David Aragón 21053
Renatto Guzmán 21646
Data Science - 10

Laboratorio 1

1. Haga una exploración rápida de sus datos. Puede usar alguna forma automatizada de hacer análisis exploratorio siempre y cuando explique los resultados que arrojan los módulos/paquetes.

Tenemos para este dataset de mujeres para la predicción del cáncer cervical, 858 registros en total, ya que todas las columnas tienen 858 valores no nulos.

La mayoría de las columnas (26 de 36) son de tipo 'object', lo que generalmente indica que contienen cadenas de texto o datos mixtos, y 10 columnas de estas son de tipo 'int64', lo que indica que contienen números enteros, como se puede ver en la siguiente imagen:

La gran cantidad de columnas 'object' sugiere que será necesario hacer una limpieza y conversión de datos significativa antes de poder realizar análisis estadísticos más avanzados, ya que es extremadamente raro que no hayan elementos nulos en ninguna columna, y luego de ver los datos lo más probable es que sea que estén formateados como '?'.

```
Data columns (total 36 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   Age                                 858 non-null    int64
 1   Number of sexual partners           858 non-null    object
 2   First sexual intercourse            858 non-null    object
 3   Num of pregnancies                  858 non-null    object
 4   Smokes                              858 non-null    object
 5   Smokes (years)                      858 non-null    object
 6   Smokes (packs/year)                 858 non-null    object
 7   Hormonal Contraceptives             858 non-null    object
 8   Hormonal Contraceptives (years)     858 non-null    object
 9   IUD                                 858 non-null    object
 10  IUD (years)                         858 non-null    object
 11  STDs                                858 non-null    object
 12  STDs (number)                       858 non-null    object
 13  STDs:condylomatosis                 858 non-null    object
 14  STDs:cervical condylomatosis        858 non-null    object
 15  STDs:vaginal condylomatosis         858 non-null    object
 16  STDs:vulvo-perineal condylomatosis  858 non-null    object
 17  STDs:syphilis                       858 non-null    object
 18  STDs:pelvic inflammatory disease    858 non-null    object
 19  STDs:genital herpes                 858 non-null    object
 20  STDs:molluscum contagiosum          858 non-null    object
 21  STDs:AIDS                           858 non-null    object
 22  STDs:HIV                            858 non-null    object
 23  STDs:Hepatitis B                    858 non-null    object
 24  STDs:HPV                            858 non-null    object
 25  STDs: Number of diagnosis           858 non-null    int64
 26  STDs: Time since first diagnosis    858 non-null    object
 27  STDs: Time since last diagnosis     858 non-null    object
 28  Dx:Cancer                           858 non-null    int64
 29  Dx:CIN                              858 non-null    int64
 30  Dx:HPV                              858 non-null    int64
 31  Dx                                  858 non-null    int64
 32  Hinselmann                          858 non-null    int64
 33  Schiller                            858 non-null    int64
 34  Citology                            858 non-null    int64
 35  Biopsy                              858 non-null    int64
```

```
Valores nulos en el dataset
Age                                   0
Number of sexual partners             0
First sexual intercourse              0
Num of pregnancies                    0
Smokes                                0
Smokes (years)                        0
Smokes (packs/year)                   0
Hormonal Contraceptives               0
Hormonal Contraceptives (years)       0
IUD                                   0
IUD (years)                           0
STDs                                  0
STDs (number)                         0
STDs:condylomatosis                   0
STDs:cervical condylomatosis          0
STDs:vaginal condylomatosis           0
STDs:vulvo-perineal condylomatosis    0
STDs:syphilis                         0
STDs:pelvic inflammatory disease      0
STDs:genital herpes                   0
STDs:molluscum contagiosum            0
STDs:AIDS                             0
STDs:HIV                              0
STDs:Hepatitis B                      0
STDs:HPV                              0
STDs: Number of diagnosis             0
STDs: Time since first diagnosis      0
STDs: Time since last diagnosis       0
Dx:Cancer                             0
Dx:CIN                                0
Dx:HPV                                0
Dx                                    0
Hinselmann                            0
Schiller                              0
Citology                              0
Biopsy                                0
```

Descripción estadística del dataset.

Edad:

- La edad mediana del dataset es de 26.8 años, es decir que el 50% de las participantes tiene esta edad o menos.
- La edad mínima es 13 años y la máxima 84 años, lo que indica un rango amplio.
- El 75% de las participantes tiene 32 años o menos, lo que sugiere que la muestra está sesgada hacia mujeres jóvenes.

Número de diagnósticos de ETS:

- La media es muy baja (0.087), lo que indica que la mayoría de las participantes no tienen diagnósticos de ETS.
- El máximo es 3, lo que significa que algunas participantes tienen hasta 3 diagnósticos de ETS.
- La mediana y los percentiles 25 y 75 son 0, lo que refuerza que la mayoría no tiene diagnósticos de ETS.

Diagnósticos (Dx:Cancer, Dx:CIN, Dx:HPV, Dx):

- Todas estas variables tienen medias muy bajas (entre 0.01 y 0.028), lo que indica que son eventos raros en la muestra.
- Los valores máximos son 1, lo que sugiere que estas son variables binarias .
- Para el cáncer, aproximadamente el 2.1% de la muestra tiene un diagnóstico positivo.
- Para CIN, aproximadamente el 1% de la muestra tiene un diagnóstico positivo.
- Para HPV, aproximadamente el 2.1% de la muestra tiene un diagnóstico positivo.
- Para Dx, que es si el paciente tiene diagnóstico, aproximadamente el 2.8% de la muestra tiene un resultado positivo.

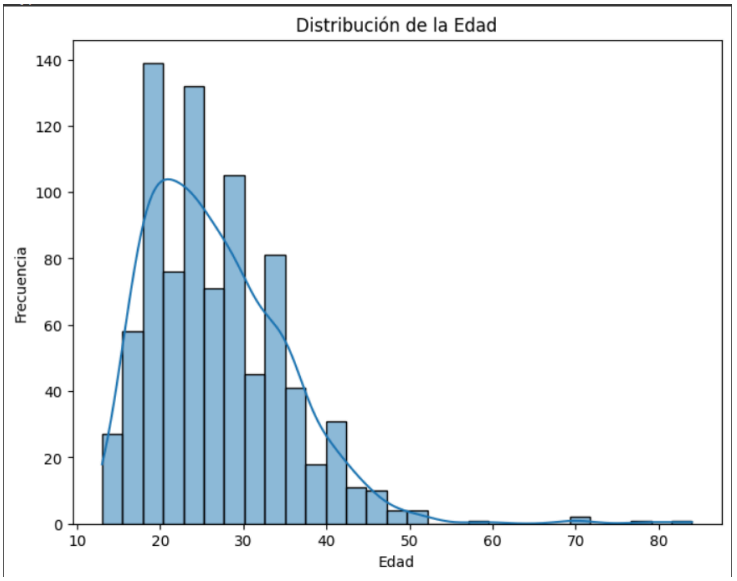
Descripción estadística del dataset

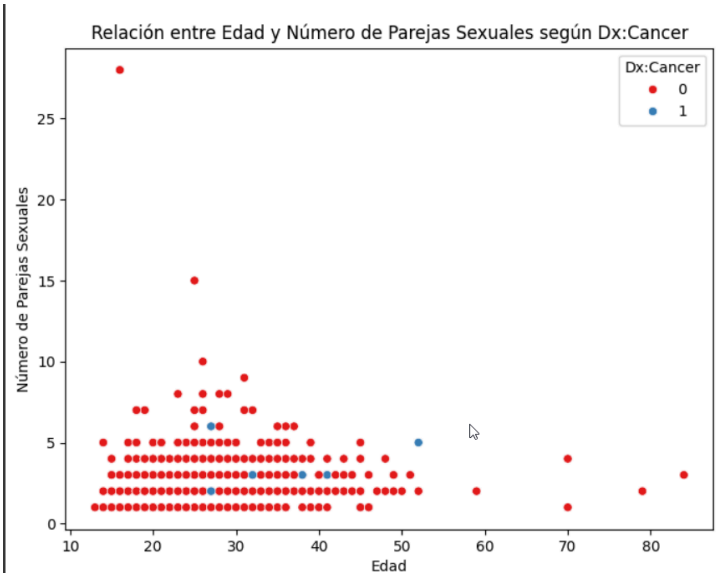
	Age	STDs: Number of diagnosis	Dx:Cancer	Dx:CIN
count	858.000000	858.000000	858.000000	858.000000
mean	26.820513	0.087413	0.020979	0.010490
std	8.497948	0.302545	0.143398	0.101939
min	13.000000	0.000000	0.000000	0.000000
25%	20.000000	0.000000	0.000000	0.000000
50%	25.000000	0.000000	0.000000	0.000000
75%	32.000000	0.000000	0.000000	0.000000
max	84.000000	3.000000	1.000000	1.000000

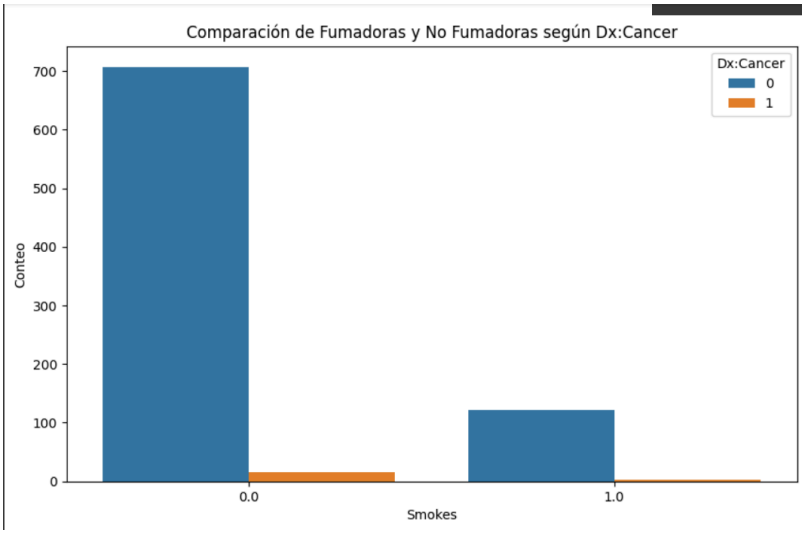
	Dx:HPV	Dx	Hinselmann	Schiller	Citology	Biopsy
count	858.000000	858.000000	858.000000	858.000000	858.000000	858.000000
mean	0.020979	0.027972	0.040793	0.086247	0.051282	0.064103
std	0.143398	0.164989	0.197925	0.280892	0.220701	0.245078
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

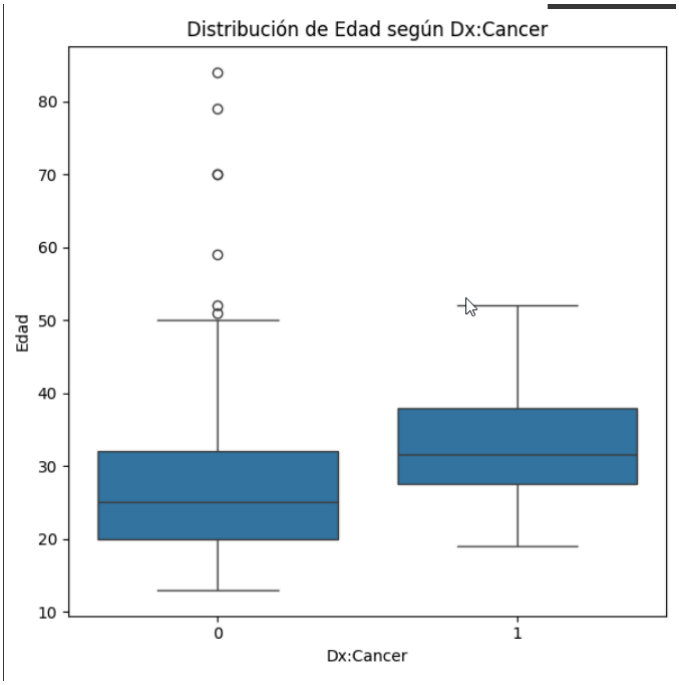
2. Diga el tipo de cada una de las variables del conjunto de datos (cualitativa o categórica, cuantitativa continua, cuantitativa discreta).
   Age: Cuantitativa continua.
   Number.of.sexual.partners: Cuantitativa discreta.
   First.sexual.intercourse: Cuantitativa continua.
   Num.of.pregnancies: Cuantitativa discreta.
   Smokes: Cualitativa.
   STDs.molluscum.contagiosum: Cualitativa.
   STDs.AIDS: Cualitativa.
   STDs.HIV: Cualitativa.
   STDs.Hepatitis.B: Cualitativa.
   STDs.HPV: Cualitativa.
   STDs.Number.of.diagnosis: Cuantitativa discreta.
   STDs.Time.since.first.diagnosis: Cuantitativa continua.
   STDs.Time.since.last.diagnosis: Cuantitativa continua.
   Dx.Cancer: Cualitativa.
   Dx.CIN: Cualitativa.
   Dx.HPV: Cualitativa.
   Dx: Cualitativa.
   Hinselmann: Cualitativa.
   Schiller: Cualitativa.
   Citology: Cualitativa.
   Biopsy: Cualitativa.
   Smokes.years: Cuantitativa continua.
   Smokes.packs.per.year: Cuantitativa continua.
   Hormonal.Contraceptives: Cualitativa.
   Hormonal.Contraceptives.years: Cuantitativa continua.
   IUD: Cualitativa.
   IUD.years: Cuantitativa continua.
   STDs: Cualitativa.
   STDs.number: Cuantitativa discreta.
   STDs.condylomatosis: Cualitativa.
   STDs.cervical.condylomatosis: Cualitativa.
   STDs.vaginal.condylomatosis: Cualitativa.
   STDs.vulvo.perineal.condylomatosis: Cualitativa.
   STDs.syphilis: Cualitativa.
   STDs.pelvic.inflammatory.disease: Cualitativa.
   STDs.genital.herpes: Cualitativa.

3. Incluya los gráficos exploratorios siendo consecuentes con el tipo de variable que están representando.


Distribución de la Edad
Frecuencia
Edad
0
20
40
60
80
100
120
140
10
20
30
40
50
60
70
80

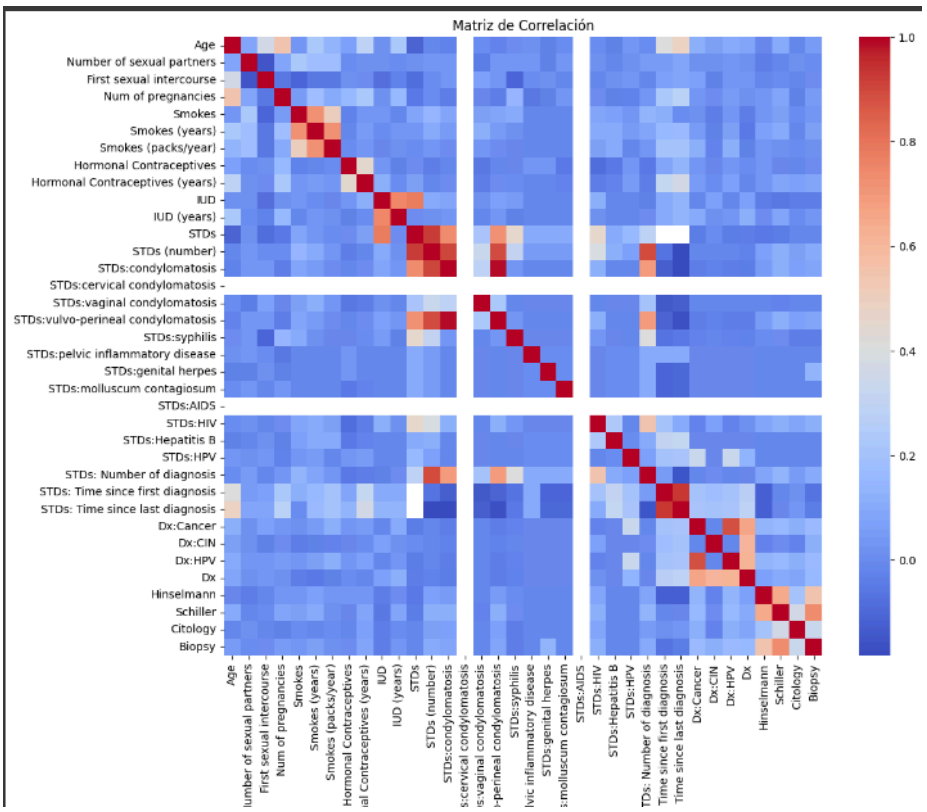

Relación entre Edad y Número de Parejas Sexuales según Dx:Cancer
Dx:Cancer
0
1
Número de Parejas Sexuales
Edad
0
5
10
15
20
25
10
20
30
40
50
60
70
80


Comparación de Fumadoras y No Fumadoras según Dx:Cancer
Dx:Cancer
0
1
Conteo
Smokes
0
100
200
300
400
500
600
700
0.0
1.0



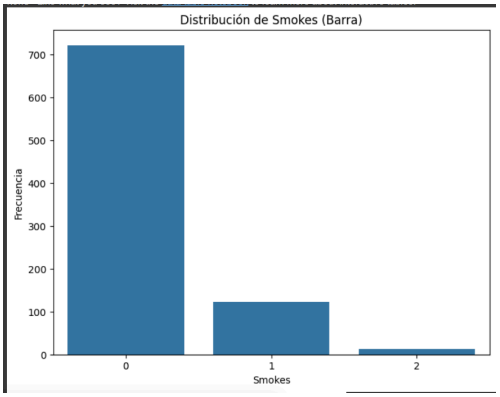
4. Aísle las variables numéricas de las categóricas, haga un análisis de correlación entre las mismas.

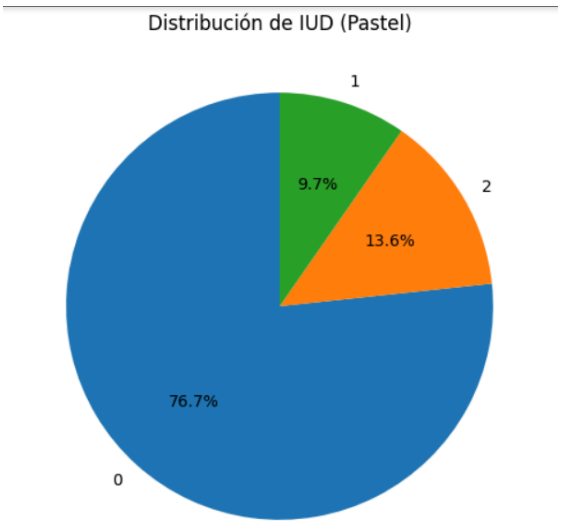
```python
# Separar variables numéricas y categóricas
numerical_vars = data.select_dtypes(include=['float64', 'int64'])
categorical_vars = data.select_dtypes(include=['object', 'bool'])

# Calcular matriz de correlación para variables numéricas
corr_matrix = numerical_vars.corr()

# Visualizar matriz de correlación
plt.figure(figsize=(12, 10))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
plt.title('Matriz de Correlación entre Variables Numéricas')
plt.show()
```

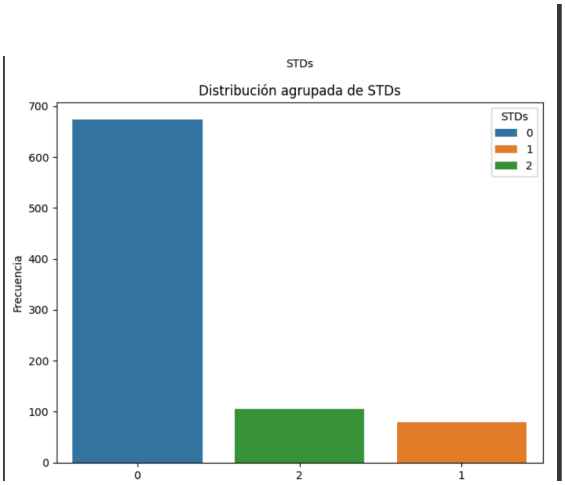


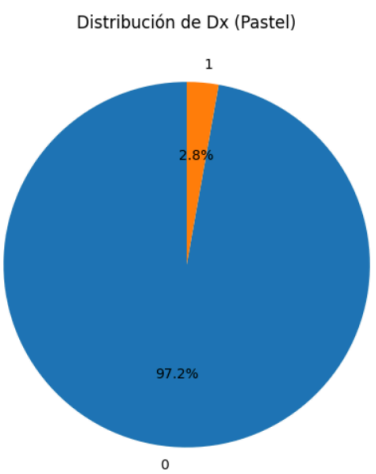
5. Utilice las variables categóricas, haga tablas de frecuencia, proporción, gráficas de barras o cualquier otra técnica que le permita explorar los datos.

```
Variables categóricas identificadas:
Index(['Smokes', 'IUD', 'STDs', 'Dx', 'Hinselmann', 'Schiller', 'Citology',
       'Biopsy'],
      dtype='object')
```









6. Determine el comportamiento a seguir con los valores faltantes. Explique si necesita remover alguna variable por la cantidad de valores faltantes que tiene. ¿Es factible eliminar todos los valores faltantes de todas las variables?

Basándonos en la información dada en las gráficas anteriores, primordialmente por la matriz de correlación nos dimos cuenta que la información de las variables ‘STDS:AIDS’ y ‘STDs:cervical condylomatosis’ no tienen datos útiles, son prácticamente columnas vacías que no aportarían información útil. Cabe destacar que la columna de condilomatosis cervical hubiera sido interesante tener datos útiles ya que es una enfermedad que se encuentra en la cérvix y asimismo podría estar relacionada con el cáncer de cérvix.

```
print(data['STDs:AIDS '].describe())
print()
print(data['STDs:cervical condylomatosis '].describe())
```

✓ 0.0s

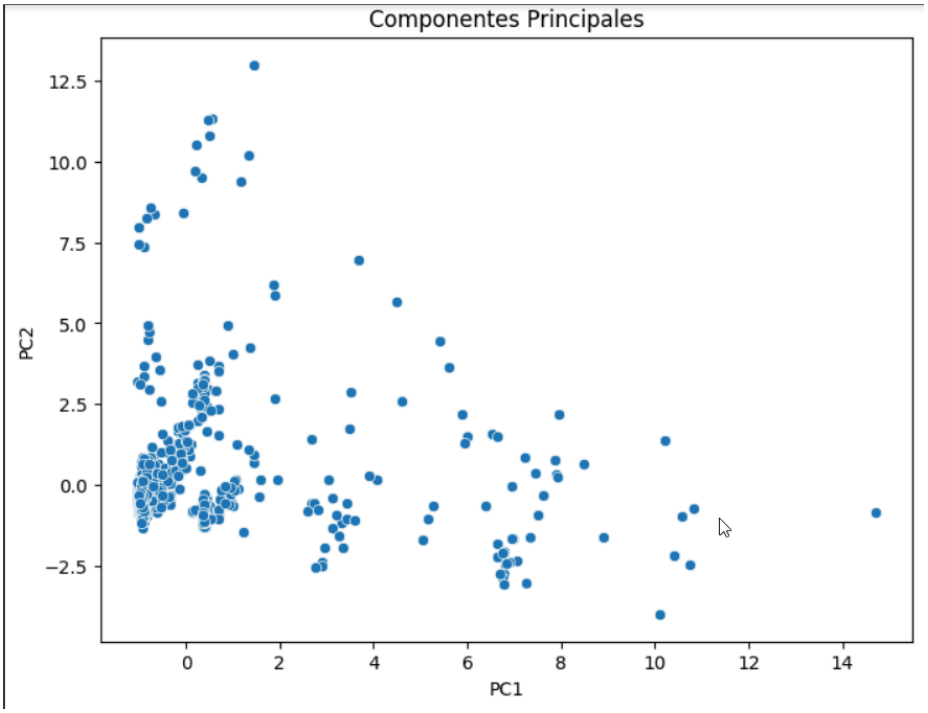
```
count           858
unique            2
top             0.0
freq            753
Name: STDs:AIDS , dtype: object

count                                858
unique                                 2
top                                  0.0
freq                                 753
Name: STDs:cervical condylomatosis , dtype: object
```

7. Estudie si es posible hacer transformaciones en las variables categóricas para incluirlas en el PCA, ¿valdrá la pena?

Decidimos utilizar el método de Codificación de Etiquetas (Label Encoding) para aplicar PCA, esto debido a que muchas de las variables categóricas que se pueden encontrar en en los datos son numéricas, por ejemplo muchas de las variables referentes a si se hicieron ciertas pruebas constan de 1 y 0 para referirse a 1 como “si” y 0 como “no”, y ya que no

todas las variables son de este tipo pero si la mayoría, decidimos hacerlo por etiquetas numéricas.

```python
# Codificación de Etiquetas
label_encoder = LabelEncoder()
for column in categorical_columns:
    if column in data.columns:
        data[column] = label_encoder.fit_transform(data[column].astype(str))

# Imputar valores faltantes
imputer = SimpleImputer(strategy='mean')  # Puedes cambiar la estrategia según sea necesario
data_imputed = pd.DataFrame(imputer.fit_transform(data), columns=data.columns)

# Estandarización de todas las variables
scaler = StandardScaler()
data_scaled = scaler.fit_transform(data_imputed)

# Aplicar PCA
pca = PCA(n_components=2)  # Ajusta el número de componentes principales según sea necesario
principal_components = pca.fit_transform(data_scaled)

# Convertir los componentes principales a un DataFrame
pca_df = pd.DataFrame(data=principal_components, columns=['PC1', 'PC2'])
```



8. Estudie si es conveniente hacer un Análisis de Componentes Principales. Recuerde que puede usar el índice KMO y el test de esfericidad de Bartlett. Haga un análisis de componentes principales con las variables numéricas, discuta los resultados e interprete los componentes.

9. Obtenga reglas de asociación interesantes del dataset. Recuerde discretizar las variables numéricas. Genere reglas con diferentes niveles de confianza y soporte. Discuta los resultados. Si considera que debe eliminar variables porque son muy frecuentes y con eso puede recibir más insigths de la generación de reglas. Hágalo y discútalo.