Setup

```
pip install ucimlrepo
```

```
Collecting ucimlrepo
  Downloading ucimlrepo-0.0.6-py3-none-any.whl (8.0 kB)
Installing collected packages: ucimlrepo
Successfully installed ucimlrepo-0.0.6
```

```
from ucimlrepo import fetch_ucirepo

# fetch dataset
cervical_cancer_risk_factors = fetch_ucirepo(id=383)

# data (as pandas dataframes)
X = cervical_cancer_risk_factors.data.features
y = cervical_cancer_risk_factors.data.targets

# metadata
print(cervical_cancer_risk_factors.metadata)

# variable information
print(cervical_cancer_risk_factors.variables)
```

```
{'uci_id': 383, 'name': 'Cervical Cancer (Risk Factors)', 'repository_url': 'https://archive.ics.uci.edu/dataset/383/cervical+cancer+
                                 name     role        type demographic  \
0                                 Age  Feature     Integer         Age
1           Number of sexual partners  Feature  Continuous       Other
2            First sexual intercourse  Feature  Continuous        None
3                  Num of pregnancies  Feature  Continuous        None
4                              Smokes  Feature  Continuous        None
5                      Smokes (years)  Feature  Continuous        None
6                 Smokes (packs/year)  Feature  Continuous        None
7             Hormonal Contraceptives  Feature  Continuous        None
8     Hormonal Contraceptives (years)  Feature  Continuous        None
9                                 IUD  Feature  Continuous        None
10                        IUD (years)  Feature  Continuous        None
11                               STDs  Feature  Continuous        None
12                      STDs (number)  Feature  Continuous        None
13                STDs:condylomatosis  Feature  Continuous        None
14       STDs:cervical condylomatosis  Feature  Continuous        None
15        STDs:vaginal condylomatosis  Feature  Continuous        None
16 STDs:vulvo-perineal condylomatosis  Feature  Continuous        None
17                      STDs:syphilis  Feature  Continuous        None
18   STDs:pelvic inflammatory disease  Feature  Continuous        None
19                STDs:genital herpes  Feature  Continuous        None
20          STDs:molluscum contagiosum  Feature  Continuous        None
21                          STDs:AIDS  Feature  Continuous        None
22                           STDs:HIV  Feature  Continuous        None
23                   STDs:Hepatitis B  Feature  Continuous        None
24                           STDs:HPV  Feature  Continuous        None
25          STDs: Number of diagnosis  Feature     Integer        None
26   STDs: Time since first diagnosis  Feature  Continuous        None
27    STDs: Time since last diagnosis  Feature  Continuous        None
28                          Dx:Cancer  Feature     Integer        None
29                             Dx:CIN  Feature     Integer        None
30                             Dx:HPV  Feature     Integer        None
31                                 Dx  Feature     Integer        None
32                         Hinselmann  Feature     Integer        None
33                           Schiller  Feature     Integer        None
34                           Citology  Feature     Integer        None
35                             Biopsy  Feature     Integer        None

   description units missing_values
0         None  None             no
1         None  None            yes
2         None  None            yes
3         None  None            yes
4         None  None            yes
5         None  None            yes
6         None  None            yes
7         None  None            yes
8         None  None            yes
9         None  None            yes
10        None  None            yes
11        None  None            yes
12        None  None            yes
```

```
13        None  None            yes
14        None  None            yes
15        None  None            yes
16        None  None            yes
```

```
import pandas as pd
import numpy as np
```

```
X
```

	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormo Contracepti
0	18	4.0	15.0	1.0	0.0	0.0	0.0	
1	15	1.0	14.0	1.0	0.0	0.0	0.0	
2	34	1.0	NaN	1.0	0.0	0.0	0.0	
3	52	5.0	16.0	4.0	1.0	37.0	37.0	
4	46	3.0	21.0	4.0	0.0	0.0	0.0	
...	...	...	...	...	...	...	...	
853	34	3.0	18.0	0.0	0.0	0.0	0.0	
854	32	2.0	19.0	1.0	0.0	0.0	0.0	
855	25	2.0	17.0	0.0	0.0	0.0	0.0	
856	33	2.0	24.0	2.0	0.0	0.0	0.0	
857	29	2.0	20.0	1.0	0.0	0.0	0.0	

858 rows × 36 columns

```
dataFrames = [X,y]
df = pd.concat(dataFrames, axis = 1)
df
```

	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormo Contracepti
0	18	4.0	15.0	1.0	0.0	0.0	0.0	
1	15	1.0	14.0	1.0	0.0	0.0	0.0	
2	34	1.0	NaN	1.0	0.0	0.0	0.0	
3	52	5.0	16.0	4.0	1.0	37.0	37.0	
4	46	3.0	21.0	4.0	0.0	0.0	0.0	
...	...	...	...	...	...	...	...	
853	34	3.0	18.0	0.0	0.0	0.0	0.0	
854	32	2.0	19.0	1.0	0.0	0.0	0.0	
855	25	2.0	17.0	0.0	0.0	0.0	0.0	
856	33	2.0	24.0	2.0	0.0	0.0	0.0	
857	29	2.0	20.0	1.0	0.0	0.0	0.0	

858 rows × 36 columns

I am more familiarized with Biopsy. So selecting this is more reasonable as it is either 0 - No Biopsy or 1 - Biopsy.

```
y = df[['Biopsy']]
y
```

	Biopsy
0	0
1	0
2	0
3	0
4	0
...	...
853	0
854	0
855	0
856	0
857	0

858 rows × 1 columns

Next steps: View recommended plots

Double-click (or enter) to edit

```
y.value_counts() # 803 no biopsy while 55 conducted biopsy
```

```
Biopsy
0         803
1          55
Name: count, dtype: int64
```

```
X.shape
```

```
(858, 36)
```

```
df.dtypes # check dtypes
```

```
Age                                     int64
Number of sexual partners             float64
First sexual intercourse              float64
Num of pregnancies                    float64
Smokes                                float64
Smokes (years)                        float64
Smokes (packs/year)                   float64
Hormonal Contraceptives               float64
Hormonal Contraceptives (years)       float64
IUD                                   float64
IUD (years)                           float64
STDs                                  float64
STDs (number)                         float64
STDs:condylomatosis                   float64
STDs:cervical condylomatosis          float64
STDs:vaginal condylomatosis           float64
STDs:vulvo-perineal condylomatosis    float64
STDs:syphilis                         float64
STDs:pelvic inflammatory disease      float64
STDs:genital herpes                   float64
STDs:molluscum contagiosum            float64
STDs:AIDS                             float64
STDs:HIV                              float64
STDs:Hepatitis B                      float64
STDs:HPV                              float64
STDs: Number of diagnosis               int64
STDs: Time since first diagnosis      float64
STDs: Time since last diagnosis       float64
Dx:Cancer                               int64
Dx:CIN                                  int64
Dx:HPV                                  int64
Dx                                      int64
Hinselmann                              int64
Schiller                                int64
Citology                                int64
```

```
    Biopsy                            int64
    dtype: object
```

```
df.isnull().sum() #check null values
```

```
    Age                                  0
    Number of sexual partners           26
    First sexual intercourse             7
    Num of pregnancies                  56
    Smokes                              13
    Smokes (years)                      13
    Smokes (packs/year)                 13
    Hormonal Contraceptives            108
    Hormonal Contraceptives (years)    108
    IUD                                117
    IUD (years)                        117
    STDs                               105
    STDs (number)                      105
    STDs:condylomatosis                105
    STDs:cervical condylomatosis       105
    STDs:vaginal condylomatosis        105
    STDs:vulvo-perineal condylomatosis 105
    STDs:syphilis                      105
    STDs:pelvic inflammatory disease   105
    STDs:genital herpes                105
    STDs:molluscum contagiosum         105
    STDs:AIDS                          105
    STDs:HIV                           105
    STDs:Hepatitis B                   105
    STDs:HPV                           105
    STDs: Number of diagnosis            0
    STDs: Time since first diagnosis   787
    STDs: Time since last diagnosis    787
    Dx:Cancer                            0
    Dx:CIN                               0
    Dx:HPV                               0
    Dx                                   0
    Hinselmann                           0
    Schiller                             0
    Citology                             0
    Biopsy                               0
    dtype: int64
```

```
def check_duplicates(df):
  if df[df.duplicated()].shape[0] != 0:
    print(df[df.duplicated()].shape[0])
  else:
    print("No existing duplicates")
check_duplicates(df)
```

```
    23
```

```
cc_df = df.copy()
```

```
df.dtypes
```

```
    Age                                  int64
    Number of sexual partners          float64
    First sexual intercourse           float64
    Num of pregnancies                 float64
    Smokes                             float64
    Smokes (years)                     float64
    Smokes (packs/year)                float64
    Hormonal Contraceptives            float64
    Hormonal Contraceptives (years)    float64
    IUD                                float64
    IUD (years)                        float64
    STDs                               float64
    STDs (number)                      float64
    STDs:condylomatosis                float64
    STDs:cervical condylomatosis       float64
    STDs:vaginal condylomatosis        float64
    STDs:vulvo-perineal condylomatosis float64
    STDs:syphilis                      float64
    STDs:pelvic inflammatory disease   float64
    STDs:genital herpes                float64
    STDs:molluscum contagiosum         float64
    STDs:AIDS                          float64
    STDs:HIV                           float64
    STDs:Hepatitis B                   float64
```

```
STDs:HPV                              float64
STDs: Number of diagnosis               int64
STDs: Time since first diagnosis      float64
STDs: Time since last diagnosis       float64
Dx:Cancer                               int64
Dx:CIN                                  int64
Dx:HPV                                  int64
Dx                                      int64
Hinselmann                              int64
Schiller                                int64
Citology                                int64
Biopsy                                  int64
dtype: object
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 858 entries, 0 to 857
Data columns (total 36 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   Age                                 858 non-null    int64
 1   Number of sexual partners           832 non-null    float64
 2   First sexual intercourse            851 non-null    float64
 3   Num of pregnancies                  802 non-null    float64
 4   Smokes                              845 non-null    float64
 5   Smokes (years)                      845 non-null    float64
 6   Smokes (packs/year)                 845 non-null    float64
 7   Hormonal Contraceptives             750 non-null    float64
 8   Hormonal Contraceptives (years)     750 non-null    float64
 9   IUD                                 741 non-null    float64
 10  IUD (years)                         741 non-null    float64
 11  STDs                                753 non-null    float64
 12  STDs (number)                       753 non-null    float64
 13  STDs:condylomatosis                 753 non-null    float64
 14  STDs:cervical condylomatosis        753 non-null    float64
 15  STDs:vaginal condylomatosis         753 non-null    float64
 16  STDs:vulvo-perineal condylomatosis  753 non-null    float64
 17  STDs:syphilis                       753 non-null    float64
 18  STDs:pelvic inflammatory disease    753 non-null    float64
 19  STDs:genital herpes                 753 non-null    float64
 20  STDs:molluscum contagiosum          753 non-null    float64
 21  STDs:AIDS                           753 non-null    float64
 22  STDs:HIV                            753 non-null    float64
 23  STDs:Hepatitis B                    753 non-null    float64
 24  STDs:HPV                            753 non-null    float64
 25  STDs: Number of diagnosis           858 non-null    int64
 26  STDs: Time since first diagnosis    71 non-null     float64
 27  STDs: Time since last diagnosis     71 non-null     float64
 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
dtypes: float64(26), int64(10)
memory usage: 241.4 KB
```

```
df.drop_duplicates(inplace=True)
```

```
cc_df.drop_duplicates(inplace=True)
```

```
check_duplicates(df)
```

```
No existing duplicates
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 835 entries, 0 to 857
Data columns (total 36 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   Age                                 835 non-null    int64
 1   Number of sexual partners           810 non-null    float64
 2   First sexual intercourse            828 non-null    float64
 3   Num of pregnancies                  779 non-null    float64
 4   Smokes                              822 non-null    float64
 5   Smokes (years)                      822 non-null    float64
 6   Smokes (packs/year)                 822 non-null    float64
 7   Hormonal Contraceptives             732 non-null    float64
 8   Hormonal Contraceptives (years)     732 non-null    float64
 9   IUD                                 723 non-null    float64
 10  IUD (years)                         723 non-null    float64
 11  STDs                                735 non-null    float64
 12  STDs (number)                       735 non-null    float64
 13  STDs:condylomatosis                 735 non-null    float64
 14  STDs:cervical condylomatosis        735 non-null    float64
 15  STDs:vaginal condylomatosis         735 non-null    float64
 16  STDs:vulvo-perineal condylomatosis  735 non-null    float64
 17  STDs:syphilis                       735 non-null    float64
 18  STDs:pelvic inflammatory disease    735 non-null    float64
 19  STDs:genital herpes                 735 non-null    float64
 20  STDs:molluscum contagiosum          735 non-null    float64
 21  STDs:AIDS                           735 non-null    float64
 22  STDs:HIV                            735 non-null    float64
 23  STDs:Hepatitis B                    735 non-null    float64
 24  STDs:HPV                            735 non-null    float64
 25  STDs: Number of diagnosis           835 non-null    int64  
 26  STDs: Time since first diagnosis    71 non-null     float64
 27  STDs: Time since last diagnosis     71 non-null     float64
 28  Dx:Cancer                           835 non-null    int64  
 29  Dx:CIN                              835 non-null    int64  
 30  Dx:HPV                              835 non-null    int64  
 31  Dx                                  835 non-null    int64  
 32  Hinselmann                          835 non-null    int64  
 33  Schiller                            835 non-null    int64  
 34  Citology                            835 non-null    int64  
 35  Biopsy                              835 non-null    int64  
dtypes: float64(26), int64(10)
memory usage: 241.4 KB
```

```
na_counts = df.isnull().sum()
```

```
columns_with_na = na_counts[na_counts > 0].index.tolist()
```

```
columns_with_na
```

```
['Number of sexual partners',
 'First sexual intercourse',
 'Num of pregnancies',
 'Smokes',
 'Smokes (years)',
 'Smokes (packs/year)',
 'Hormonal Contraceptives',
 'Hormonal Contraceptives (years)',
 'IUD',
 'IUD (years)',
 'STDs',
 'STDs (number)',
 'STDs:condylomatosis',
 'STDs:cervical condylomatosis',
 'STDs:vaginal condylomatosis',
 'STDs:vulvo-perineal condylomatosis',
 'STDs:syphilis',
 'STDs:pelvic inflammatory disease',
 'STDs:genital herpes',
 'STDs:molluscum contagiosum',
 'STDs:AIDS',
 'STDs:HIV',
 'STDs:Hepatitis B',
 'STDs:HPV',
 'STDs: Time since first diagnosis',
 'STDs: Time since last diagnosis']
```

```
def fill_missing_values(df, columns):
    for col in columns:
        df[col] = df[col].fillna(df[col].median())
    return df
```

```
cc_df = fill_missing_values(df, columns_with_na)
```

```
cc_df.isnull().sum()
```

```
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
dtype: int64
```

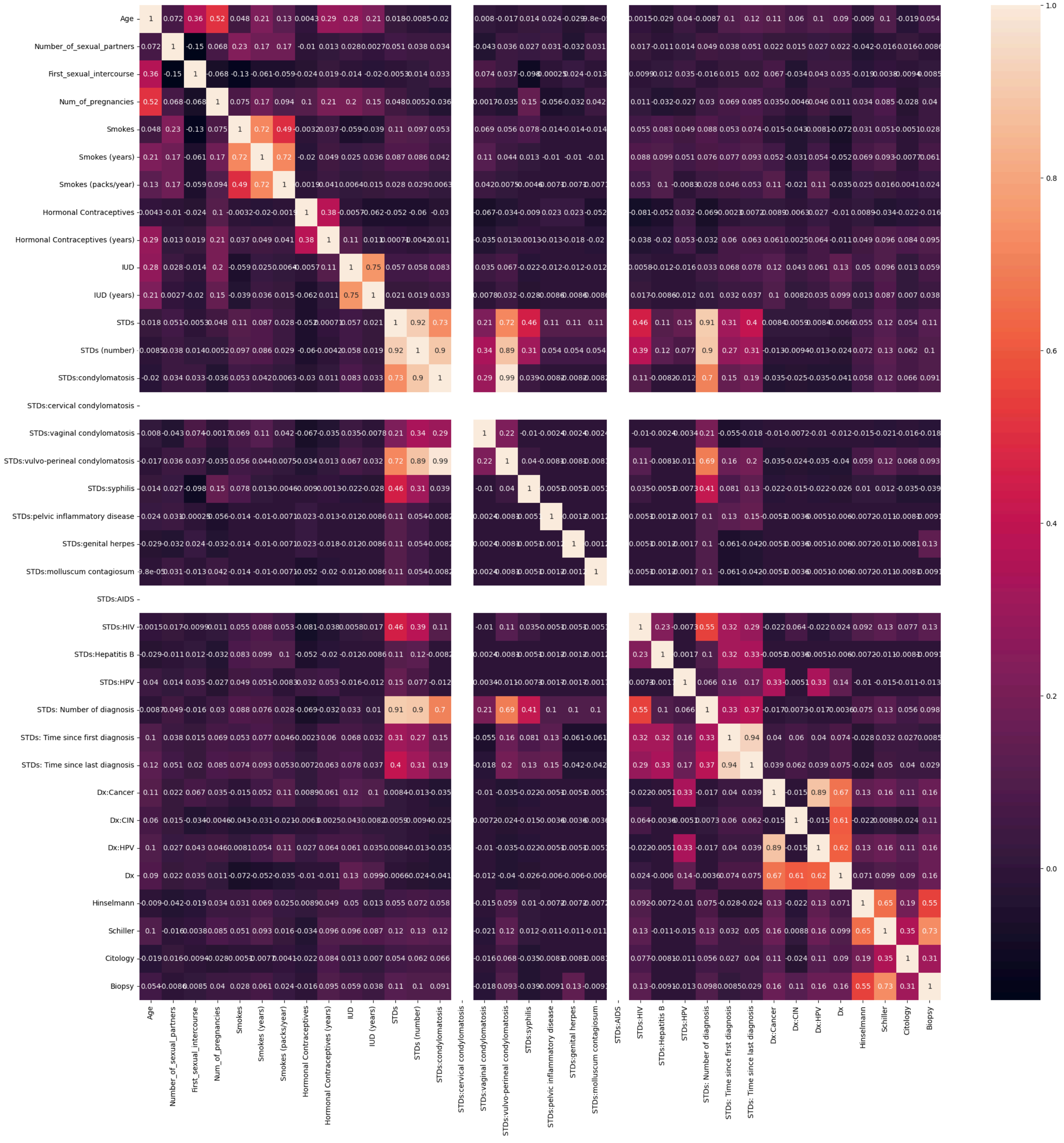
```
cc_df.rename(columns={'Number of sexual partners': 'Number_of_sexual_partners',
                      'First sexual intercourse': 'First_sexual_intercourse',
                      'Num of pregnancies': 'Num_of_pregnancies'},inplace=True)
```

```
bio_df = cc_df.copy()
```

```
%matplotlib inline
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(25, 25))
sns.heatmap(cc_df.corr(), annot=True)
```

```
<Axes: >
```


Age
Number_of_sexual_partners
First_sexual_intercourse
Num_of_pregnancies
Smokes
Smokes (years)
Smokes (packs/year)
Hormonal Contraceptives
Hormonal Contraceptives (years)
IUD
IUD (years)
STDs
STDs (number)
STDs:condylomatosis
STDs:cervical condylomatosis
STDs:vaginal condylomatosis
STDs:vulvo-perineal condylomatosis
STDs:syphilis
STDs:pelvic inflammatory disease
STDs:genital herpes
STDs:molluscum contagiosum
STDs:AIDS
STDs:HIV
STDs:Hepatitis B
STDs:HPV
STDs: Number of diagnosis
STDs: Time since first diagnosis
STDs: Time since last diagnosis
Dx:Cancer
Dx:CIN
Dx:HPV
Dx
Hinselmann
Schiller
Citology
Biopsy
1.0
0.8
0.6
0.4
0.2
0.0

```
cc_df.shape
```

```
(835, 36)
```

```
cc_df.dtypes
```

```
Age                                    int64
Number_of_sexual_partners            float64
First_sexual_intercourse             float64
Num_of_pregnancies                   float64
Smokes                               float64
Smokes (years)                       float64
Smokes (packs/year)                  float64
Hormonal Contraceptives              float64
Hormonal Contraceptives (years)      float64
IUD                                  float64
IUD (years)                          float64
STDs                                 float64
STDs (number)                        float64
STDs:condylomatosis                  float64
STDs:cervical condylomatosis         float64
STDs:vaginal condylomatosis          float64
STDs:vulvo-perineal condylomatosis   float64
STDs:syphilis                        float64
STDs:pelvic inflammatory disease     float64
STDs:genital herpes                  float64
STDs:molluscum contagiosum           float64
STDs:AIDS                            float64
STDs:HIV                             float64
STDs:Hepatitis B                     float64
STDs:HPV                             float64
STDs: Number of diagnosis              int64
STDs: Time since first diagnosis     float64
STDs: Time since last diagnosis      float64
Dx:Cancer                              int64
Dx:CIN                                 int64
Dx:HPV                                 int64
Dx                                     int64
Hinselmann                             int64
Schiller                               int64
Citology                               int64
Biopsy                                 int64
dtype: object
```

```
numerical = [var for var in cc_df.columns if cc_df[var].dtype!='O']
print(numerical)
```

```
['Age', 'Number_of_sexual_partners', 'First_sexual_intercourse', 'Num_of_pregnancies', 'Smokes', 'Smokes (years)', 'Smokes (packs/year)'
```

Outliers in our dataset

```
print(round(cc_df[numerical].describe()),2)
```

```
std      8.0                        2.0                       3.0
min     13.0                        1.0                      10.0
25%     21.0                        2.0                      15.0
50%     26.0                        2.0                      17.0
75%     32.0                        3.0                      18.0
max     84.0                       28.0                      32.0

       Num_of_pregnancies  Smokes  Smokes (years)  Smokes (packs/year)  \
count               835.0   835.0           835.0                835.0
```

```
25%                1.0    0.0           0.0               0.0
50%                2.0    0.0           0.0               0.0
75%                3.0    0.0           0.0               0.0
max               11.0    1.0          37.0              37.0

       Hormonal Contraceptives  Hormonal Contraceptives (years)    IUD  ...  \
count                    835.0                            835.0  835.0  ...
mean                       1.0                              2.0    0.0  ...
std                        0.0                              4.0    0.0  ...
min                        0.0                              0.0    0.0  ...
25%                        0.0                              0.0    0.0  ...
50%                        1.0                              0.0    0.0  ...
75%                        1.0                              3.0    0.0  ...
max                        1.0                             30.0    1.0  ...

       STDs: Time since first diagnosis  STDs: Time since last diagnosis  \
count                             835.0                            835.0
mean                                4.0                              3.0
std                                 2.0                              2.0
min                                 1.0                              1.0
25%                                 4.0                              3.0
50%                                 4.0                              3.0
75%                                 4.0                              3.0
max                                22.0                             22.0

       Dx:Cancer  Dx:CIN  Dx:HPV     Dx  Hinselmann  Schiller  Citology  \
count      835.0   835.0   835.0  835.0       835.0     835.0     835.0
mean         0.0     0.0     0.0    0.0         0.0       0.0       0.0
std          0.0     0.0     0.0    0.0         0.0       0.0       0.0
min          0.0     0.0     0.0    0.0         0.0       0.0       0.0
25%          0.0     0.0     0.0    0.0         0.0       0.0       0.0
50%          0.0     0.0     0.0    0.0         0.0       0.0       0.0
75%          0.0     0.0     0.0    0.0         0.0       0.0       0.0
max          1.0     1.0     1.0    1.0         1.0       1.0       1.0

       Biopsy
count   835.0
mean      0.0
std       0.0
min       0.0
25%       0.0
50%       0.0
75%       0.0
max       1.0

[8 rows x 36 columns] 2
```

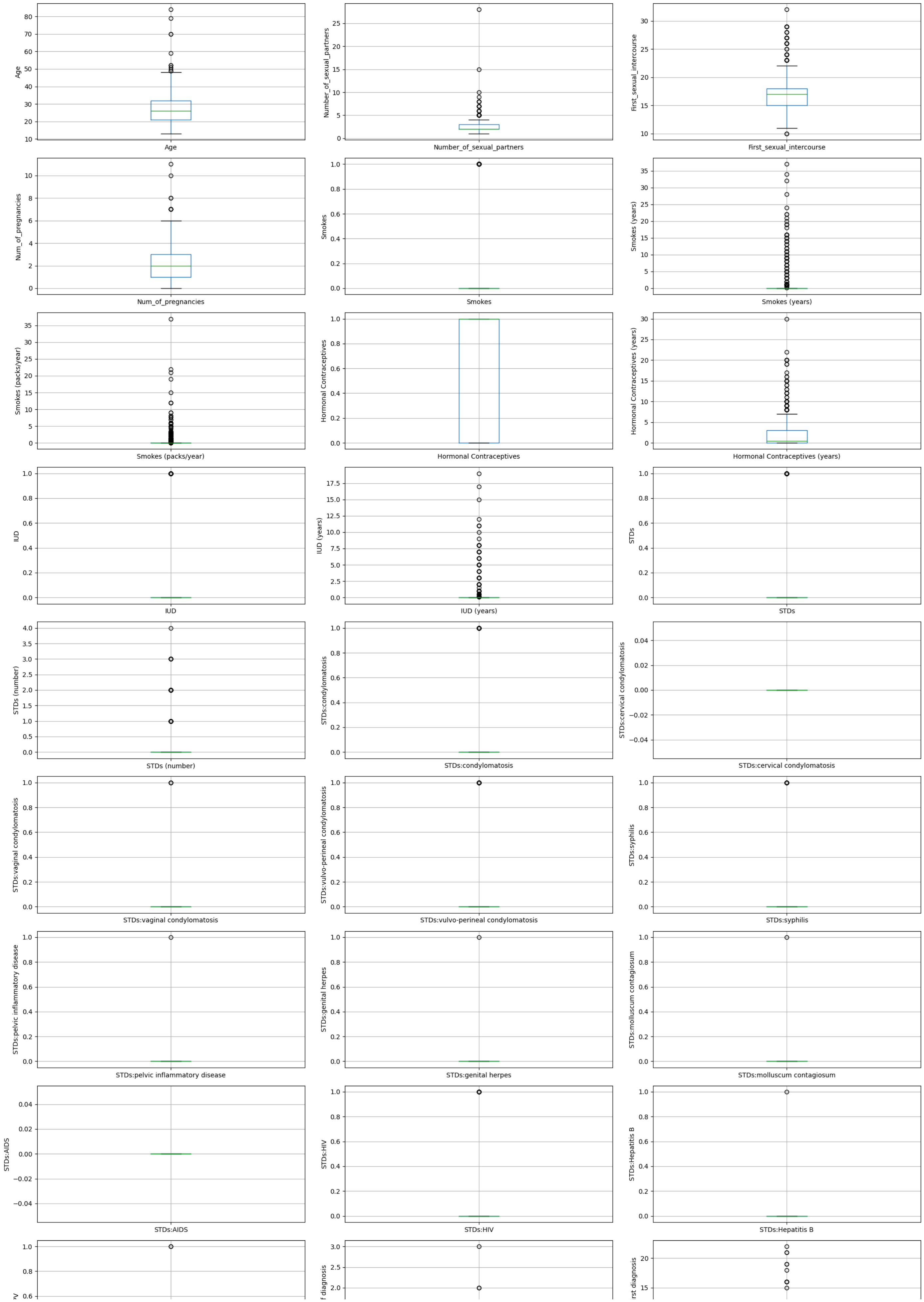
Subplotting box plots to select columns

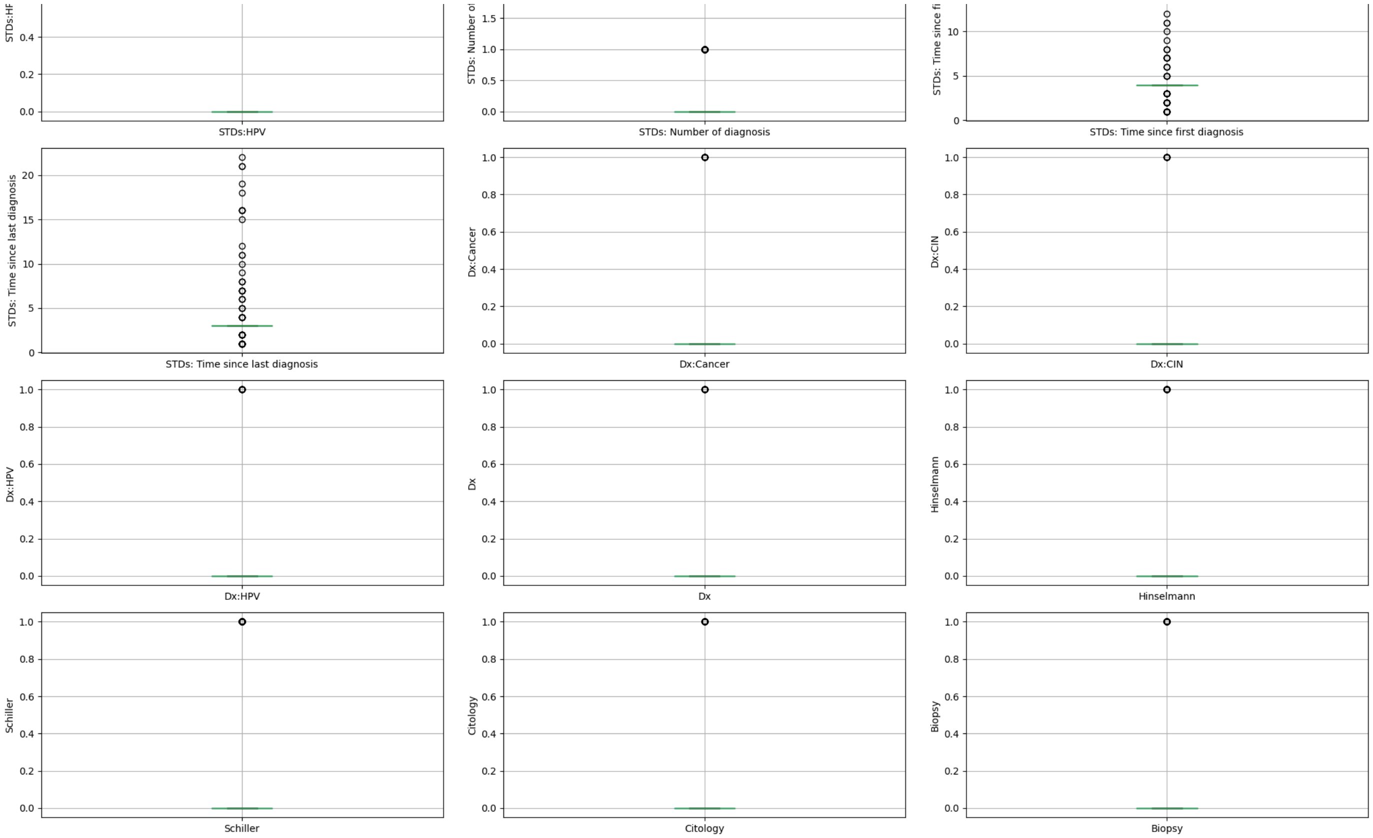
```
num_rows = 12
num_cols = 3
total_plots = num_rows * num_cols

# Create a new figure
plt.figure(figsize=(20, 40))

# Iterate through each column and create boxplots
for i, column in enumerate(cc_df.columns[:total_plots], 1):
    plt.subplot(num_rows, num_cols, i)
    fig = cc_df.boxplot(column=column)
    fig.set_title('')
    fig.set_ylabel(column)

# Adjust layout
plt.tight_layout()
```


Age
10 20 30 40 50 60 70 80
Number_of_sexual_partners
0 5 10 15 20 25
First_sexual_intercourse
10 15 20 25 30
Num_of_pregnancies
0 2 4 6 8 10
Smokes
0.0 0.2 0.4 0.6 0.8 1.0
Smokes (years)
0 5 10 15 20 25 30 35
Smokes (packs/year)
0 5 10 15 20 25 30 35
Hormonal Contraceptives
0.0 0.2 0.4 0.6 0.8 1.0
Hormonal Contraceptives (years)
0 5 10 15 20 25 30
IUD
0.0 0.2 0.4 0.6 0.8 1.0
IUD (years)
0.0 2.5 5.0 7.5 10.0 12.5 15.0 17.5
STDs
0.0 0.2 0.4 0.6 0.8 1.0
STDs (number)
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
STDs:condylomatosis
0.0 0.2 0.4 0.6 0.8 1.0
STDs:cervical condylomatosis
−0.04 −0.02 0.00 0.02 0.04
STDs:vaginal condylomatosis
0.0 0.2 0.4 0.6 0.8 1.0
STDs:vulvo-perineal condylomatosis
0.0 0.2 0.4 0.6 0.8 1.0
STDs:syphilis
0.0 0.2 0.4 0.6 0.8 1.0
STDs:pelvic inflammatory disease
0.0 0.2 0.4 0.6 0.8 1.0
STDs:genital herpes
0.0 0.2 0.4 0.6 0.8 1.0
STDs:molluscum contagiosum
0.0 0.2 0.4 0.6 0.8 1.0
STDs:AIDS
−0.04 −0.02 0.00 0.02 0.04
STDs:HIV
0.0 0.2 0.4 0.6 0.8 1.0
STDs:Hepatitis B
0.0 0.2 0.4 0.6 0.8 1.0
V
0.6 0.8 1.0
f diagnosis
2.0 2.5 3.0
rst diagnosis
15 20


STDs:HF
0.4
0.2
0.0
STDs:HPV
STDs: Number of
1.5
1.0
0.5
0.0
STDs: Number of diagnosis
STDs: Time since fi
10
5
0
STDs: Time since first diagnosis
STDs: Time since last diagnosis
20
15
10
5
0
STDs: Time since last diagnosis
Dx:Cancer
Dx:CIN
Dx:HPV
Dx
Hinselmann
Schiller
Citology
Biopsy
1.0
0.8
0.6
0.4
0.2
0.0

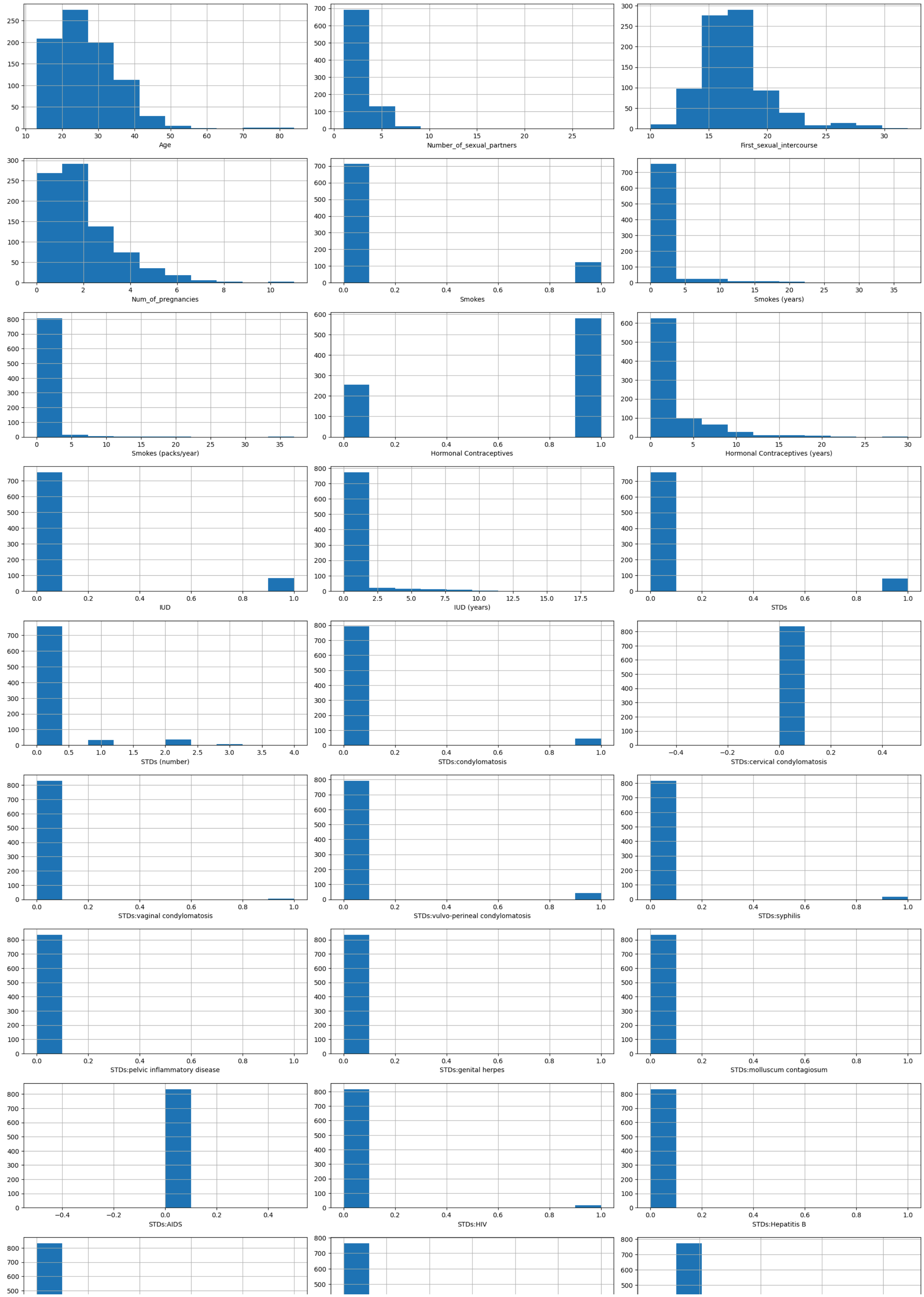
Subplotting histograms to check skewed distributions

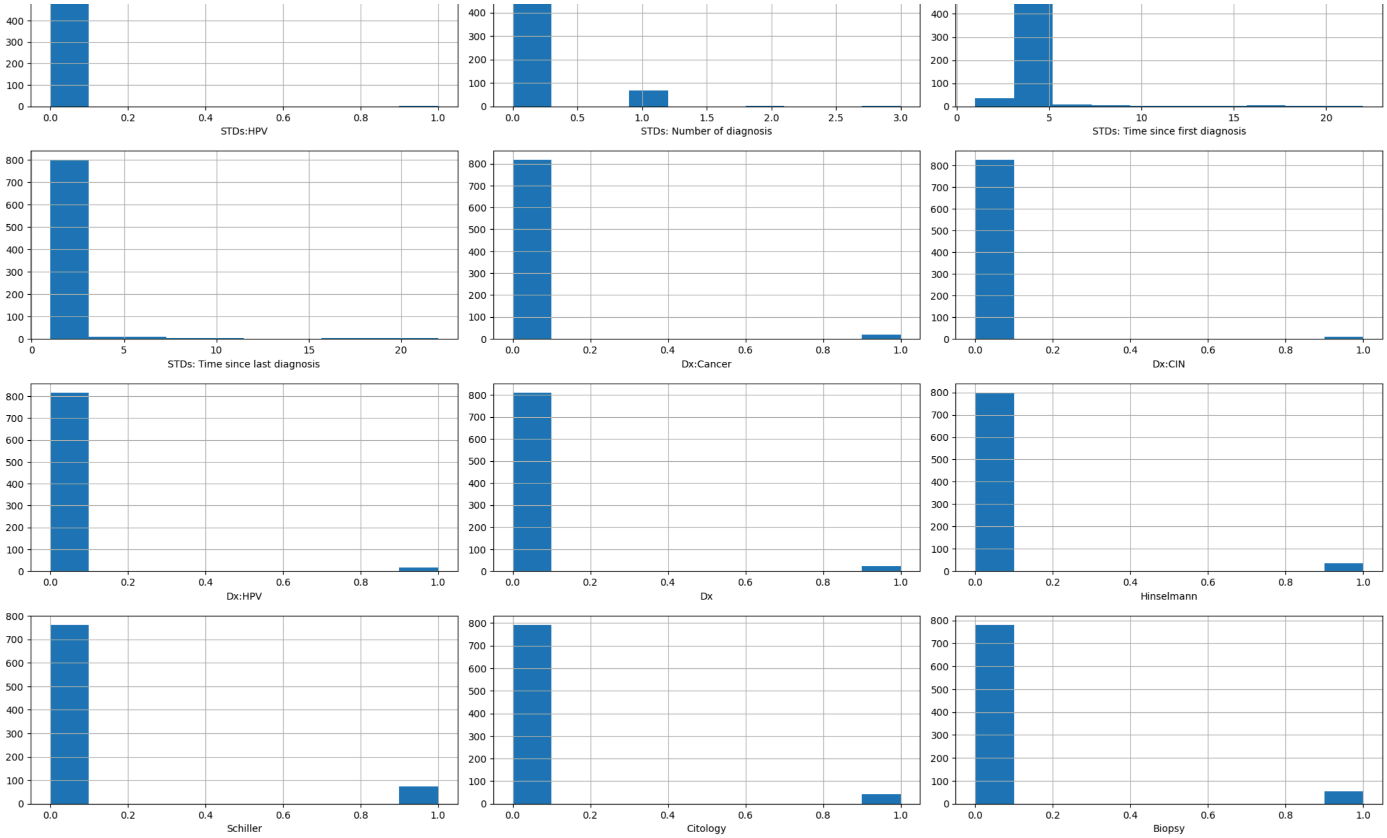
```
num_rows = 12
num_cols = 3
total_plots = num_rows * num_cols

# Create a new figure
plt.figure(figsize=(20, 40))

# Iterate through each column and create boxplots
for i, column in enumerate(cc_df.columns[:total_plots], 1):
    plt.subplot(num_rows, num_cols, i)
    cc_df[column].hist()
    plt.xlabel(column)
    plt.ylabel('')

# Adjust layout
plt.tight_layout()
```


Age
Number_of_sexual_partners
First_sexual_intercourse
Num_of_pregnancies
Smokes
Smokes (years)
Smokes (packs/year)
Hormonal Contraceptives
Hormonal Contraceptives (years)
IUD
IUD (years)
STDs
STDs (number)
STDs:condylomatosis
STDs:cervical condylomatosis
STDs:vaginal condylomatosis
STDs:vulvo-perineal condylomatosis
STDs:syphilis
STDs:pelvic inflammatory disease
STDs:genital herpes
STDs:molluscum contagiosum
STDs:AIDS
STDs:HIV
STDs:Hepatitis B


STDs:HPV
STDs: Number of diagnosis
STDs: Time since first diagnosis
STDs: Time since last diagnosis
Dx:Cancer
Dx:CIN
Dx:HPV
Dx
Hinselmann
Schiller
Citology
Biopsy
0
100
200
300
400
500
600
700
800
0.0
0.2
0.4
0.6
0.8
1.0
0.5
1.5
2.0
2.5
3.0
5
10
15
20

```
IQR = cc_df['Age'].quantile(0.75) - cc_df['Age'].quantile(0.25)
Lower_fence = cc_df['Age'].quantile(0.25) - (IQR * 1.5)
Upper_fence = cc_df['Age'].quantile(0.75) + (IQR * 1.5)
print(f"Age outliers are values < {Lower_fence}  or > {Upper_fence}")
```

```
Age outliers are values < 4.5  or > 48.5
```

```
IQR = cc_df['Number_of_sexual_partners'].quantile(0.75) - cc_df['Number_of_sexual_partners'].quantile(0.25)
Lower_fence = cc_df['Number_of_sexual_partners'].quantile(0.25) - (IQR * 1.5)
Upper_fence = cc_df['Number_of_sexual_partners'].quantile(0.75) + (IQR * 1.5)
print(f"Number_of_sexual_partners outliers are values < {Lower_fence}  or > {Upper_fence}")
```

```
Number_of_sexual_partners outliers are values < 0.5  or > 4.5
```

```
IQR = cc_df['First_sexual_intercourse'].quantile(0.75) - cc_df['First_sexual_intercourse'].quantile(0.25)
Lower_fence = cc_df['First_sexual_intercourse'].quantile(0.25) - (IQR * 1.5)
Upper_fence = cc_df['First_sexual_intercourse'].quantile(0.75) + (IQR * 1.5)
print(f"First_sexual_intercourse outliers are values < {Lower_fence}  or > {Upper_fence}")
```

```
First_sexual_intercourse outliers are values < 10.5  or > 22.5
```

```
IQR = cc_df['Num_of_pregnancies'].quantile(0.75) - cc_df['Num_of_pregnancies'].quantile(0.25)
Lower_fence = cc_df['Num_of_pregnancies'].quantile(0.25) - (IQR * 1.5)
Upper_fence = cc_df['Num_of_pregnancies'].quantile(0.75) + (IQR * 1.5)
print(f"Num_of_pregnancies outliers are values < {Lower_fence}  or > {Upper_fence}")
```

```
Num_of_pregnancies outliers are values < -2.0  or > 6.0
```