Hands-on Activity 11.2 Classification using Logistic Regression

Name: Xander Sam E. Galapia

Section: CPE22S3

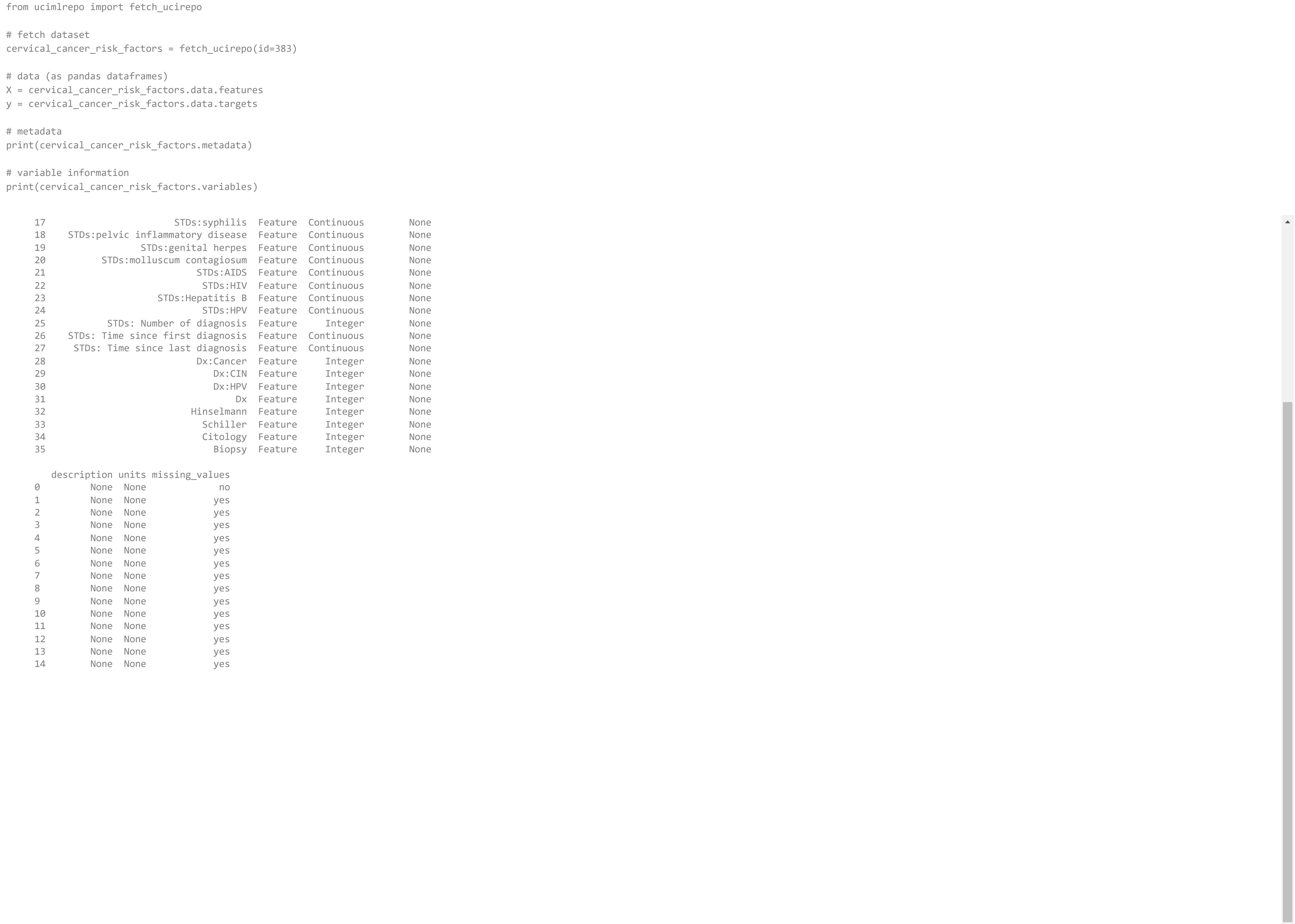
```
pip install ucimlrepo
```

```
     Requirement already satisfied: ucimlrepo in /usr/local/lib/python3.10/dist-packages (0.0.6)
```

```
!pip install hvplot
```

```
     Requirement already satisfied: hvplot in /usr/local/lib/python3.10/dist-packages (0.9.2)
     Requirement already satisfied: bokeh>=1.0.0 in /usr/local/lib/python3.10/dist-packages (from hvplot) (3.3.4)
     Requirement already satisfied: colorcet>=2 in /usr/local/lib/python3.10/dist-packages (from hvplot) (3.1.0)
     Requirement already satisfied: holoviews>=1.11.0 in /usr/local/lib/python3.10/dist-packages (from hvplot) (1.17.1)
     Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-packages (from hvplot) (2.0.3)
     Requirement already satisfied: numpy>=1.15 in /usr/local/lib/python3.10/dist-packages (from hvplot) (1.25.2)
     Requirement already satisfied: packaging in /usr/local/lib/python3.10/dist-packages (from hvplot) (24.0)
     Requirement already satisfied: panel>=0.11.0 in /usr/local/lib/python3.10/dist-packages (from hvplot) (1.3.8)
     Requirement already satisfied: param<3.0,>=1.12.0 in /usr/local/lib/python3.10/dist-packages (from hvplot) (2.1.0)
     Requirement already satisfied: Jinja2>=2.9 in /usr/local/lib/python3.10/dist-packages (from bokeh>=1.0.0->hvplot) (3.1.3)
     Requirement already satisfied: contourpy>=1 in /usr/local/lib/python3.10/dist-packages (from bokeh>=1.0.0->hvplot) (1.2.1)
     Requirement already satisfied: pillow>=7.1.0 in /usr/local/lib/python3.10/dist-packages (from bokeh>=1.0.0->hvplot) (9.4.0)
     Requirement already satisfied: PyYAML>=3.10 in /usr/local/lib/python3.10/dist-packages (from bokeh>=1.0.0->hvplot) (6.0.1)
     Requirement already satisfied: tornado>=5.1 in /usr/local/lib/python3.10/dist-packages (from bokeh>=1.0.0->hvplot) (6.3.3)
     Requirement already satisfied: xyzservices>=2021.09.1 in /usr/local/lib/python3.10/dist-packages (from bokeh>=1.0.0->hvplot) (2024.4.0)
     Requirement already satisfied: pyviz-comms>=0.7.4 in /usr/local/lib/python3.10/dist-packages (from holoviews>=1.11.0->hvplot) (3.0.2)
     Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.10/dist-packages (from pandas->hvplot) (2.8.2)
     Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas->hvplot) (2023.4)
     Requirement already satisfied: tzdata>=2022.1 in /usr/local/lib/python3.10/dist-packages (from pandas->hvplot) (2024.1)
     Requirement already satisfied: markdown in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->hvplot) (3.6)
     Requirement already satisfied: markdown-it-py in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->hvplot) (3.0.0)
     Requirement already satisfied: linkify-it-py in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->hvplot) (2.0.3)
     Requirement already satisfied: mdit-py-plugins in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->hvplot) (0.4.0)
     Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->hvplot) (2.31.0)
     Requirement already satisfied: tqdm>=4.48.0 in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->hvplot) (4.66.2)
     Requirement already satisfied: bleach in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->hvplot) (6.1.0)
     Requirement already satisfied: typing-extensions in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->hvplot) (4.11.0)
     Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.10/dist-packages (from Jinja2>=2.9->bokeh>=1.0.0->hvplot) (2.1.5)
     Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.8.2->pandas->hvplot) (1.16.0)
     Requirement already satisfied: webencodings in /usr/local/lib/python3.10/dist-packages (from bleach->panel>=0.11.0->hvplot) (0.5.1)
     Requirement already satisfied: uc-micro-py in /usr/local/lib/python3.10/dist-packages (from linkify-it-py->panel>=0.11.0->hvplot) (1.0.3)
     Requirement already satisfied: mdurl~=0.1 in /usr/local/lib/python3.10/dist-packages (from markdown-it-py->panel>=0.11.0->hvplot) (0.1.2)
     Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests->panel>=0.11.0->hvplot) (3.3.2)
     Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests->panel>=0.11.0->hvplot) (3.7)
     Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests->panel>=0.11.0->hvplot) (2.0.7)
     Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests->panel>=0.11.0->hvplot) (2024.2.2)
```

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
%matplotlib inline

from sklearn.model_selection import train_test_split
from sklearn import metrics
from sklearn.linear_model import LinearRegression
```



```
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)
```

```
    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
```

```
    34        None  None              no
    35        None  None              no
```

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

	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormonal Contraceptives	Hormonal Contraceptives (years)	IUD	...	STDs: Time since first diagnosis	STDs: Time since last diagnosis	Dx:Cancer	Dx:CIN	Dx:HPV	Dx
0	18	4.0	15.0	1.0	0.0	0.0	0.0	0.0	0.00	0.0	...	NaN	NaN	0	0	0	(
1	15	1.0	14.0	1.0	0.0	0.0	0.0	0.0	0.00	0.0	...	NaN	NaN	0	0	0	(
2	34	1.0	NaN	1.0	0.0	0.0	0.0	0.0	0.00	0.0	...	NaN	NaN	0	0	0	(
3	52	5.0	16.0	4.0	1.0	37.0	37.0	1.0	3.00	0.0	...	NaN	NaN	1	0	1	(
4	46	3.0	21.0	4.0	0.0	0.0	0.0	1.0	15.00	0.0	...	NaN	NaN	0	0	0	(
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	..
853	34	3.0	18.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	...	NaN	NaN	0	0	0	(
854	32	2.0	19.0	1.0	0.0	0.0	0.0	1.0	8.00	0.0	...	NaN	NaN	0	0	0	(
855	25	2.0	17.0	0.0	0.0	0.0	0.0	1.0	0.08	0.0	...	NaN	NaN	0	0	0	(
856	33	2.0	24.0	2.0	0.0	0.0	0.0	1.0	0.08	0.0	...	NaN	NaN	0	0	0	(
857	29	2.0	20.0	1.0	0.0	0.0	0.0	1.0	0.50	0.0	...	NaN	NaN	0	0	0	(

```
CC.isna().sum()
```

```
    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
```

```
no_val = []
for x in CC.columns:
  if CC[x].isnull().any():
    no_val.append(x)
no_val
```

```
     ['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']
```

```
for x in no_val:
  CC[x].fillna(CC[x].mode()[0], inplace=True)
```

```
CC.isna().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
```

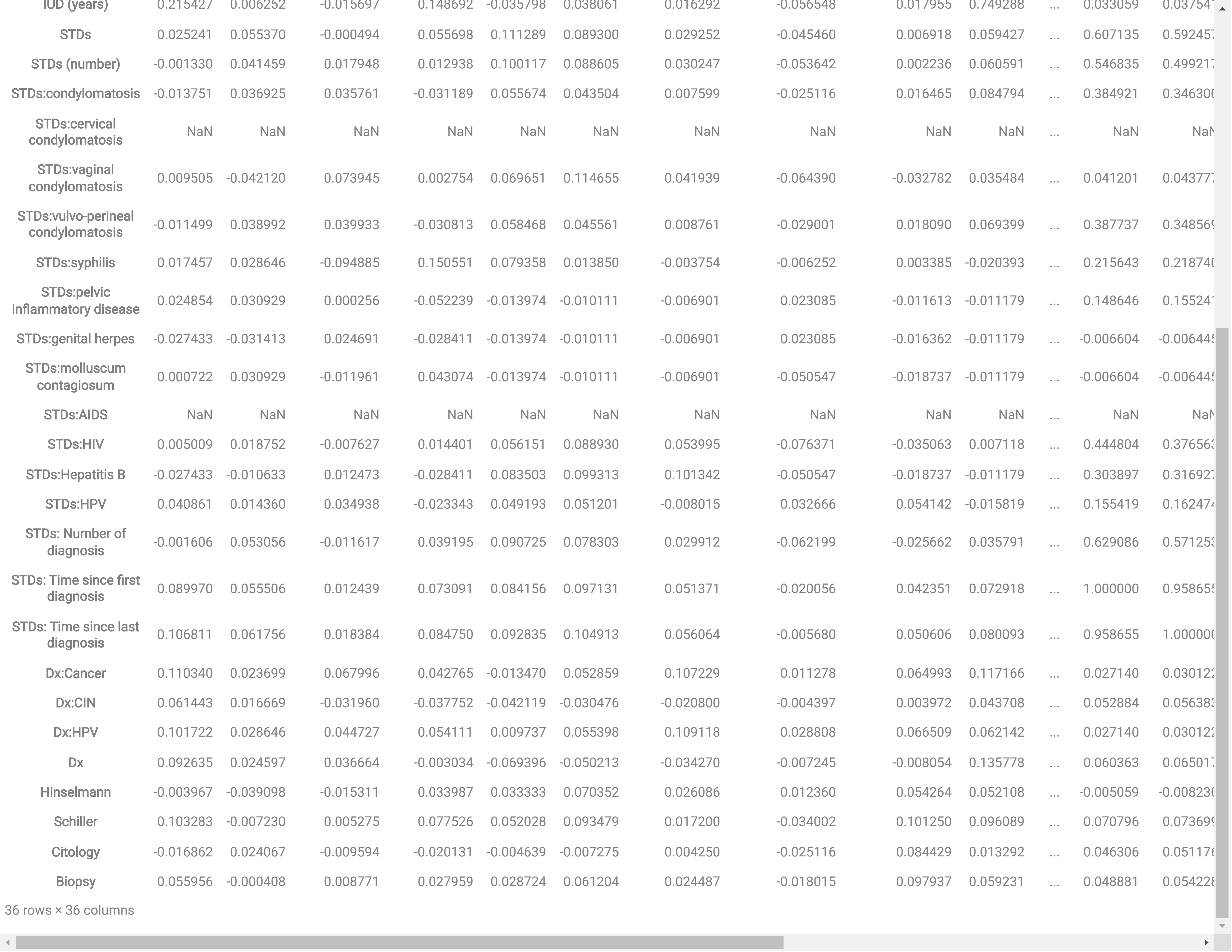
```
Dx:HPV                                0
Dx                                    0
Hinselmann                            0
Schiller                              0
Citology                              0
Biopsy                                0
dtype: int64
```

```
CC.describe()
```

	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormonal Contraceptives	Hormonal Contraceptives (years)	IUD	...	STDs: Time since first diagnosis	STDs: Time since last diagnosis	Dx:
count	858.000000	858.000000	858.000000	858.000000	858.000000	858.000000	858.000000	858.000000	858.000000	858.000000	...	858.000000	858.000000	858.
mean	26.820513	2.511655	16.979021	2.192308	0.143357	1.201241	0.446278	0.686480	1.972394	0.096737	...	1.425408	1.398601	0.
std	8.497948	1.644759	2.797653	1.434395	0.350641	4.060623	2.210351	0.464194	3.597888	0.295771	...	2.201550	2.113931	0.
min	13.000000	1.000000	10.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	1.000000	1.000000	0.
25%	20.000000	2.000000	15.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	1.000000	1.000000	0.
50%	25.000000	2.000000	17.000000	2.000000	0.000000	0.000000	0.000000	1.000000	0.250000	0.000000	...	1.000000	1.000000	0.
75%	32.000000	3.000000	18.000000	3.000000	0.000000	0.000000	0.000000	1.000000	2.000000	0.000000	...	1.000000	1.000000	0.
max	84.000000	28.000000	32.000000	11.000000	1.000000	37.000000	37.000000	1.000000	30.000000	1.000000	...	22.000000	22.000000	1.

8 rows × 36 columns

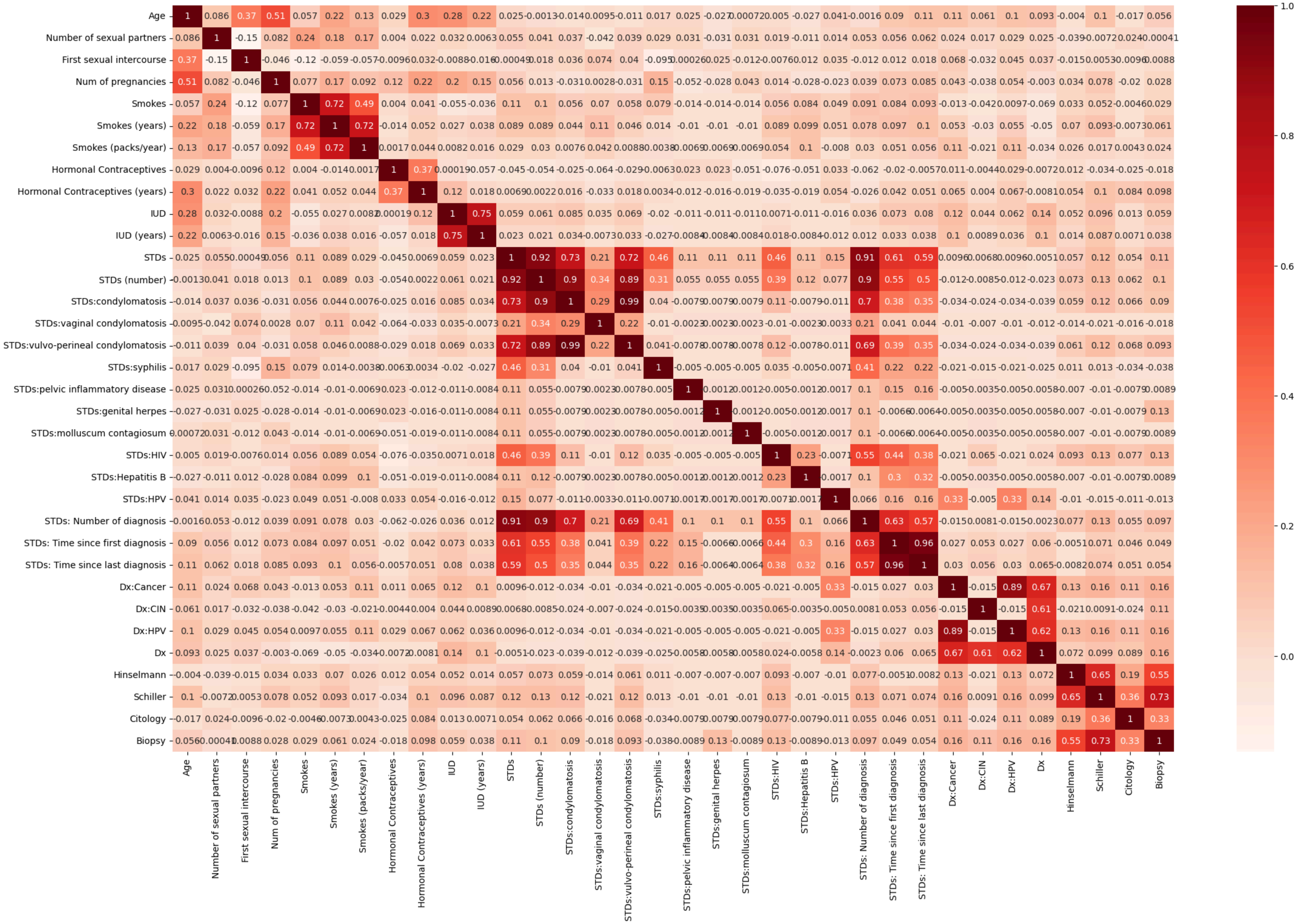
```
CC.corr()
```



IUD (years)	0.215427	0.006252	-0.015697	0.148692	-0.035798	0.038061	0.016292	-0.056548	0.017955	0.749288	...	0.033059	0.03754
STDs	0.025241	0.055370	-0.000494	0.055698	0.111289	0.089300	0.029252	-0.045460	0.006918	0.059427	...	0.607135	0.59245
STDs (number)	-0.001330	0.041459	0.017948	0.012938	0.100117	0.088605	0.030247	-0.053642	0.002236	0.060591	...	0.546835	0.49921
STDs:condylomatosis	-0.013751	0.036925	0.035761	-0.031189	0.055674	0.043504	0.007599	-0.025116	0.016465	0.084794	...	0.384921	0.34630
STDs:cervical condylomatosis	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	Na
STDs:vaginal condylomatosis	0.009505	-0.042120	0.073945	0.002754	0.069651	0.114655	0.041939	-0.064390	-0.032782	0.035484	...	0.041201	0.04377
STDs:vulvo-perineal condylomatosis	-0.011499	0.038992	0.039933	-0.030813	0.058468	0.045561	0.008761	-0.029001	0.018090	0.069399	...	0.387737	0.34856
STDs:syphilis	0.017457	0.028646	-0.094885	0.150551	0.079358	0.013850	-0.003754	-0.006252	0.003385	-0.020393	...	0.215643	0.21874
STDs:pelvic inflammatory disease	0.024854	0.030929	0.000256	-0.052239	-0.013974	-0.010111	-0.006901	0.023085	-0.011613	-0.011179	...	0.148646	0.15524
STDs:genital herpes	-0.027433	-0.031413	0.024691	-0.028411	-0.013974	-0.010111	-0.006901	0.023085	-0.016362	-0.011179	...	-0.006604	-0.00644
STDs:molluscum contagiosum	0.000722	0.030929	-0.011961	0.043074	-0.013974	-0.010111	-0.006901	-0.050547	-0.018737	-0.011179	...	-0.006604	-0.00644
STDs:AIDS	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	Na
STDs:HIV	0.005009	0.018752	-0.007627	0.014401	0.056151	0.088930	0.053995	-0.076371	-0.035063	0.007118	...	0.444804	0.37656
STDs:Hepatitis B	-0.027433	-0.010633	0.012473	-0.028411	0.083503	0.099313	0.101342	-0.050547	-0.018737	-0.011179	...	0.303897	0.31692
STDs:HPV	0.040861	0.014360	0.034938	-0.023343	0.049193	0.051201	-0.008015	0.032666	0.054142	-0.015819	...	0.155419	0.16247
STDs: Number of diagnosis	-0.001606	0.053056	-0.011617	0.039195	0.090725	0.078303	0.029912	-0.062199	-0.025662	0.035791	...	0.629086	0.57125
STDs: Time since first diagnosis	0.089970	0.055506	0.012439	0.073091	0.084156	0.097131	0.051371	-0.020056	0.042351	0.072918	...	1.000000	0.95865
STDs: Time since last diagnosis	0.106811	0.061756	0.018384	0.084750	0.092835	0.104913	0.056064	-0.005680	0.050606	0.080093	...	0.958655	1.00000
Dx:Cancer	0.110340	0.023699	0.067996	0.042765	-0.013470	0.052859	0.107229	0.011278	0.064993	0.117166	...	0.027140	0.03012
Dx:CIN	0.061443	0.016669	-0.031960	-0.037752	-0.042119	-0.030476	-0.020800	-0.004397	0.003972	0.043708	...	0.052884	0.05638
Dx:HPV	0.101722	0.028646	0.044727	0.054111	0.009737	0.055398	0.109118	0.028808	0.066509	0.062142	...	0.027140	0.03012
Dx	0.092635	0.024597	0.036664	-0.003034	-0.069396	-0.050213	-0.034270	-0.007245	-0.008054	0.135778	...	0.060363	0.06501
Hinselmann	-0.003967	-0.039098	-0.015311	0.033987	0.033333	0.070352	0.026086	0.012360	0.054264	0.052108	...	-0.005059	-0.00823
Schiller	0.103283	-0.007230	0.005275	0.077526	0.052028	0.093479	0.017200	-0.034002	0.101250	0.096089	...	0.070796	0.07369
Citology	-0.016862	0.024067	-0.009594	-0.020131	-0.004639	-0.007275	0.004250	-0.025116	0.084429	0.013292	...	0.046306	0.05117
Biopsy	0.055956	-0.000408	0.008771	0.027959	0.028724	0.061204	0.024487	-0.018015	0.097937	0.059231	...	0.048881	0.05422

36 rows × 36 columns

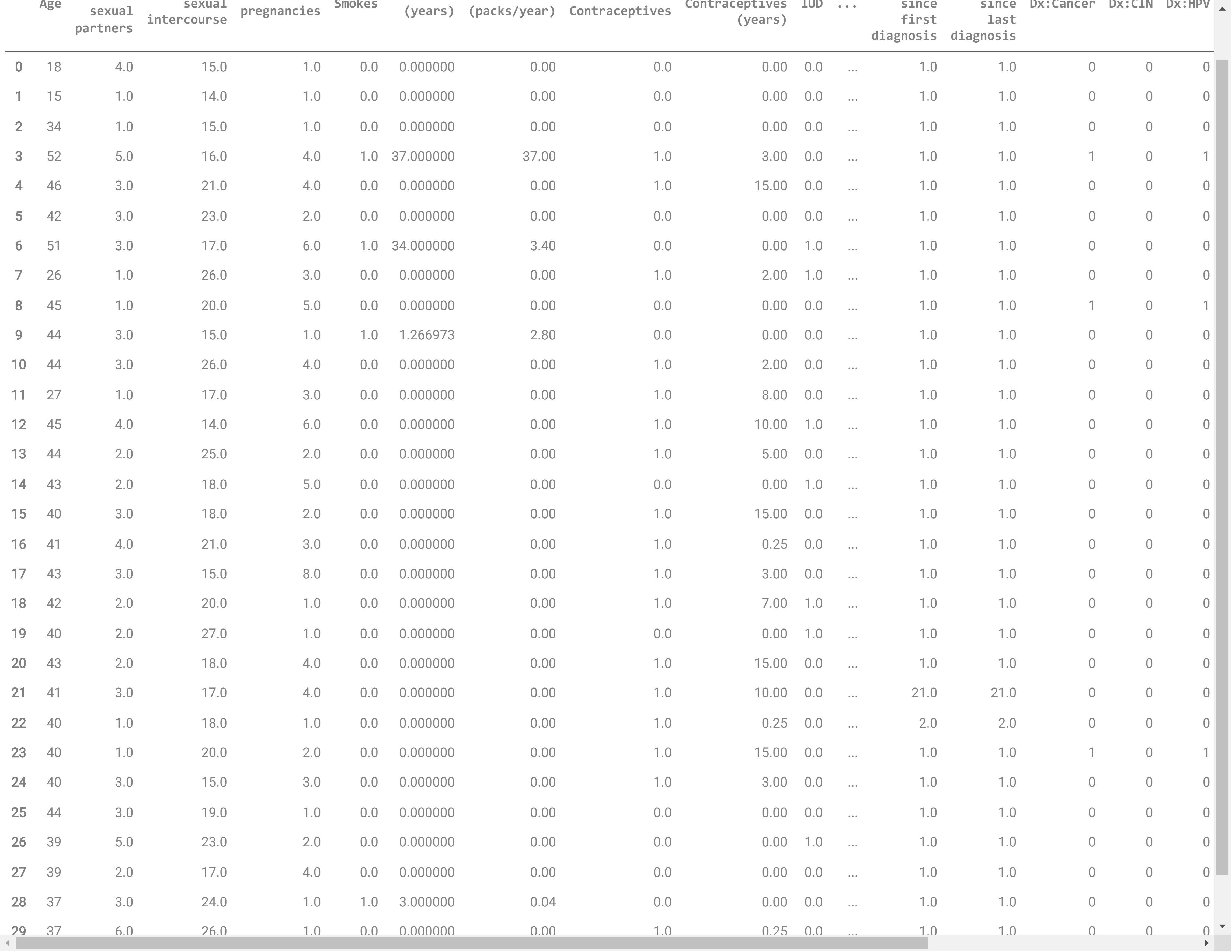
```
CC = CC.drop(['STDs:AIDS',
              'STDs:cervical condylomatosis'], axis = 1)


plt.figure(figsize =(25,15))
ax = sns.heatmap(CC.corr(), annot = True, cmap = 'Reds')
```


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: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
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.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: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: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
```

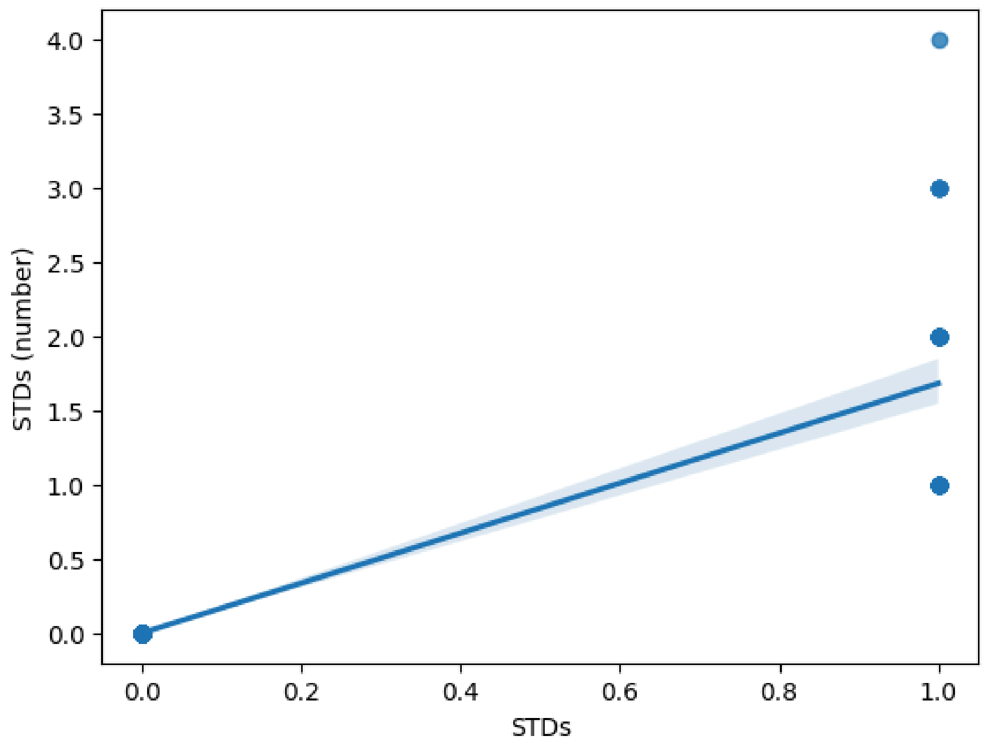
```
CC.head(30)
```



	Age	sexual partners	sexual intercourse	pregnancies	Smokes	(years)	(packs/year)	Contraceptives	Contraceptives (years)	IUD	...	since first diagnosis	since last diagnosis	Dx:Cancer	Dx:CIN	Dx:HPV
0	18	4.0	15.0	1.0	0.0	0.000000	0.00	0.0	0.00	0.0	...	1.0	1.0	0	0	0
1	15	1.0	14.0	1.0	0.0	0.000000	0.00	0.0	0.00	0.0	...	1.0	1.0	0	0	0
2	34	1.0	15.0	1.0	0.0	0.000000	0.00	0.0	0.00	0.0	...	1.0	1.0	0	0	0
3	52	5.0	16.0	4.0	1.0	37.000000	37.00	1.0	3.00	0.0	...	1.0	1.0	1	0	1
4	46	3.0	21.0	4.0	0.0	0.000000	0.00	1.0	15.00	0.0	...	1.0	1.0	0	0	0
5	42	3.0	23.0	2.0	0.0	0.000000	0.00	0.0	0.00	0.0	...	1.0	1.0	0	0	0
6	51	3.0	17.0	6.0	1.0	34.000000	3.40	0.0	0.00	1.0	...	1.0	1.0	0	0	0
7	26	1.0	26.0	3.0	0.0	0.000000	0.00	1.0	2.00	1.0	...	1.0	1.0	0	0	0
8	45	1.0	20.0	5.0	0.0	0.000000	0.00	0.0	0.00	0.0	...	1.0	1.0	1	0	1
9	44	3.0	15.0	1.0	1.0	1.266973	2.80	0.0	0.00	0.0	...	1.0	1.0	0	0	0
10	44	3.0	26.0	4.0	0.0	0.000000	0.00	1.0	2.00	0.0	...	1.0	1.0	0	0	0
11	27	1.0	17.0	3.0	0.0	0.000000	0.00	1.0	8.00	0.0	...	1.0	1.0	0	0	0
12	45	4.0	14.0	6.0	0.0	0.000000	0.00	1.0	10.00	1.0	...	1.0	1.0	0	0	0
13	44	2.0	25.0	2.0	0.0	0.000000	0.00	1.0	5.00	0.0	...	1.0	1.0	0	0	0
14	43	2.0	18.0	5.0	0.0	0.000000	0.00	0.0	0.00	1.0	...	1.0	1.0	0	0	0
15	40	3.0	18.0	2.0	0.0	0.000000	0.00	1.0	15.00	0.0	...	1.0	1.0	0	0	0
16	41	4.0	21.0	3.0	0.0	0.000000	0.00	1.0	0.25	0.0	...	1.0	1.0	0	0	0
17	43	3.0	15.0	8.0	0.0	0.000000	0.00	1.0	3.00	0.0	...	1.0	1.0	0	0	0
18	42	2.0	20.0	1.0	0.0	0.000000	0.00	1.0	7.00	1.0	...	1.0	1.0	0	0	0
19	40	2.0	27.0	1.0	0.0	0.000000	0.00	0.0	0.00	1.0	...	1.0	1.0	0	0	0
20	43	2.0	18.0	4.0	0.0	0.000000	0.00	1.0	15.00	0.0	...	1.0	1.0	0	0	0
21	41	3.0	17.0	4.0	0.0	0.000000	0.00	1.0	10.00	0.0	...	21.0	21.0	0	0	0
22	40	1.0	18.0	1.0	0.0	0.000000	0.00	1.0	0.25	0.0	...	2.0	2.0	0	0	0
23	40	1.0	20.0	2.0	0.0	0.000000	0.00	1.0	15.00	0.0	...	1.0	1.0	1	0	1
24	40	3.0	15.0	3.0	0.0	0.000000	0.00	1.0	3.00	0.0	...	1.0	1.0	0	0	0
25	44	3.0	19.0	1.0	0.0	0.000000	0.00	0.0	0.00	0.0	...	1.0	1.0	0	0	0
26	39	5.0	23.0	2.0	0.0	0.000000	0.00	0.0	0.00	1.0	...	1.0	1.0	0	0	0
27	39	2.0	17.0	4.0	0.0	0.000000	0.00	0.0	0.00	0.0	...	1.0	1.0	0	0	0
28	37	3.0	24.0	1.0	1.0	3.000000	0.04	0.0	0.00	0.0	...	1.0	1.0	0	0	0
29	37	6.0	26.0	1.0	0.0	0.000000	0.00	1.0	0.25	0.0	...	1.0	1.0	0	0	0

```
sns.regplot(x = CC['STDs'], y = CC['STDs (number)'])
```

```
<Axes: xlabel='STDs', ylabel='STDs (number)'>
```



```
CC.corr()

AX = CC.drop('Dx:Cancer', axis = 1)
AY = CC['Number of sexual partners']

print("X = ",AX.shape, "\nY = ", AY.shape)
```

```
X =  (858, 33)
Y =  (858,)
```

```
AX_train, AX_test, AY_train, AY_test = train_test_split(AX, AY, test_size = 0.3, random_state = 101)
```

```
AX_train.shape
```

```
(600, 33)
```

```
AY_train.shape
```

```
(600,)
```

```
model1 = LinearRegression()
```

```
model1.fit(AX_train, AY_train)
```

```
▾ LinearRegression
LinearRegression()
```

```
model1.coef_
```

```
array([ 3.33080316e-16,  1.00000000e+00, -2.88423056e-16, -4.00164546e-16,
       -4.64422615e-15,  3.05222863e-16, -2.77110936e-16,  1.09266336e-15,
       -4.14819107e-17, -5.13183030e-15,  5.95855059e-16,  2.63540623e-15,
       -6.75706770e-15,  5.10884151e-15, -5.62762244e-15,  3.30942626e-15,
       -2.20432056e-15,  3.46944695e-18, -1.88096556e-15, -1.73472348e-18,
        1.35785772e-15, -1.22743345e-14,  5.33725057e-15,  4.18966208e-15,
       -1.40958362e-16,  2.65004674e-16,  7.93483498e-15,  1.83527187e-15,
       -6.24551880e-15,  8.53607961e-16,  9.09769183e-15,  5.06142163e-15,
       -1.17494850e-14])
```

```
pd.DataFrame(model1.coef_, AX.columns, columns = ['Coedicients'])
```

Number of sexual partners	1.000000e+00
First sexual intercourse	-2.884231e-16
Num of pregnancies	-4.001645e-16
Smokes	-4.644226e-15
Smokes (years)	3.052229e-16
Smokes (packs/year)	-2.771109e-16
Hormonal Contraceptives	1.092663e-15
Hormonal Contraceptives (years)	-4.148191e-17
IUD	-5.131830e-15
IUD (years)	5.958551e-16
STDs	2.635406e-15
STDs (number)	-6.757068e-15
STDs:condylomatosis	5.108842e-15
STDs:vaginal condylomatosis	-5.627622e-15
STDs:vulvo-perineal condylomatosis	3.309426e-15
STDs:syphilis	-2.204321e-15
STDs:pelvic inflammatory disease	3.469447e-18
STDs:genital herpes	-1.880966e-15
STDs:molluscum contagiosum	-1.734723e-18
STDs:HIV	1.357858e-15
STDs:Hepatitis B	-1.227433e-14
STDs:HPV	5.337251e-15
STDs: Number of diagnosis	4.189662e-15
STDs: Time since first diagnosis	-1.409584e-16
STDs: Time since last diagnosis	2.650047e-16
Dx:CIN	7.934835e-15
Dx:HPV	1.835272e-15
Dx	-6.245519e-15
Hinselmann	8.536080e-16
Schiller	9.097692e-15
Citology	5.061422e-15
Biopsy	-1.174949e-14

```
AY_pred = model1.predict(AX_test)
```

```
MAE = metrics.mean_absolute_error(AY_test, AY_pred)
MSE = metrics.mean_squared_error(AY_test, AY_pred)
RMSE = np.sqrt(MSE)
print("MAE: ",MAE)
print("MSE: ",MSE)
print("RMSE: ",RMSE)
```

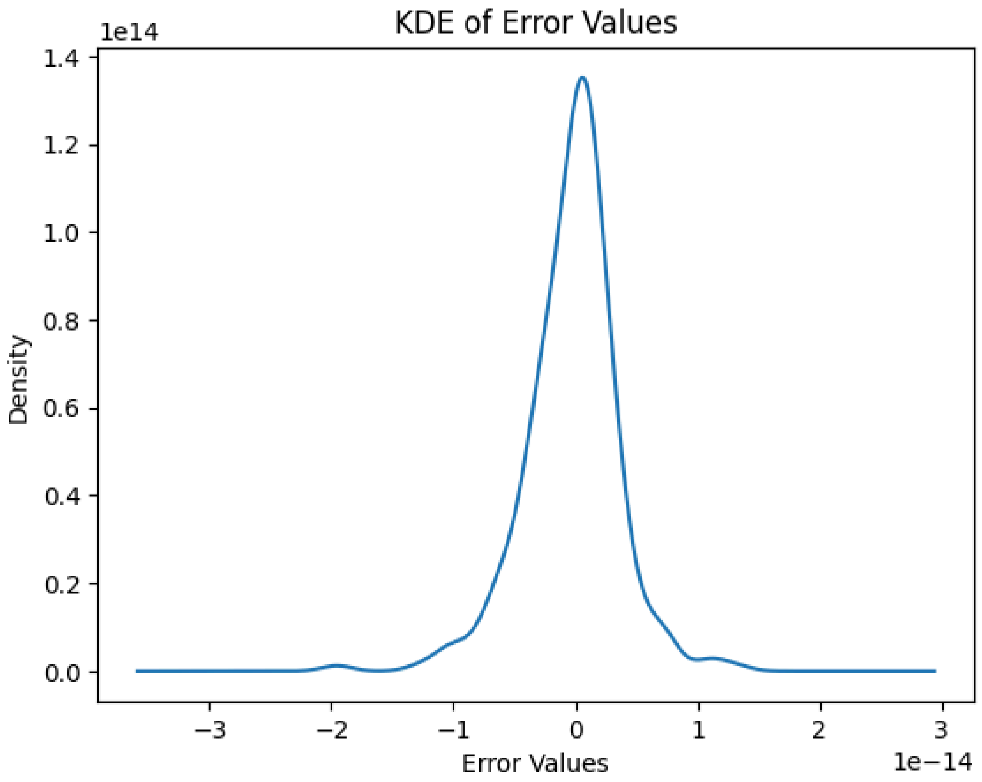
```
     MAE:  2.6503347320412168e-15
     MSE:  1.4147373762025927e-29
     RMSE:  3.7612994778435185e-15
```

```
CC['Dx:Cancer'].mean()
```

```
     0.02097902097902098
```

```
test_residual = AY_test - AY_pred
```

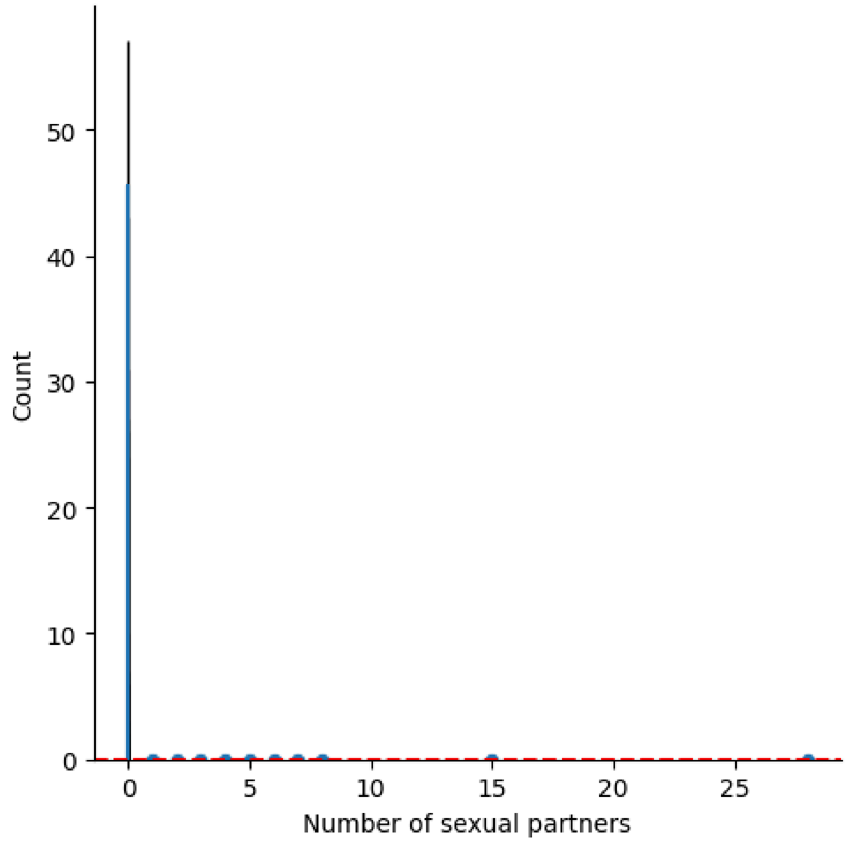
```
test_residual.plot(kind='kde')
plt.xlabel('Error Values')
plt.ylabel('Density')
plt.title('KDE of Error Values')
plt.show()
```



```
sns.displot(test_residual, bins = 25, kde = True)
```

```
sns.scatterplot(x = AY_test, y=test_residual)
plt.axhline(y=0, color='red', ls='--')
```

```
<matplotlib.lines.Line2D at 0x7cc07bd9dcc0>
```



Logistic regression

```
XCan = CC.drop(columns = 'Dx:Cancer')
```

```
YCan = CC['Dx:Cancer']
```

```
XCan_train, XCan_test, YCan_train, YCan_test = train_test_split(XCan, YCan, test_size = 0.3, random_state = 101)
```

```
from sklearn.preprocessing import StandardScaler
```

```
scaler = StandardScaler()
```

```
XCan_train_scaled = scaler.fit_transform(XCan_train)
```

```
XCan_test_scaled = scaler.transform(XCan_test)
```

```
XCan_train_scaled
```

```
array([[ 1.14018916,  2.58717906,  3.54934575, ..., -0.30151134,
        -0.23735633, -0.25635788],
       [ 0.23572577, -1.08605198,  0.82506721, ..., -0.30151134,
        -0.23735633, -0.25635788],
       [-1.23402724, -1.08605198, -1.12084603, ...,  3.31662479,
        -0.23735633, -0.25635788],
       ...,
       [ 0.91407331,  0.38324044,  0.43588457, ..., -0.30151134,
        -0.23735633, -0.25635788],
       [ 2.38382633, -0.35140577, -0.34248073, ..., -0.30151134,
        -0.23735633, -0.25635788],
       [ 0.57489954,  1.11788665,  0.04670192, ..., -0.30151134,
        -0.23735633, -0.25635788]])
```

```
XCan_test_scaled
```

```
     array([[ 0.3487837 , -0.35140577,  4.7168937 , ..., -0.30151134,
             -0.23735633, -0.25635788],
            [-0.89485347, -0.35140577, -0.73166338, ..., -0.30151134,
             -0.23735633, -0.25635788],
            [-1.23402724, -0.35140577, -0.73166338, ..., -0.30151134,
              4.21307489, -0.25635788],
            ...,
            [ 0.12266785,  0.38324044,  1.60343251, ..., -0.30151134,
             -0.23735633, -0.25635788],
            [-0.66873762, -0.35140577,  0.82506721, ..., -0.30151134,
             -0.23735633, -0.25635788],
            [ 0.00960992,  1.85253285,  0.82506721, ..., -0.30151134,
              4.21307489, -0.25635788]])
```

```
from sklearn.linear_model import LogisticRegression
```

```
from sklearn.metrics import accuracy_score
```

```
log_reg = LogisticRegression(random_state = 0).fit(XCan_train_scaled, YCan_train)
```

```
log_reg.predict(XCan_train_scaled)
```

```
     array([0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0])
```

```
log_reg.score(XCan_train_scaled, YCan_train)
```

```
     1.0
```

```
log_reg.score(XCan_test_scaled, YCan_test)
```

```
     0.9844961240310077
```