Exploratory Data Analysis

- Cervical cancer-

2019056462 임단비

Index

General Description : Background

- 매년 미국에서는 자궁경부암으로 약 4,000명의 여성이 사망하고 전 세계적으로는 약 300,000명의 여성이 사망한다.

- 자궁경부암 사망의 90%는 자원이 부족한 환경(ex. 개발도상국)에서 발생한다.

(개발 도상국에서는 의료기기 부족과 검진 비용 부담이 존재하고, 예방 접종 여건도 미약하다)

- 자궁경부암은 조기 진단 시 가장 예방할 수 있는 암으로, 자궁경부암을 조기 발견하여 사망률을 낮추는데 기여하고자 한다.

General Description : Analytics Objectives

- 자궁경부암을 유발하는 주요 요인들을 알아내고, 자궁경부암이 발병했을 것 같은 사람들에게 자궁경부암 검사를 권하여 자궁경부암의 조기 발견을 돕는다.

  (자궁경부암 검진 및 HPV 백신에 대한 자원이 부족한 환경에서 유용하게 이용될 수 있다)

- 자궁경부암의 위험 및 보호 요인을 탐색하여 자궁경부암 예방 캠페인에 이용할 수 있다.

- 자궁경부암에 영향을 미치는 요소에는 무엇이 있을까?

  가설 1) 성병을 가진 사람이 자궁경부암에 걸릴 확률이 높을 것이다.

  가설 2) 흡연 시 자궁경부암 발병률이 더 높아질 것이다.

  가설 3) 성적 파트너가 많을수록 자궁경부암에 걸릴 확률이 높을 것이다.

General Description: Data Set (Data Preprocessing)

```
# ? 값을 포함하고 있는 컬럼 찾기
for feature in data.columns:
    numq = len(data[data[feature]=='?'])
    if numq!=0:
        print(feature, numq)
```

```
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: Time since first diagnosis 787
STDs: Time since last diagnosis 787
```

```
# 통일성을 위해 ? => median으로 대체
for feature in data.columns:
    data[feature].replace('?',np.nan,inplace=True )
    data[feature].fillna(value=0,inplace=True)
for feature in data.columns:
    data[feature].replace(0,data[feature].median(),inplace=True)
```

- 환자의 개인 정보 보호 문제로 누락된 데이터에는 “?” 값이 저장되어 있다.
- 왼쪽은 “?” 값을 포함하고 있는 feature가 몇개의 “?” 값을 가지고 있는지를 출력한 결과이다.
- 데이터 분석을 위해 데이터가 수집되지 않아 “?” 로 표기된 값들은 중앙값(median)으로 대체하였다.

General Description : Data Set

- 총 858개의 환자 데이터 존재
- 36개의 속성 존재

(32개의 위험인자 + 4개의 목표 변수)

속성	속성 정보
Age	환자의 나이
Number of sexual partners	성적 파트너 수
First sexual intercourse	첫 성경험 나이
Num of pregnancies	임신 횟수
Smokes	흡연 여부
Smokes (years)	흡연 기간
Smokes (packs/year)	1년당 피는 담배 팩 수
Hormonal Contraceptives	호르몬 피임약 복용 여부
Hormonal Contraceptives (years)	호르몬 피임약 복용 기간
IUD	자궁내 장치 여부
IUD (years)	자궁내 장치 사용 기간
STDs	성병 여부
STDs (number)	성병 개수
STDs:condylomatosis	성병(condylomatosis) 여부
STDs:cervical condylomatosis	성병(cervical condylomatosis) 여부
STDs:vaginal condylomatosis	성병(vaginal condylomatosis) 여부
STDs:vulvo-perineal condylomatosis	성병(vulvo-perineal condylomatosis) 여부
STDs:syphilis	성병(syphilis) 여부
STDs:pelvic inflammatory disease	성병(pelvic inflammatory disease) 여부
STDs:genital herpes	성병(genital herpes) 여부
STDs:molluscum contagiosum	성병(molluscum contagiosum) 여부
STDs:AIDS	성병(AIDS) 여부
STDs:HIV	성병(HIV) 여부
STDs:Hepatitis B	성병(Hepatitis B) 여부
STDs:HPV	성병(HPV) 여부
STDs: Number of diagnosis	성병 진단 횟수
STDs: Time since first diagnosis	첫 성병 진단 후 경과시간
STDs: Time since last diagnosis	마지막 성병 진단 후 경과시간
Dx:Cancer	암 진단 여부
Dx:CIN	CIN 진단 여부
Dx:HPV	HPV 진단 여부
Dx	진단 여부
Hinselmann	Hinselmann 검사 필요 여부 (목표 변수)
Schiller	Schiller 검사 필요 여부 (목표 변수)
Cytology	Cytology 검사 필요 여부 (목표 변수)
Biopsy	Biopsy 검사 필요 여부 (목표 변수)

Descriptive Statics

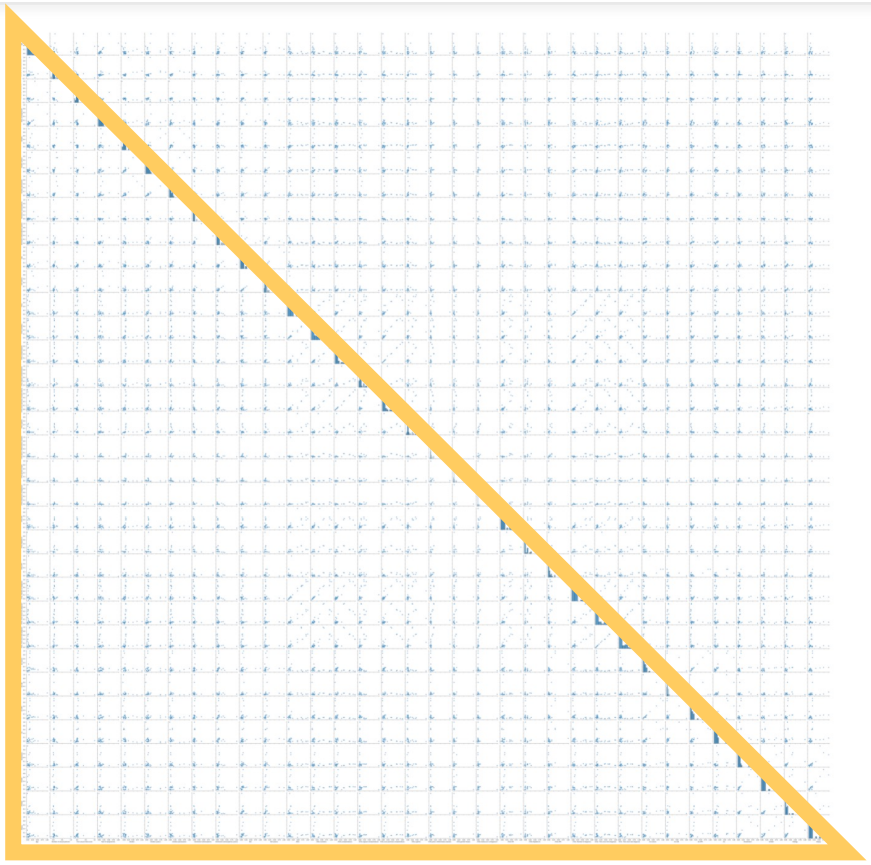
Cervica_cancer의 기술 통계를 정리한 결과 정수 및 수치형 속성은 12개, 명목형 속성은 24개이다.

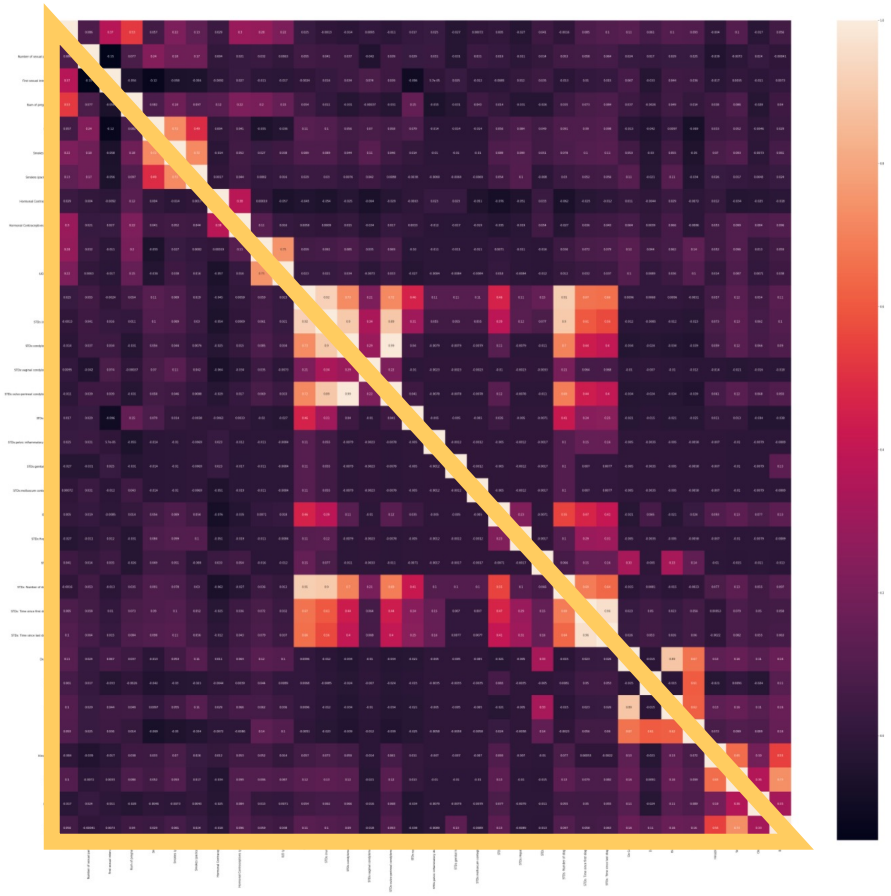
ID	속성	Min	Max	Avg/최빈값	Std
Age	정수형	13	84	26.82051282	8.49794807
Number of sexual partners	정수형	1	28	2.511655	1.644759
First sexual intercourse	정수형	10	32	16.995338	2.791883
Num of pregnancies	정수형	0	11	2.257576	1.400981
Smokes	명목형	0	1	0	
Smokes (years)	수치형	0	37	0.446278	2.210351
Smokes (packs/year)	수치형	0	37	1.201241	4.060623
Hormonal Contraceptives	명목형	0	1	1	
Hormonal Contraceptives (years)	수치형	0	30	1.911422	3.623672
IUD	명목형	0	1	0	
IUD (years)	수치형	0	19	0.444604	1.814218
STDs	명목형	0	1	0	
STDs (number)	정수형	0	4	0.155012	0.529617
STDs:condylomatosis	명목형	0	1	0	
STDs:cervical condylomatosis	명목형	0	1	0	
STDs:vaginal condylomatosis	명목형	0	1	0	
STDs:vulvo-perineal condylomatosis	명목형	0	1	0	
STDs:syphilis	명목형	0	1	0	
STDs:pelvic inflammatory disease	명목형	0	1	0	
STDs:genital herpes	명목형	0	1	0	
STDs:molluscum contagiosum	명목형	0	1	0	
STDs:AIDS	명목형	0	1	0	
STDs:HIV	명목형	0	1	0	
STDs:Hepatitis B	명목형	0	1	0	
STDs:HPV	명목형	0	1	0	
STDs: Number of diagnosis	정수형	0	3	0.087413	0.302545
STDs: Time since first diagnosis	정수형	0	22	0.508159	2.388333
STDs: Time since last diagnosis	정수형	0	22	0.481352	2.297125
Dx:Cancer	명목형	0	1	0	
Dx:CIN	명목형	0	1	0	
Dx:HPV	명목형	0	1	0	
Dx	명목형	0	1	0	
Hinselmann	명목형	0	1	0	
Schiller	명목형	0	1	0	
Cytology	명목형	0	1	0	
Biopsy	명목형	0	1	0	

Data Visualization

- 먼저 각 속성 간의 관계를 파악하기 위해 단변량 분석 및 다변량 분석을 시행하였다.
- 특징이 많기 때문에 이를 통해 얻어낸 결과는 뒤에서 하나씩 자세히 살펴보도록 한다.

* 변수가 너무 많아 분석을 통해 얻은 유의미한 결과는 뒤에서 다시 한번 크게 시각화 하였습니다.

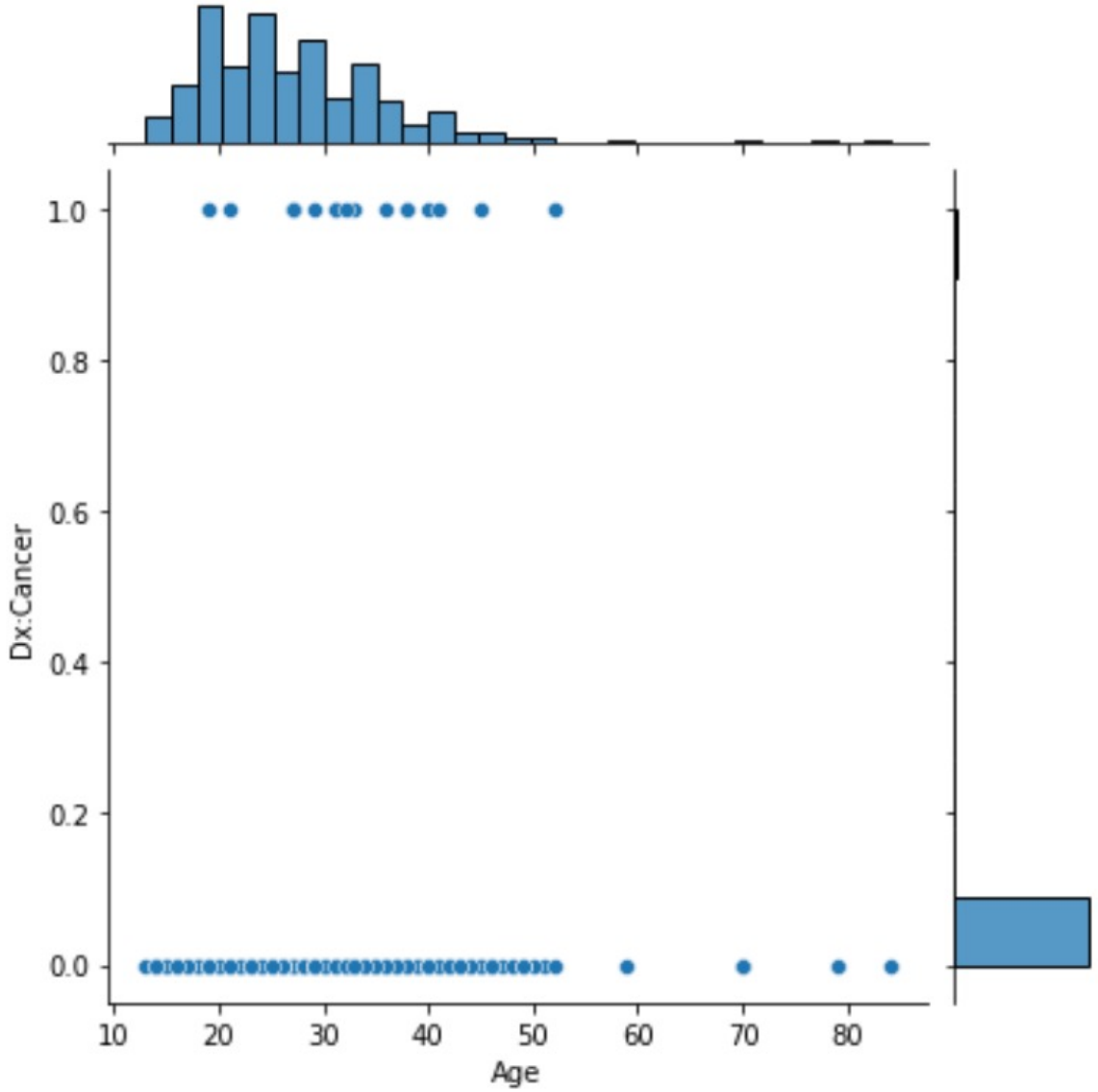




Data Visualization

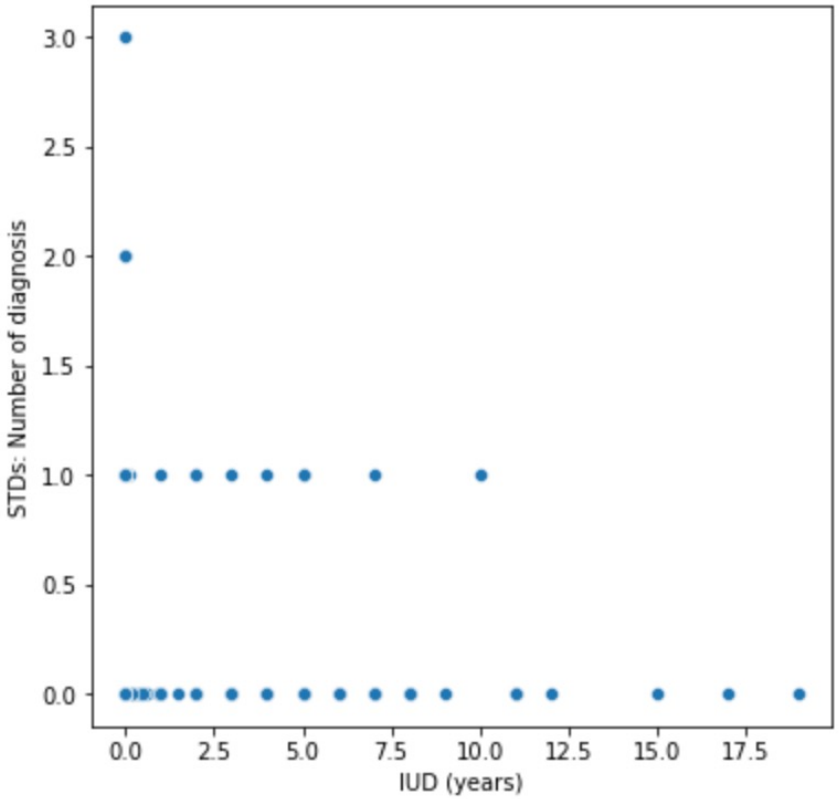
- 나이에 따른 자궁경부암의 분포를 시각화한 것이다.

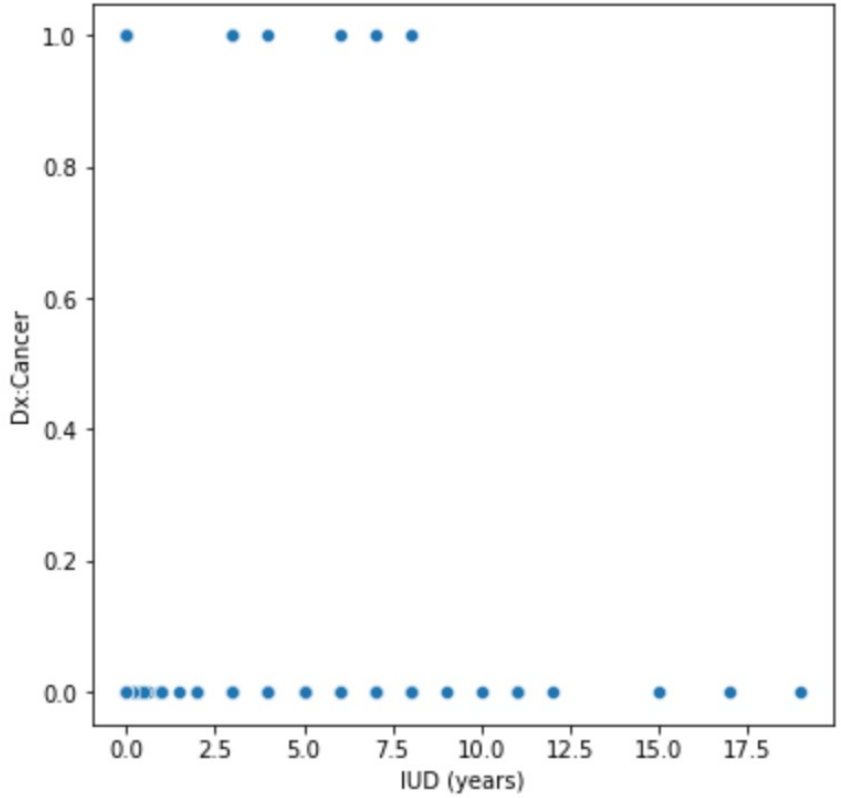
- 데이터 셋 내에서 자궁경부암은 19세에서 52세 사이 환자들에게 발생한 것을 확인할 수 있었다.



Data Visualization

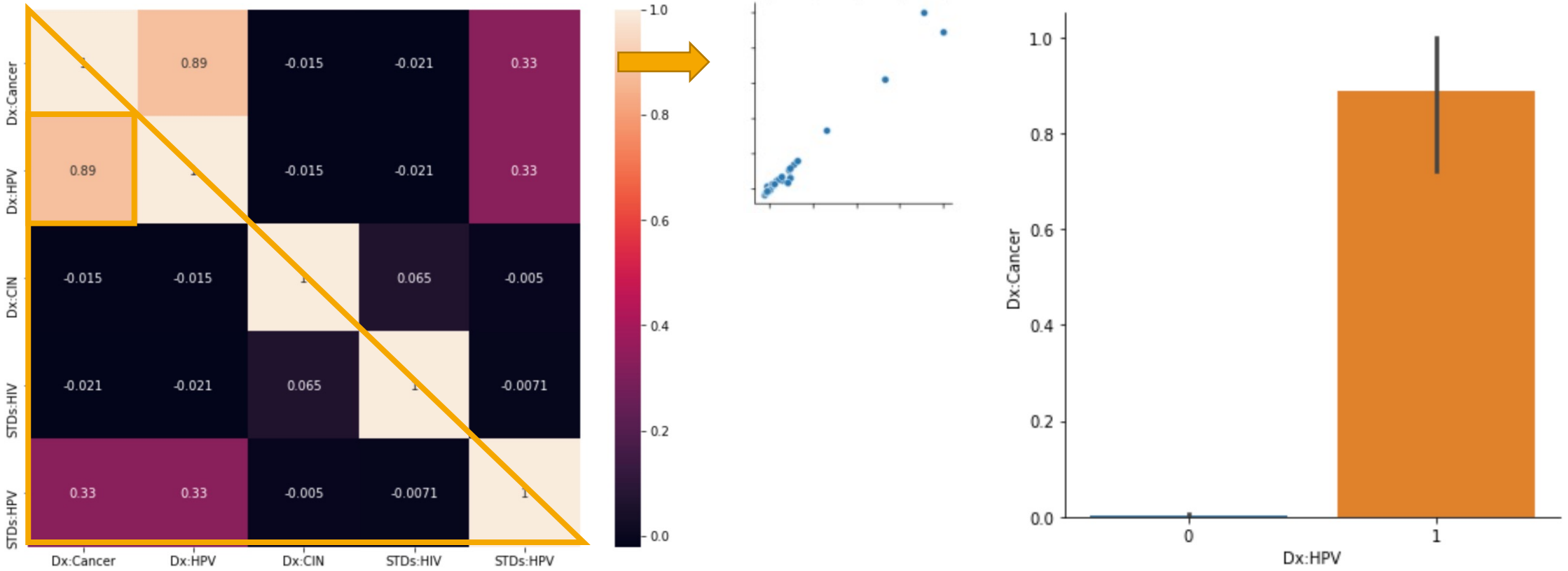
- IUD(자궁 내 장치)를 오래 사용할수록 STD(성병)이 더 적게 나타났다.
- 또한, IUD를 10년이상 사용한 경우 자궁 경부암이 걸린 사례가 없었다.





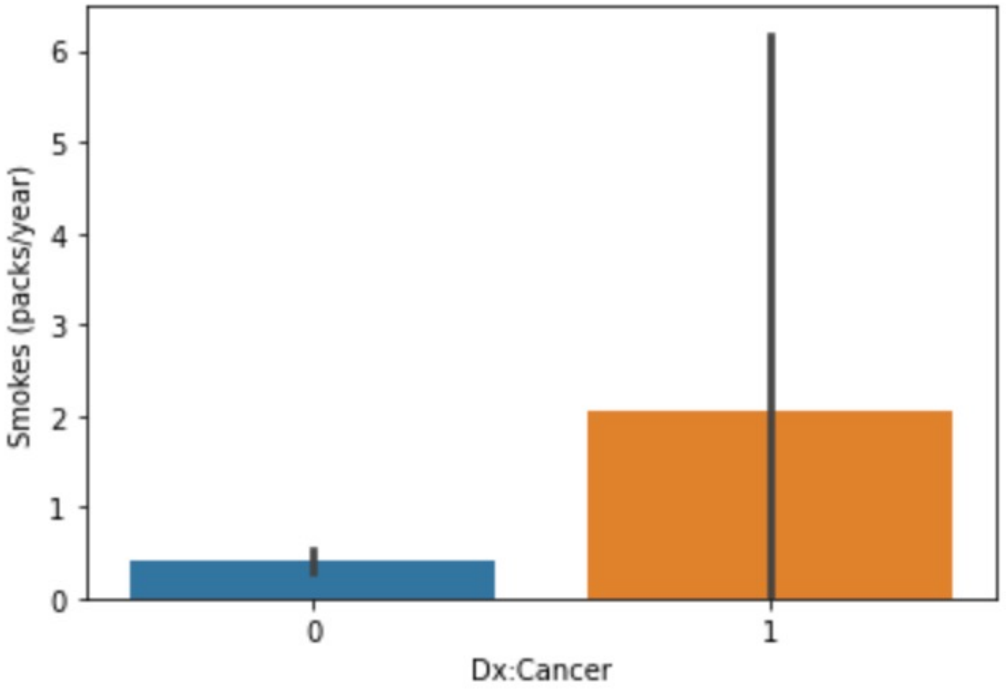
Data Visualization

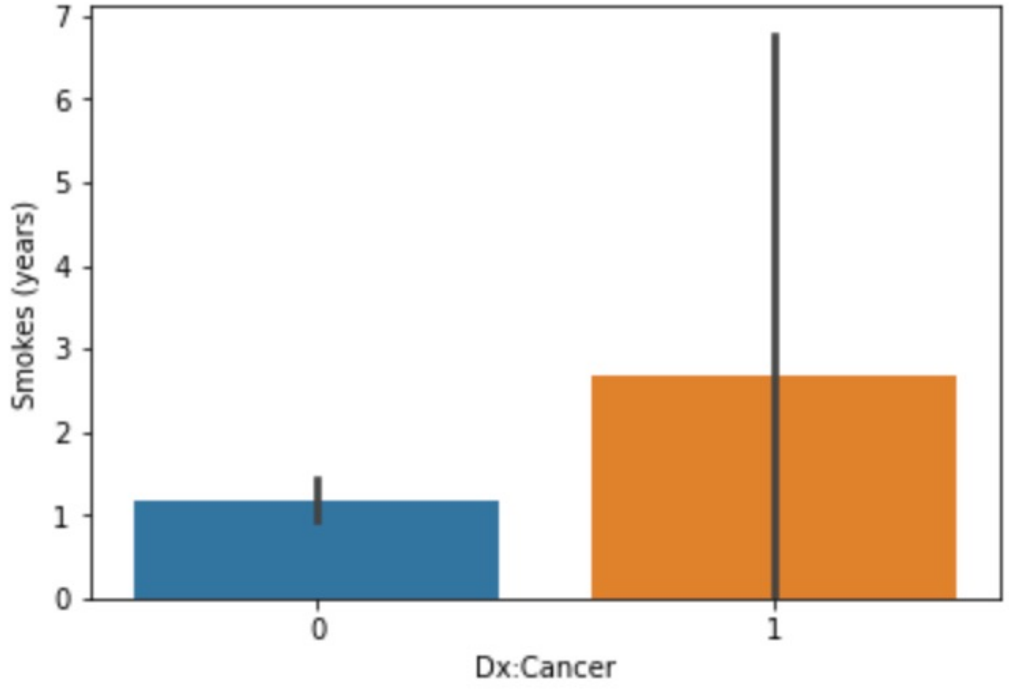
- **가설1**. “성병을 가진 사람이 자궁경부암에 걸릴 확률이 높을 것이다.”는 가설을 확인하기 위해 상관분석을 이용했다.
- 결과적으로 <u>자궁경부암</u>을 진단 받은 사람들은 여러 성병 중 <u>HPV 성병</u>을 진단 받은 경우가 많았음을 확인할 수 있었다.



Data Visualization

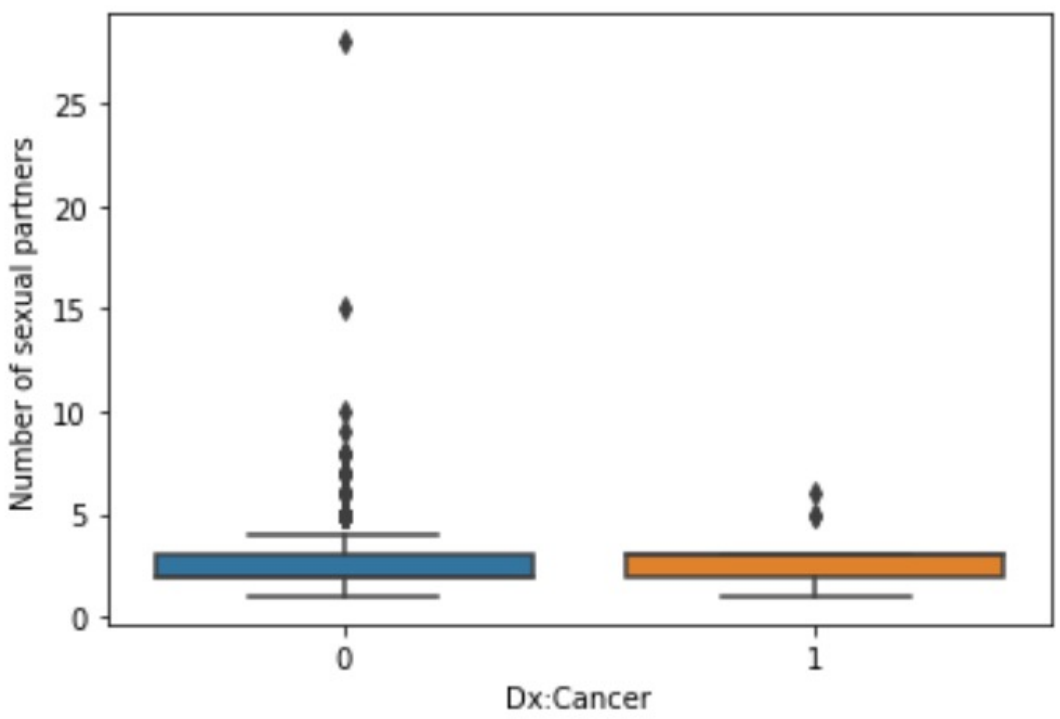
- **가설2.** “흡연 시 자궁경부암 발병률이 더 높아질 것이다." 를 확인하기 위해 다음과 같이 데이터를 시각화 하였다.
- 자궁경부암 발병률은 Smokes(years) 와 Smokes (packs/year) 이 길수록 높았다.

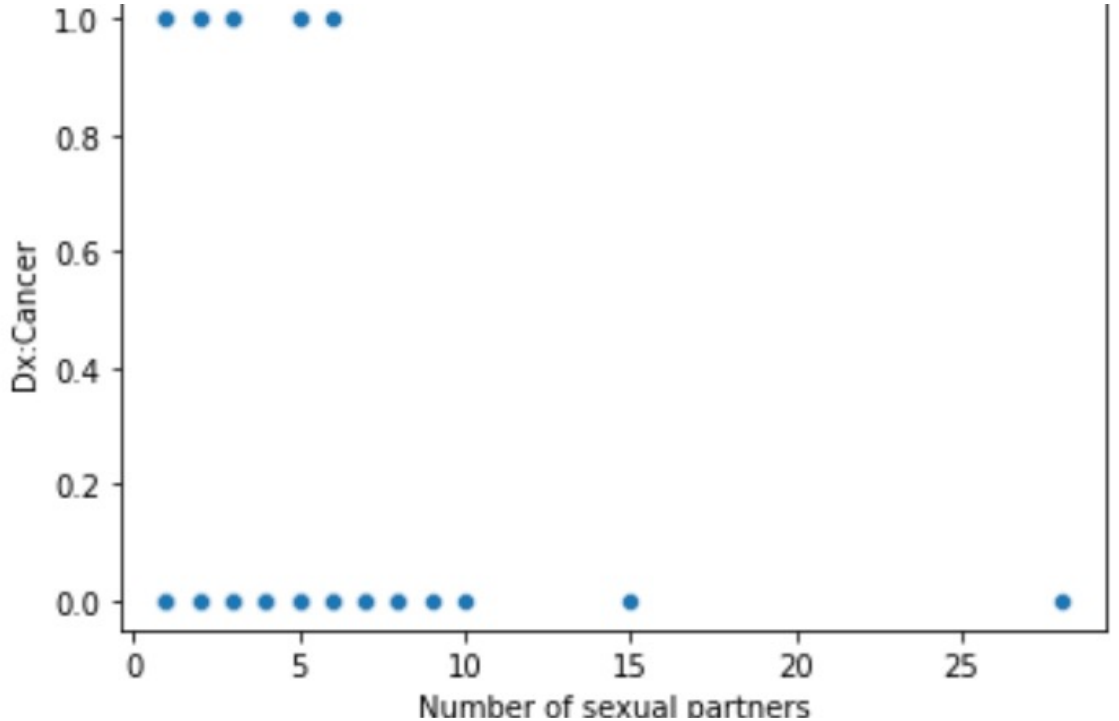




Data Visualization

- **가설3.** “성적 파트너가 많을수록 자궁경부암에 걸릴 확률이 높을 것이다.”을 확인하기 위해 데이터를 시각화 했다.
- 성적 파트너가 많을수록 자궁 경부암에 걸릴 확률이 높을 것이라는 가설과는 달리 자궁 경부암은 성적 파트너수에 관계 없이 나타났고, 오히려 가장 많은 성적 파트너를 가진 사람은 자궁경부암이 발병하지 않았다.





Result

- 앞서 시각화한 내용을 총 정리해보면 자궁경부암에 영향을 미치는 요소는 다음과 같다.

1) HPV 성병
2) 흡연
3) IUD 장치

- 위 결과를 토대로 HPV 성병이 걸린 사람에게 자궁경부암 검사를 권유하는 등의 방식으로 자궁경부암을 조기 진단할 수 있을 것이다.
- 또한, 자궁경부암과 흡연 예방 캠페인을 위한 자료로 흡연과 자궁 경부암의 연관성을 사용할 수도 있을 것이다.

감사합니다 ☺