

```
In [ ]: 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
```

```
In [ ]: 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+risk+factors', 'data_url': 'https://archive.ics.uci.edu/static/public/383/data.csv', 'abstract': 'This dataset focuses on the prediction of indicators/diagnosis of cervical cancer. The features cover demographic information, habits, and historic medical records.', 'area': 'Health and Medicine', 'tasks': ['Classification'], 'characteristics': ['Multivariate'], 'num_instances': 858, 'num_features': 36, 'feature_types': ['Integer', 'Real'], 'demographics': ['Age', 'Other'], 'target_col': None, 'index_col': None, 'has_missing_values': 'yes', 'missing_values_symbol': 'NaN', 'year_of_dataset_creation': 2017, 'last_updated': 'Sun Mar 10 2024', 'dataset_doi': '10.24432/C5Z310', 'creators': ['Kelwin Fernandes', 'Jaime Cardoso', 'Jessica Fernandes'], 'intro_paper': {'title': 'Transfer Learning with Partial Observability Applied to Cervical Cancer Screening', 'authors': 'Kelwin Fernandes, Jaime S. Cardoso, Jessica C. Fernandes', 'published_in': 'Iberian Conference on Pattern Recognition and Image Analysis', 'year': 2017, 'url': 'https://www.semanticscholar.org/paper/Transfer-Learning-with-Partial-Observability-to-Fernandes-Cardoso/1c02438ba4dfa775399ba414508e9cd335b69012', 'doi': None}, 'additional_info': {'summary': "The dataset was collected at 'Hospital Universitario de Caracas' in Caracas, Venezuela. The dataset comprises demographic information, habits, and historic medical records of 858 patients. Several patients decided not to answer some of the questions because of privacy concerns (missing values).", 'purpose': None, 'funded_by': None, 'instances_represent': None, 'recommended_data_splits': None, 'sensitive_data': None, 'preprocessing_description': None, 'variable_info': '(int) Age\r\n(int) Number of sexual partners\r\n(int) First sexual intercourse (age)\r\n(int) Num of pregnancies\r\n(bool) Smokes\r\n(bool) Smokes (years)\r\n(bool) Smokes (packs/year)\r\n(bool) Hormonal Contraceptives\r\n(int) Hormonal Contraceptives (years)\r\n(bool) IUD\r\n(int) IUD (years)\r\n(bool) STDs\r\n(int) STDs (number)\r\n(bool) STDs:condylomatosis\r\n(bool) STDs:cervical condylomatosis\r\n(bool) STDs:vaginal condylomatosis\r\n(bool) STDs:vulvo-perineal condylomatosis\r\n(bool) STDs:syphilis\r\n(bool) STDs:pelvic inflammatory disease\r\n(bool) STDs:genital herpes\r\n(bool) STDs:molluscum contagiosum\r\n(bool) STDs:AIDS\r\n(bool) STDs:HIV\r\n(bool) STDs:Hepatitis B\r\n(bool) STDs:HPV\r\n(int) STDs: Number of diagnosis\r\n(int) STDs: Time since first diagnosis\r\n(int) STDs: Time since last diagnosis\r\n(bool) Dx:Cancer\r\n(bool) Dx:CIN\r\n(bool) Dx:HPV\r\n(bool) Dx\r\n(bool) Hinselmann: target variable\r\n(bool) Schiller: target variable\r\n(bool) Cytology: target variable\r\n(bool) Biopsy: target variable', 'citation': None}}
```

	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
17	None	None	yes
18	None	None	yes
19	None	None	yes

20	None	None	yes
21	None	None	yes
22	None	None	yes
23	None	None	yes
24	None	None	yes
25	None	None	no
26	None	None	yes
27	None	None	yes
28	None	None	no
29	None	None	no
30	None	None	no
31	None	None	no
32	None	None	no
33	None	None	no
34	None	None	no
35	None	None	no

In []: X.head()

Out[3]:

	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	STD Time since l2 diagnosis
0	18	4.0	15.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	NaN	N:
1	15	1.0	14.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	NaN	N:
2	34	1.0	NaN	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	NaN	N:
3	52	5.0	16.0	4.0	1.0	37.0	37.0	1.0	3.0	0.0	...	NaN	N:
4	46	3.0	21.0	4.0	0.0	0.0	0.0	1.0	15.0	0.0	...	NaN	N:

5 rows × 36 columns



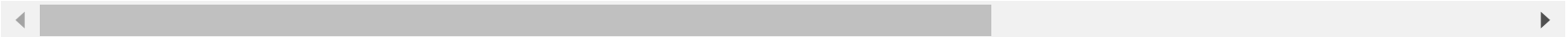
In []: y

```
In [ ]: cervix_df = X.copy()
cervix_df.head()
```

Out[5]:

	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	STD Time since l:
0	18	4.0	15.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	NaN	N:
1	15	1.0	14.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	NaN	N:
2	34	1.0	NaN	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	NaN	N:
3	52	5.0	16.0	4.0	1.0	37.0	37.0	1.0	3.0	0.0	...	NaN	N:
4	46	3.0	21.0	4.0	0.0	0.0	0.0	1.0	15.0	0.0	...	NaN	N:

5 rows × 36 columns



```
In [ ]: import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

EDA and Data Wrangling

```
In [ ]: cervix_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

```
In [ ]: cervix_df.describe()
```

Out[8]:

	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormonal Contraceptives	Hormonal Contraceptives (years)	II
count	858.000000	832.000000	851.000000	802.000000	845.000000	845.000000	845.000000	750.000000	750.000000	741.000000
mean	26.820513	2.527644	16.995300	2.275561	0.145562	1.219721	0.453144	0.641333	2.256419	0.112000
std	8.497948	1.667760	2.803355	1.447414	0.352876	4.089017	2.226610	0.479929	3.764254	0.315500
min	13.000000	1.000000	10.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	20.000000	2.000000	15.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	25.000000	2.000000	17.000000	2.000000	0.000000	0.000000	0.000000	1.000000	0.500000	0.000000
75%	32.000000	3.000000	18.000000	3.000000	0.000000	0.000000	0.000000	1.000000	3.000000	0.000000
max	84.000000	28.000000	32.000000	11.000000	1.000000	37.000000	37.000000	1.000000	30.000000	1.000000

8 rows × 36 columns



Changing each columns to their corresponding datatypes.

```
In [ ]: column_data_types = {
    "Smokes": bool,
    "Smokes (years)": bool,
    "Smokes (packs/year)": bool,
    "Hormonal Contraceptives": bool,
    "IUD": bool,
    "STDs": bool,
    "STDs:condylomatosis": bool,
    "STDs:cervical condylomatosis": bool,
    "STDs:vaginal condylomatosis": bool,
    "STDs:vulvo-perineal condylomatosis": bool,
    "STDs:syphilis": bool,
    "STDs:pelvic inflammatory disease": bool,
    "STDs:genital herpes": bool,
    "STDs:molluscum contagiosum": bool,
    "STDs:AIDS": bool,
    "STDs:HIV": bool,
    "STDs:Hepatitis B": bool,
    "STDs:HPV": bool,
    "Dx:Cancer": bool,
    "Dx:CIN": bool,
    "Dx:HPV": bool,
    "Dx": bool,
    "HinseImann": bool,
    "Schiller": bool,
    "Citology": bool,
    "Biopsy": bool
}

# converting datatypes
cervix_df = cervix_df.astype(column_data_types)
```

```
In [ ]: cervix_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	858 non-null	bool
5	Smokes (years)	858 non-null	bool
6	Smokes (packs/year)	858 non-null	bool
7	Hormonal Contraceptives	858 non-null	bool
8	Hormonal Contraceptives (years)	750 non-null	float64
9	IUD	858 non-null	bool
10	IUD (years)	741 non-null	float64
11	STDs	858 non-null	bool
12	STDs (number)	753 non-null	float64
13	STDs:condylomatosis	858 non-null	bool
14	STDs:cervical condylomatosis	858 non-null	bool
15	STDs:vaginal condylomatosis	858 non-null	bool
16	STDs:vulvo-perineal condylomatosis	858 non-null	bool
17	STDs:syphilis	858 non-null	bool
18	STDs:pelvic inflammatory disease	858 non-null	bool
19	STDs:genital herpes	858 non-null	bool
20	STDs:molluscum contagiosum	858 non-null	bool
21	STDs:AIDS	858 non-null	bool
22	STDs:HIV	858 non-null	bool
23	STDs:Hepatitis B	858 non-null	bool
24	STDs:HPV	858 non-null	bool
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	bool
29	Dx:CIN	858 non-null	bool
30	Dx:HPV	858 non-null	bool
31	Dx	858 non-null	bool
32	Hinselmann	858 non-null	bool
33	Schiller	858 non-null	bool
34	Citology	858 non-null	bool
35	Biopsy	858 non-null	bool

dtypes: bool(26), float64(8), int64(2)
memory usage: 88.9 KB

```
In [ ]: cervix_df.head()
```

Out[15]:

	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	ST 1 since diagn
0	18	4.0	15.0	1.0	False	False	False	False	0.0	False	...	NaN	
1	15	1.0	14.0	1.0	False	False	False	False	0.0	False	...	NaN	
2	34	1.0	NaN	1.0	False	False	False	False	0.0	False	...	NaN	
3	52	5.0	16.0	4.0	True	True	True	True	3.0	False	...	NaN	
4	46	3.0	21.0	4.0	False	False	False	True	15.0	False	...	NaN	

5 rows × 36 columns



Handling missing values

```
In [ ]: cervix_df.isnull().sum()
```

```
Out[16]: Age                                0
Number of sexual partners                  26
First sexual intercourse                   7
Num of pregnancies                        56
Smokes                                    0
Smokes (years)                           0
Smokes (packs/year)                      0
Hormonal Contraceptives                   0
Hormonal Contraceptives (years)          108
IUD                                        0
IUD (years)                              117
STDs                                      0
STDs (number)                            105
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           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
```

```
In [ ]: # STDs: Time since first diagnosis and STDs: Time since last diagnosis have the highest amount of missing val
cervix_df = cervix_df.drop(columns = ['STDs: Time since first diagnosis', 'STDs: Time since last diagnosis'])
cervix_df.head()
```

Out[17]:

	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormonal Contraceptives	Hormonal Contraceptives (years)	IUD	...	STDs:HPV	S' Nur diagn
0	18	4.0	15.0	1.0	False	False	False	False	0.0	False	...	False	
1	15	1.0	14.0	1.0	False	False	False	False	0.0	False	...	False	
2	34	1.0	NaN	1.0	False	False	False	False	0.0	False	...	False	
3	52	5.0	16.0	4.0	True	True	True	True	3.0	False	...	False	
4	46	3.0	21.0	4.0	False	False	False	True	15.0	False	...	False	

5 rows × 34 columns

```
In [ ]: cervix_df.mean()
```

```
Out[18]: Age                26.820513
          Number of sexual partners    2.527644
          First sexual intercourse    16.995300
          Num of pregnancies          2.275561
          Smokes                    0.158508
          Smokes (years)             0.158508
          Smokes (packs/year)        0.158508
          Hormonal Contraceptives    0.686480
          Hormonal Contraceptives (years)  2.256419
          IUD                        0.233100
          IUD (years)                0.514804
          STDs                       0.214452
          STDs (number)              0.176627
          STDs:condylomatosis        0.173660
          STDs:cervical condylomatosis  0.122378
          STDs:vaginal condylomatosis  0.127040
          STDs:vulvo-perineal condylomatosis  0.172494
          STDs:syphilis              0.143357
          STDs:pelvic inflammatory disease  0.123543
          STDs:genital herpes        0.123543
          STDs:molluscum contagiosum  0.123543
          STDs:AIDS                  0.122378
          STDs:HIV                   0.143357
          STDs:Hepatitis B           0.123543
          STDs:HPV                   0.124709
          STDs: Number of diagnosis  0.087413
          Dx:Cancer                  0.020979
          Dx:CIN                     0.010490
          Dx:HPV                     0.020979
          Dx                          0.027972
          Hinselmann                 0.040793
          Schiller                   0.086247
          Citology                   0.051282
          Biopsy                     0.064103
          dtype: float64
```



```
In [ ]: # replacing NaN values with the mean value of each column
cervix_df = cervix_df.fillna(cervix_df.mean())
cervix_df.head()
```

Out[19]:

	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormonal Contraceptives	Hormonal Contraceptives (years)	IUD	...	STDs:HPV	S Nur diagn
0	18	4.0	15.0000	1.0	False	False	False	False	0.0	False	...	False	
1	15	1.0	14.0000	1.0	False	False	False	False	0.0	False	...	False	
2	34	1.0	16.9953	1.0	False	False	False	False	0.0	False	...	False	
3	52	5.0	16.0000	4.0	True	True	True	True	3.0	False	...	False	
4	46	3.0	21.0000	4.0	False	False	False	True	15.0	False	...	False	

5 rows × 34 columns

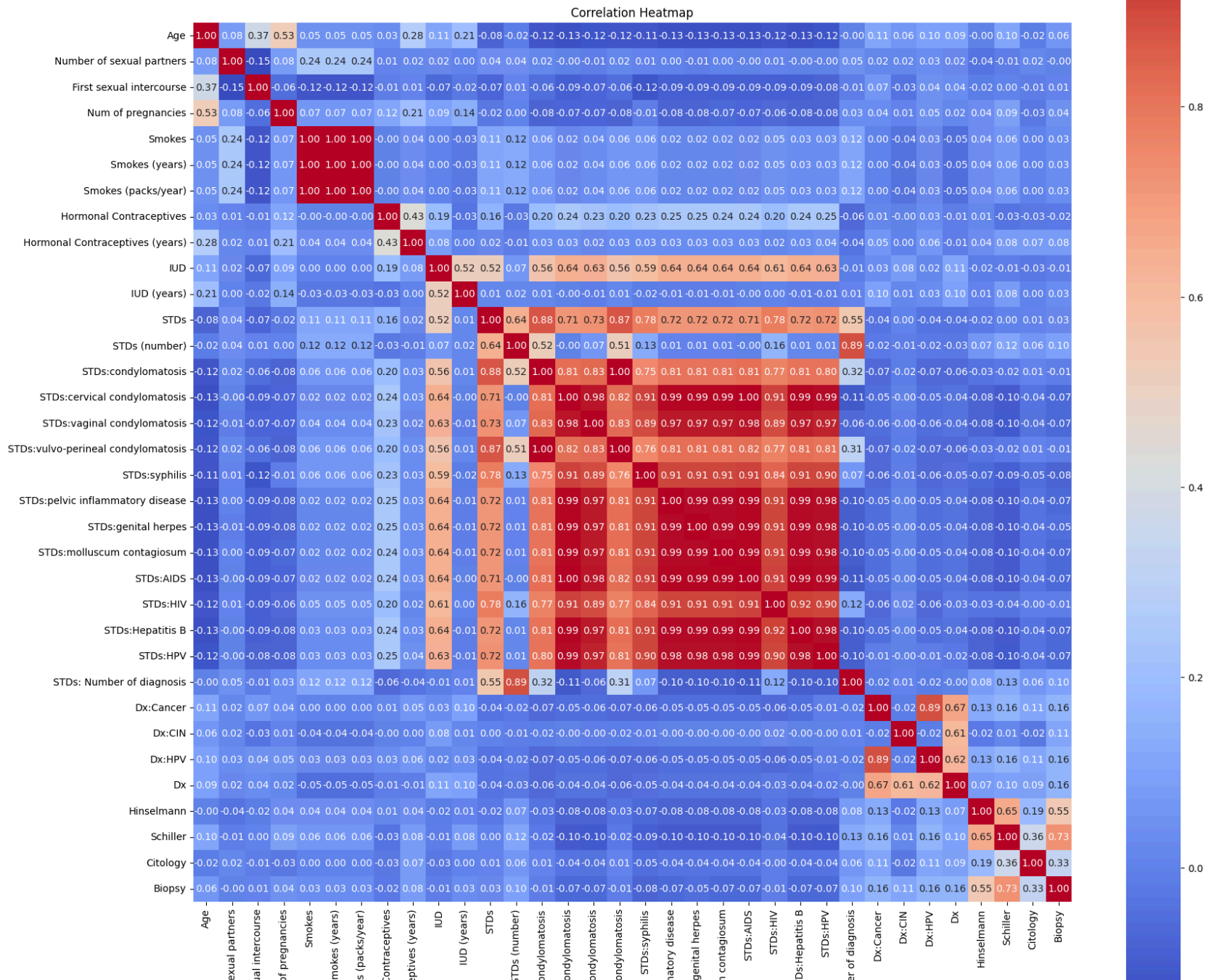


```
In [ ]: cervix_df.isnull().sum()
```

```
Out[20]: 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
          Dx:Cancer                         0
          Dx:CIN                            0
          Dx:HPV                            0
          Dx                                0
          Hinselmann                        0
          Schiller                          0
          Citology                          0
          Biopsy                            0
          dtype: int64
```

```
In [ ]: # correlation of variables
correlation_matrix = cervix_df.corr()

plt.figure(figsize=(20, 20))
sns.heatmap(correlation_matrix, annot=True, fmt='.2f', cmap='coolwarm', square=True)
plt.title('Correlation Heatmap')
plt.show()
```

Number of s
First sex
Num (

S
Smoke
Hormonal C
Hormonal Contrace

s
STDs:ci
STDs:cervical c
STDs:vaginal c
STDs:vulvo-perineal c
STDs:pelvic inflam
STDs:
STDs:molluscun

ST

STDs: Numb



```
In [ ]: correlation_matrix
```

Out[22]:

	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormonal Contraceptives	Hormonal Contraceptives (years)
Age	1.000000	0.084896	0.369168	0.526137	0.045244	0.045244	0.045244	0.029201	0.277181
Number of sexual partners	0.084896	1.000000	-0.147937	0.076719	0.241673	0.241673	0.241673	0.005771	0.018552
First sexual intercourse	0.369168	-0.147937	1.000000	-0.058223	-0.119364	-0.119364	-0.119364	-0.009233	0.008000
Num of pregnancies	0.526137	0.076719	-0.058223	1.000000	0.068466	0.068466	0.068466	0.118860	0.207839
Smokes	0.045244	0.241673	-0.119364	0.068466	1.000000	1.000000	1.000000	-0.002485	0.036849
Smokes (years)	0.045244	0.241673	-0.119364	0.068466	1.000000	1.000000	1.000000	-0.002485	0.036849
Smokes (packs/year)	0.045244	0.241673	-0.119364	0.068466	1.000000	1.000000	1.000000	-0.002485	0.036849
Hormonal Contraceptives	0.029201	0.005771	-0.009233	0.118860	-0.002485	-0.002485	-0.002485	1.000000	0.433573
Hormonal Contraceptives (years)	0.277181	0.018552	0.008000	0.207839	0.036849	0.036849	0.036849	0.433573	1.000000
IUD	0.107725	0.023182	-0.070207	0.091457	0.002252	0.002252	0.002252	0.194324	0.079862
IUD (years)	0.205886	0.004215	-0.024803	0.143642	-0.026532	-0.026532	-0.026532	-0.033394	0.000455
STDs	-0.084916	0.035939	-0.073416	-0.022581	0.107566	0.107566	0.107566	0.163352	0.024324
STDs (number)	-0.015488	0.039359	0.006487	0.001706	0.123340	0.123340	0.123340	-0.027045	-0.006468
STDs:condylomatosis	-0.119277	0.018207	-0.057674	-0.083053	0.062190	0.062190	0.062190	0.197063	0.031559
STDs:cervical condylomatosis	-0.128618	-0.002320	-0.089871	-0.074901	0.022948	0.022948	0.022948	0.244691	0.033223
STDs:vaginal condylomatosis	-0.124629	-0.011049	-0.073367	-0.073965	0.035673	0.035673	0.035673	0.227629	0.024709
STDs:vulvo-perineal condylomatosis	-0.118207	0.019247	-0.055621	-0.082667	0.063695	0.063695	0.063695	0.195504	0.032746
STDs:syphilis	-0.113142	0.008980	-0.123280	-0.010497	0.059224	0.059224	0.059224	0.226274	0.031918
STDs:pelvic inflammatory disease	-0.125518	0.000866	-0.089500	-0.080359	0.021318	0.021318	0.021318	0.246090	0.031571
STDs:genital herpes	-0.130940	-0.005608	-0.086961	-0.077826	0.021318	0.021318	0.021318	0.246090	0.031068

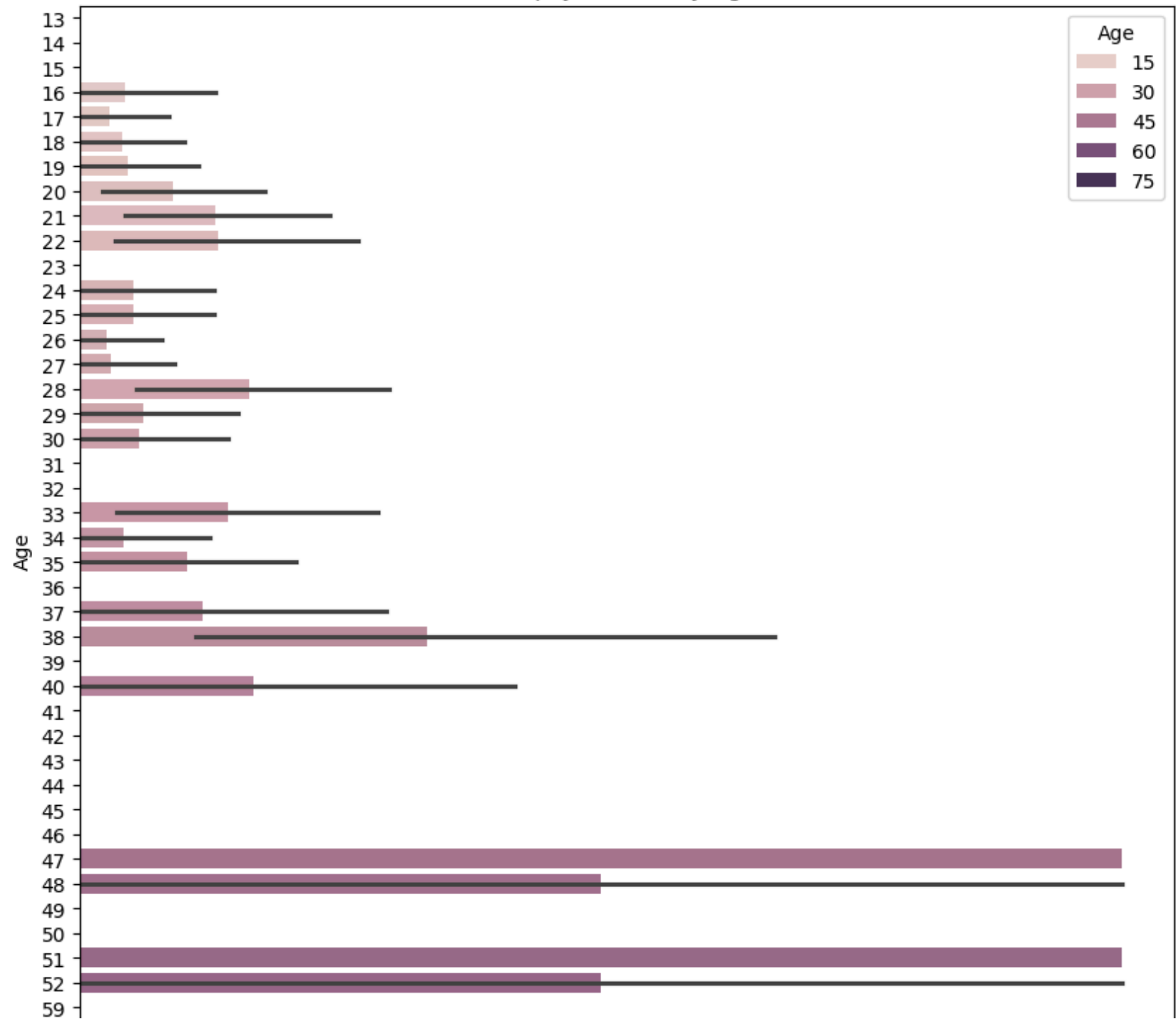
	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormonal Contraceptives	Hormonal Contraceptives (years)
STDs:molluscum contagiosum	-0.128020	0.000866	-0.090769	-0.070229	0.021318	0.021318	0.021318	0.238455	0.030816
STDs:AIDS	-0.128618	-0.002320	-0.089871	-0.074901	0.022948	0.022948	0.022948	0.244691	0.033223
STDs:HIV	-0.118233	0.005996	-0.087521	-0.064539	0.050118	0.050118	0.050118	0.197598	0.015842
STDs:Hepatitis B	-0.130940	-0.003450	-0.088231	-0.077826	0.031016	0.031016	0.031016	0.238455	0.030816
STDs:HPV	-0.121616	-0.000271	-0.084077	-0.078209	0.029360	0.029360	0.029360	0.247484	0.040465
STDs: Number of diagnosis	-0.001606	0.051559	-0.013327	0.033514	0.117277	0.117277	0.117277	-0.062199	-0.037219
Dx:Cancer	0.110340	0.022309	0.067283	0.035123	0.003270	0.003270	0.003270	0.011278	0.054627
Dx:CIN	0.061443	0.015691	-0.032626	0.007344	-0.044686	-0.044686	-0.044686	-0.004397	0.003086
Dx:HPV	0.101722	0.027264	0.043966	0.046753	0.025538	0.025538	0.025538	0.028808	0.061394
Dx	0.092635	0.022982	0.035750	0.019025	-0.054271	-0.054271	-0.054271	-0.007245	-0.012865
Hinselmann	-0.003967	-0.039273	-0.016546	0.038685	0.039562	0.039562	0.039562	0.012360	0.038825
Schiller	0.103283	-0.008899	0.003493	0.087687	0.059913	0.059913	0.059913	-0.034002	0.078707
Citology	-0.016862	0.021839	-0.010971	-0.029656	0.000371	0.000371	0.000371	-0.025116	0.074324
Biopsy	0.055956	-0.001429	0.007262	0.043460	0.029733	0.029733	0.029733	-0.018015	0.078995

34 rows × 34 columns

```
In [ ]: plt.figure(figsize=(10, 10))
sns.barplot(x="Biopsy", y="Age", data=cervix_df, orient="h", hue="Age", dodge=False)

plt.xlabel("Total Instances")
plt.ylabel("Age")
plt.title("Biopsy Results by Age")
plt.show()
```


Biopsy Results by Age





```
In [ ]: # histogram
cervix_df.hist(bins = 10, figsize = (30,30), color = 'b')
```

```
Out[24]: array([[<Axes: title={'center': 'Age'}>,
  <Axes: title={'center': 'Number of sexual partners'}>,
  <Axes: title={'center': 'First sexual intercourse'}>],
  [<Axes: title={'center': 'Num of pregnancies'}>,
  <Axes: title={'center': 'Hormonal Contraceptives (years)}>,
  <Axes: title={'center': 'IUD (years)}>],
  [<Axes: title={'center': 'STDs (number)}>,
  <Axes: title={'center': 'STDs: Number of diagnosis'}>, <Axes: >]],
  dtype=object)
```