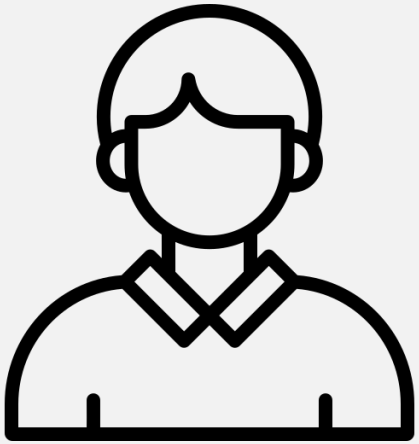
폐암 원인 추론

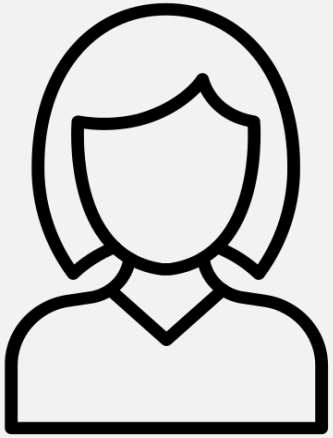
팀 원
방 은호
윤 송이
이 지성

팀원 소개



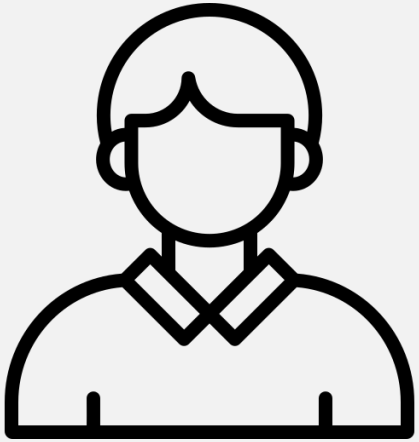
방 은 호

- 데이터 수집
- 데이터 전처리
- 시각화



윤 송 이

- 데이터 수집
- 데이터 전처리
- 시각화



이 지 성

- 데이터 수집
- 데이터 전처리
- 시각화

사용 데이터

- API key 활용

```
import requests


my_url = 'http://apis.data.go.kr/B551172/Lung09/luTuberchoIsisByType?'
params ={'serviceKey' : my_service,
         'pageNo' : '1',
         'numOfRows' : '100',
         'centerNm' : '국립암센터',
         'fromYear' : '2010',
         'toYear' : '2019',
         'type' : 'json'}

response = requests.get(my_url, params= params)
print(response.content)
```

```
import pandas as pd
import json
```

```
df = pd.read_json(response.content)
```

```
df
```

```
from pandas.io.json import json_normalize
```

```
js_file = json.loads(response.content)
js_file
```

```
js_file['response']['items']['item']
```

```
jsdf = pd.DataFrame(js_file['response']['items']['item'])
```

```
jsdf
```

	centerNm	critYr	ptAge	ptSexCd	statsTrgtNm	ncsNmvl	wholNcsDnmvl	ptCntNmvl	wholPtCntDnmvl
0	국립암센터	2010	69	M	Y	3	16	3	15
1	국립암센터	2019	76	F	N	2	3	2	3
2	국립암센터	2011	63	M	Y	5	24	5	23
3	국립암센터	2017	78	F	N	6	8	4	5
4	국립암센터	2017	78	F	Y	2	8	1	5
5	국립암센터	2017	70	M	N	24	31	18	22

선정 배경

20년간 암사망률 1위 '폐암'.. 5대 증상 알아두세요

이금숙 헬스조선 기자

2020/08/14 17:26

암 사망률 가장 높은 폐암... 가장 확실한 예방법은?

이금숙 헬스조선 기자

흡연이 주원인, 발병 위험 13배 높여... 흡연 양과 기간도 연관

담배도 안 피우는데... 폐암 유발하는 원인 5

룽지혜 헬스조선 인턴기자

요리 매연

음식 조리 시 발생하는 연기는 폐암의 원인이 된다
이 발생한다. 발암물질이 섞인 연기나 그을음은 폐
지 후드 같은 환기 장치를 켜고 창문을 열어놓는 습

라돈

라돈은 암석이나 토양 속의 우라늄이 붕괴되는 과
호흡을 통해 우리 몸에 들어오며 원소가 쪼개지면
석이나 토양에서 자연적으로 발생하고 건물 벽 내
실 등에 들어가는 것을 삼가야 한다.

간접흡연

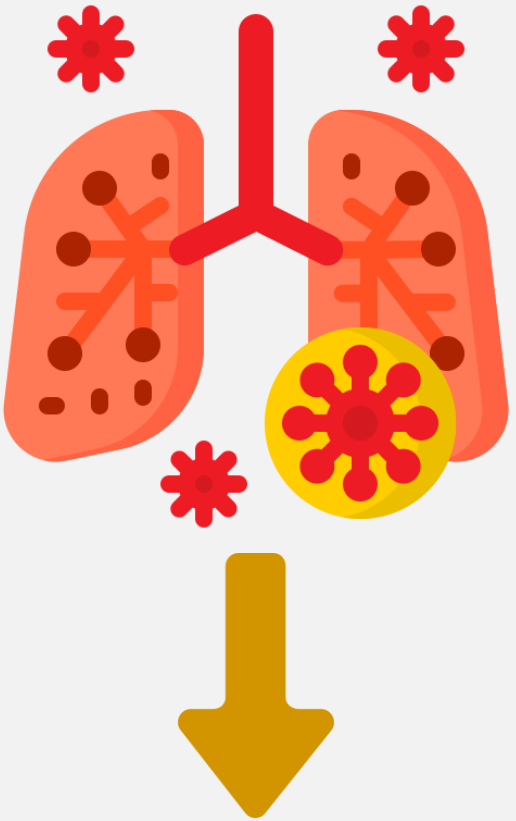
담배 연기에 들어 있는 성분으로 따져보면 직접흡
연'이 있다. 간접흡연시 주로 흡입되는 부류연은 주

석면

석면은 건물을 지을 때 단열재 등으
이는 폐가 딱딱해지는 섬유화를 일
도나 오래된 건물 등에서 생활하는
주소와 위해성 등급을 찾을 수 있다

기존 폐질환

폐렴이나 폐결핵, 만성폐쇄성폐질환
등으로 숨길이 좁아지고 허파꽈리
다.

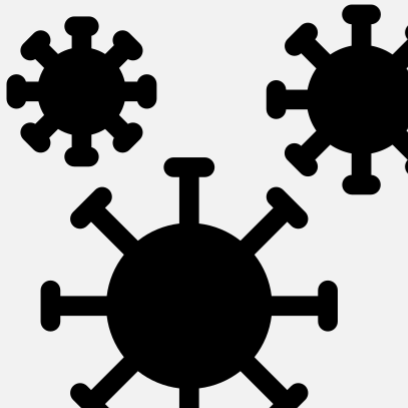


폐암의 직접적인 원인은?

가설

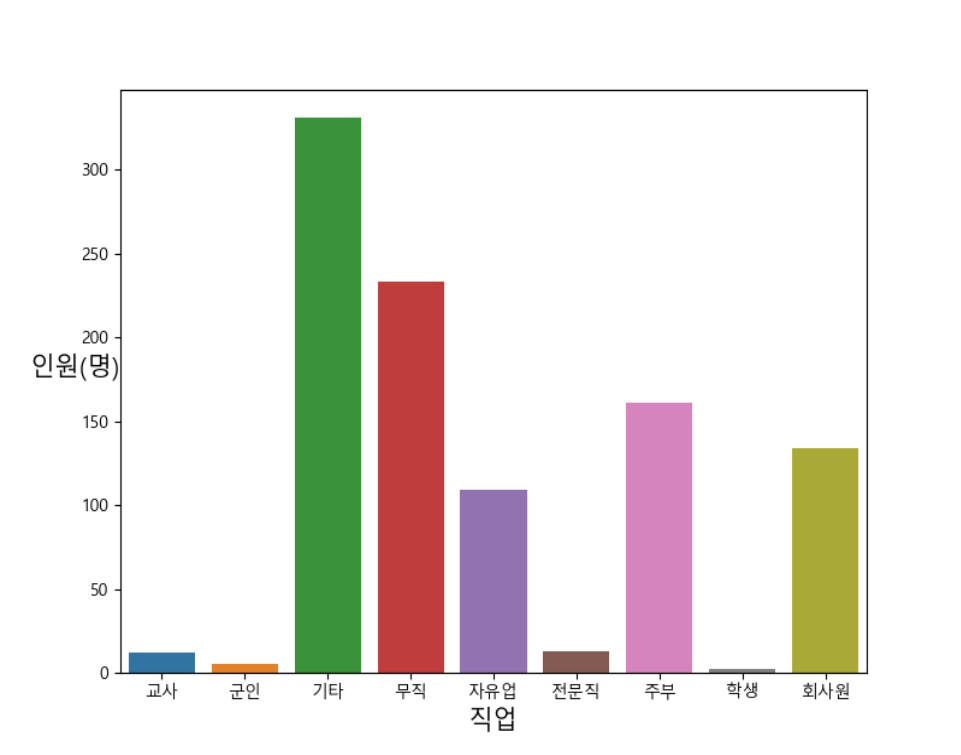
- 연구 결과 가장 큰 원인은 생활환경? or 흡연?

- 폐암과 가장 연관성이 있는 질병은 결핵 ?



1. 직업

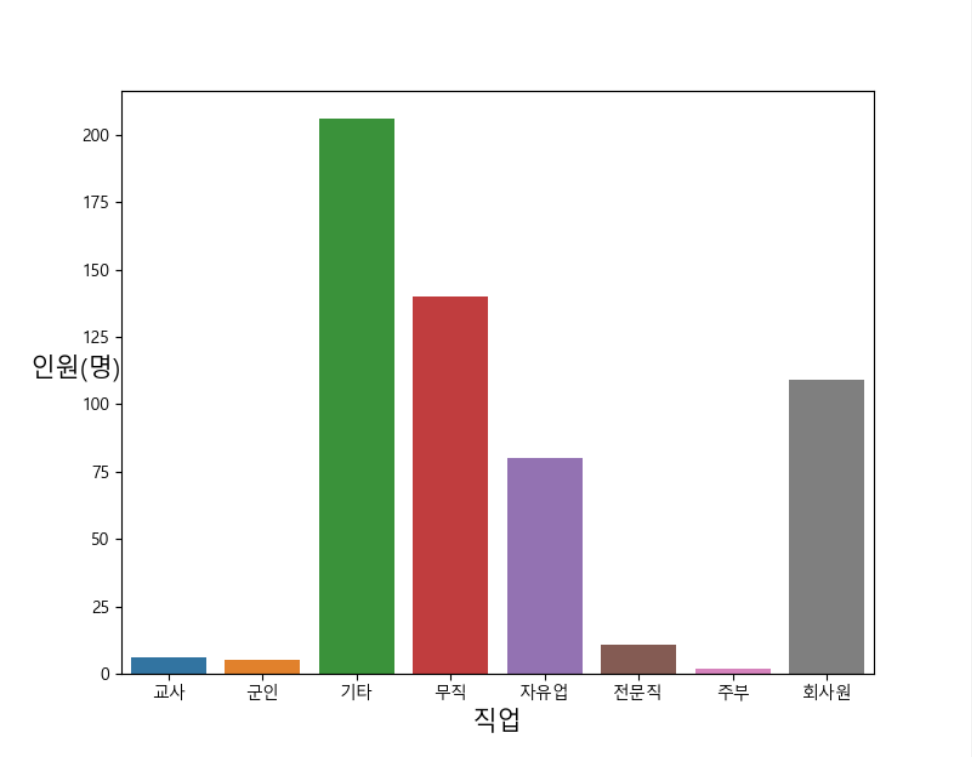
1) 전체 폐암 환자



- 기타, 무직 , 주부 상위 차지
- 기타, 무직에 현장 업종이 포함되어 있음을 추론
- 성별 환자수 구분을 위해 남녀로 분할

1. 직업

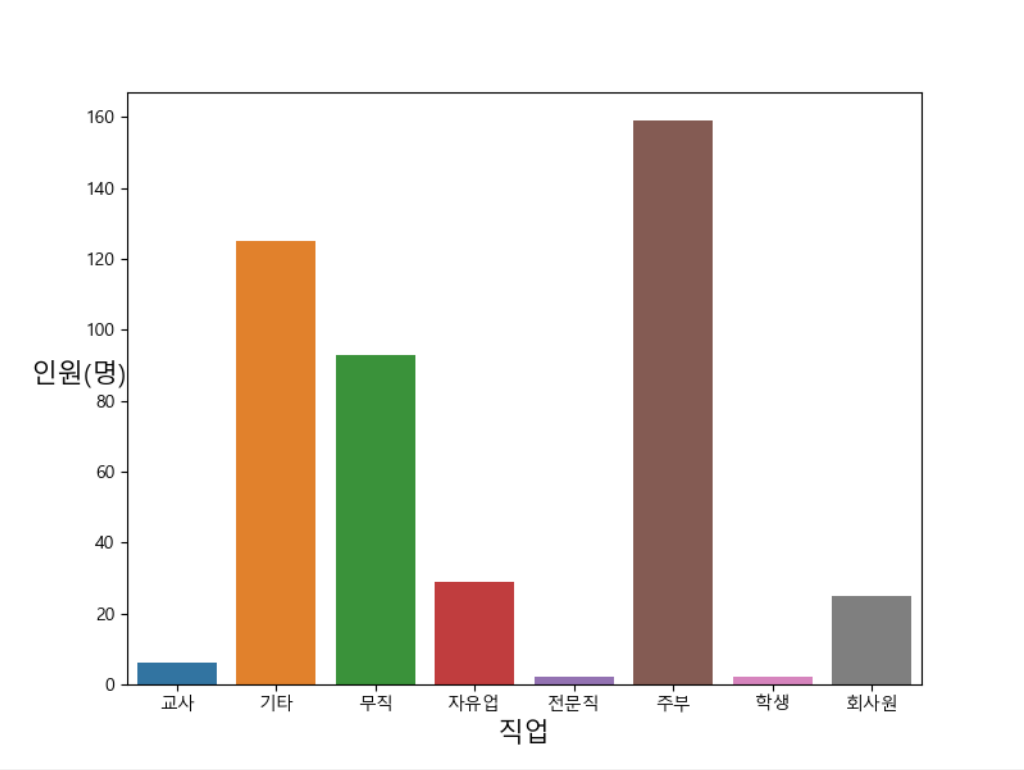
2) 남성



- 기타, 무직 , 자유업 상위 차지
- 기타, 무직에 현장 업종이 포함되어 있음을 추론(확신 추론 향상)
- 주부의 수치 감소

1. 직업

3) 여성



- 주부, 기타 , 무직 상위에 위치
- 기타, 무직 수치 상당수
- **주부**와 비슷한 조리업 종사자로 추론

중간 추론

“학교 급식 노동자 폐암 발병률, 일반인의 24배...환경 개선 필요”

0

입력 : 2021-10-14 16:53 | 수정 : 2021-1

학비노조 ‘급식
급식 노동자 56
직업성 암 전수

여성 폐암 환자의 증가는 이같은 간접흡연과 더불어 음식을 조리할 때 생기는 주방 내 유해연기, 방사성 유해물질 노출, 노령화에 따른 암 발병 자체의 증가 등이 요인으로 추정된다

담배도 안 피우는데 폐암?

여성들 '요리매연' 조심을

2021-11-12 11:27:05 게재

노동안전

‘급식실 조리실무사 폐암 사망’ 산재 최초 인정

한 학기 조리일 81% 튀김·볶음·구이 ... “폐암 위험 높은 조리흄에 노출”

어고은 기자 입력 2021.04.06 07:30

흡연, 음주 여부

1) 흡연 여부

```
lr.score(scaled_minmax_test, y_test_val)
```

0.11158296326569894

```
from sklearn.metrics import mean_squared_error
a_pred = lr.predict(X_test_val)

a = mean_squared_error(y_test_val,a_pred)**0.5
a
```

0.5702095684301874

OLS

```
import statsmodels.api as sm

model = sm.OLS(y, X)
result = model.fit()
result.summary()
```

OLS Regression Results

Dep. Variable:	statsTrgtNm	**R-squared (uncentered):**	0.298
Model:	OLS	**Adj. R-squared (uncentered):**	0.297
Method:	Least Squares	**F-statistic:**	423.4
Date:	Wed, 15 Dec 2021	**Prob (F-statistic):**	1.03e-78
Time:	20:35:22	**Log-Likelihood:**	-827.21
No. Observations:	1000	**AIC:**	1656.
Df Residuals:	999	**BIC:**	1661.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
ptSexCd	0.2313	0.011	20.577	0.000	0.209	0.253

Omnibus:	4325.976	**Durbin-Watson:**	1.870
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	145.622
Skew:	0.234	**Prob(JB):**	2.39e-32
Kurtosis:	1.190	**Cond. No.**	1.00

흡연, 음주 여부

2) 음주 여부

```
lr.score(scaled_minmax_test, y_test_val)
```

0.029424695060832562

```
from sklearn.metrics import mean_squared_error
a_pred = lr.predict(X_test_val)

a = mean_squared_error(y_test_val,a_pred)**0.5
a
```

0.5115734565248449

OLS

```
import statsmodels.api as sm

model = sm.OLS(y, X)
result = model.fit()
result.summary()
```

OLS Regression Results

Dep. Variable:	statsTrgtNm	**R-squared (uncentered):**	0.360
Model:	OLS	**Adj. R-squared (uncentered):**	0.359
Method:	Least Squares	**F-statistic:**	562.1
Date:	Wed, 15 Dec 2021	**Prob (F-statistic):**	6.22e-99
Time:	14:37:45	**Log-Likelihood:**	-815.04
No. Observations:	1000	**AIC:**	1632.
Df Residuals:	999	**BIC:**	1637.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
ptSexCd	0.2665	0.011	23.708	0.000	0.244	0.289

Omnibus:	4402.855	**Durbin-Watson:**	1.981
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	135.401
Skew:	0.112	**Prob(JB):**	3.96e-30
Kurtosis:	1.211	**Cond. No.**	1.00

다른 질병과 선형회귀(OLS 활용)

1) 결핵

OLS Regression Results

Dep. Variable:	statsTrgtNm	R-squared (uncentered):	0.191
Model:	OLS	Adj. R-squared (uncentered):	0.191
Method:	Least Squares	F-statistic:	228.7
Date:	Wed, 15 Dec 2021	Prob (F-statistic):	1.54e-46
Time:	10:36:54	Log-Likelihood:	-621.08
No. Observations:	967	AIC:	1244.
Df Residuals:	966	BIC:	1249.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
ptSexCd	0.1477	0.010	15.124	0.000	0.129	0.167

Omnibus:	201.951	Durbin-Watson:	1.900
Prob(Omnibus):	0.000	Jarque-Bera (JB):	199.394
Skew:	1.038	Prob(JB):	5.04e-44
Kurtosis:	2.200	Cond. No.	1.00

2) 간질환

OLS Regression Results

Dep. Variable:	statsTrgtNm	R-squared (uncentered):	0.073
Model:	OLS	Adj. R-squared (uncentered):	0.072
Method:	Least Squares	F-statistic:	75.35
Date:	Wed, 15 Dec 2021	Prob (F-statistic):	1.68e-17
Time:	10:36:53	Log-Likelihood:	-211.02
No. Observations:	958	AIC:	424.0
Df Residuals:	957	BIC:	428.9
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
ptSexCd	0.0545	0.006	8.680	0.000	0.042	0.067

Omnibus:	511.186	Durbin-Watson:	1.852
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2220.495
Skew:	2.668	Prob(JB):	0.00
Kurtosis:	8.210	Cond. No.	1.00

다른 질병과 선형회귀(OLS 활용)

3) 당뇨병

OLS Regression Results

Dep. Variable:	statsTrgtNm	R-squared (uncentered):	0.292
Model:	OLS	Adj. R-squared (uncentered):	0.292
Method:	Least Squares	F-statistic:	400.7
Date:	Tue, 14 Dec 2021	Prob (F-statistic):	6.93e-75
Time:	21:09:26	Log-Likelihood:	-720.03
No. Observations:	971	AIC:	1442.
Df Residuals:	970	BIC:	1447.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
ptSexCd	0.2134	0.011	20.018	0.000	0.192	0.234

Omnibus:	6034.541	Durbin-Watson:	1.979
Prob(Omnibus):	0.000	Jarque-Bera (JB):	141.779
Skew:	0.520	Prob(JB):	1.63e-31
Kurtosis:	1.444	Cond. No.	1.00

4) 심장질환

OLS Regression Results

Dep. Variable:	statsTrgtNm	R-squared (uncentered):	0.153
Model:	OLS	Adj. R-squared (uncentered):	0.152
Method:	Least Squares	F-statistic:	173.5
Date:	Wed, 15 Dec 2021	Prob (F-statistic):	1.46e-36
Time:	10:36:53	Log-Likelihood:	-530.04
No. Observations:	965	AIC:	1062.
Df Residuals:	964	BIC:	1067.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
ptSexCd	0.1160	0.009	13.173	0.000	0.099	0.133

Omnibus:	188.148	Durbin-Watson:	1.833
Prob(Omnibus):	0.000	Jarque-Bera (JB):	315.589
Skew:	1.400	Prob(JB):	2.96e-69
Kurtosis:	3.071	Cond. No.	1.00

다른 질병과 선형회귀(OLS 활용)

5) 고혈압

OLS Regression Results

Dep. Variable:	statsTrgtNm	**R-squared (uncentered):**	0.401
Model:	OLS	**Adj. R-squared (uncentered):**	0.400
Method:	Least Squares	**F-statistic:**	646.5
Date:	Wed, 15 Dec 2021	**Prob (F-statistic):**	1.35e-109
Time:	10:36:53	**Log-Likelihood:**	-740.36
No. Observations:	967	**AIC:**	1483.
Df Residuals:	966	**BIC:**	1488.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
ptSexCd	0.2768	0.011	25.426	0.000	0.255	0.298

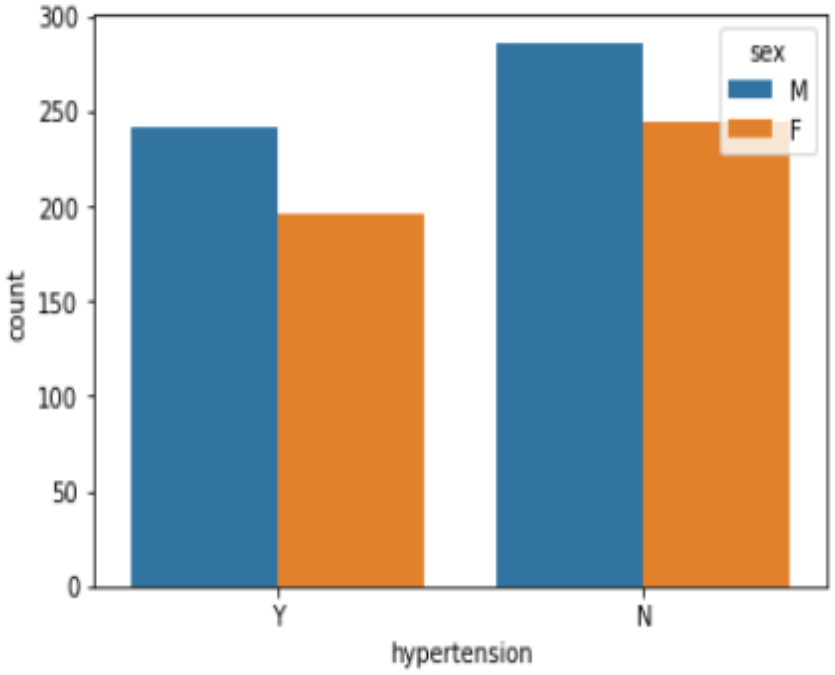
Omnibus:	4667.385	**Durbin-Watson:**	1.915
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	121.928
Skew:	0.169	**Prob(JB):**	3.34e-27
Kurtosis:	1.294	**Cond. No.**	1.00

성별과 고혈압의 상관관계(Phi 상관분석)

고혈압

```
1 from sklearn.metrics import matthews_corrcoef
2
3 y_true = hypertension.gender
4 y_pred = hypertension.hyp
5 matthews_corrcoef(y_true, y_pred)
```

```
0.005928931511501603
```



중간 추론 2

- 폐암과 흡연의 관계성이 높지 않음
  (오히려 음주가 더 높음)
- 폐암은 결핵보다 고혈압과
  더 관계성이 높음
- 고혈압의 원인 중 하나인 음주와
  흡연의 관계성을 추론해봄

음주, 흡연과 고혈압의 관계분석

1) 음주 + 고혈압

```
lr.score(scaled_minmax_test, y_test_val)
```

0.029424695060832562

```
from sklearn.metrics import mean_squared_error
a_pred = lr.predict(X_test_val)

a = mean_squared_error(y_test_val,a_pred)**0.5
a
```

0.5115734565248449

OLS

```
import statsmodels.api as sm

model = sm.OLS(y, X)
result = model.fit()
result.summary()
```

OLS Regression Results

Dep. Variable:	statsTrgtNm	**R-squared (uncentered):**	0.360
Model:	OLS	**Adj. R-squared (uncentered):**	0.359
Method:	Least Squares	**F-statistic:**	562.1
Date:	Wed, 15 Dec 2021	**Prob (F-statistic):**	6.22e-99
Time:	14:37:45	**Log-Likelihood:**	-815.04
No. Observations:	1000	**AIC:**	1632.
Df Residuals:	999	**BIC:**	1637.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
ptSexCd	0.2665	0.011	23.708	0.000	0.244	0.289

Omnibus:	4402.855	**Durbin-Watson:**	1.981
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	135.401
Skew:	0.112	**Prob(JB):**	3.96e-30
Kurtosis:	1.211	**Cond. No.**	1.00

음주, 흡연과 고혈압의 관계분석

2) 흡연 + 고혈압

```
lr.score(X_test_val, y_test_val)
```

```
-0.011691344783902 88
```

```
from sklearn.metrics import mean_squared_error
a_pred = lr.predict(X_test_val)

a = mean_squared_error(y_test_val,a_pred)**0.5
a
```

```
0.49632968663479937
```

OLS

```
import statsmodels.api as sm

model = sm.OLS(y, X)
result = model.fit()
result.summary()
```

OLS Regression Results

Dep. Variable:	hype	**R-squared (uncentered):**	0.178
Model:	OLS	**Adj. R-squared (uncentered):**	0.177
Method:	Least Squares	**F-statistic:**	208.7
Date:	Wed, 15 Dec 2021	**Prob (F-statistic):**	5.69e-43
Time:	20:15:32	**Log-Likelihood:**	-893.52
No. Observations:	967	**AIC:**	1789.
Df Residuals:	966	**BIC:**	1794.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
smoke	0.4289	0.030	14.446	0.000	0.371	0.487

Omnibus:	7532.706	**Durbin-Watson:**	1.608
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	91.109
Skew:	0.132	**Prob(JB):**	1.64e-20
Kurtosis:	1.520	**Cond. No.**	1.00

연구 자료

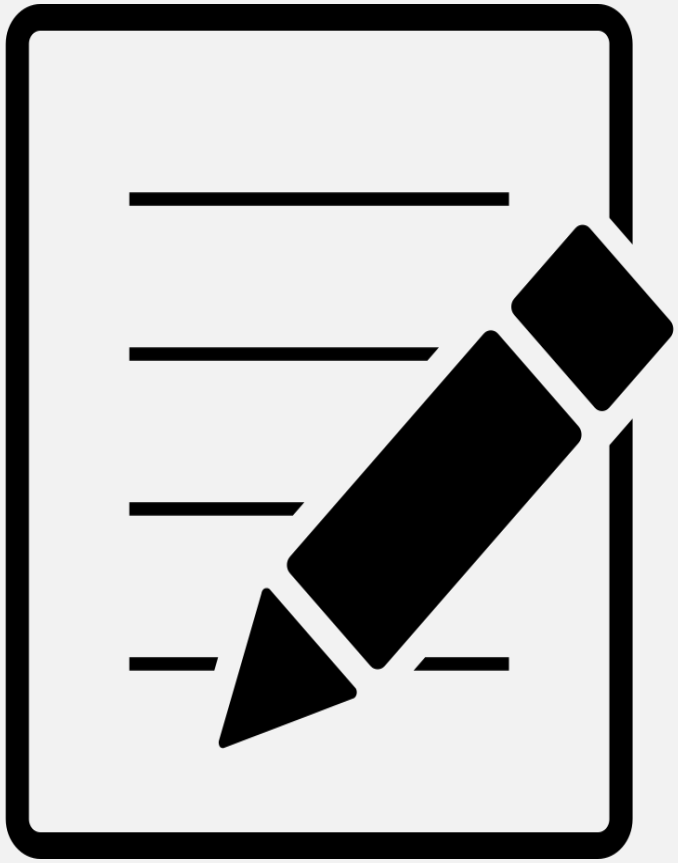
고혈압약이 폐암 유발?..."10년 복용 시 31%↑" 연구결과

김진구 헬스조선 기자 | 정선유 헬스조선 인턴기자

고혈압약으로 흔히 사용되는 '안지오텐신전환효소 억제제(이하 ACE억제제)'가 폐암 위험을 증가시킨다는 연구결과가 나왔다.
캐나다 맥길대 연구팀은 1995~2015년에 새롭게 항고혈압제를 복용하기 시작한 100만 명가량의 환자를 대상으로
혈압약과 폐암 발병의 상관관계를 연구했다. 환자는 이전에 암이 없었고, 평균 6.4년의 추적 조사 기간에 7962건의 폐암이 확인됐다.
성별, 체중(BMI), 흡연 상태, 알코올 관련 질환, 폐 질환의 병력을 포함해 결과에 잠재적으로 영향을 미칠 수 있는 요인을 고려한 결과,
ACE억제제 사용은 안지오텐신 수용체 차단제(ARBs)에 비해 폐암 위험이 14% 증가했다.
특히 10년 이상 ACE억제제를 복용한 환자는 폐암 위험이 31% 높았다.

고혈압 발병 원인보다 고혈압 치료약이
폐암의 발병에 더 직접적인 원인 가능성 추론

최종 추론



- 고혈압의 원인 중 하나인 음주와 흡연의 관계를 파악(음주 > 흡연)
- 폐암은 흡연, 고혈압은 음주와 관계도가 높음
- 각 객체마다 관계성이 뚜렷하지 않아 직접적인 원인은 직업, 생활환경일 것으로 추론
- 데이터 분석을 통해 관계가 없음을 추론했지만, 개인적인 건강을 위해 하지 않는 것이 가장 좋음

아쉬운 점

- 수집한 데이터가 국립암센터 기준으로 정리가 되어 있어 해석이 어려움
- 다른 분야(유방암, 대장암 등)암 데이터는 많이 존재, 폐암 관련 데이터는 찾기 어렵거나 있어도 해석이 어려움
- 딥러닝을 적용시켰다면 더 자세한 결과를 얻을 수 있을 것으로 추론

Thank you for your attention