Final Project Notebook

Cervical Cancer Risk Factor Analysis

Cervical Cancer Risk Factors for Biopsy: This Dataset is Obtained from UCI Repository and kindly acknowledged!

This file contains a list of Risk Factors for cervical cancer that collected before Biopsy Examination!

About 11,000 new cases of invasive cervical cancer are diagnosed each year in the U.S. However, the number of new cervical cancer cases has been declining steadily over the past decades. Although it is the most preventable type of cancer, each year cervical cancer kills about 4,000 women in the U.S. and about 300,000 women worldwide. In the United States, cervical cancer mortality rates plunged by 74% from 1955 - 1992 thanks to increased screening and early detection with the Pap test. **Age** Fifty percent of cervical cancer diagnoses occur in women ages 35 - 54, and about 20% occur in women over 65 years of age. The median age of diagnosis is 48 years. About 15% of women develop cervical cancer between the ages of 20 - 30. Cervical cancer is extremely rare in women younger than age 20. However, many young women become infected with multiple types of human papilloma virus, which then can increase their risk of getting cervical cancer in the future. Young women with early abnormal changes who do not have regular examinations are at high risk for localized cancer by the time they are age 40, and for invasive cancer by age 50. **SOCIOECONOMIC AND ETHNIC FACTORS** Although the rate of cervical cancer has declined among both Caucasian and African-American women over the past decades, it remains much more prevalent in African-Americans -- whose death rates are twice as high as Caucasian women. Hispanic American women have more than twice the risk of invasive cervical cancer as Caucasian women, also due to a lower rate of screening. These differences, however, are almost certainly due to social and economic differences. Numerous studies report that high poverty levels are linked with low screening rates. In addition, lack of health insurance, limited transportation, and language difficulties hinder a poor woman’s access to screening services. **HIGH SEXUAL ACTIVITY** Human papilloma virus (HPV) is the main risk factor for cervical cancer. In adults, the most important risk factor for HPV is sexual activity with an infected person. Women most at risk for cervical cancer are those with a history of multiple sexual partners, sexual intercourse at age 17 years or younger, or both. A woman who has never been sexually active has a very low risk for developing cervical cancer. Sexual activity with multiple partners increases the likelihood of many other sexually transmitted infections (chlamydia, gonorrhea, syphilis).Studies have found an association between chlamydia and cervical cancer risk, including the possibility that chlamydia may prolong HPV infection. **FAMILY HISTORY** Women have a higher risk of cervical cancer if they have a first-degree relative (mother, sister) who has had cervical cancer. **USE OF ORAL CONTRACEPTIVES** Studies have reported a strong association between cervical cancer and long-term use of oral contraception (OC). Women who take birth control pills for more than 5 - 10 years appear to have a much higher risk HPV infection (up to four times higher) than those who do not use OCs. (Women taking OCs for fewer than 5 years do not have a significantly higher risk.) The reasons for this risk from OC use are not entirely clear. Women who use OCs may be less likely to use a diaphragm, condoms, or other methods that offer some protection against sexual transmitted diseases, including HPV. Some research also suggests that the hormones in OCs might help the virus enter the genetic material of cervical cells. **HAVING MANY CHILDREN** Studies indicate that having many children increases the risk for developing cervical cancer, particularly in women infected with HPV. SMOKING Smoking is associated with a higher risk for precancerous changes (dysplasia) in the cervix and for progression to invasive cervical cancer, especially for women infected with HPV. **IMMUNOSUPPRESSION** Women with weak immune systems, (such as those with HIV / AIDS), are more susceptible to acquiring HPV. Immunocompromised patients are also at higher risk for having cervical precancer develop rapidly into invasive cancer. DIETHYLSTILBESTROL (DES) From 1938 - 1971, diethylstilbestrol (DES), an estrogen-related drug, was widely prescribed to pregnant women to help prevent miscarriages. The daughters of these women face a higher risk for cervical cancer. DES is no longer prsecribed.

Firstly, Let us import the necessary libraries to explore and analyze the data set...

In [1]:
```
import pandas as pd
import numpy as np
import random
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

Now, let us load the dataset....

In [2]:
```
df = pd.read_csv("/home/shahir/data/Risk_factors_cervical_cancer.csv")
```

Next, we will check how the data frame look like by displaying first 5 rows...

Exploring Data

In [3]:
```
df.head()
```

Out[3]:

	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormona Contrace
0	18	4.0	15.0	1.0	0.0	0.0	0.0	0.0
1	15	1.0	14.0	1.0	0.0	0.0	0.0	0.0
2	34	1.0	?	1.0	0.0	0.0	0.0	0.0
3	52	5.0	16.0	4.0	1.0	37.0	37.0	1.0
4	46	3.0	21.0	4.0	0.0	0.0	0.0	1.0

5 rows × 36 columns

Since we could not see the complete column, we will check all the columns in the data frame...

In [4]:
```
df.columns.values
```

Out[4]:
```
array(['Age', 'Number of sexual partners', 'First sexual interco
urse',
       '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 condylomato
sis',
       '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 diagnosi
s',
       'STDs: Time since first diagnosis',
       'STDs: Time since last diagnosis', 'Dx:Cancer', 'Dx:CIN',
'Dx:HPV',
       'Dx', 'Hinselmann', 'Schiller', 'Citology', 'Biopsy'], dt
ype=object)
```

As we see, there are many columns that are important for our analysis.... So let us see the first 10 rows for all columns by using df.head(10).transpose() command...

In [5]:
```
df.head(10).transpose()
```

Out[5]:

	0	1	2	3	4	5	6	7	8	9
Age	18	15	34	52	46	42	51	26	45	44
Number of sexual partners	4.0	1.0	1.0	5.0	3.0	3.0	3.0	1.0	1.0	3.0
First sexual intercourse	15.0	14.0	?	16.0	21.0	23.0	17.0	26.0	20.0	15.0
Num of pregnancies	1.0	1.0	1.0	4.0	4.0	2.0	6.0	3.0	5.0	?
Smokes	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0
Smokes (years)	0.0	0.0	0.0	37.0	0.0	0.0	34.0	0.0	0.0	1.266972909
Smokes (packs/year)	0.0	0.0	0.0	37.0	0.0	0.0	3.4	0.0	0.0	2.8
Hormonal Contraceptives	0.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0
Hormonal Contraceptives (years)	0.0	0.0	0.0	3.0	15.0	0.0	0.0	2.0	0.0	0.0
IUD	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	?
IUD (years)	0.0	0.0	0.0	0.0	0.0	0.0	7.0	7.0	0.0	?
STDs	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STDs (number)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STDs:condylomatosis	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STDs:cervical condylomatosis	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STDs:vaginal condylomatosis	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STDs:vulvo-perineal condylomatosis	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STDs:syphilis	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STDs:pelvic inflammatory disease	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STDs:genital herpes	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STDs:molluscum contagiosum	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STDs:AIDS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STDs:HIV	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STDs:Hepatitis B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STDs:HPV	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STDs: Number of diagnosis	0	0	0	0	0	0	0	0	0	0
STDs: Time since first diagnosis	?	?	?	?	?	?	?	?	?	?

In [6]:
```
df.shape
```

Out[6]:
```
(858, 36)
```

Next, let us explore the descriptive statistics on our data set to get more insight in to the data.

In [7]:
```
df.describe().transpose()
```

Out[7]:

	count	mean	std	min	25%	50%	75%	max
Age	858.0	26.820513	8.497948	13.0	20.0	25.0	32.0	84.0
STDs: Number of diagnosis	858.0	0.087413	0.302545	0.0	0.0	0.0	0.0	3.0
Dx:Cancer	858.0	0.020979	0.143398	0.0	0.0	0.0	0.0	1.0
Dx:CIN	858.0	0.010490	0.101939	0.0	0.0	0.0	0.0	1.0
Dx:HPV	858.0	0.020979	0.143398	0.0	0.0	0.0	0.0	1.0
Dx	858.0	0.027972	0.164989	0.0	0.0	0.0	0.0	1.0
Hinselmann	858.0	0.040793	0.197925	0.0	0.0	0.0	0.0	1.0
Schiller	858.0	0.086247	0.280892	0.0	0.0	0.0	0.0	1.0
Citology	858.0	0.051282	0.220701	0.0	0.0	0.0	0.0	1.0
Biopsy	858.0	0.064103	0.245078	0.0	0.0	0.0	0.0	1.0

We could see details of only 10 columns. This means that other columns are non numeric.... So we will use dataFrame.describe(include="all") method to get the details of all columns.

In [8]:
```
df.describe(include="all").transpose()
```

Out[8]:

	count	unique	top	freq	mean	std	min	25%	50%
Age	858	NaN	NaN	NaN	26.8205	8.49795	13	20	25
Number of sexual partners	858	13	2.0	272	NaN	NaN	NaN	NaN	Na
First sexual intercourse	858	22	15.0	163	NaN	NaN	NaN	NaN	Na
Num of pregnancies	858	12	1.0	270	NaN	NaN	NaN	NaN	Na
Smokes	858	3	0.0	722	NaN	NaN	NaN	NaN	Na
Smokes (years)	858	31	0.0	722	NaN	NaN	NaN	NaN	Na
Smokes (packs/year)	858	63	0.0	722	NaN	NaN	NaN	NaN	Na
Hormonal Contraceptives	858	3	1.0	481	NaN	NaN	NaN	NaN	Na
Hormonal Contraceptives (years)	858	41	0.0	269	NaN	NaN	NaN	NaN	Na
IUD	858	3	0.0	658	NaN	NaN	NaN	NaN	Na
IUD (years)	858	27	0.0	658	NaN	NaN	NaN	NaN	Na
STDs	858	3	0.0	674	NaN	NaN	NaN	NaN	Na
STDs (number)	858	6	0.0	674	NaN	NaN	NaN	NaN	Na
STDs:condylomatosis	858	3	0.0	709	NaN	NaN	NaN	NaN	Na
STDs:cervical condylomatosis	858	2	0.0	753	NaN	NaN	NaN	NaN	Na
STDs:vaginal condylomatosis	858	3	0.0	749	NaN	NaN	NaN	NaN	Na
STDs:vulvo-perineal condylomatosis	858	3	0.0	710	NaN	NaN	NaN	NaN	Na
STDs:syphilis	858	3	0.0	735	NaN	NaN	NaN	NaN	Na
STDs:pelvic inflammatory disease	858	3	0.0	752	NaN	NaN	NaN	NaN	Na
STDs:genital herpes	858	3	0.0	752	NaN	NaN	NaN	NaN	Na
STDs:molluscum contagiosum	858	3	0.0	752	NaN	NaN	NaN	NaN	Na
STDs:AIDS	858	2	0.0	753	NaN	NaN	NaN	NaN	Na
STDs:HIV	858	3	0.0	735	NaN	NaN	NaN	NaN	Na
STDs:Hepatitis B	858	3	0.0	752	NaN	NaN	NaN	NaN	Na
STDs:HPV	858	3	0.0	751	NaN	NaN	NaN	NaN	Na
STDs: Number of diagnosis	858	NaN	NaN	NaN	0.0874126	0.302545	0	0	0
STDs: Time since first diagnosis	858	19	?	787	NaN	NaN	NaN	NaN	Na

In [9]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 858 entries, 0 to 857
Data columns (total 36 columns):
Age                                   858 non-null int64
Number of sexual partners             858 non-null object
First sexual intercourse              858 non-null object
Num of pregnancies                    858 non-null object
Smokes                                858 non-null object
Smokes (years)                        858 non-null object
Smokes (packs/year)                   858 non-null object
Hormonal Contraceptives               858 non-null object
Hormonal Contraceptives (years)       858 non-null object
IUD                                   858 non-null object
IUD (years)                           858 non-null object
STDs                                  858 non-null object
STDs (number)                         858 non-null object
STDs:condylomatosis                   858 non-null object
STDs:cervical condylomatosis          858 non-null object
STDs:vaginal condylomatosis           858 non-null object
STDs:vulvo-perineal condylomatosis    858 non-null object
STDs:syphilis                         858 non-null object
STDs:pelvic inflammatory disease      858 non-null object
STDs:genital herpes                   858 non-null object
STDs:molluscum contagiosum            858 non-null object
STDs:AIDS                             858 non-null object
STDs:HIV                              858 non-null object
STDs:Hepatitis B                      858 non-null object
STDs:HPV                              858 non-null object
STDs: Number of diagnosis             858 non-null int64
STDs: Time since first diagnosis      858 non-null object
STDs: Time since last diagnosis       858 non-null object
Dx:Cancer                             858 non-null int64
Dx:CIN                                858 non-null int64
Dx:HPV                                858 non-null int64
Dx                                    858 non-null int64
Hinselmann                            858 non-null int64
Schiller                              858 non-null int64
Citology                              858 non-null int64
Biopsy                                858 non-null int64
dtypes: int64(10), object(26)
memory usage: 241.4+ KB
```

Data Cleaning: Handling Missing Data

As we can see that there are 26 object type data types in our data set....We need to convert all the objective data type columns to numeric data type for our analysis. We will do it later.

Now let us examine how many null values are there...

In [12]:
```
df.isnull().any()
```

Out[12]:
```
Age                                    False
Number of sexual partners               True
First sexual intercourse                True
Num of pregnancies                      True
Smokes                                  True
Smokes (years)                          True
Smokes (packs/year)                     True
Hormonal Contraceptives                 True
Hormonal Contraceptives (years)         True
IUD                                     True
IUD (years)                             True
STDs                                    True
STDs (number)                           True
STDs:condylomatosis                     True
STDs:cervical condylomatosis            True
STDs:vaginal condylomatosis             True
STDs:vulvo-perineal condylomatosis      True
STDs:syphilis                           True
STDs:pelvic inflammatory disease        True
STDs:genital herpes                     True
STDs:molluscum contagiosum              True
STDs:AIDS                               True
STDs:HIV                                True
STDs:Hepatitis B                        True
STDs:HPV                                True
STDs: Number of diagnosis              False
STDs: Time since first diagnosis        True
STDs: Time since last diagnosis         True
Dx:Cancer                              False
Dx:CIN                                 False
Dx:HPV                                 False
Dx                                     False
Hinselmann                             False
Schiller                               False
Citology                               False
Biopsy                                 False
dtype: bool
```

seems to be that non value columns are filled with "?" character. So we need covert those values to "NaN" character...

In [13]:
```
df = df.replace('?', np.nan)
```

In [14]:
```
df.isnull().any()
```

Out[14]:
```
Age                                    False
Number of sexual partners               True
First sexual intercourse                True
Num of pregnancies                      True
Smokes                                  True
Smokes (years)                          True
Smokes (packs/year)                     True
Hormonal Contraceptives                 True
Hormonal Contraceptives (years)         True
IUD                                     True
IUD (years)                             True
STDs                                    True
STDs (number)                           True
STDs:condylomatosis                     True
STDs:cervical condylomatosis            True
STDs:vaginal condylomatosis             True
STDs:vulvo-perineal condylomatosis      True
STDs:syphilis                           True
STDs:pelvic inflammatory disease        True
STDs:genital herpes                     True
STDs:molluscum contagiosum              True
STDs:AIDS                               True
STDs:HIV                                True
STDs:Hepatitis B                        True
STDs:HPV                                True
STDs: Number of diagnosis              False
STDs: Time since first diagnosis        True
STDs: Time since last diagnosis         True
Dx:Cancer                              False
Dx:CIN                                 False
Dx:HPV                                 False
Dx                                     False
Hinselmann                             False
Schiller                               False
Citology                               False
Biopsy                                 False
dtype: bool
```

In [15]:
```
df.isnull().sum()
```

Out[15]:
```
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
```

Now, we can see that there are many NaN values are in objective data type columns....

Also, two columns "STDs: Time since first diagnosis" and "STDs: Time since last diagnosis" are having majority of the values as NaN. So we will drop these two columns.

In [16]:
```
df = df.drop(['STDs: Time since first diagnosis', 'STDs: Time si
nce last diagnosis'], axis=1)
```

In [17]:
```
df.shape
```

Out[17]:
```
(858, 34)
```

Fixing Null Values by Deleting Them

For the easy analysis, we will drop all the rows containing Nan.

In [18]:
```
df = df.dropna(axis=0)
```

In [19]:
```
df.shape
```

Out[19]:
```
(668, 34)
```

Now only 34 columns are available in our data set.

Now let us convert all columns to numeric data type....

In [20]:
```
df = df.convert_objects(convert_numeric=True)
```

```
/home/shahir/anaconda2/lib/python2.7/site-packages/ipykernel/__m
ain__.py:1: FutureWarning: convert_objects is deprecated.  Use t
he data-type specific converters pd.to_datetime, pd.to_timedelta
and pd.to_numeric.
  if __name__ == '__main__':
```

In [21]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 668 entries, 0 to 857
Data columns (total 34 columns):
Age                                   668 non-null int64
Number of sexual partners             668 non-null float64
First sexual intercourse              668 non-null float64
Num of pregnancies                    668 non-null float64
Smokes                                668 non-null float64
Smokes (years)                        668 non-null float64
Smokes (packs/year)                   668 non-null float64
Hormonal Contraceptives               668 non-null float64
Hormonal Contraceptives (years)       668 non-null float64
IUD                                   668 non-null float64
IUD (years)                           668 non-null float64
STDs                                  668 non-null float64
STDs (number)                         668 non-null float64
STDs:condylomatosis                   668 non-null float64
STDs:cervical condylomatosis          668 non-null float64
STDs:vaginal condylomatosis           668 non-null float64
STDs:vulvo-perineal condylomatosis    668 non-null float64
STDs:syphilis                         668 non-null float64
STDs:pelvic inflammatory disease      668 non-null float64
STDs:genital herpes                   668 non-null float64
STDs:molluscum contagiosum            668 non-null float64
STDs:AIDS                             668 non-null float64
STDs:HIV                              668 non-null float64
STDs:Hepatitis B                      668 non-null float64
STDs:HPV                              668 non-null float64
STDs: Number of diagnosis             668 non-null int64
Dx:Cancer                             668 non-null int64
Dx:CIN                                668 non-null int64
Dx:HPV                                668 non-null int64
Dx                                    668 non-null int64
Hinselmann                            668 non-null int64
Schiller                              668 non-null int64
Citology                              668 non-null int64
Biopsy                                668 non-null int64
dtypes: float64(24), int64(10)
memory usage: 182.7 KB
```

In [22]:
```
df.head()
```

Out[22]:

	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormona Contrace
0	18	4.0	15.0	1.0	0.0	0.0	0.0	0.0
1	15	1.0	14.0	1.0	0.0	0.0	0.0	0.0
3	52	5.0	16.0	4.0	1.0	37.0	37.0	1.0
4	46	3.0	21.0	4.0	0.0	0.0	0.0	1.0
5	42	3.0	23.0	2.0	0.0	0.0	0.0	0.0

5 rows × 34 columns

In [23]:
```
df.head(10).transpose()
```

Out[23]:

	0	1	3	4	5	6	7	8	10	11
Age	18.0	15.0	52.0	46.0	42.0	51.0	26.0	45.0	44.0	27.0
Number of sexual partners	4.0	1.0	5.0	3.0	3.0	3.0	1.0	1.0	3.0	1.0
First sexual intercourse	15.0	14.0	16.0	21.0	23.0	17.0	26.0	20.0	26.0	17.0
Num of pregnancies	1.0	1.0	4.0	4.0	2.0	6.0	3.0	5.0	4.0	3.0
Smokes	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
Smokes (years)	0.0	0.0	37.0	0.0	0.0	34.0	0.0	0.0	0.0	0.0
Smokes (packs/year)	0.0	0.0	37.0	0.0	0.0	3.4	0.0	0.0	0.0	0.0
Hormonal Contraceptives	0.0	0.0	1.0	1.0	0.0	0.0	1.0	0.0	1.0	1.0
Hormonal Contraceptives (years)	0.0	0.0	3.0	15.0	0.0	0.0	2.0	0.0	2.0	8.0
IUD	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0
IUD (years)	0.0	0.0	0.0	0.0	0.0	7.0	7.0	0.0	0.0	0.0
STDs	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STDs (number)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STDs:condylomatosis	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STDs:cervical condylomatosis	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STDs:vaginal condylomatosis	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STDs:vulvo-perineal condylomatosis	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STDs:syphilis	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STDs:pelvic inflammatory disease	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STDs:genital herpes	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STDs:molluscum contagiosum	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STDs:AIDS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STDs:HIV	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STDs:Hepatitis B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STDs:HPV	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STDs: Number of diagnosis	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Dx:Cancer	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
Dx:CIN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Dx:HPV	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
Dx	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
Hinselmann	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0

In [24]:
```
df.describe().transpose()
```

Out[24]:

	count	mean	std	min	25%	50%	75%	max
Age	668.0	27.264970	8.727432	13.0	21.0	26.0	33.0	84.0
Number of sexual partners	668.0	2.523952	1.640299	1.0	2.0	2.0	3.0	28.0
First sexual intercourse	668.0	17.142216	2.852046	10.0	15.0	17.0	18.0	32.0
Num of pregnancies	668.0	2.323353	1.465319	0.0	1.0	2.0	3.0	11.0
Smokes	668.0	0.143713	0.351061	0.0	0.0	0.0	0.0	1.0
Smokes (years)	668.0	1.235524	4.193611	0.0	0.0	0.0	0.0	37.0
Smokes (packs/year)	668.0	0.458953	2.336308	0.0	0.0	0.0	0.0	37.0
Hormonal Contraceptives	668.0	0.643713	0.479260	0.0	0.0	1.0	1.0	1.0
Hormonal Contraceptives (years)	668.0	2.290037	3.724400	0.0	0.0	0.5	3.0	22.0
IUD	668.0	0.112275	0.315942	0.0	0.0	0.0	0.0	1.0
IUD (years)	668.0	0.530030	2.001308	0.0	0.0	0.0	0.0	19.0
STDs	668.0	0.097305	0.296595	0.0	0.0	0.0	0.0	1.0
STDs (number)	668.0	0.166168	0.551073	0.0	0.0	0.0	0.0	4.0
STDs:condylomatosis	668.0	0.055389	0.228910	0.0	0.0	0.0	0.0	1.0
STDs:cervical condylomatosis	668.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
STDs:vaginal condylomatosis	668.0	0.005988	0.077208	0.0	0.0	0.0	0.0	1.0
STDs:vulvo-perineal condylomatosis	668.0	0.053892	0.225974	0.0	0.0	0.0	0.0	1.0
STDs:syphilis	668.0	0.022455	0.148269	0.0	0.0	0.0	0.0	1.0
STDs:pelvic inflammatory disease	668.0	0.001497	0.038691	0.0	0.0	0.0	0.0	1.0
STDs:genital herpes	668.0	0.001497	0.038691	0.0	0.0	0.0	0.0	1.0
STDs:molluscum contagiosum	668.0	0.001497	0.038691	0.0	0.0	0.0	0.0	1.0
STDs:AIDS	668.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
STDs:HIV	668.0	0.019461	0.138242	0.0	0.0	0.0	0.0	1.0
STDs:Hepatitis B	668.0	0.001497	0.038691	0.0	0.0	0.0	0.0	1.0
STDs:HPV	668.0	0.002994	0.054677	0.0	0.0	0.0	0.0	1.0
STDs: Number of diagnosis	668.0	0.092814	0.310355	0.0	0.0	0.0	0.0	3.0
Dx:Cancer	668.0	0.025449	0.157603	0.0	0.0	0.0	0.0	1.0
Dx:CIN	668.0	0.004491	0.066915	0.0	0.0	0.0	0.0	1.0
Dx:HPV	668.0	0.023952	0.153015	0.0	0.0	0.0	0.0	1.0
Dx	668.0	0.023952	0.153015	0.0	0.0	0.0	0.0	1.0

Feature Correlation Analysis

How these features are correlated to Dx:Cancer (Pearson's correlation coefficient)

We see that Pearson's Correlation Coefficient for these two columns is 0.39.

Pearson goes from -1 to +1. A value of 0 would have told there is no correlation, so we shouldn't bother looking at that attribute.

As we the column Dx:Cancer indicate the diagnosed cervical cancer patients. Here, we will explore how other features are correlated with this feature. So that we can understand that how each of these features are contributing cervical cancer.

In [25]:
```
for f in df:
    related = df['Dx:Cancer'].corr(df[f])
    print("%s: %f" % (f, related))
```

```
Age: 0.105179
Number of sexual partners: 0.023736
First sexual intercourse: 0.051974
Num of pregnancies: 0.035725
Smokes: -0.012007
Smokes (years): 0.061238
Smokes (packs/year): 0.119537
Hormonal Contraceptives: 0.020978
Hormonal Contraceptives (years): 0.056982
IUD: 0.093078
IUD (years): 0.076003
STDs: 0.011091
STDs (number): -0.014239
STDs:condylomatosis: -0.039131
STDs:cervical condylomatosis: nan
STDs:vaginal condylomatosis: -0.012542
STDs:vulvo-perineal condylomatosis: -0.038568
STDs:syphilis: -0.024492
STDs:pelvic inflammatory disease: -0.006257
STDs:genital herpes: -0.006257
STDs:molluscum contagiosum: -0.006257
STDs:AIDS: nan
STDs:HIV: -0.022766
STDs:Hepatitis B: -0.006257
STDs:HPV: 0.339113
STDs: Number of diagnosis: -0.017712
Dx:Cancer: 1.000000
Dx:CIN: -0.010854
Dx:HPV: 0.907229
Dx: 0.782890
Hinselmann: 0.148549
Schiller: 0.143002
Citology: 0.121929
Biopsy: 0.184112
```

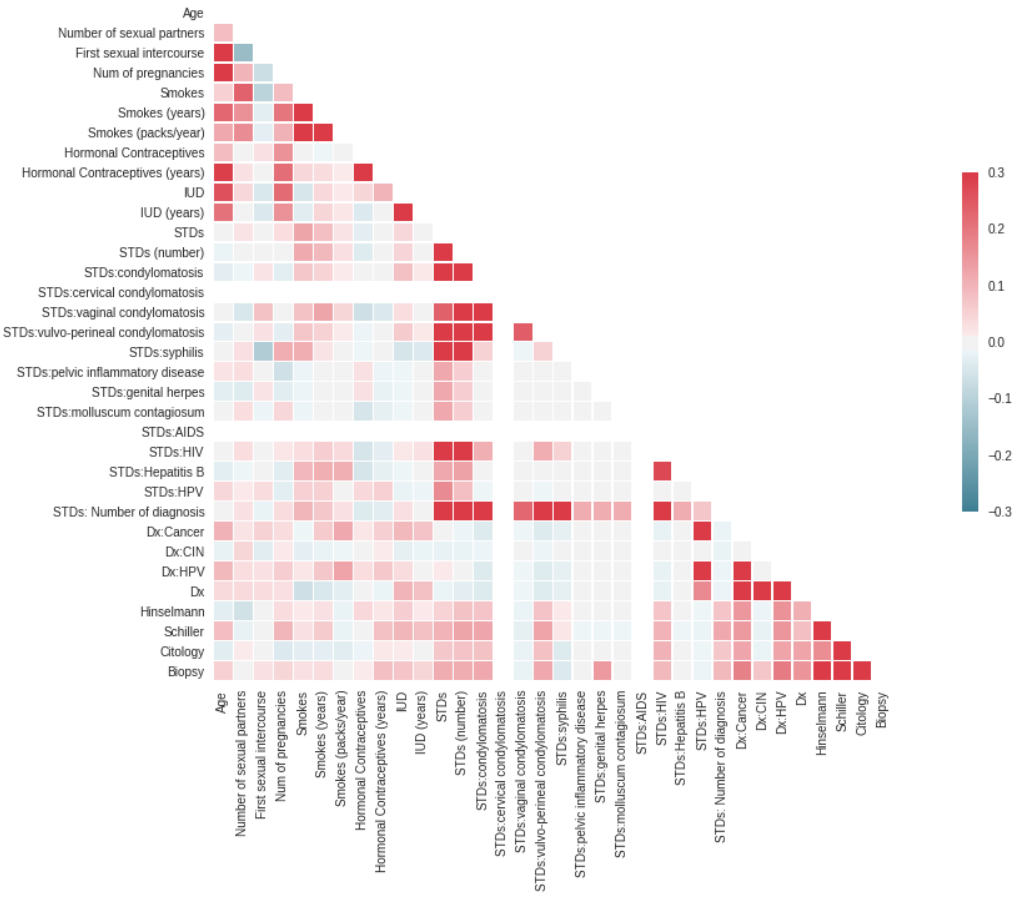
"Dx:HPV" columns is higly positively correlated with "Dx:Cancer" column (0.907229). Net higly correlated column is "STDs:HPV"(0.339113). That means HPV diagnosed patients are more prone to have cervical cancer.

Data Visualization:

We will use heat map from seaborn library to plot diagnal correlation matrix

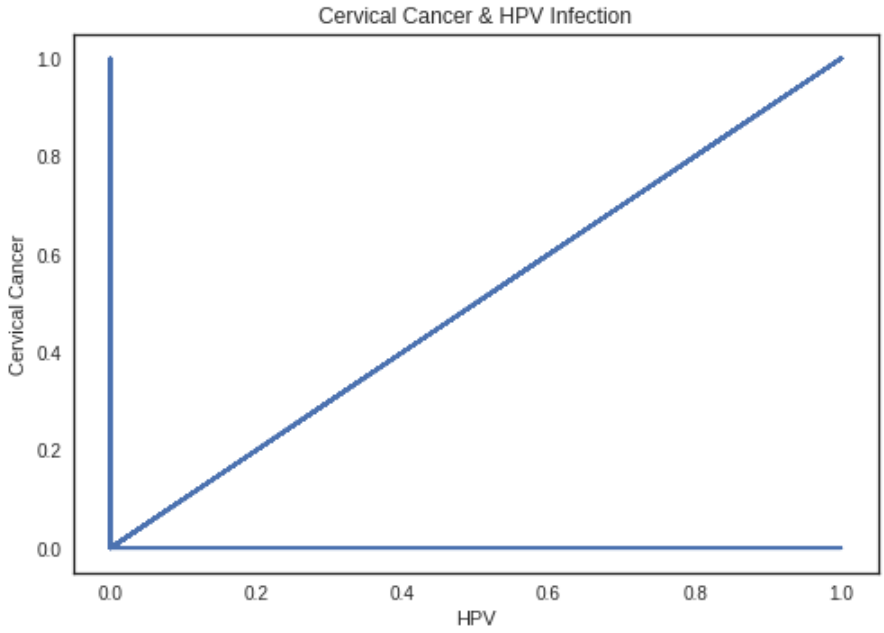
In [26]:
```
sns.set(style="white")

# Compute the correlation matrix
corr = df.corr()

# Generate a mask for the upper triangle
mask = np.zeros_like(corr, dtype=np.bool)
mask[np.triu_indices_from(mask)] = True

# Set up the matplotlib figure
f, ax = plt.subplots(figsize=(20, 10))

# Generate a custom diverging colormap
cmap = sns.diverging_palette(220, 10, as_cmap=True)

# Draw the heatmap with the mask and correct aspect ratio
sns.heatmap(corr, mask=mask, cmap=cmap, vmax= .3, center=0,
            square=True, linewidths=.5, cbar_kws={"shrink": .5})
```

Out[26]: <matplotlib.axes._subplots.AxesSubplot at 0x7fbf7ecf1f10>



Let us examin the linear relationship between Dx:Cancer and Dx:HPV columns byploting both columns.

In [27]:
```
plt.plot(df['Dx:HPV'].values, df['Dx:Cancer'].values)
plt.xlabel('HPV')
plt.ylabel('Cervical Cancer')

plt.title('Cervical Cancer & HPV Infection')



plt.show()
```



In [ ]:
```

```