gression-analysis-cervical-cancer

April 28, 2024

1 Hands-on Activity 11.1 Linear Regression Analysis

Course: CPE 311	Program: BSCpE
Course Title: Computational Thinking with Python	**Date Performed:** April 27 , 2024
Section: BSCPE22S3	**Date Submitted:** April 28, 2024
Student Name: John Louie V. Adornado	**Instructor's Name:** Engr. Roman Richard

[1]:
```
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
```

[2]:
```
pip install hvplot
```

```
Collecting hvplot
  Downloading hvplot-0.9.2-py2.py3-none-any.whl (1.8 MB)
                           1.8/1.8 MB
8.9 MB/s eta 0:00:00
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)
Installing collected packages: hvplot
Successfully installed hvplot-0.9.2
```

[3]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import hvplot.pandas
from sklearn.model_selection import train_test_split
from sklearn import metrics
from sklearn.linear_model import LinearRegression
%matplotlib inline
```

[4]:
```
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
```

[5]:
```
df = pd.concat([X, y], axis = 1)
df
```

[5]:
```
      Age  Number of sexual partners  First sexual intercourse  \
0      18                        4.0                      15.0
1      15                        1.0                      14.0
2      34                        1.0                       NaN
3      52                        5.0                      16.0
4      46                        3.0                      21.0
..    …                         …                        …
853    34                        3.0                      18.0
854    32                        2.0                      19.0
855    25                        2.0                      17.0
856    33                        2.0                      24.0
857    29                        2.0                      20.0

      Num of pregnancies  Smokes  Smokes (years)  Smokes (packs/year)  \
0                    1.0     0.0             0.0                  0.0
1                    1.0     0.0             0.0                  0.0
2                    1.0     0.0             0.0                  0.0
3                    4.0     1.0            37.0                 37.0
4                    4.0     0.0             0.0                  0.0
..                   …     …               …                    …
853                  0.0     0.0             0.0                  0.0
854                  1.0     0.0             0.0                  0.0
855                  0.0     0.0             0.0                  0.0
856                  2.0     0.0             0.0                  0.0
857                  1.0     0.0             0.0                  0.0

      Hormonal Contraceptives  Hormonal Contraceptives (years)  IUD  …  \
0                         0.0                             0.00  0.0  …
1                         0.0                             0.00  0.0  …
2                         0.0                             0.00  0.0  …
3                         1.0                             3.00  0.0  …
4                         1.0                            15.00  0.0  …
```

```
..                        …                                …  …  …
853                       0.0                               0.00  0.0  …
854                       1.0                               8.00  0.0  …
855                       1.0                               0.08  0.0  …
856                       1.0                               0.08  0.0  …
857                       1.0                               0.50  0.0  …

     STDs: Time since first diagnosis  STDs: Time since last diagnosis  \
0                                 NaN                               NaN
1                                 NaN                               NaN
2                                 NaN                               NaN
3                                 NaN                               NaN
4                                 NaN                               NaN
..                                …                                 …
853                               NaN                               NaN
854                               NaN                               NaN
855                               NaN                               NaN
856                               NaN                               NaN
857                               NaN                               NaN

     Dx:Cancer  Dx:CIN  Dx:HPV  Dx  Hinselmann  Schiller  Citology  Biopsy
0            0       0       0   0           0         0         0       0
1            0       0       0   0           0         0         0       0
2            0       0       0   0           0         0         0       0
3            1       0       1   0           0         0         0       0
4            0       0       0   0           0         0         0       0
..         …     …     …   ..          …         …         …     …
853          0       0       0   0           0         0         0       0
854          0       0       0   0           0         0         0       0
855          0       0       0   0           0         0         1       0
856          0       0       0   0           0         0         0       0
857          0       0       0   0           0         0         0       0

[858 rows x 36 columns]
```

[6]:
```
df.head(20)
```

[6]:
```
    Age  Number of sexual partners  First sexual intercourse  \
0    18                        4.0                      15.0
1    15                        1.0                      14.0
2    34                        1.0                       NaN
3    52                        5.0                      16.0
4    46                        3.0                      21.0
5    42                        3.0                      23.0
6    51                        3.0                      17.0
7    26                        1.0                      26.0
8    45                        1.0                      20.0
```

9	44	3.0	15.0
10	44	3.0	26.0
11	27	1.0	17.0
12	45	4.0	14.0
13	44	2.0	25.0
14	43	2.0	18.0
15	40	3.0	18.0
16	41	4.0	21.0
17	43	3.0	15.0
18	42	2.0	20.0
19	40	2.0	27.0

	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	\
0	1.0	0.0	0.000000	0.0	
1	1.0	0.0	0.000000	0.0	
2	1.0	0.0	0.000000	0.0	
3	4.0	1.0	37.000000	37.0	
4	4.0	0.0	0.000000	0.0	
5	2.0	0.0	0.000000	0.0	
6	6.0	1.0	34.000000	3.4	
7	3.0	0.0	0.000000	0.0	
8	5.0	0.0	0.000000	0.0	
9	NaN	1.0	1.266973	2.8	
10	4.0	0.0	0.000000	0.0	
11	3.0	0.0	0.000000	0.0	
12	6.0	0.0	0.000000	0.0	
13	2.0	0.0	0.000000	0.0	
14	5.0	0.0	0.000000	0.0	
15	2.0	0.0	0.000000	0.0	
16	3.0	0.0	0.000000	0.0	
17	8.0	0.0	0.000000	0.0	
18	NaN	0.0	0.000000	0.0	
19	NaN	0.0	0.000000	0.0	

	Hormonal Contraceptives	Hormonal Contraceptives (years)	IUD	…	\
0	0.0	0.00	0.0	…	
1	0.0	0.00	0.0	…	
2	0.0	0.00	0.0	…	
3	1.0	3.00	0.0	…	
4	1.0	15.00	0.0	…	
5	0.0	0.00	0.0	…	
6	0.0	0.00	1.0	…	
7	1.0	2.00	1.0	…	
8	0.0	0.00	0.0	…	
9	0.0	0.00	NaN	…	
10	1.0	2.00	0.0	…	
11	1.0	8.00	0.0	…	

12	1.0	10.00	1.0	…
13	1.0	5.00	0.0	…
14	0.0	0.00	1.0	…
15	1.0	15.00	0.0	…
16	1.0	0.25	0.0	…
17	1.0	3.00	0.0	…
18	1.0	7.00	1.0	…
19	0.0	0.00	1.0	…

	STDs: Time since first diagnosis	STDs: Time since last diagnosis	\
0	NaN	NaN	
1	NaN	NaN	
2	NaN	NaN	
3	NaN	NaN	
4	NaN	NaN	
5	NaN	NaN	
6	NaN	NaN	
7	NaN	NaN	
8	NaN	NaN	
9	NaN	NaN	
10	NaN	NaN	
11	NaN	NaN	
12	NaN	NaN	
13	NaN	NaN	
14	NaN	NaN	
15	NaN	NaN	
16	NaN	NaN	
17	NaN	NaN	
18	NaN	NaN	
19	NaN	NaN	

	Dx:Cancer	Dx:CIN	Dx:HPV	Dx	Hinselmann	Schiller	Citology	Biopsy
0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0
3	1	0	1	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	1	1	0	1
7	0	0	0	0	0	0	0	0
8	1	0	1	1	0	0	0	0
9	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0

```
15          0       0       0   0           0         0         0       0
16          0       0       0   0           0         0         0       0
17          0       0       0   0           0         0         0       0
18          0       0       0   0           0         0         0       0
19          0       0       0   0           0         0         0       0

[20 rows x 36 columns]
```

[7]:
```
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
```

[8]:
```
df.describe()
```

[8]:

	Age	Number of sexual partners	First sexual intercourse	\
count	858.000000	832.000000	851.000000	
mean	26.820513	2.527644	16.995300	
std	8.497948	1.667760	2.803355	
min	13.000000	1.000000	10.000000	
25%	20.000000	2.000000	15.000000	
50%	25.000000	2.000000	17.000000	
75%	32.000000	3.000000	18.000000	
max	84.000000	28.000000	32.000000	

	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	\
count	802.000000	845.000000	845.000000	845.000000	
mean	2.275561	0.145562	1.219721	0.453144	
std	1.447414	0.352876	4.089017	2.226610	
min	0.000000	0.000000	0.000000	0.000000	
25%	1.000000	0.000000	0.000000	0.000000	
50%	2.000000	0.000000	0.000000	0.000000	
75%	3.000000	0.000000	0.000000	0.000000	
max	11.000000	1.000000	37.000000	37.000000	

	Hormonal Contraceptives	Hormonal Contraceptives (years)	IUD	\
count	750.000000	750.000000	741.000000	
mean	0.641333	2.256419	0.112011	
std	0.479929	3.764254	0.315593	
min	0.000000	0.000000	0.000000	
25%	0.000000	0.000000	0.000000	
50%	1.000000	0.500000	0.000000	
75%	1.000000	3.000000	0.000000	
max	1.000000	30.000000	1.000000	

	…	STDs: Time since first diagnosis	STDs: Time since last diagnosis	\
count	…	71.000000	71.000000	
mean	…	6.140845	5.816901	
std	…	5.895024	5.755271	
min	…	1.000000	1.000000	
25%	…	2.000000	2.000000	
50%	…	4.000000	3.000000	
75%	…	8.000000	7.500000	

```
max    …                          22.000000                          22.000000

        Dx:Cancer      Dx:CIN      Dx:HPV          Dx  Hinselmann    Schiller  \
count  858.000000  858.000000  858.000000  858.000000  858.000000  858.000000
mean     0.020979    0.010490    0.020979    0.027972    0.040793    0.086247
std      0.143398    0.101939    0.143398    0.164989    0.197925    0.280892
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

         Citology      Biopsy
count  858.000000  858.000000
mean     0.051282    0.064103
std      0.220701    0.245078
min      0.000000    0.000000
25%      0.000000    0.000000
50%      0.000000    0.000000
75%      0.000000    0.000000
max      1.000000    1.000000

[8 rows x 36 columns]
```

[9]:
```
df.isnull().sum()
```

[9]:
```
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
```

[11]:
```
small_missing_cols = ['Number of sexual partners', 'First sexual intercourse',␣
↪'Num of pregnancies', 'Smokes',
                      'Smokes (years)', 'Smokes (packs/year)']
for col in small_missing_cols:
    if df[col].dtype == 'object':
        # For categorical columns, fill missing values with the mode
        df[col].fillna(df[col].mode()[0], inplace=True)
    else:
        # For numerical columns, fill missing values with the median
        df[col].fillna(df[col].median(), inplace=True)
```

[13]:
```
small_missing_cols = ['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']
for col in small_missing_cols:
    if df[col].dtype == 'object':
        # For categorical columns, fill missing values with the mode
        df[col].fillna(df[col].mode()[0], inplace=True)
    else:
        # For numerical columns, fill missing values with the median
        df[col].fillna(df[col].median(), inplace=True)
```

[15]:
```
# Drop columns with a large number of missing values
large_missing_cols = ['STDs: Time since first diagnosis', 'STDs: Time since␣
 ↪last diagnosis']
df.drop(large_missing_cols, axis=1, inplace=True)
```

[16]:
```
print(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
Dx:Cancer                             0
Dx:CIN                                0
Dx:HPV                                0
Dx                                    0
Hinselmann                            0
Schiller                              0
Citology                              0
Biopsy                                0
dtype: int64
```

[17]:
```
df.columns
```

[17]:
```
Index(['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',
       'Dx:Cancer', 'Dx:CIN', 'Dx:HPV', 'Dx', 'Hinselmann', 'Schiller',
       'Citology', 'Biopsy'],
      dtype='object')
```

[28]:
```
df
```

[28]:

	Age	Number of sexual partners	First sexual intercourse	\
0	18	4.0	15.0	
1	15	1.0	14.0	
2	34	1.0	17.0	
3	52	5.0	16.0	
4	46	3.0	21.0	
..	…	…	…	
853	34	3.0	18.0	
854	32	2.0	19.0	
855	25	2.0	17.0	
856	33	2.0	24.0	
857	29	2.0	20.0	

	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	\
0	1.0	0.0	0.0	0.0	
1	1.0	0.0	0.0	0.0	
2	1.0	0.0	0.0	0.0	
3	4.0	1.0	37.0	37.0	
4	4.0	0.0	0.0	0.0	
..	…	…	…	…	
853	0.0	0.0	0.0	0.0	
854	1.0	0.0	0.0	0.0	
855	0.0	0.0	0.0	0.0	
856	2.0	0.0	0.0	0.0	
857	1.0	0.0	0.0	0.0	

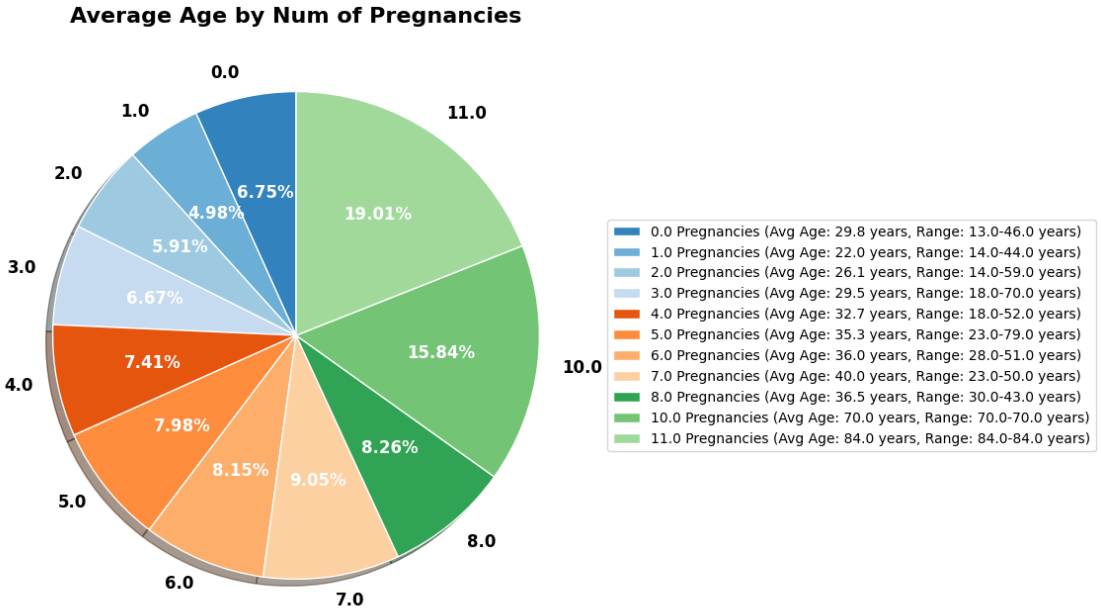
	Hormonal Contraceptives	Hormonal Contraceptives (years)	IUD	…	\
0	0.0	0.00	0.0	…	
1	0.0	0.00	0.0	…	
2	0.0	0.00	0.0	…	
3	1.0	3.00	0.0	…	
4	1.0	15.00	0.0	…	
..	…	…	…	…	
853	0.0	0.00	0.0	…	

```
854                    1.0                              8.00  0.0  …
855                    1.0                              0.08  0.0  …
856                    1.0                              0.08  0.0  …
857                    1.0                              0.50  0.0  …

     STDs:HPV  STDs: Number of diagnosis  Dx:Cancer  Dx:CIN  Dx:HPV  Dx  \
0         0.0                         0          0       0       0   0
1         0.0                         0          0       0       0   0
2         0.0                         0          0       0       0   0
3         0.0                         0          1       0       1   0
4         0.0                         0          0       0       0   0
..        …                         …        …      …     …  ..
853       0.0                         0          0       0       0   0
854       0.0                         0          0       0       0   0
855       0.0                         0          0       0       0   0
856       0.0                         0          0       0       0   0
857       0.0                         0          0       0       0   0

     Hinselmann  Schiller  Citology  Biopsy
0             0         0         0       0
1             0         0         0       0
2             0         0         0       0
3             0         0         0       0
4             0         0         0       0
..          …       …       …     …
853           0         0         0       0
854           0         0         0       0
855           0         0         1       0
856           0         0         0       0
857           0         0         0       0

[858 rows x 34 columns]
```

2 Exploratory Data Analysis(EDA)

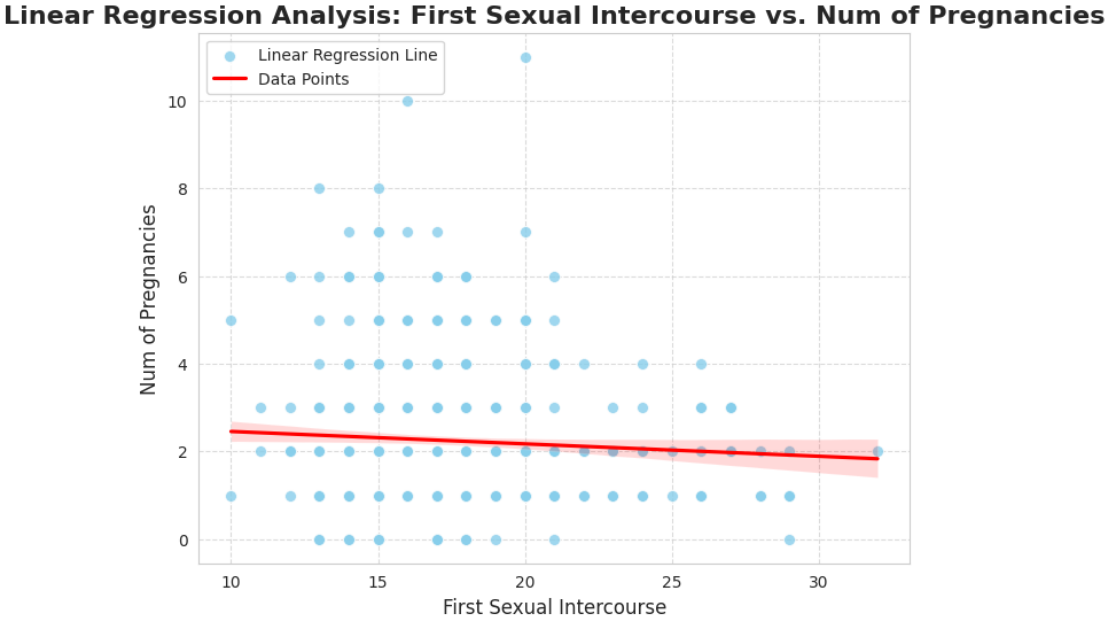
[47]:
```
import pandas as pd
import matplotlib.pyplot as plt

grouped_data = df.groupby('Num of pregnancies')['Age'].agg(['mean', 'min',␣
 ↪'max']).reset_index()
colors = plt.cm.tab20c(range(len(grouped_data)))
fig, ax = plt.subplots(figsize=(8, 8))
wedges, texts, autotexts = ax.pie(grouped_data['mean'],␣
 ↪labels=grouped_data['Num of pregnancies'], colors=colors,
                                  autopct='%.2f%%', shadow=True, startangle=90)
plt.setp(texts, size=12, weight="bold", color='black')
```

```
plt.setp(autotexts, size=12, weight="bold", color='white')
for wedge in wedges:
    wedge.set_edgecolor('white')
legend_labels = [f'{preg} Pregnancies (Avg Age: {age:.1f} years, Range:␣
↪{min_age:.1f}-{max_age:.1f} years)'
                 for preg, age, min_age, max_age in zip(grouped_data['Num of␣
↪pregnancies'],
                                                        grouped_data['mean'],
                                                        grouped_data['min'],
                                                        grouped_data['max'])]
plt.legend(legend_labels, loc='center left', fontsize=10, bbox_to_anchor=(1, 0.
↪5))
plt.title('Average Age by Num of Pregnancies', fontsize=16, weight='bold')
ax.set_aspect('equal')
plt.show()
```

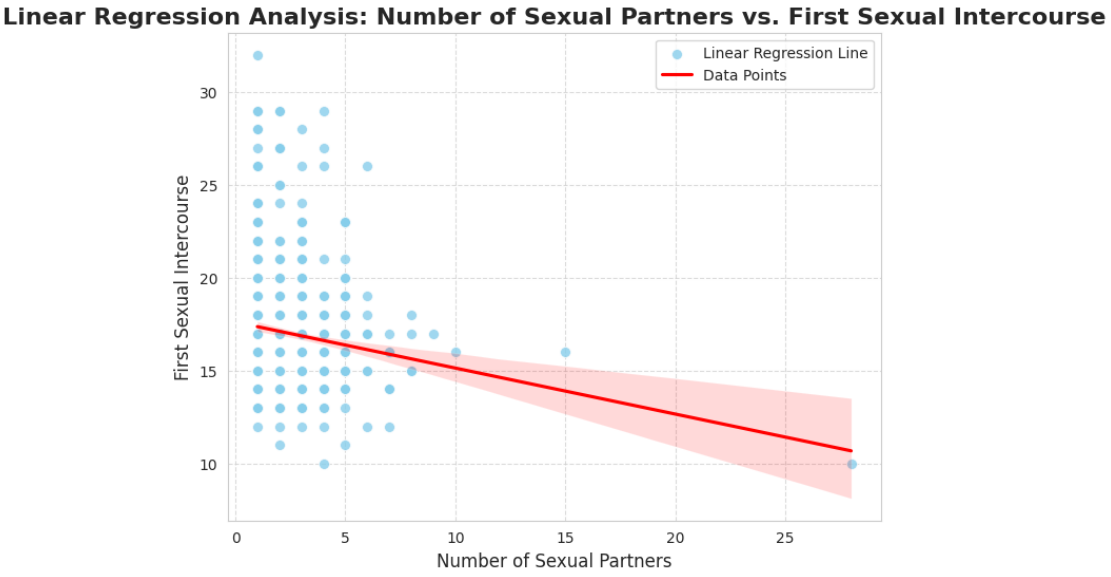


[48]:
```
import seaborn as sns
from matplotlib import pyplot as plt
sns.set_style("whitegrid")
fig, ax = plt.subplots(figsize=(8, 6))
sns.scatterplot(data=df, x='First sexual intercourse', y='Num of pregnancies',␣
↪s=50, alpha=0.8, color='skyblue', ax=ax)
sns.regplot(data=df, x='First sexual intercourse', y='Num of pregnancies',␣
↪scatter=False, color='red', ax=ax)
```

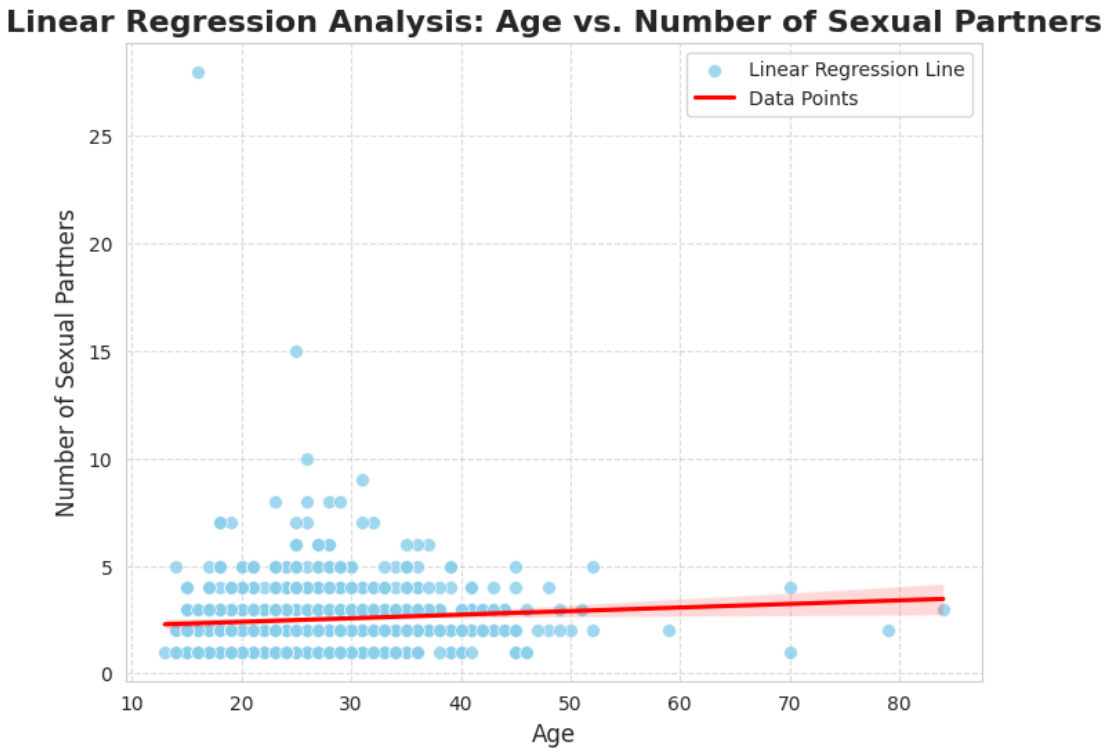
```
plt.title('Linear Regression Analysis: First Sexual Intercourse vs. Num of␣
 ↪Pregnancies', fontsize=16, weight='bold')
plt.xlabel('First Sexual Intercourse', fontsize=12)
plt.ylabel('Num of Pregnancies', fontsize=12)

plt.grid(True, linestyle='--', alpha=0.7)
plt.legend(['Linear Regression Line', 'Data Points'], loc='upper left')
plt.show()
```



[49]:
```
import seaborn as sns
from matplotlib import pyplot as plt
sns.set_style("whitegrid")
fig, ax = plt.subplots(figsize=(8, 6))
sns.scatterplot(data=df, x='Number of sexual partners', y='First sexual␣
 ↪intercourse', s=50, alpha=0.8, color='skyblue', ax=ax)
sns.regplot(data=df, x='Number of sexual partners', y='First sexual␣
 ↪intercourse', scatter=False, color='red', ax=ax)
plt.title('Linear Regression Analysis: Number of Sexual Partners vs. First␣
 ↪Sexual Intercourse', fontsize=16, weight='bold')
plt.xlabel('Number of Sexual Partners', fontsize=12)
plt.ylabel('First Sexual Intercourse', fontsize=12)

plt.grid(True, linestyle='--', alpha=0.7)
plt.legend(['Linear Regression Line', 'Data Points'], loc='upper right')
plt.show()
```

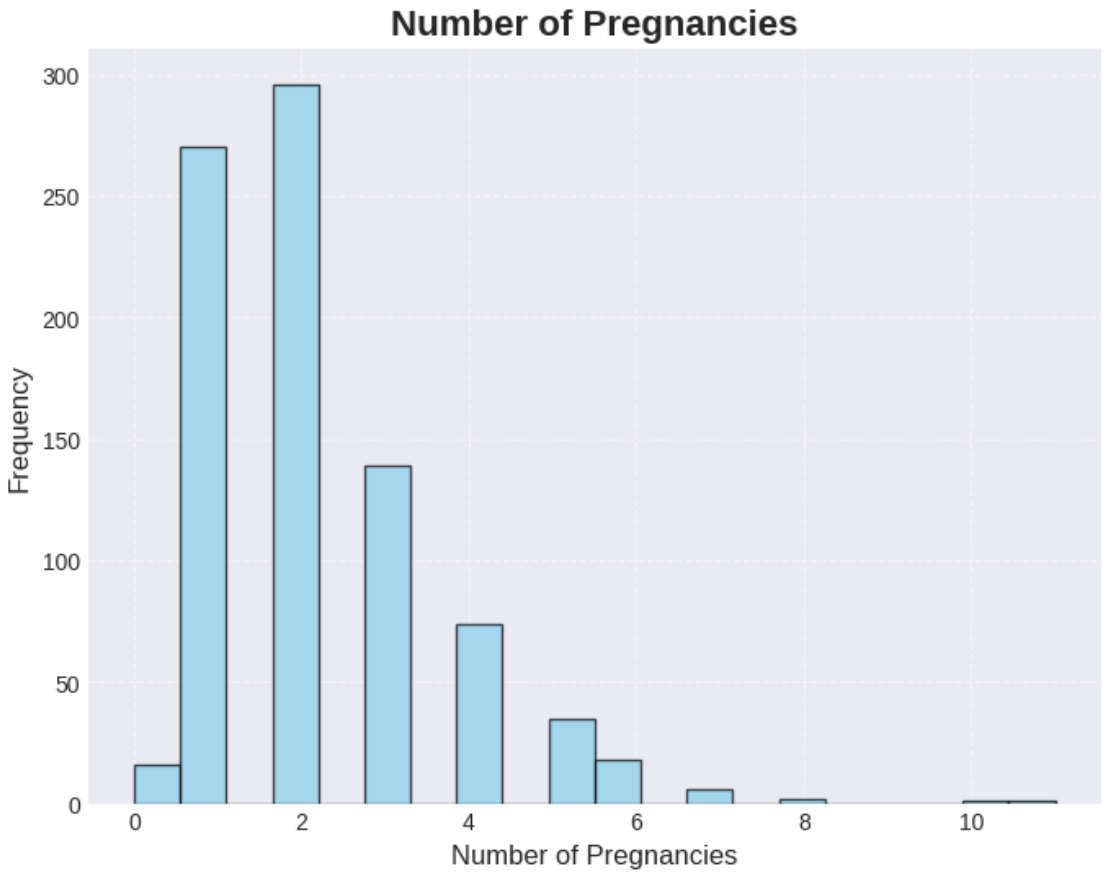


[50]:
```
import seaborn as sns
from matplotlib import pyplot as plt

sns.set_style("whitegrid")
fig, ax = plt.subplots(figsize=(8, 6))
sns.scatterplot(data=df, x='Age', y='Number of sexual partners', s=50, alpha=0.
 ↪8, color='skyblue', ax=ax)
sns.regplot(data=df, x='Age', y='Number of sexual partners', scatter=False,␣
 ↪color='red', ax=ax)

plt.title('Linear Regression Analysis: Age vs. Number of Sexual Partners',␣
 ↪fontsize=16, weight='bold')
plt.xlabel('Age', fontsize=12)
plt.ylabel('Number of Sexual Partners', fontsize=12)

plt.grid(True, linestyle='--', alpha=0.7)
plt.legend(['Linear Regression Line', 'Data Points'], loc='upper right')
plt.show()
```



[63]:
```
from matplotlib import pyplot as plt

plt.style.use('seaborn-darkgrid')
fig, ax = plt.subplots(figsize=(8, 6))
df['Num of pregnancies'].plot(kind='hist', bins=20, alpha=0.7, color='skyblue',␣
 ↪edgecolor='black', ax=ax)

plt.title('Number of Pregnancies', fontsize=16, weight='bold')
plt.xlabel('Number of Pregnancies', fontsize=12)
plt.ylabel('Frequency', fontsize=12)

ax.spines['top'].set_visible(False)
ax.spines['right'].set_visible(False)

plt.grid(True, linestyle='--', alpha=0.5)

plt.show()
```

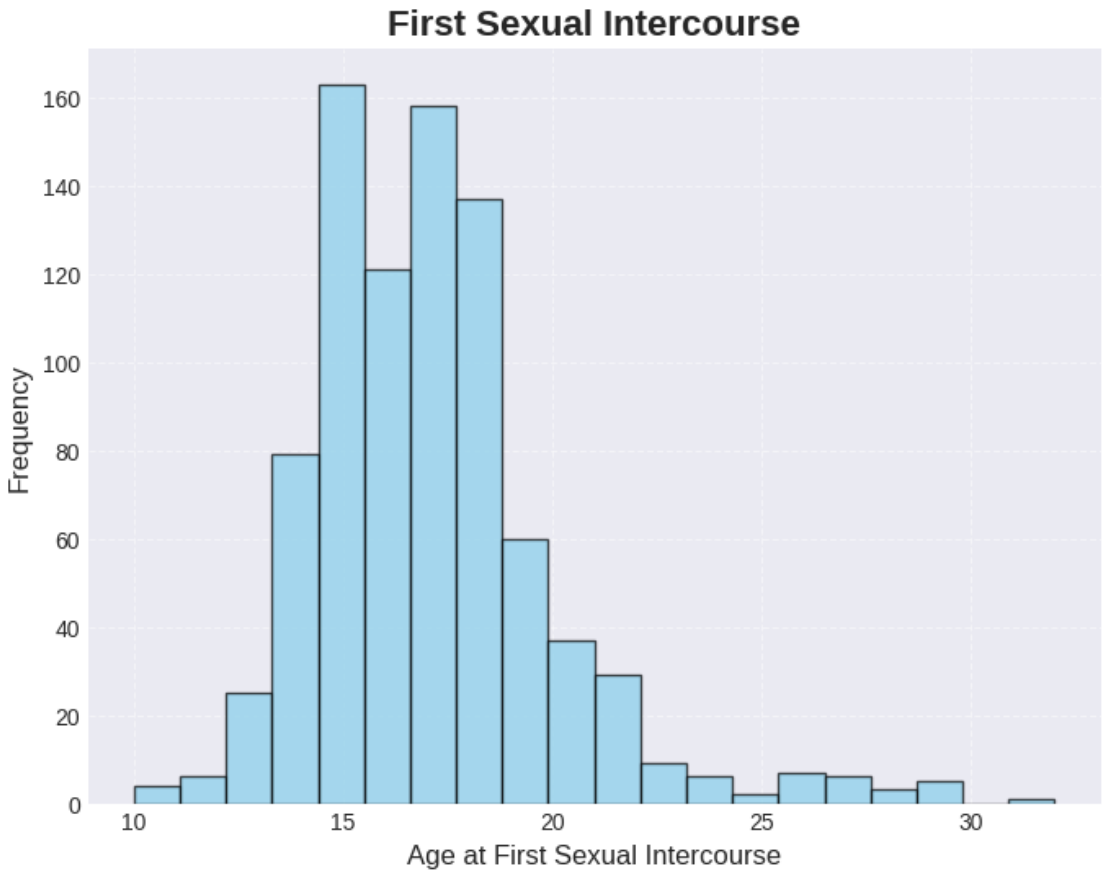
```
<ipython-input-63-ffeb28b4eb0e>:3: MatplotlibDeprecationWarning: The seaborn
styles shipped by Matplotlib are deprecated since 3.6, as they no longer
correspond to the styles shipped by seaborn. However, they will remain available
as 'seaborn-v0_8-<style>'. Alternatively, directly use the seaborn API instead.
  plt.style.use('seaborn-darkgrid')
```



[64]:
```
from matplotlib import pyplot as plt

plt.style.use('seaborn-darkgrid')
fig, ax = plt.subplots(figsize=(8, 6))
df['First sexual intercourse'].plot(kind='hist', bins=20, alpha=0.7,␣
 ↪color='skyblue', edgecolor='black', ax=ax)

plt.title('First Sexual Intercourse', fontsize=16, weight='bold')
plt.xlabel('Age at First Sexual Intercourse', fontsize=12)
plt.ylabel('Frequency', fontsize=12)

ax.spines['top'].set_visible(False)
ax.spines['right'].set_visible(False)

plt.grid(True, linestyle='--', alpha=0.5)
plt.show()
```

```
<ipython-input-64-b3e75e74aec8>:3: MatplotlibDeprecationWarning: The seaborn
styles shipped by Matplotlib are deprecated since 3.6, as they no longer
correspond to the styles shipped by seaborn. However, they will remain available
```

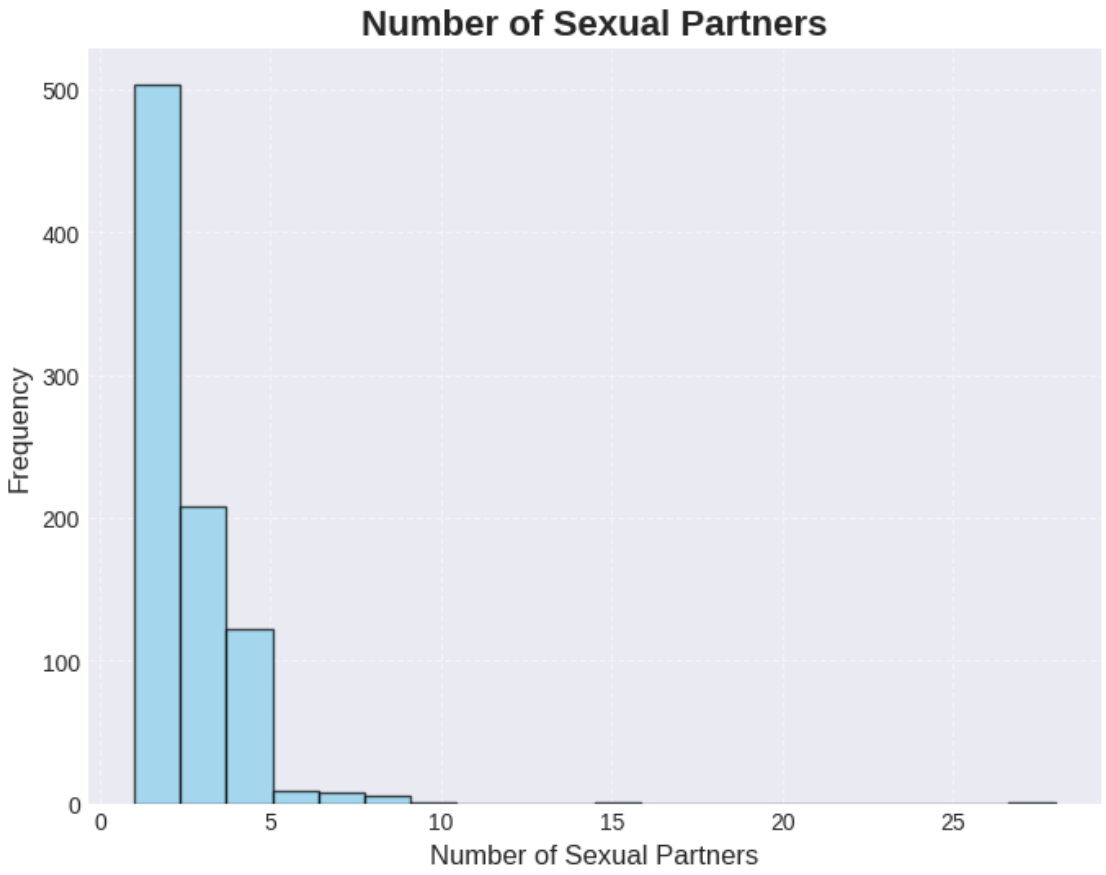
```
as 'seaborn-v0_8-<style>'. Alternatively, directly use the seaborn API instead.
  plt.style.use('seaborn-darkgrid')
```



[65]:
```
from matplotlib import pyplot as plt

plt.style.use('seaborn-darkgrid')
fig, ax = plt.subplots(figsize=(8, 6))
df['Number of sexual partners'].plot(kind='hist', bins=20, alpha=0.7,␣
 ↪color='skyblue', edgecolor='black', ax=ax)

plt.title('Number of Sexual Partners', fontsize=16, weight='bold')
plt.xlabel('Number of Sexual Partners', fontsize=12)
plt.ylabel('Frequency', fontsize=12)


ax.spines['top'].set_visible(False)
ax.spines['right'].set_visible(False)


plt.grid(True, linestyle='--', alpha=0.5)
```

```
plt.show()
```

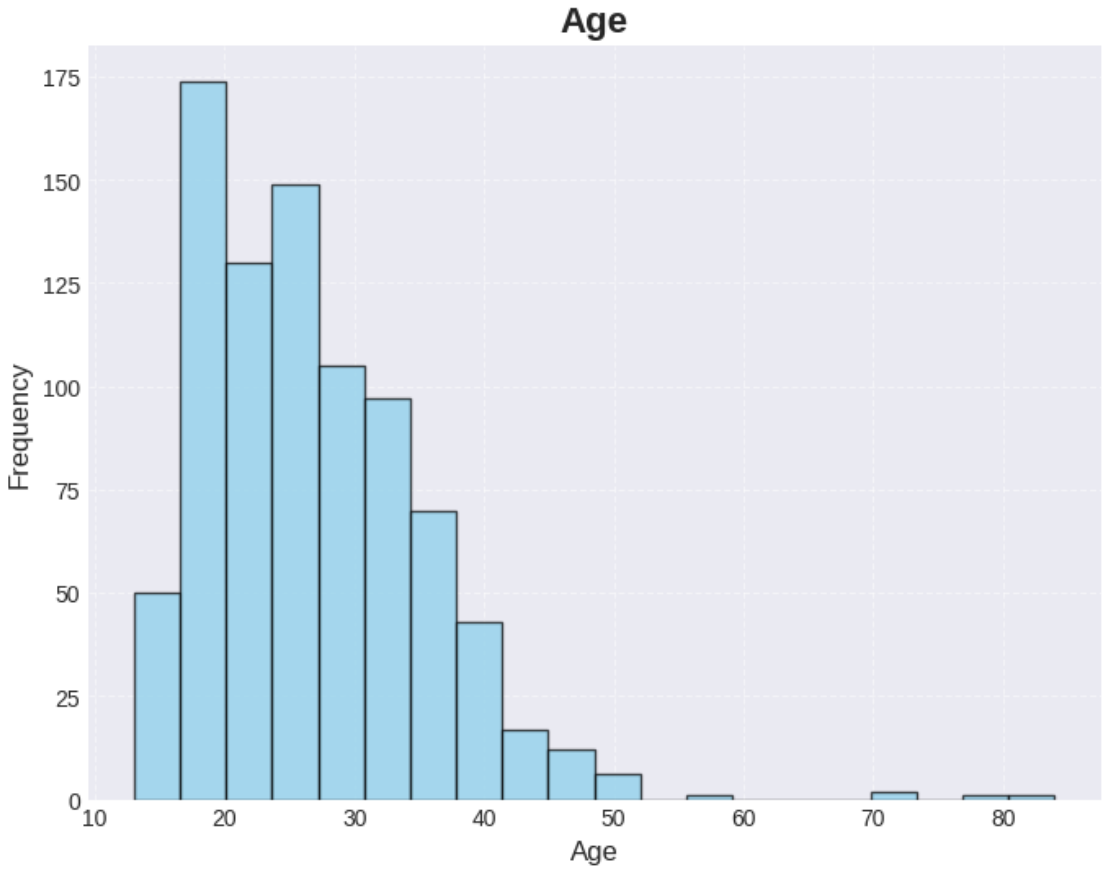
```
<ipython-input-65-80185e5cf9b6>:3: MatplotlibDeprecationWarning: The seaborn
styles shipped by Matplotlib are deprecated since 3.6, as they no longer
correspond to the styles shipped by seaborn. However, they will remain available
as 'seaborn-v0_8-<style>'. Alternatively, directly use the seaborn API instead.
  plt.style.use('seaborn-darkgrid')
```



[66]:
```
from matplotlib import pyplot as plt

plt.style.use('seaborn-darkgrid')
fig, ax = plt.subplots(figsize=(8, 6))
df['Age'].plot(kind='hist', bins=20, alpha=0.7, color='skyblue',␣
 ↪edgecolor='black', ax=ax)

plt.title('Age', fontsize=16, weight='bold')
plt.xlabel('Age', fontsize=12)
plt.ylabel('Frequency', fontsize=12)
```

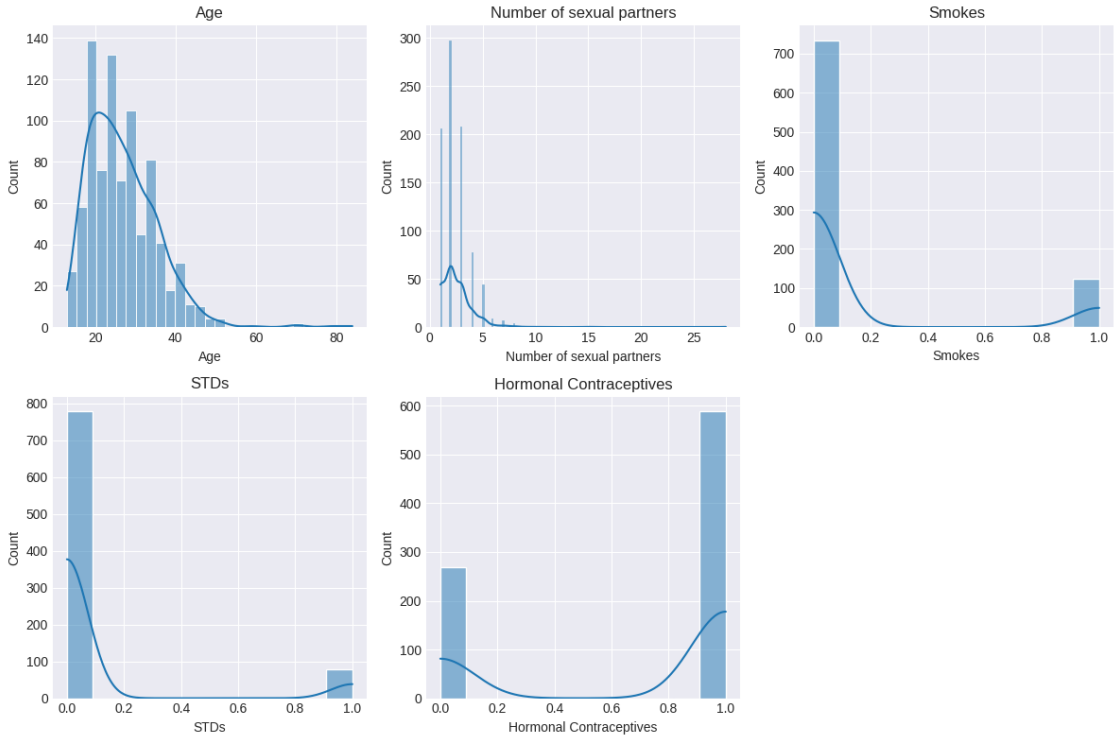
```
ax.spines['top'].set_visible(False)
ax.spines['right'].set_visible(False)

plt.grid(True, linestyle='--', alpha=0.5)
plt.show()
```

```
<ipython-input-66-98095f1409d0>:3: MatplotlibDeprecationWarning: The seaborn
styles shipped by Matplotlib are deprecated since 3.6, as they no longer
correspond to the styles shipped by seaborn. However, they will remain available
as 'seaborn-v0_8-<style>'. Alternatively, directly use the seaborn API instead.
  plt.style.use('seaborn-darkgrid')
```

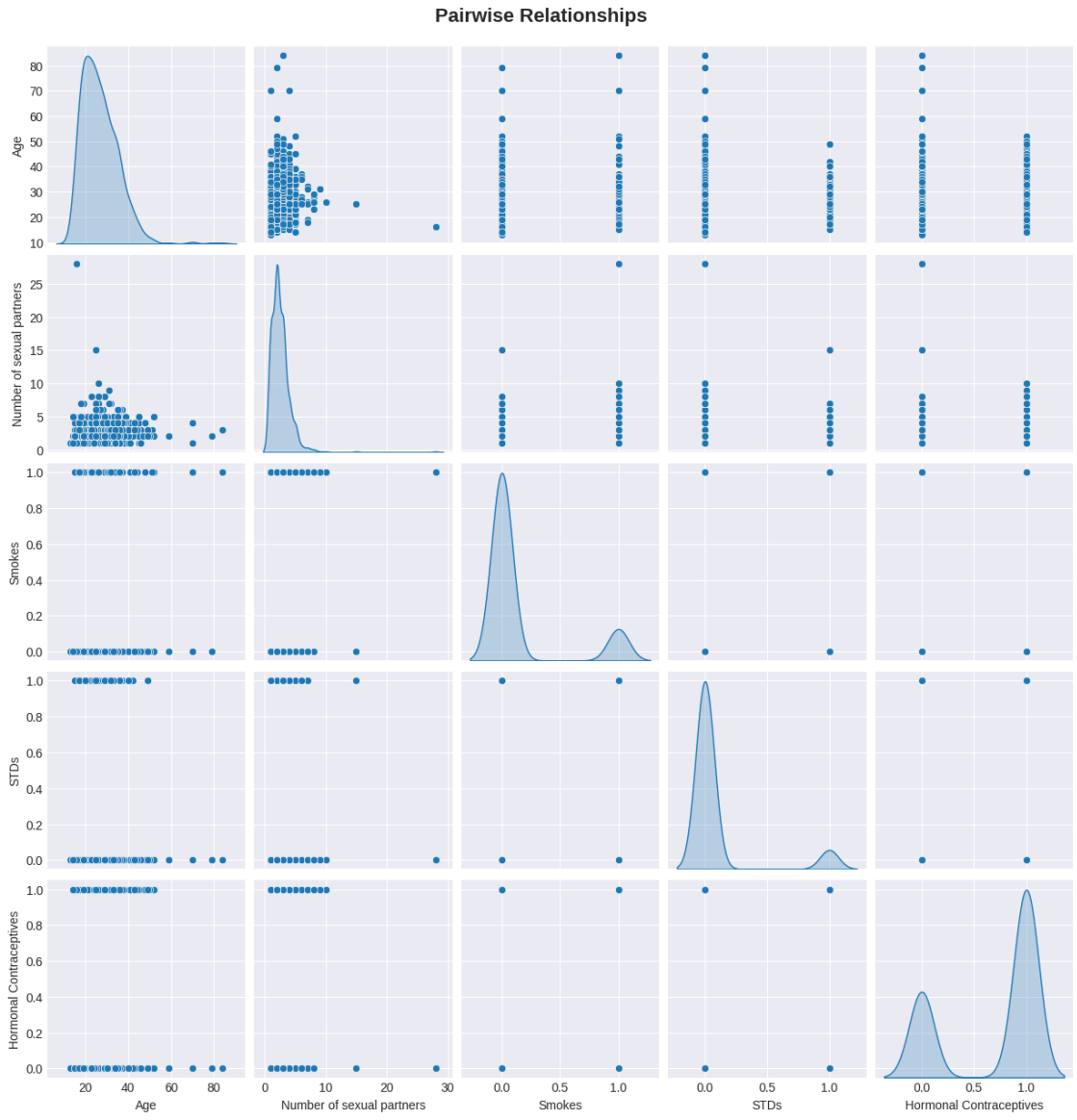


[72]:
```
import seaborn as sns
import matplotlib.pyplot as plt

# Select the important variables
important_variables = ['Age', 'Number of sexual partners', 'Smokes', 'STDs',␣
 ↪'Hormonal Contraceptives']
```

```
# Subset the dataframe
selected_df = df[important_variables]

# Plot histograms for each variable
plt.figure(figsize=(12, 8))
for i, variable in enumerate(selected_df.columns):
    plt.subplot(2, 3, i+1)
    sns.histplot(selected_df[variable], kde=True)
    plt.title(variable)
plt.tight_layout()
plt.show()
```

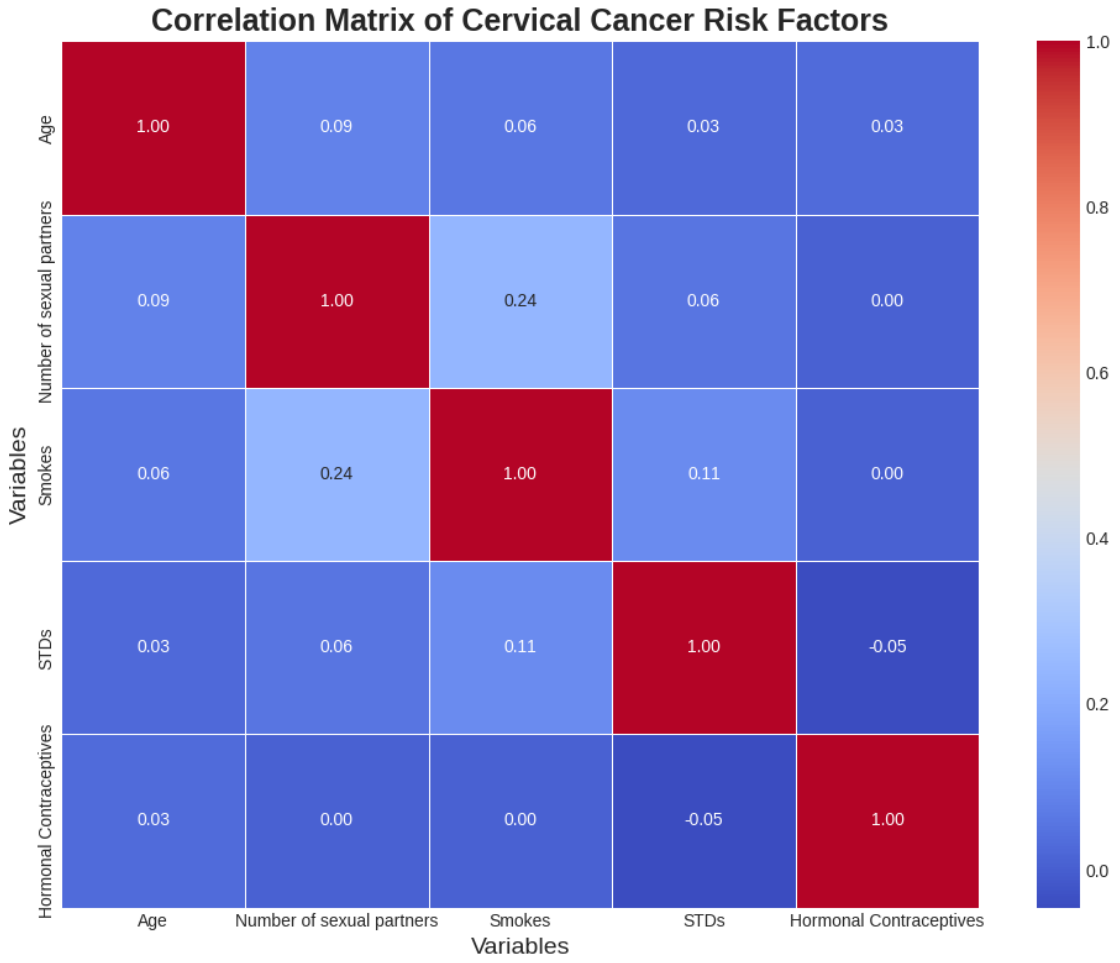


[73]:
```
# Plot scatter plots for pairwise relationships
sns.pairplot(selected_df, diag_kind='kde')
plt.suptitle('Pairwise Relationships', y=1.02, fontsize=16, weight='bold')
plt.show()
```



[79]:
```
import seaborn as sns
import matplotlib.pyplot as plt

correlation_matrix = selected_df.corr()
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f",␣
 ↪linewidths=0.5)
plt.title('Correlation Matrix of Cervical Cancer Risk Factors', fontsize=18,␣
 ↪weight='bold')

plt.xlabel('Variables', fontsize=14)
plt.ylabel('Variables', fontsize=14)
```

```
plt.tight_layout()
plt.show()
```



[81]:
```
print("Summary of Correlations:")
for i in range(len(correlation_matrix)):
    for j in range(i+1, len(correlation_matrix)):
        variable1 = correlation_matrix.index[i]
        variable2 = correlation_matrix.columns[j]
        correlation = correlation_matrix.iloc[i, j]
        if correlation > 0:
            print(f"The correlation between '{variable1}' and '{variable2}' is␣
 ↪positive: {correlation:.2f}")
        elif correlation < 0:
            print(f"The correlation between '{variable1}' and '{variable2}' is␣
 ↪negative: {correlation:.2f}")
```

```
Summary of Correlations:
The correlation between 'Age' and 'Number of sexual partners' is positive: 0.09
The correlation between 'Age' and 'Smokes' is positive: 0.06
The correlation between 'Age' and 'STDs' is positive: 0.03
The correlation between 'Age' and 'Hormonal Contraceptives' is positive: 0.03
The correlation between 'Number of sexual partners' and 'Smokes' is positive:
0.24
The correlation between 'Number of sexual partners' and 'STDs' is positive: 0.06
The correlation between 'Number of sexual partners' and 'Hormonal
Contraceptives' is positive: 0.00
The correlation between 'Smokes' and 'STDs' is positive: 0.11
The correlation between 'Smokes' and 'Hormonal Contraceptives' is positive: 0.00
The correlation between 'STDs' and 'Hormonal Contraceptives' is negative: -0.05
```

3 Exploratory Data Analysis (EDA) in Linear Regression Analysis of Cervical Cancer and Correlation Summary

1. **Age and Sexual Behavior**:
   - A weak positive correlation is observed between Age and the Number of sexual partners (0.09). This implies that, on average, older individuals tend to have slightly more sexual partners. Means that this is not that significant
2. **Age and Lifestyle Factors**:
   - There is a weak positive correlation between Age and Smokes (0.06). Additionally, weak positive correlations exist between Age and the STDs (0.03) and Hormonal Contraceptives (0.03). These correlations suggest that between Age and lifestyle factors are not highly influential/significant.
3. **Sexual Behavior and Lifestyle Choices**:
   - A moderate positive correlation is found between the Number of sexual partners and Smokes (0.24). This suggests relationship with more sexual partners are more likely to smoke, and vice versa.
4. **STDs and Contraceptive Use**:
   - A weak positive correlation is observed between Smokes and STDs (0.11). however in the opposite there is a weak negative correlation exists between STDs and Hormonal Contraceptives (-0.05). These findings between STD status, contraceptive choices, and lifestyle factors gave us insight on how impactful it is in cervical cancer risk.

##**indicate better model performance**

[121]:
```
from sklearn.datasets import load_iris
from sklearn.linear_model import LinearRegression
import matplotlib.pyplot as plt
```

[122]:
```
iris = load_iris()
X = iris.data[:, 2].reshape(-1, 1)
y = iris.data[:, 3]
```

[123]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,␣
↪random_state=42)
```

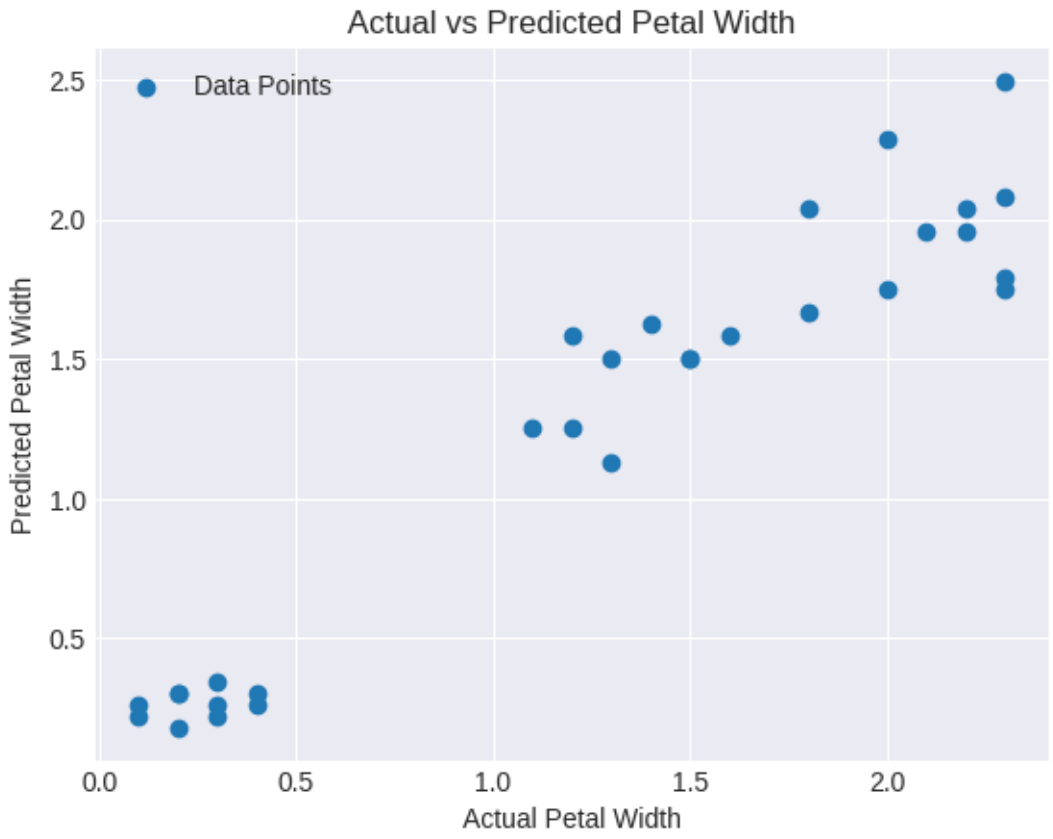
[124]:
```
model = LinearRegression()
```

[125]:
```
model.fit(X_train, y_train)
```

[125]:
```
LinearRegression()
```

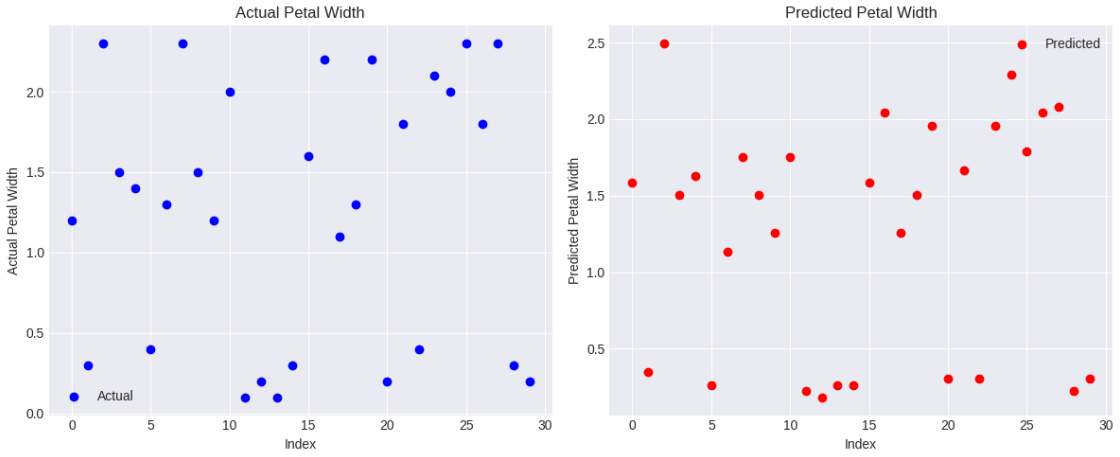
[126]:
```
y_pred = model.predict(X_test)
```

[127]:
```
import matplotlib.pyplot as plt


plt.scatter(y_test, y_pred, label='Data Points')
plt.xlabel('Actual Petal Width')
plt.ylabel('Predicted Petal Width')
plt.title('Actual vs Predicted Petal Width')


plt.legend()

plt.show()
```

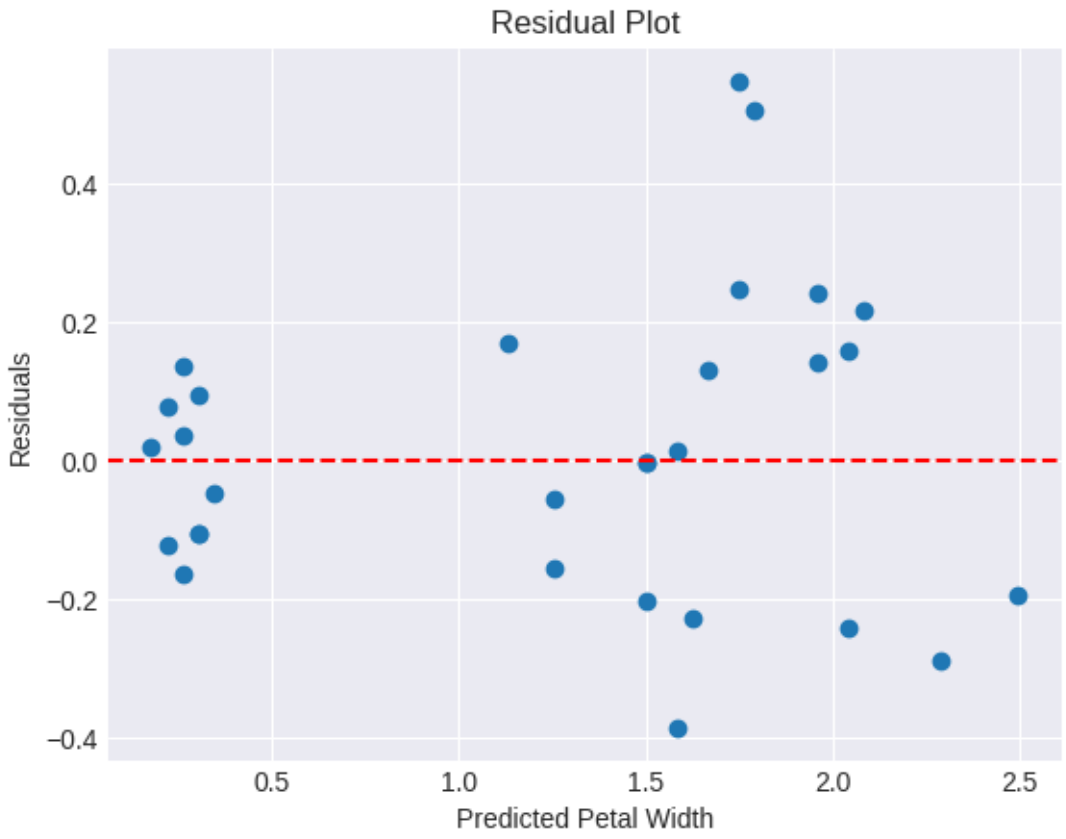


[128]:
```
import matplotlib.pyplot as plt

# Create subplots
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 5))

# Plot for Actual Petal Width
ax1.scatter(range(len(y_test)), y_test, color='blue', label='Actual')
ax1.set_xlabel('Index')
ax1.set_ylabel('Actual Petal Width')
ax1.set_title('Actual Petal Width')
ax1.legend()

# Plot for Predicted Petal Width
ax2.scatter(range(len(y_pred)), y_pred, color='red', label='Predicted')
ax2.set_xlabel('Index')
ax2.set_ylabel('Predicted Petal Width')
ax2.set_title('Predicted Petal Width')
ax2.legend()

plt.tight_layout()
```

```
plt.show()
```



[129]:
```
residuals = y_test - y_pred
plt.scatter(y_pred, residuals)
plt.xlabel('Predicted Petal Width')
plt.ylabel('Residuals')
plt.title('Residual Plot')
plt.axhline(y=0, color='r', linestyle='--')
plt.show()
```



3.1 Residual

- A well-performing model will have residuals scattered randomly around zero (the red dashed line).

[130]:
```
from sklearn.metrics import mean_squared_error, r2_score

# Calculate Mean Squared Error
mse = mean_squared_error(y_test, y_pred)

# Calculate R-squared
r2 = r2_score(y_test, y_pred)

# Print summary
print("Summary of Actual vs. Predicted Values:")
print(f"Mean Squared Error (MSE): {mse:.2f}")
print(f"R-squared: {r2:.2f}")
```

```
Summary of Actual vs. Predicted Values:
Mean Squared Error (MSE): 0.05
R-squared: 0.93
```

3.2 Mean Squared Error (MSE):

- A lower MSE indicates better accuracy.

3.3 R-squared (R^2):

- Higher R-squared values closer to 1 signify a better fit of the model to the data.