**UNIVERSIDAD DEL VALLE DE GUATEMALA**

CC3084 - Data Science

Sección 11

Ing. Lynette García Pérez


# Avance Laboratorio No. 1


| | |
|---|---|
| José Pablo Orellana | 21970 |
| Diego Alberto Leiva | 21752 |

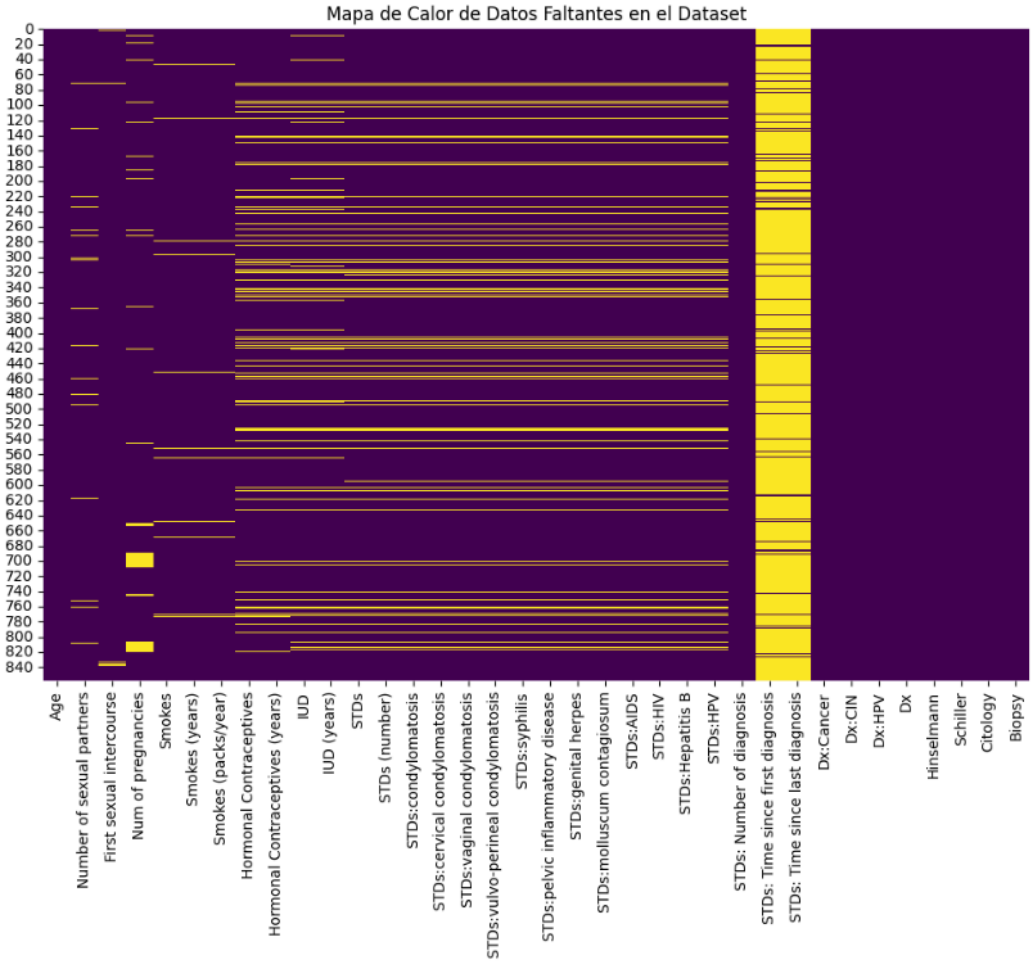

**GUATEMALA, 18 de julio del 2024**

## Descripción de los datos

- **Age:** Edad de la paciente.
- **Number.of.sexual.partners:** Cantidad de parejas sexuales que ha tenido la paciente desde que inició su vida sexual.
- **First.sexual.intercourse:** Edad a la que tuvo el primer encuentro sexual.
- **Num.of.pregnancies:** Cantidad de embarazos.
- **Smokes:** Si fuma o no.
- **Smokes.years:** Años que lleva fumando.
- **Smokes.packs.per.year:** Cajetillas de cigarrillos por año que fuma la paciente.
- **Hormonal.Contraceptives:** Si usa anticonceptivos hormonales o no.
- **Hormonal.Contraceptives.years:** Años que lleva usando. anticonceptivos hormonales.
- **IUD:** Si tiene colocado algún. dispositivo intrauterino (DIU).
- **IUD.years:** Años que lleva usando un DIU.
- **STDs:** Si ha tiene enfermedades de transmisión sexual (ETS).
- **STDs.number:** Cuantas ETS ha tenido.
- **STDs.condylomatosis:** si ha tenido condilomatosis.
- **STDs.cervical.condylomatosis:** si ha tenido condilomatosis cervical.
- **STDs.vaginal.condylomatosis:** Si ha tenido condilomatosis vaginal.
- **STDs.vulvo.perineal.condylomatosis:** Si ha tenido condilomatosis vulvo perineal.
- **STDs.syphilis:** Si ha tenido Sífilis.
- **STDs.pelvic.inflammatory.disease:** Si ha tenido inflamaciones pélvicas.
- **STDs.genital.herpes:** si ha tenido herpes genital.
- **STDs.molluscum.contagiosum:** Si ha tenido molusco contagioso.
- **STDs.AIDS:** Si tiene SIDA.
- **STDs.HIV:** Si tiene VIH.
- **STDs.Hepatitis.B:** si ha tenido o tiene hepatitis B.
- **STDs.HPV:** Si ha tenido o tiene Virus del Papiloma Humano (VPH).
- **STDs.Number.of.diagnosis:** Cantidad de diagnósticos de ETS.
- **STDs.Time.since.first.diagnosis:** Tiempo desde el primer diagnóstico.
- **STDs.Time.since.last.diagnosis:** Tiempo desde el último diagnóstico.
- **Dx.Cancer:** Si tiene diagnóstico de cáncer o no.
- **Dx.CIN:** Si tiene diagnóstico de NIC (Neoplasia Intraepitelial Cervical).
- **Dx.HPV:** Si tiene diagnóstico de Virus del Papiloma Humano.
- **Dx:** Si tiene diagnóstico.
- Pruebas para diagnosticar.
- **Hinselmann:** Si hicieron Colposcopía.
- **Schiller:** Sí hicieron la prueba de Schiller.
- **Citology:** Si hicieron citología o no.
- **Biopsy:** Si hicieron Biopsia o no.

## Procesamiento y Limpieza de Datos:

El primer paso en el procesamiento y limpieza de datos es la verificación y manejo de valores nulos o faltantes. Para visualizar la presencia de datos faltantes en nuestro conjunto de datos, se creó un mapa de calor que muestra las variables con valores nulos (ver Figura 1). Este mapa de calor revela que algunas variables tienen una cantidad significativa de datos faltantes, especialmente las variables relacionadas con el tiempo desde el diagnóstico de ETS (STDs: Time since first diagnosis y STDs: Time since last diagnosis), que carecen de más del 90% de sus datos.




(Figura 1).

Se calculó el porcentaje de valores nulos para cada variable numérica, y se encontró que, además de las dos variables mencionadas anteriormente, otras variables como IUD (years), Hormonal Contraceptives (years), y varias relacionadas con las ETS también tienen un alto porcentaje de valores faltantes. En total, hay 18 variables adicionales con más del 10% de valores faltantes.

Para evaluar el impacto de eliminar los registros con valores nulos, se eliminó temporalmente cualquier fila que contuviera valores nulos. Esto dejó apenas un 6.88% de los datos originales, lo que determinó que esta no es una acción viable para el análisis.

En consecuencia, se adoptó una estrategia combinada de eliminación e imputación de datos para manejar los valores nulos. Las variables con más del 10% de valores faltantes fueron eliminadas, mientras que para las variables con menos del 10% de valores faltantes se realizó una imputación de la media o la mediana para evitar la pérdida de datos y prevenir el sesgo. Esta nueva estrategia tuvo un impacto mucho menor, conservando cerca del 90% de los datos originales, lo que asegura la confiabilidad y la integridad del análisis.

## Análisis Exploratorio:

### Variables Cuantitativas

Las variables cuantitativas son aquellas que se expresan numéricamente y permiten realizar operaciones aritméticas. Estas variables pueden ser discretas, cuando toman valores enteros, o continuas, cuando pueden tomar cualquier valor dentro de un rango. Dentro del conjunto de datos, se logramos identificar las siguientes variables cuantitativas:

- La edad de la paciente (Age), que se mide en años.
- El número de parejas sexuales (Number of sexual partners), que representa una cuenta de las parejas sexuales que ha tenido la paciente desde que inició su vida sexual.
- La edad en el primer encuentro sexual (First sexual intercourse), también medida en años
- El número de embarazos (Num of pregnancies), que cuenta cuántas veces ha estado embarazada la paciente
- Los años fumando (Smokes (years)), que indica el tiempo en años que la paciente ha fumado; las cajetillas de cigarrillos por año (Smokes (packs/year)), que mide la cantidad de cajetillas fumadas anualmente
- Los años usando anticonceptivos hormonales (Hormonal Contraceptives (years)), que refleja el tiempo en años de uso de anticonceptivos hormonales
- Los años usando DIU (IUD (years)), que indica el tiempo en años de uso de dispositivo intrauterino
- El número de ETS (STDs (number)), que cuenta cuántas enfermedades de transmisión sexual ha tenido la paciente
- El tiempo desde el primer diagnóstico de ETS (STDs: Time since first diagnosis) y el tiempo desde el último diagnóstico de ETS (STDs: Time since last diagnosis), ambos medidos en años
- Resultados de diversas pruebas de diagnóstico (Hinselmann, Schiller, Citology, Biopsy), que son variables binarias que indican si se realizó o no una prueba diagnóstica específica.

En el conjunto de datos se identificaron un total de 34 variables cuantitativas. Estas se dividen en variables cuantitativas binarias, discretas y continuas.

**Variables Cuantitativas Binarias**

- Las variables cuantitativas binarias son aquellas que solo pueden tomar uno de dos valores posibles, generalmente representando la presencia o ausencia de una condición específica. En nuestro conjunto de datos, se identificaron 22 variables cuantitativas binarias, que incluyen:

  - Smokes
  - Hormonal Contraceptives
  - IUD
  - STDs
  - STDs: condylomatosis
  - STDs: vaginal condylomatosis
  - STDs: vulvo-perineal condylomatosis
  - STDs: syphilis
  - STDs: pelvic inflammatory disease
  - STDs: genital herpes
  - STDs: molluscum contagiosum
  - STDs: HIV
  - STDs: Hepatitis B
  - STDs: HPV
  - Dx: Cancer
  - Dx: CIN
  - Dx: HPV
  - Dx
  - Hinselmann
  - Schiller
  - Citology
  - Biopsy
  -

  Estas variables son críticas para determinar la presencia de factores de riesgo y diagnósticos específicos en las pacientes.

**Variables Cuantitativas Discretas**

- Las variables cuantitativas discretas pueden tomar valores enteros y representan conteos de eventos o características. En el conjunto de datos, se identificaron 8 variables cuantitativas discretas:

  - Age
  - Number of sexual partners
  - First sexual intercourse
  - Num of pregnancies
  - STDs (number)
  - STDs: cervical condylomatosis
  - STDs: AIDS
  - STDs: Number of diagnosis

  Estas variables permiten analizar la frecuencia de eventos como el número de parejas sexuales, embarazos y diagnósticos de ETS en las pacientes.

**Variables Cuantitativas Continuas**

- Las variables cuantitativas continuas pueden tomar cualquier valor dentro de un rango y generalmente se usan para medir magnitudes. En nuestro conjunto de datos, se identificaron 4 variables cuantitativas continuas:

    - Smokes (years)
    - Smokes (packs/year)
    - Hormonal Contraceptives (years)
    - IUD (years)

    Estas variables proporcionan información detallada sobre la duración y cantidad de ciertos comportamientos y tratamientos, como los años fumando y el uso de anticonceptivos hormonales.

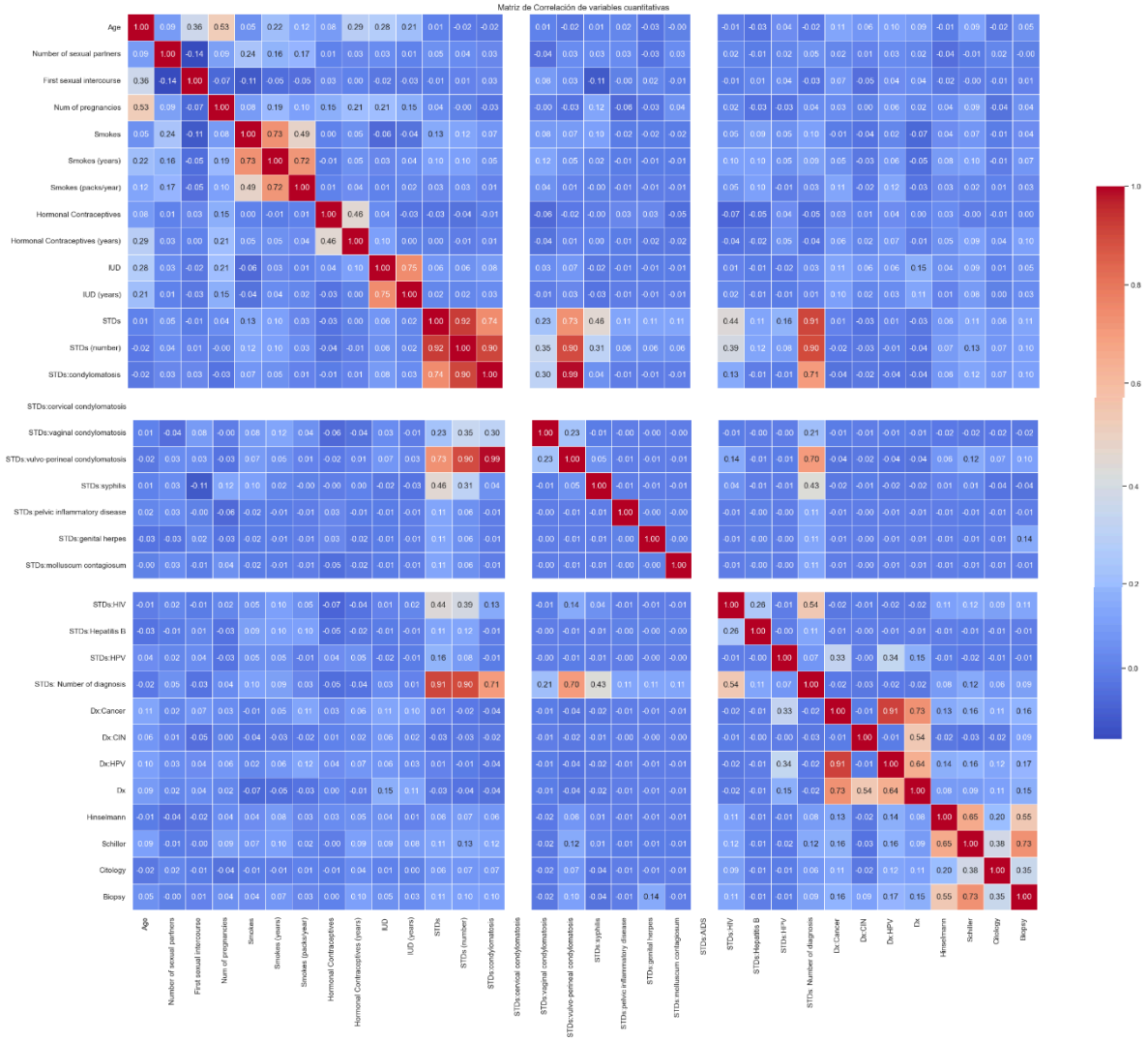
**Variables Categóricas**

- No se encontraron variables categóricas en el dataset. Las variables categóricas potenciales, como el año y el código en el ejemplo anterior, están representadas por números en este dataset, lo que indica que no hay variables categóricas estrictamente definidas.

**Estadística Descriptiva**

- El análisis de estadística descriptiva reveló las siguientes observaciones sobre nuestras variables cuantitativas: se observó un número variable de observaciones para cada variable, con la mayoría de las variables cuantitativas teniendo alrededor de 753 observaciones no nulas. La media y la mediana de las variables como la edad y el número de parejas sexuales varían, lo que indica una diversidad en el comportamiento de los pacientes. Los valores mínimos y máximos de variables como la edad y el número de parejas sexuales muestran la amplitud de los datos recogidos. La desviación estándar proporciona una medida de la variabilidad dentro de las variables, con algunas variables mostrando mayor variabilidad que otras.

## Matriz de Correlación de Variables Continuas



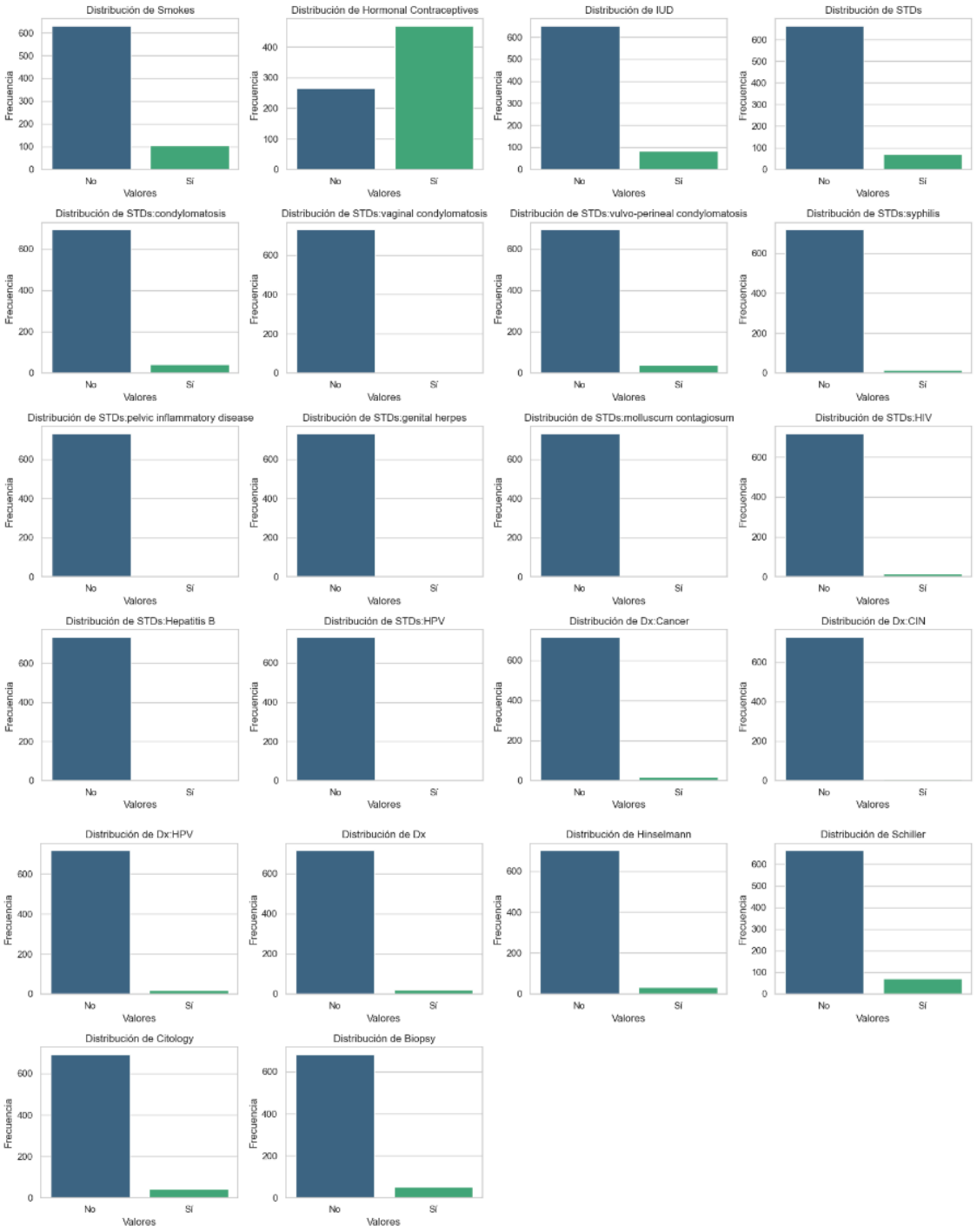

La mayoría de las correlaciones presentes en la matriz son bastante intuitivas. Por ejemplo, la variable STDs indica si el paciente tiene o no enfermedades de transmisión sexual (ETS) y, por lo tanto, tiene alta correlación con el número de diagnósticos (STDs: Number of diagnosis) y con otras variables relacionadas con las ETS que la paciente ha tenido.

Otro ejemplo claro es la correlación entre Number of sexual partners y First sexual intercourse. Una edad más temprana en el primer encuentro sexual tiende a estar relacionada con un mayor número de parejas sexuales a lo largo del tiempo.

# Análisis General

## Distribuciones de Variables Cuantitativas

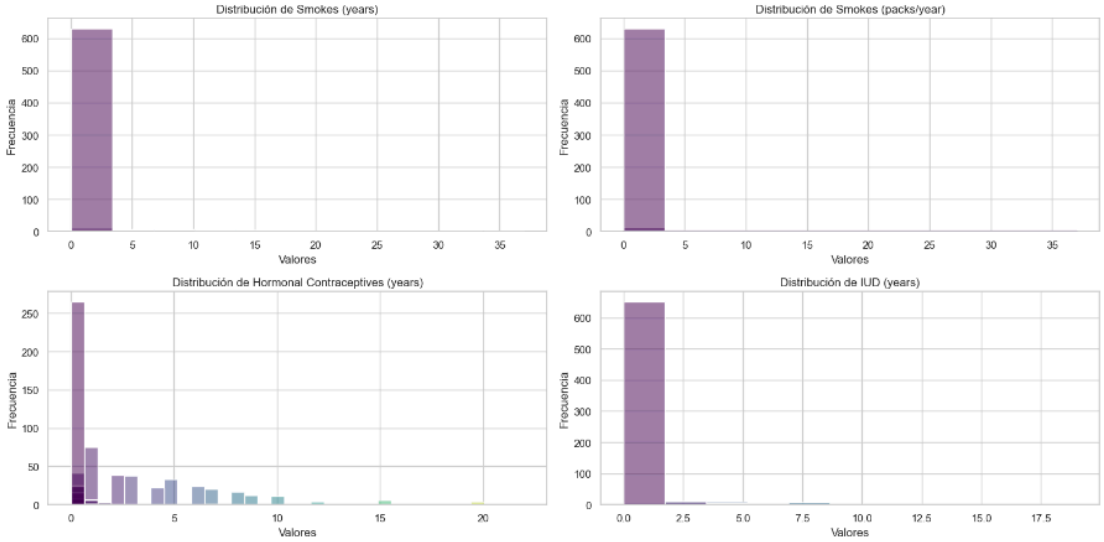
- **Variables Cuantitativas Binarias**



En base a las gráficas obtenidas en este análisis podemos deducir lo siguiente:

- Distribución de Smokes: La mayoría de los pacientes no fuman.
- Distribución de Hormonal Contraceptives: Un número significativo de pacientes usa anticonceptivos hormonales.
- Distribución de IUD: La mayoría de los pacientes no utilizan DIU.
- Distribución de STDs: La mayoría de los pacientes no tienen ETS.
- Distribución de STDs: condylomatosis, vaginal condylomatosis, vulvo-perineal condylomatosis, syphilis, pelvic inflammatory disease, genital herpes, molluscum contagiosum, HIV, Hepatitis B, HPV: La prevalencia de estas condiciones es muy baja entre los pacientes.

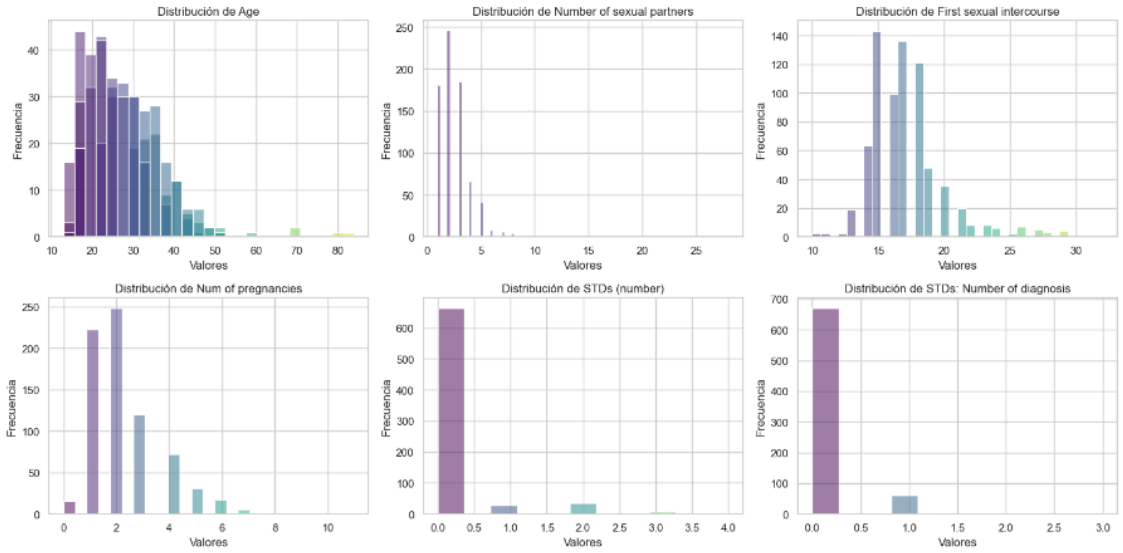
- **Variables Cuantitativas Continuas**




En base a las gráficas obtenidas en este análisis podemos deducir lo siguiente:

- Distribución de Smokes (years): La mayoría de los pacientes han fumado durante un corto periodo.
- Distribución de Smokes (packs/year): La cantidad de cigarrillos fumados por año también es baja.
- Distribución de Hormonal Contraceptives (years): La mayoría de los pacientes que usan anticonceptivos hormonales lo han hecho por un tiempo relativamente corto.
- Distribución de IUD (years): Similarmente, el uso de DIU también es por un periodo corto.

- **Variables Cuantitativas Discretas**




En base a las gráficas obtenidas en este análisis podemos deducir lo siguiente:

- Distribución de Age: La distribución muestra que la mayoría de los pacientes están en el rango de 20 a 30 años.
- Distribución de Number of sexual partners: La mayoría de los pacientes han tenido entre 1 y 3 parejas sexuales.
- Distribución de First sexual intercourse: La mayoría de los pacientes tuvieron su primer encuentro sexual entre los 15 y 20 años.
- Distribución de Num of pregnancies: La mayoría de las pacientes han tenido entre 1 y 3 embarazos.
- Distribución de STDs (number): La mayoría de los pacientes no han tenido ETS.
- Distribución de STDs: Number of diagnosis: La mayoría de los pacientes han tenido entre 0 y 1 diagnósticos de ETS.

## Enlaces

- **GitHub: https://github.com/LeivaDiego/DataScience-Lab1**
- **Drive: https://acortar.link/Lq7WJb**