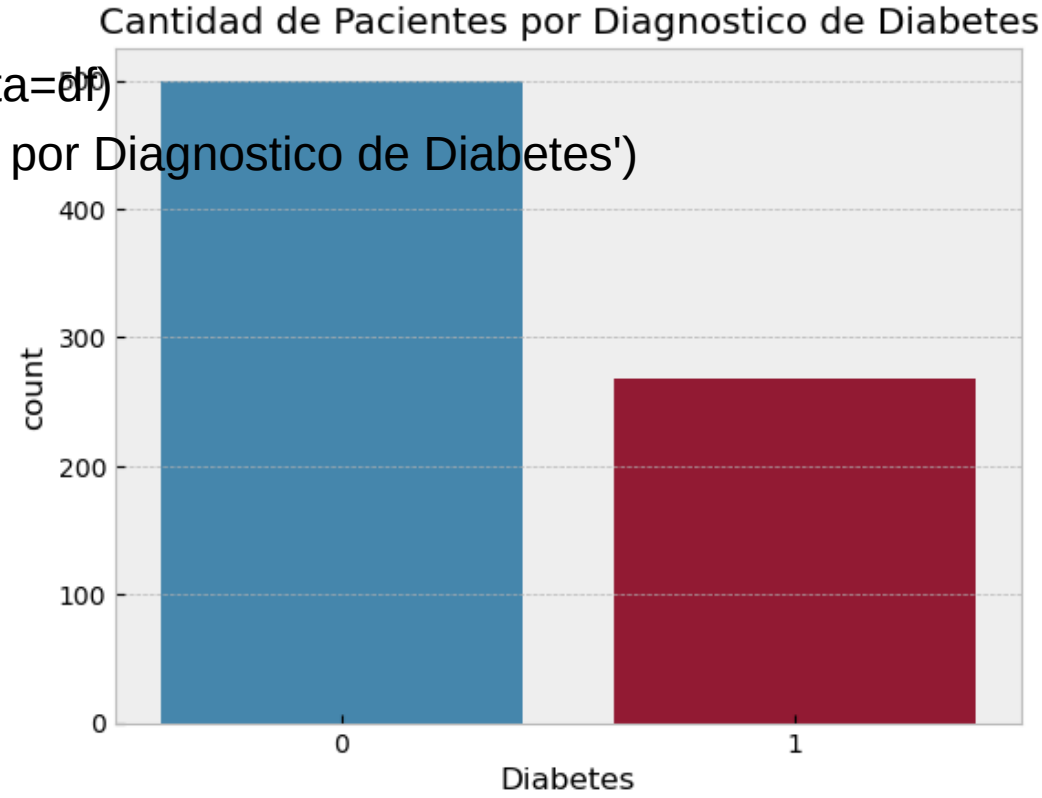


Uso de Librerías Gráficas: Matplotlib y Seaborn

Los principales gráficos con Seaborn:

Diagramas de Recuento

```
sns.countplot(x='Diabetes', data=df)
plt.title('Cantidad de Pacientes por Diagnostico de Diabetes')
plt.show()
```

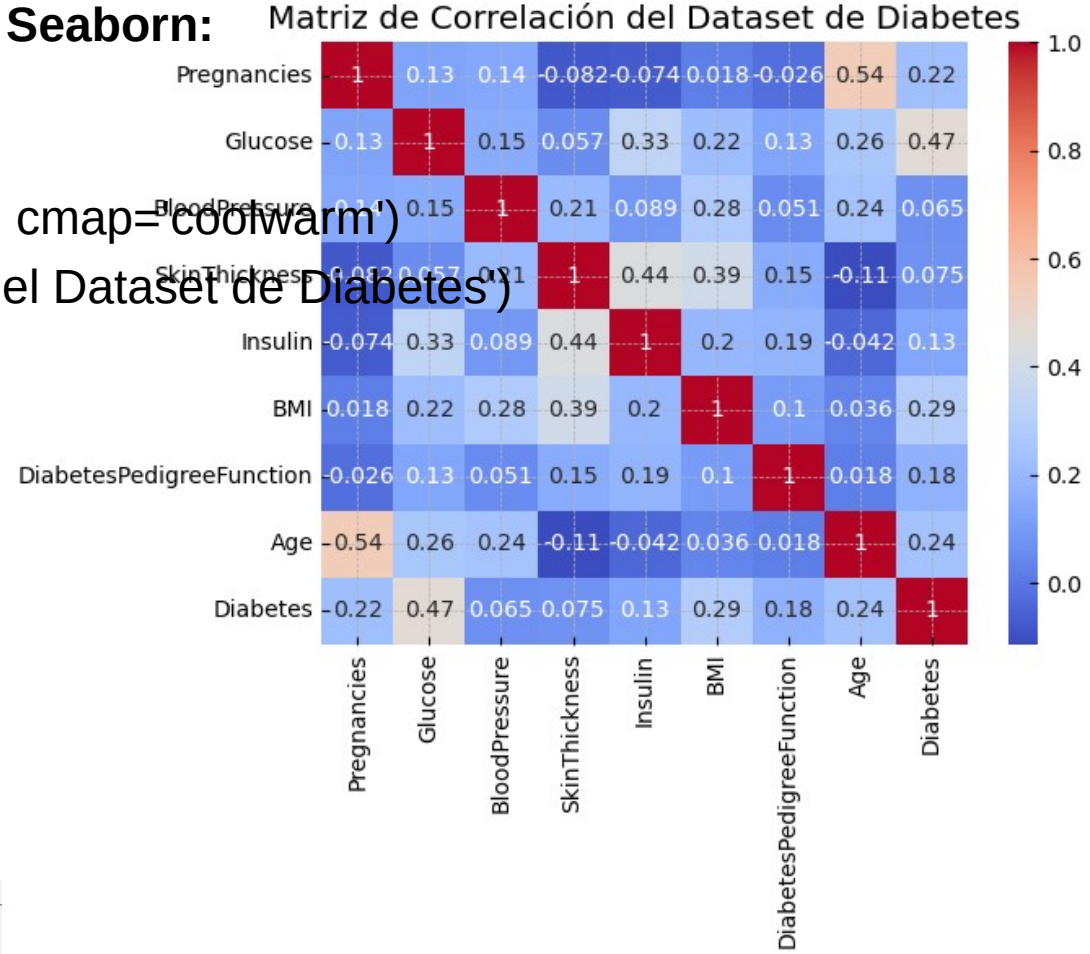


Uso de Librerías Gráficas: Matplotlib y Seaborn

Los principales gráficos con Seaborn:

Mapas de Calor

```
corr = df.corr()
sns.heatmap(corr, annot=True, cmap='coolwarm')
plt.title('Matriz de Correlación del Dataset de Diabetes')
plt.show()
```



Uso de Librerías Gráficas: Matplotlib y Seaborn

Los principales gráficos con Seaborn:

Gráficos de Pares (Pair plots)

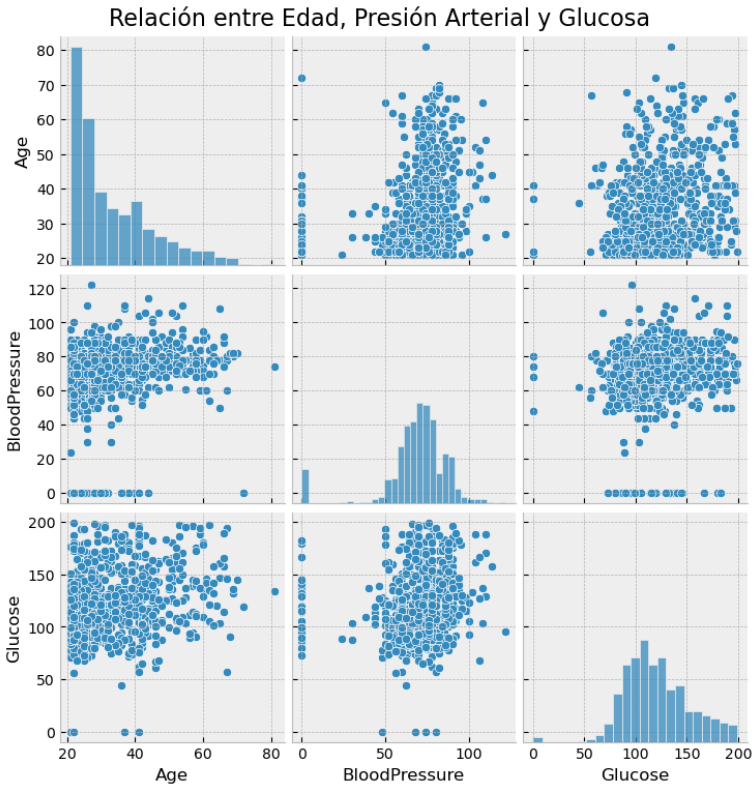
Muestra gráficos de dispersión y distribuciones para todas las combinaciones posibles de un conjunto de variables numéricas, como la relación entre la edad, la presión arterial y los niveles de glucosa.

```
# Genera el Pair Plot
sns.pairplot(df[['Age', 'BloodPressure', 'Glucose']])
# Agrega un título arriba del gráfico
plt.suptitle('Relación entre Edad, Presión Arterial y Glucosa', y=1.02, fontsize=16)
# Muestra el gráfico
plt.show()
```

Uso de Librerías Gráficas: Matplotlib y Seaborn

Los principales gráficos con Seaborn:

Gráficos de Pares (Pair plots)



Uso de Librerías Gráficas: Matplotlib y Seaborn

Los principales gráficos con Seaborn:

Gráficas de Violín

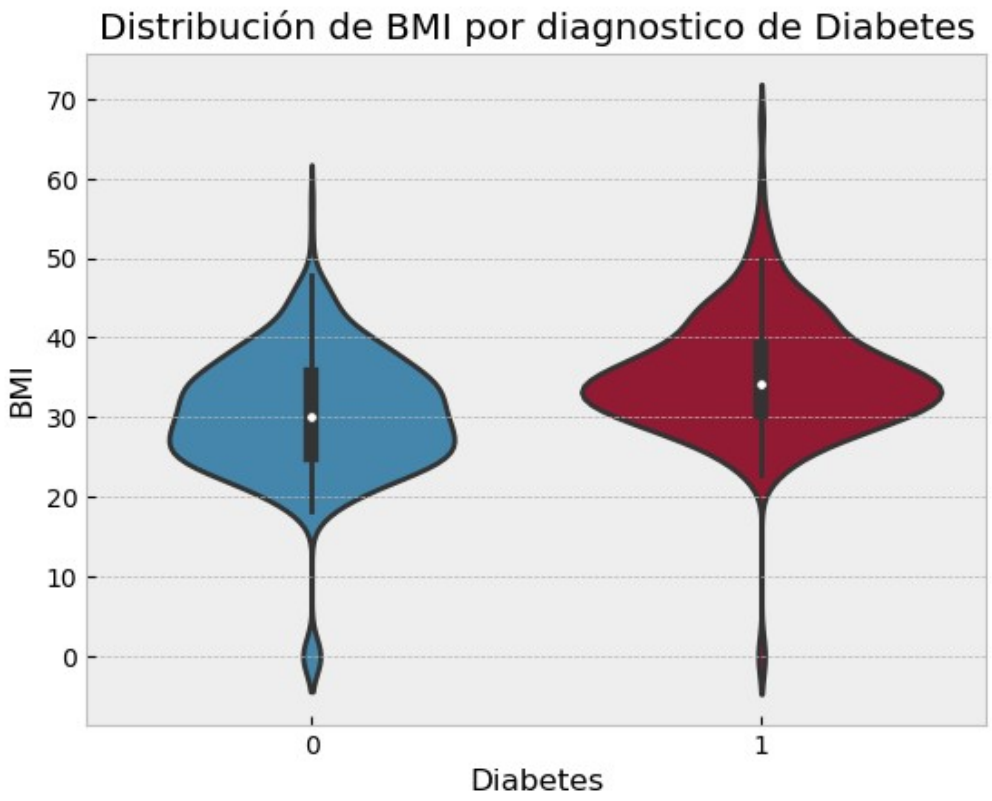
Combina el gráfico de cajas y el gráfico de densidad para mostrar la distribución de una variable numérica y su densidad, por ejemplo, la distribución del índice de masa corporal (BMI) por resultado.

```
sns.violinplot(x='Outcome', y='BMI', data=Diabetes)
plt.title('Distribución de BMI por Outcome')
plt.show()
```

Uso de Librerías Gráficas: Matplotlib y Seaborn

Los principales gráficos con Seaborn:

Gráficas de Violín



Uso de Librerías Gráficas: Matplotlib y Seaborn

Los principales gráficos con Seaborn:

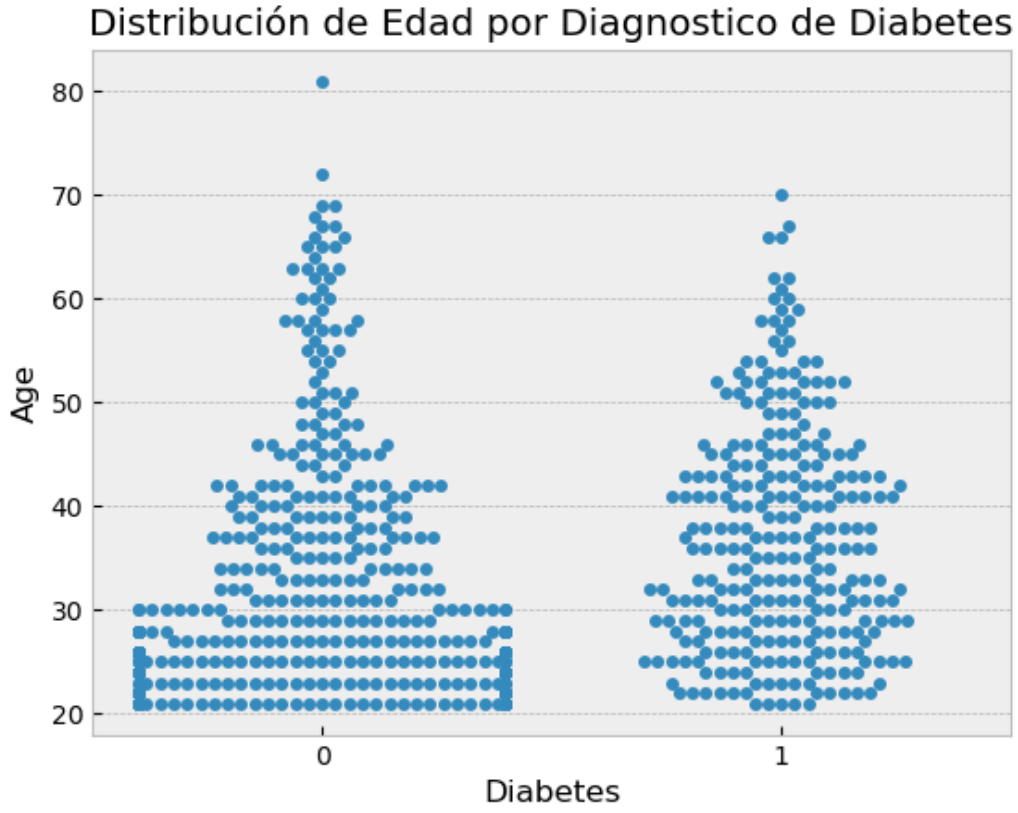
Gráficas de Enjambre

Muestra la distribución de puntos de datos a lo largo de una variable categórica, similar a un gráfico de dispersión pero con una separación de los puntos para evitar la superposición, por ejemplo, la distribución de la edad por resultado.

Uso de Librerías Gráficas: Matplotlib y Seaborn

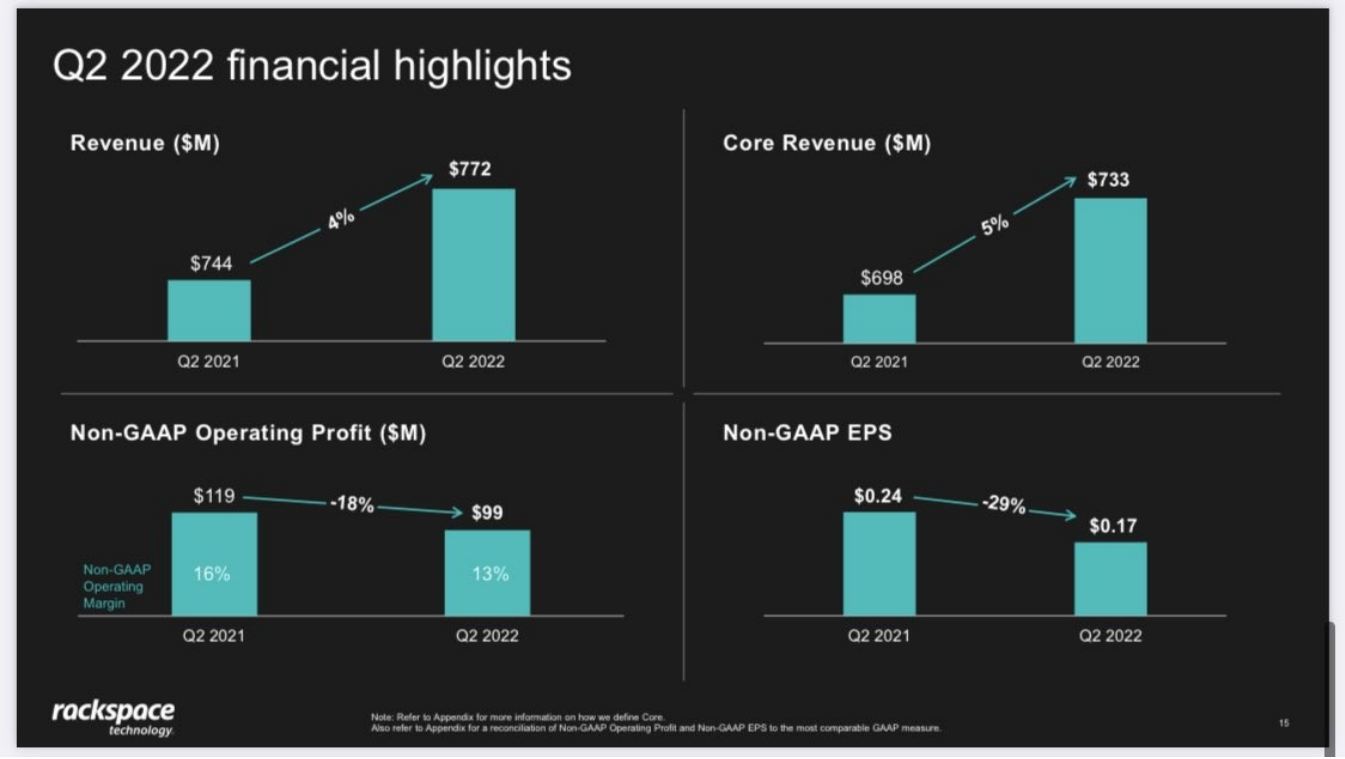
Los principales gráficos con Seaborn:

Gráficas de Enjambre



Gráficos con errores

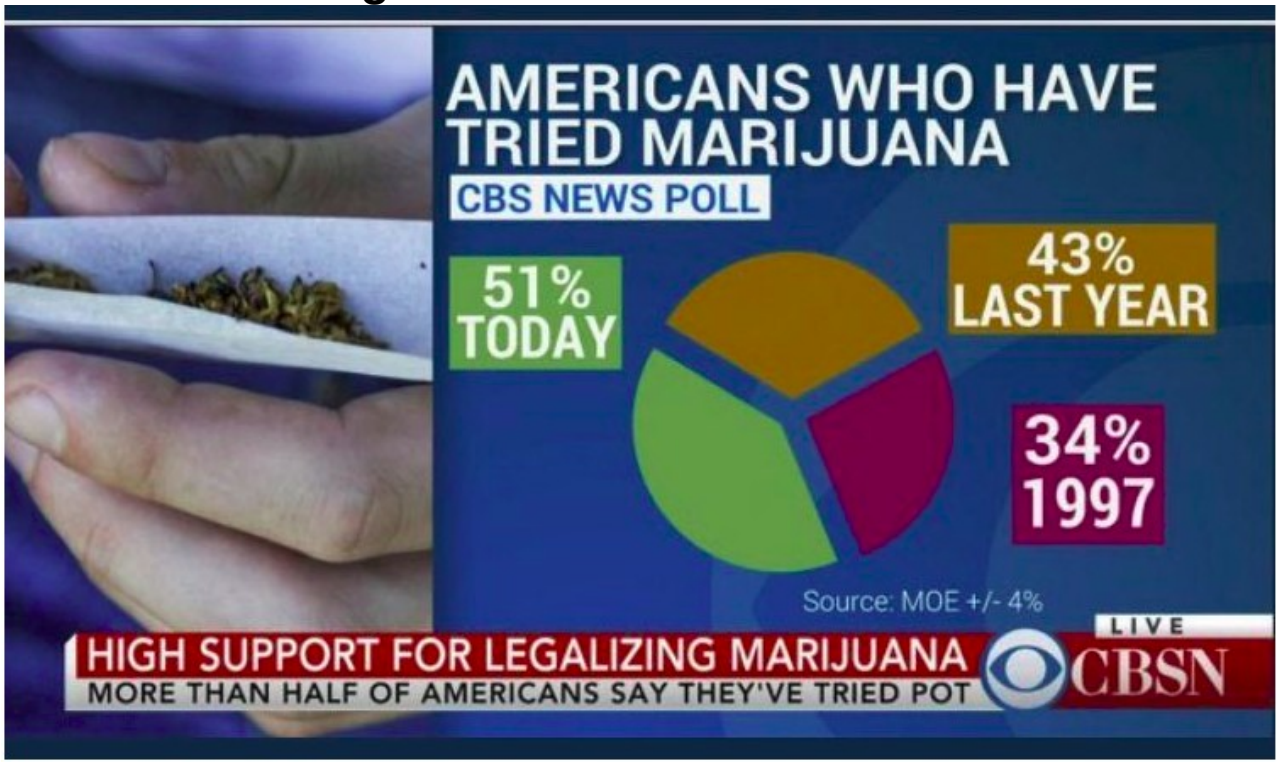
¿Qué errores ve en esta gráfica?



Fuente: https://x.com/mpfix1/status/1828109846912372940?t=rT5mJHrWh0cfG5YES3M0ug&s=08

Gráficos con errores

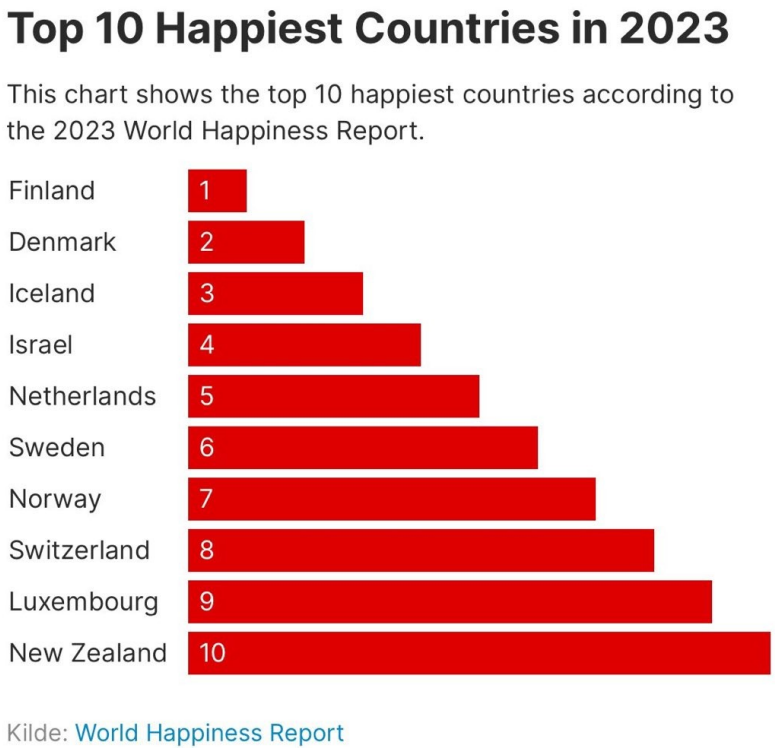
¿Qué errores ve en esta gráfica?



Fuente: https://x.com/mpfix1/status/1828109846912372940?t=rT5mJHrWh0cfG5YES3M0ug&s=08

Gráficos con errores

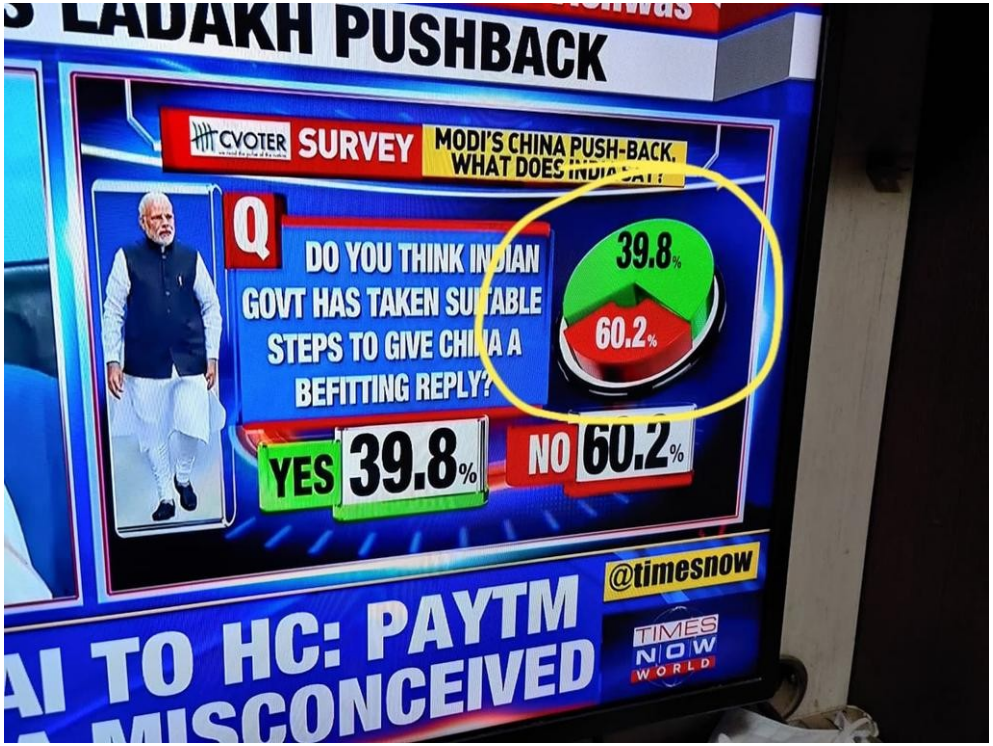
¿Qué errores ve en esta gráfica?



Fuente: https://x.com/mpfix1/status/1828109846912372940?t=rT5mJHrWh0cfG5YES3M0ug&s=08

Gráficos con errores

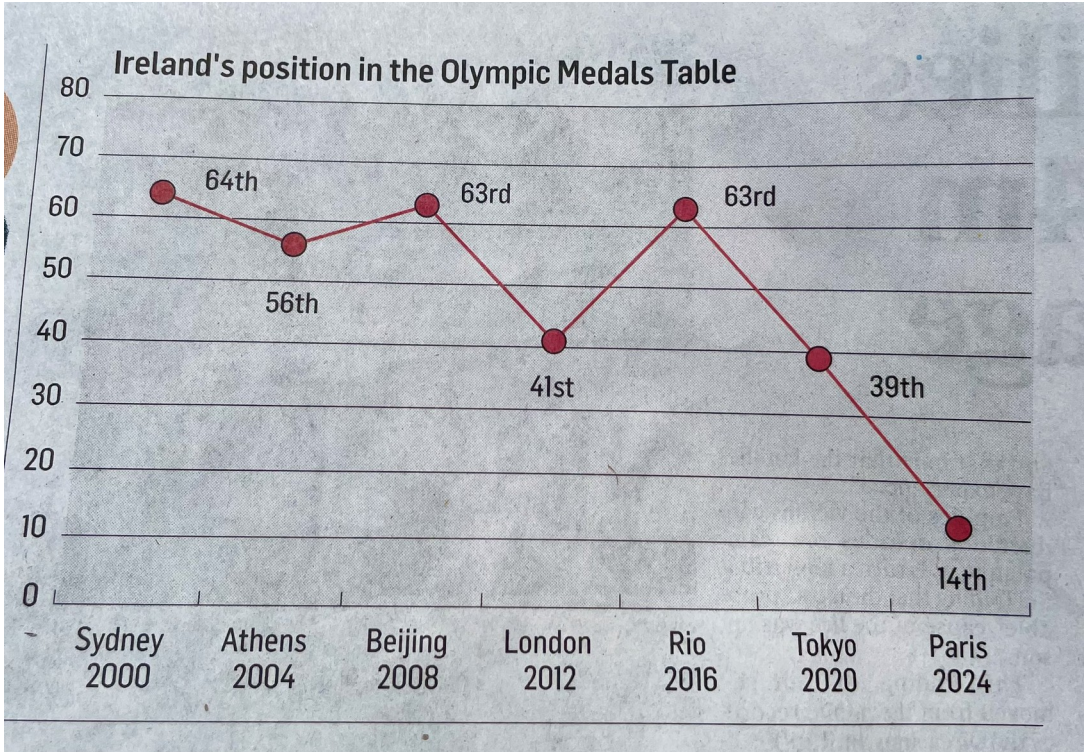
¿Qué errores ve en esta gráfica?



Fuente: https://x.com/mpfix1/status/1828109846912372940?t=rT5mJHrWh0cfG5YES3M0ug&s=08

Gráficos con errores

¿Qué errores ve en esta gráfica?



Fuente: https://x.com/mpfix1/status/1828109846912372940?t=rT5mJHrWh0cfG5YES3M0ug&s=08

Gráficos con errores

¿Qué errores ve en esta gráfica?

Fuente: https://x.com/mpfix1/status/1828109846912372940?t=rT5mJHrWh0cfG5YES3M0ug&s=08

Conclusión

Conclusión

Iniciamos esta presentación haciendo un breve repaso acerca de indicadores y estadísticos.

Luego presentamos cómo podemos obtener los indicadores mencionados utilizando Pyhon y Pandas

Presentamos ejemplos de los gráficos más comunes utilizando librerías muy usadas en Python: Matplotlib y Seaborn.

Y por último, un breve compendio de gráficas erróneas.