

Introducción

En este trabajo nos proponemos analizar, mediante el lenguaje de programación Python, las principales características de una muestra de observaciones desde un punto de vista estadístico. Esta muestra contiene las siguientes variables que la caracterizan: **ID**, **edad**, **años_educ**, **en_pareja**, **num_hijos** y **bajo_socioecon**.

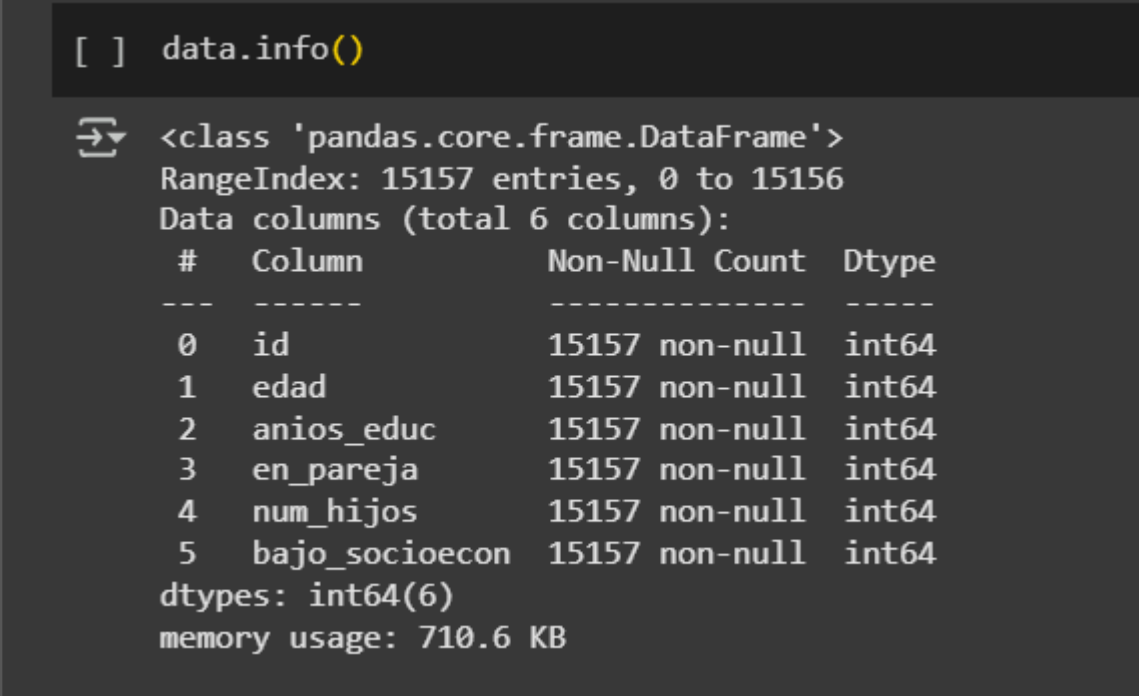
ID es un identificador para cada individuo de la muestra, **edad** representa la edad de la persona, **años_educ** son los años de educación y **num_hijos** la cantidad de hijos. En tanto las variables **en_pareja** y **bajo_socioecon** son dos variables categóricas que indican si la persona tiene pareja o no, y si su nivel socioeconómico es bajo o no.

Desarrollo del trabajo

Funciones principales utilizadas

1) La primera función que utilizamos para analizar nuestro dataset es `.info()`. Ésta nos da un pantallazo general:

```
[ ] data.info()
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15157 entries, 0 to 15156
Data columns (total 6 columns):
#   Column          Non-Null Count  Dtype
---  -
0   id              15157 non-null  int64
1   edad            15157 non-null  int64
2   años_educ       15157 non-null  int64
3   en_pareja       15157 non-null  int64
4   num_hijos       15157 non-null  int64
5   bajo_socioecon  15157 non-null  int64
dtypes: int64(6)
memory usage: 710.6 KB
```

Observamos que nuestro dataset no tiene problemas de valores faltantes, un problema usual que nos podemos encontrar en las bases de datos. En este caso observamos que todas las variables tienen el total de observaciones de la muestra (15.157).

También observamos que todas las variables son *int64*, es decir son números enteros (sin decimales).

2) Pasamos ahora a la función `.describe()`. Esta función nos muestra las principales **estadísticas descriptivas** de nuestro dataset:

```
[ ] data.describe()
```

	id	edad	anios_educ	en_pareja	num_hijos	bajo_socioecon
count	15157.00000	15157.000000	15157.000000	15157.000000	15157.000000	15157.000000
mean	7579.00000	16.962064	8.506763	0.161707	0.150887	0.250049
std	4375.59335	1.413214	1.176005	0.368194	0.409855	0.433056
min	1.00000	15.000000	6.000000	0.000000	0.000000	0.000000
25%	3790.00000	16.000000	8.000000	0.000000	0.000000	0.000000
50%	7579.00000	17.000000	9.000000	0.000000	0.000000	0.000000
75%	11368.00000	18.000000	9.000000	0.000000	0.000000	1.000000
max	15157.00000	19.000000	12.000000	1.000000	3.000000	1.000000

Observamos que todas las variables tienen un comportamiento lógico, en cuanto a que no observamos estadísticas disparatadas o “extrañas”. Las variables que más prestamos atención aquí son el mínimo, el máximo, la media y la mediana.

- 3) Otra función útil es la función `.head()`. Esta nos muestra las primeras observaciones de nuestra muestra:

```
data.head()
```

	id	edad	anios_educ	en_pareja	num_hijos	bajo_socioecon
0	1	18	9	0	0	1
1	2	16	7	0	0	1
2	3	15	9	0	0	1
3	4	17	9	1	0	0
4	5	18	9	1	0	0

Como no pusimos ningún número entre los () por defecto nos arroja las observaciones para los primeros 5 individuos.

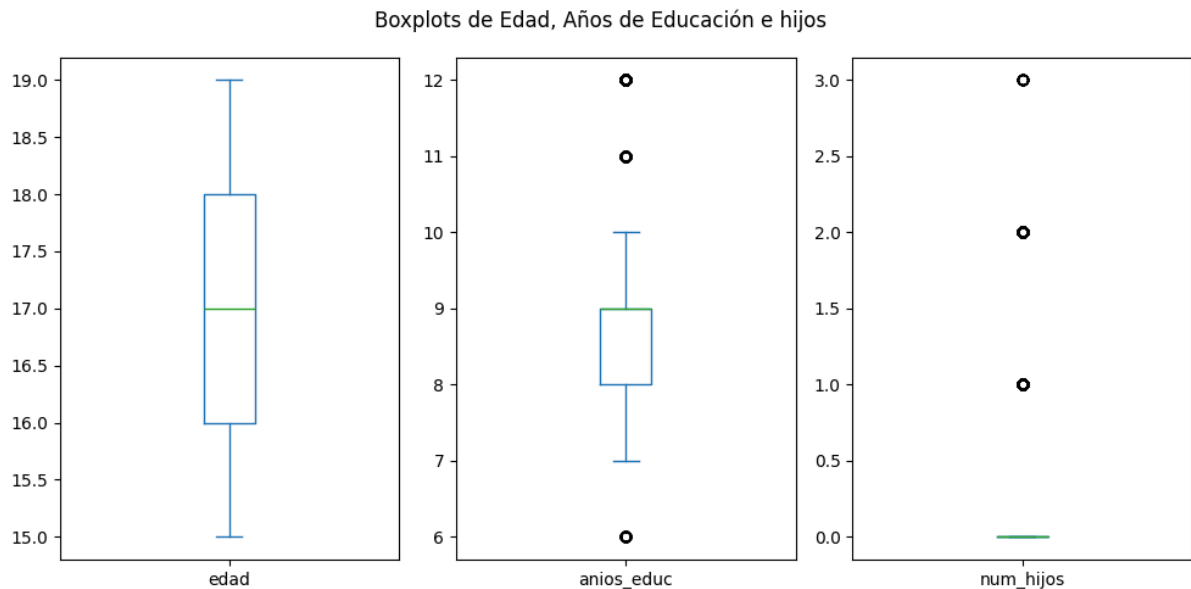
- 4) Para las 2 variables categóricas hacemos un `.value_counts()`. Esta función nos devuelve cuántas observaciones tenemos con el valor 0 (que es como un “No”) y cuántas con 1 (que es como un “Sí”).

Tenemos en nuestro dataset 12.706 personas que no están en pareja y las restantes 2.451 si lo están. En cuanto al nivel socioeconómico, tenemos 11.367 personas que no están clasificadas en nivel bajo, y 3.790 que sí.

Gráficos

1) Empezamos realizando un **boxplot** o diagrama de caja.

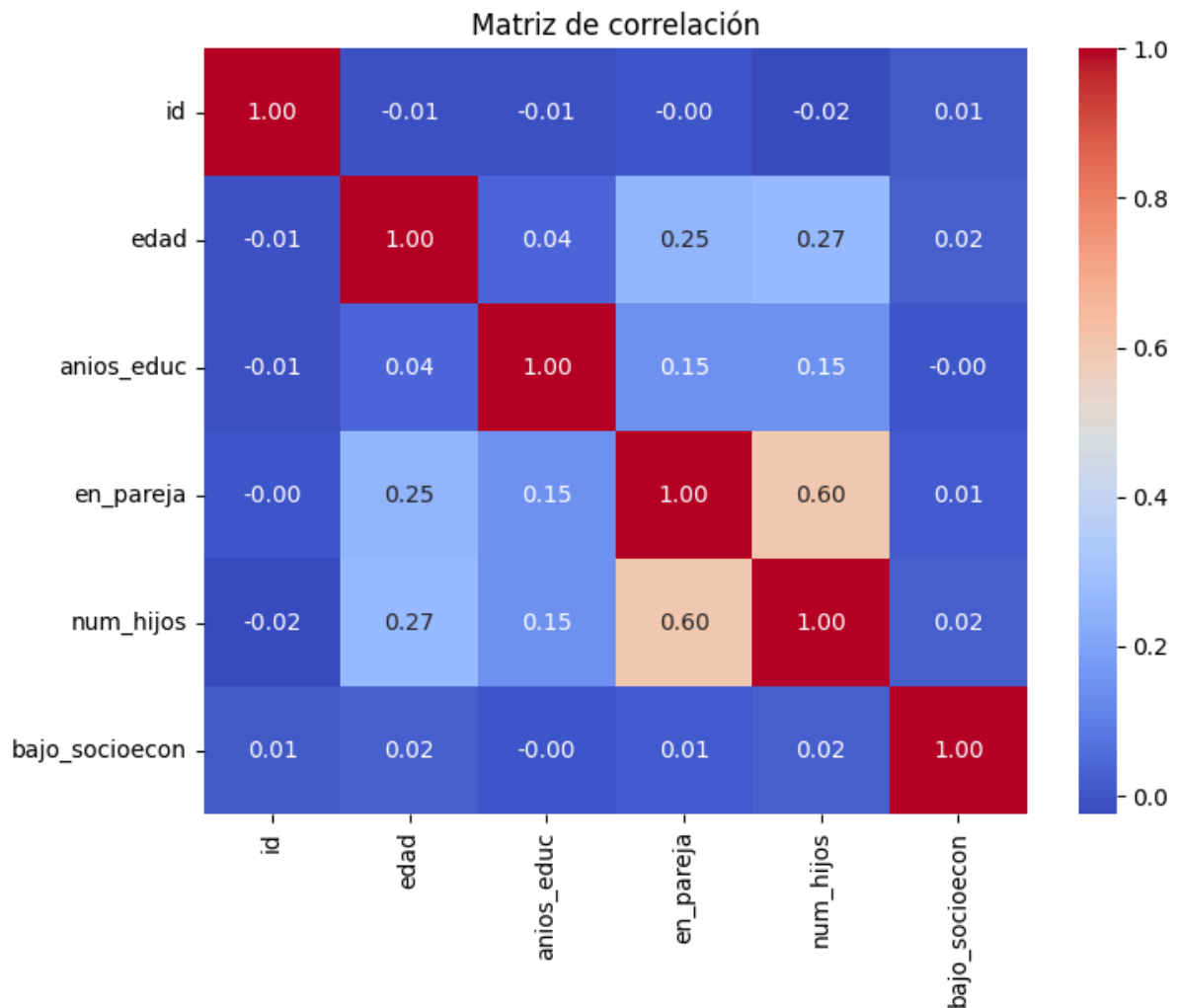
Observamos que la variable `anios_educ` tiene 3 valores atípicos, es decir, valores extremos que se encuentran lejos de la mayoría de las observaciones representada por la caja y los bigotes. Estos son 6, 11 y 12.



Para la variable `edad`, no tenemos valores atípicos. El máximo, se encuentra en el valor 19 y el mínimo en 15, por lo que podemos afirmar que las personas de nuestra muestra tienen entre 15 y 19 años. En tanto la mediana, representada por la línea que divide la caja a la mitad, es igual a 17. El primer cuartil, que representa el valor cuya cantidad de observaciones acumula el 25% de la muestra, es igual a 16 y el tercer cuartil, que acumula el 75%, es igual a 18.

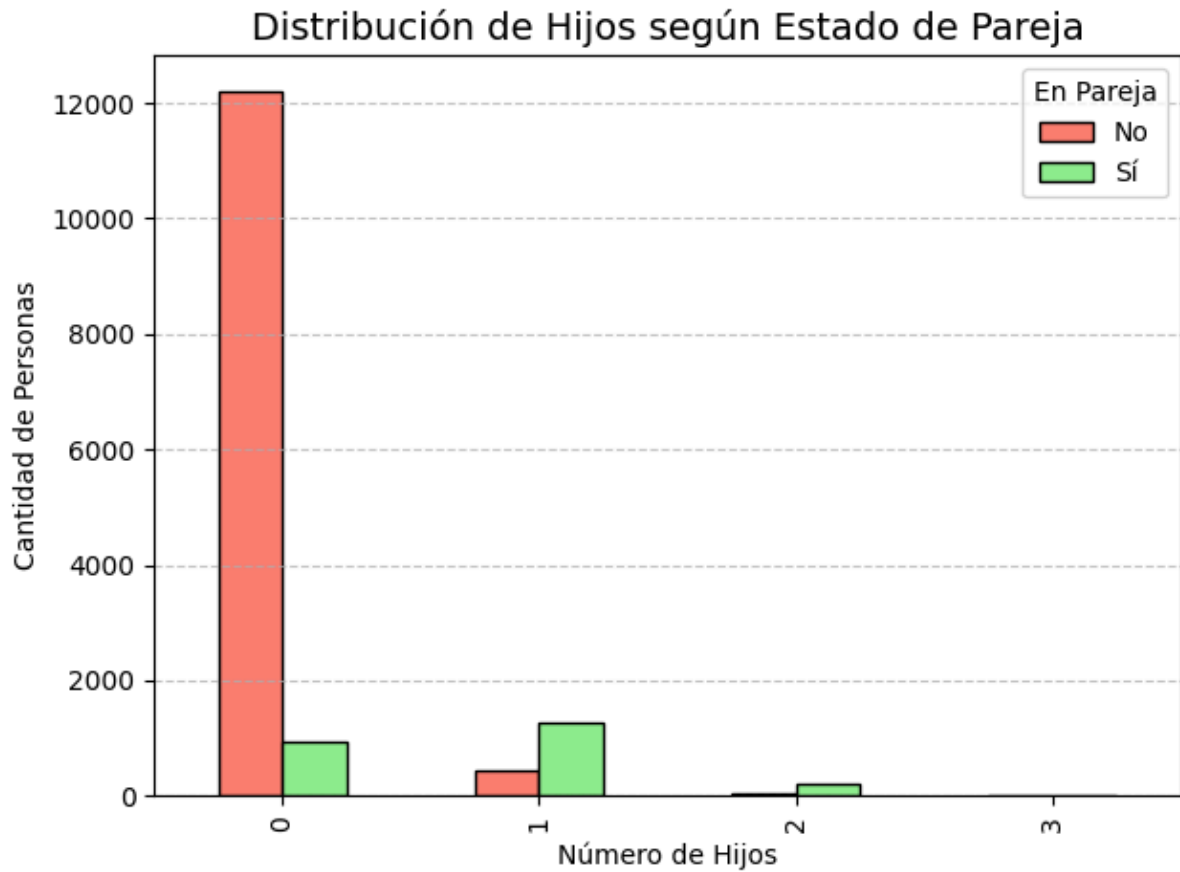
Por último, la variable `num_hijos`, tiene como mínimo 0 y máximo 3. La mediana, primer y tercer cuartil valen 0.

2) Hicimos también una **matriz de correlación**, para ver las relaciones entre las variables de nuestro dataset.



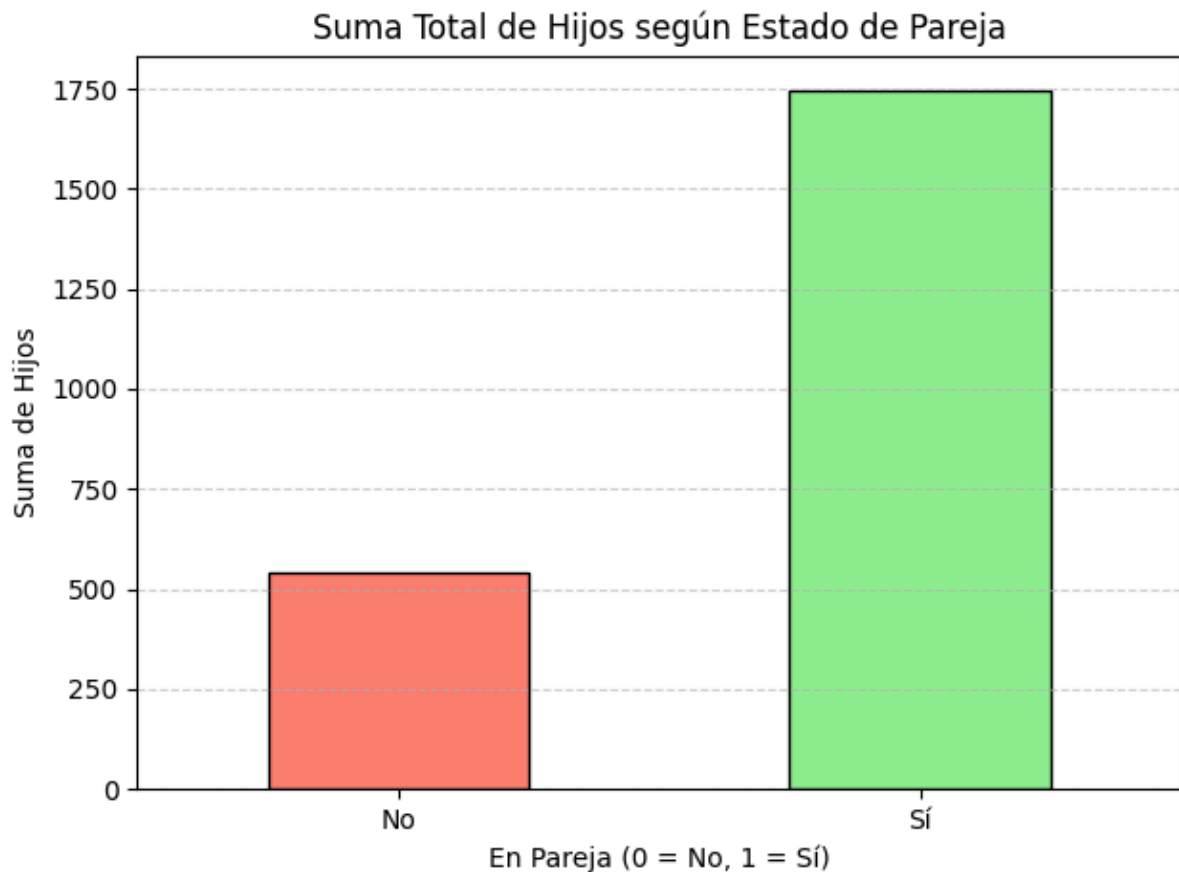
Como se puede ver en la matriz, las variables que se relacionan más fuertemente son num_hijos y en_pareja. Esta relación es **positiva** y vale 0.6.

Nos apoyamos aquí con un gráfico de barras para bajar a números esta relación:



Como se puede ver, las personas que no están en pareja representan **casi en su totalidad** las que no tienen hijos. Esta tendencia se revierte a partir de 1 hijo, donde vemos que aquí son más las personas que están en pareja.

Vemos ahora esta relación en números, con una tabla donde definimos la variable `sum_hijos` y con otro gráfico de barras:



```
#Variable suma de hijos  
sum_hijos = data.groupby('en_pareja')['num_hijos'].sum()  
sum_hijos
```

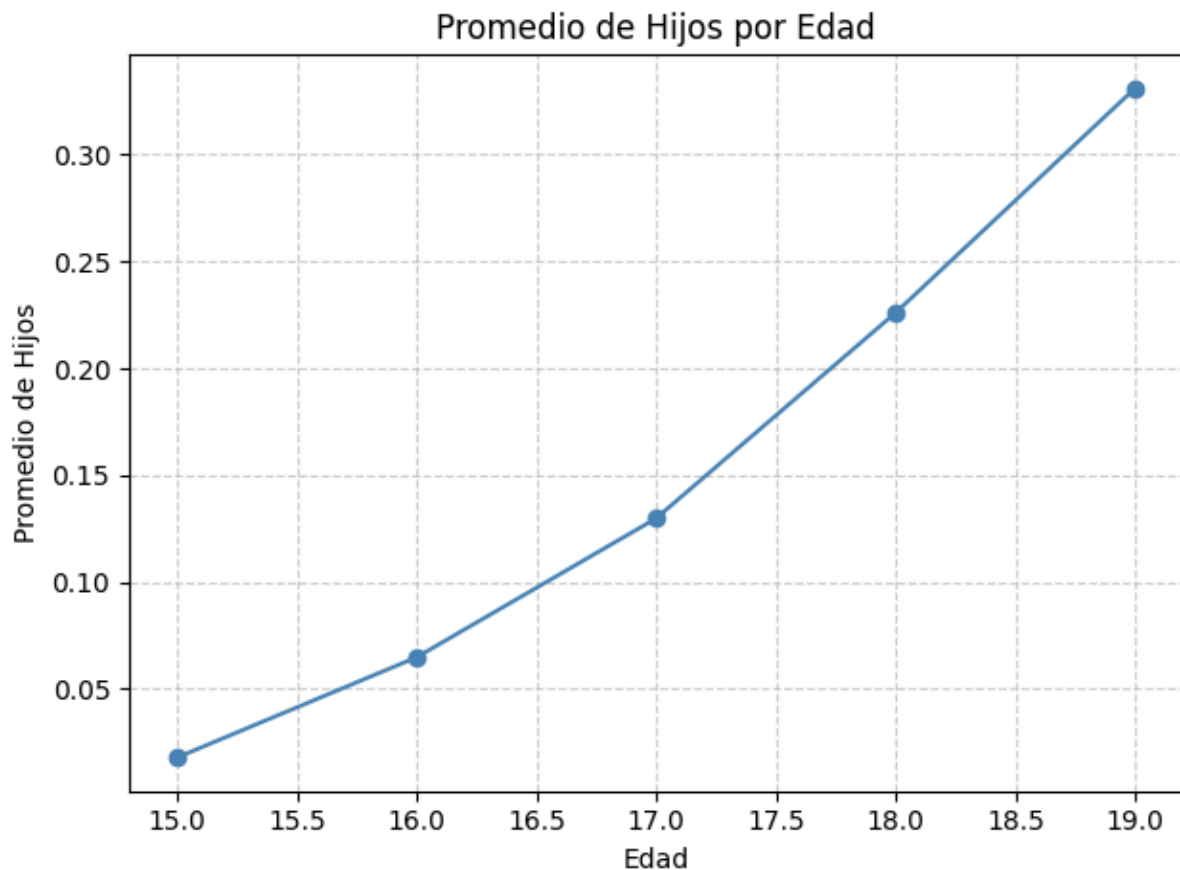
```
num_hijos  
  
en_pareja  
0         542  
1        1745
```

```
dtype: int64
```

Son 542 los hijos de las personas que no están en pareja vs. 1745 hijos de las personas que tienen pareja, por lo que vemos una clara relación positiva entre las variables estudiadas. Es decir, hay más probabilidad que 1 persona de nuestra muestra tenga hijo/a(s) si se encuentra en pareja.

3) Relación edad con cantidad de hijos

Como se puede observar en la matriz de correlación, existe una relación moderada entre la edad y la cantidad de hijos, con un coeficiente de 0,27. Para analizar mejor esta relación, se elaboró un gráfico de líneas que muestra el promedio de hijos por grupo de edad.



Las conclusiones que obtuvimos son las siguientes:

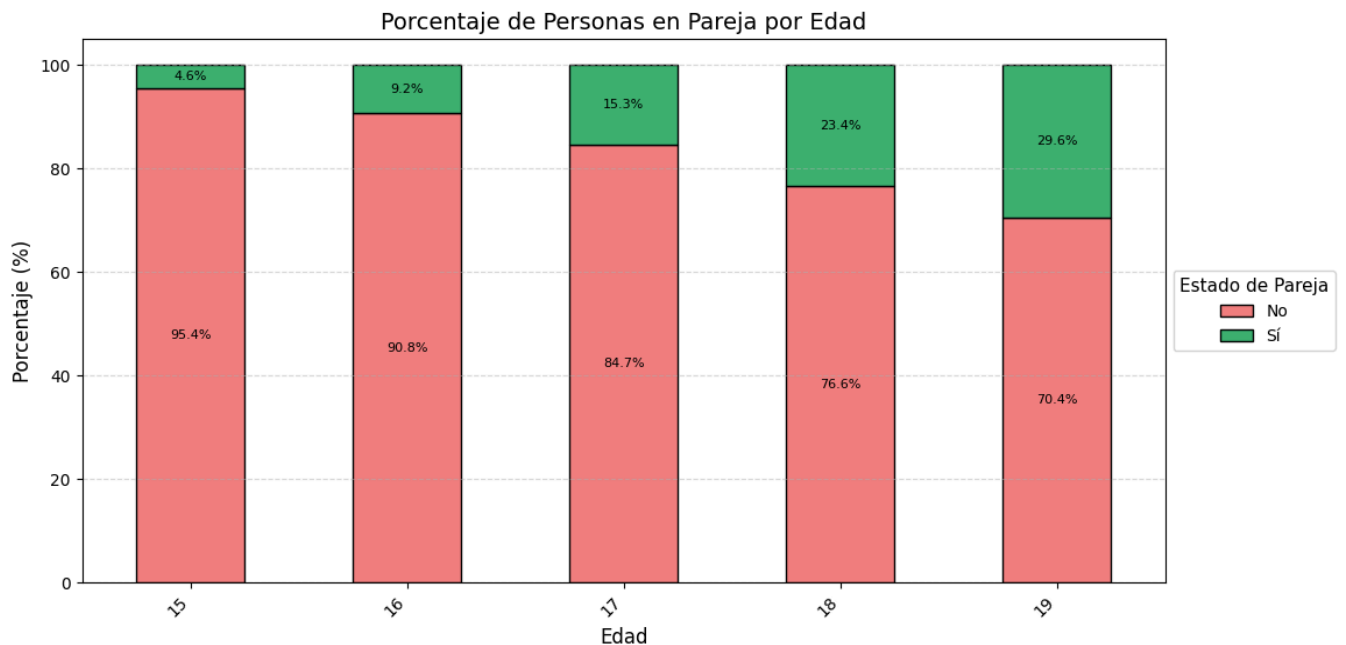
- Hay una relación positiva entre edad y cantidad de hijos: A medida que la edad aumenta, también lo hace el promedio de hijos.
- A partir de los 17 años el crecimiento en el promedio de hijos se acelera.

Estos hallazgos respaldan lo observado en la matriz de correlación, confirmando una relación moderada entre ambas variables.

4) Relación en pareja y edad

Otra relación moderada que se desprende de la matriz de correlación es En_pareja y edad con un coeficiente de 0,25.

Para estudiar a fondo esta relación realizamos una gráfica de columnas apiladas:



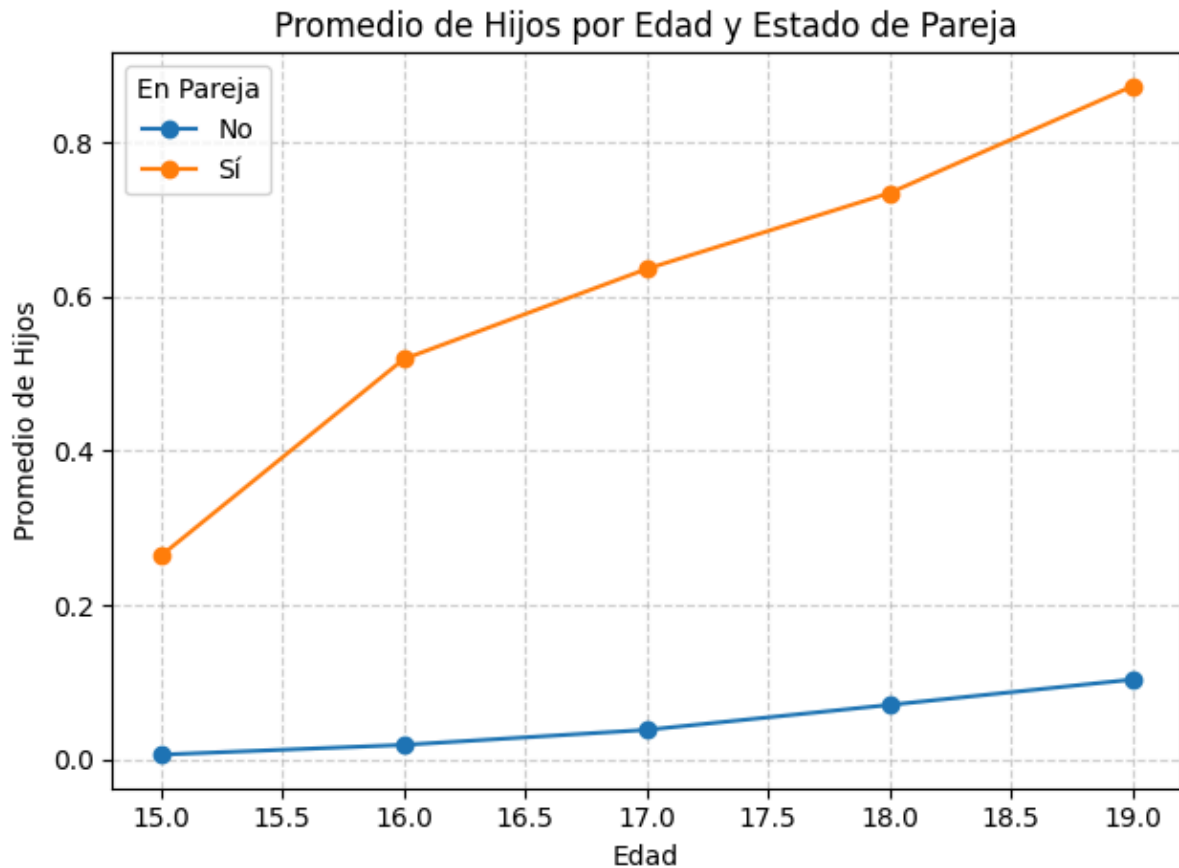
Las conclusiones que logramos obtener de dicha gráfica son las siguientes:

- Aumenta el porcentaje en pareja conforme a la edad: a mayor edad, mayor porcentaje de personas en pareja. Por ejemplo, mientras que solo el 4,6 % de los jóvenes de 15 años declara estar en pareja, este porcentaje asciende al 29,6 % entre los encuestados de 19 años.
- Mayoría aún sin pareja: A pesar del aumento, incluso a los 19 años, la mayoría (70,4%) no está en pareja.

Dichas conclusiones refuerzan la correlación moderada de las variables.

5) Relación edad, en pareja y número de hijos.

Una relación multivariable que nos pareció especialmente interesante fue la que vincula la cantidad de hijos con el estado de pareja y la edad.



Las conclusiones que logramos obtener son las siguientes:

- A mayor edad, mayor cantidad promedio de hijos, independientemente del estado de pareja. Esto sugiere una relación positiva entre edad y maternidad/paternidad.
- Las personas que están en pareja tienen, en promedio, significativamente más hijos que quienes no lo están en todos los rangos de edad.
- La diferencia en el promedio de hijos entre quienes están en pareja y quienes no se amplía con la edad, siendo especialmente notoria a los 18 y 19 años.
- A los 19 años, el promedio de hijos en jóvenes en pareja se acerca a 0.9, mientras que en quienes no están en pareja apenas supera 0.1. Esto refuerza el vínculo entre el hecho de estar en pareja y la probabilidad de tener hijos.

Conclusión

1) Podemos decir que nuestro dataset está limpio y que no hay valores faltantes ni duplicados. Podemos decir también que la población es joven, ya que el rango etario va de los 15 años a los 19. Los datos también arrojan que la mayoría de las personas de nuestro dataset son solteros, y la mayor cantidad de las personas cuenta con 9 años de estudio (esto es, la moda de `anios_educ` es 9). Otro dato revelador es que las personas que están en pareja son las que tienen más hijos.

2) A partir de la visualización de los diagramas de caja, extraemos las siguientes conclusiones para las variables **edad**, **anios_educ** y **num_hijos**:

- **Edad:** No se detectan outliers. La distribución es relativamente simétrica y la media y la mediana son similares, lo que sugiere una distribución aproximadamente normal.
- **Años de educación:** Se observan outliers tanto por debajo como por encima del rango intercuartílico. Los valores bajos (6 y 7 años) hacen que la media (8,51) se ubique por debajo de la mediana. La distribución no es normal; presenta asimetría negativa (sesgo a la izquierda), con mayor concentración de observaciones en los valores bajos.
- **Cantidad de hijos:** La caja del boxplot está concentrada en 0 hijos, indicando que la mayoría de los casos tienen ese valor. Existen outliers hacia valores superiores (1, 2 y 3 hijos). En este caso, la media es mayor que la mediana, debido a la influencia de esas observaciones con más hijos.

3) En cuanto a la matriz de correlación, la principal conclusión que destacamos es la correlación positiva entre las variables **en_pareja** y **num_hijos**. Esta correlación se refleja en los datos de nuestra muestra: las personas que están en pareja son las que tienen más hijos.

También destacamos la correlación positiva y moderada entre **edad** y **num_hijos**. Si bien nuestro dataset es de edades muy jóvenes (15 a 19 años), observamos que la mayor cantidad de hijos se da en personas de los 17 años en adelante.