

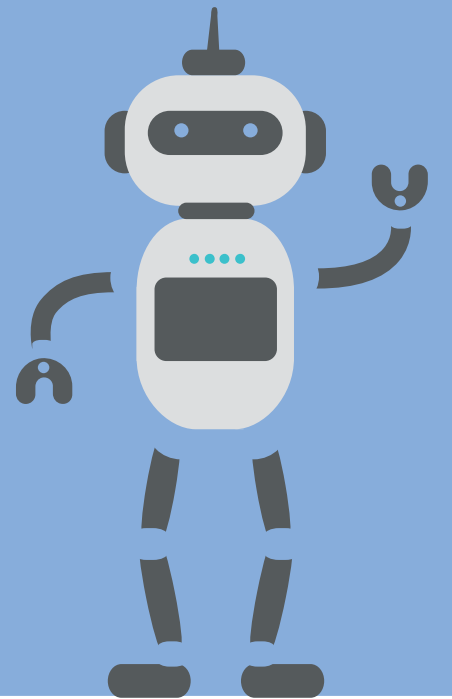
# 꾸준히 성장하는 데이터 분석가 방은호 입니다.

PORTFOLIO

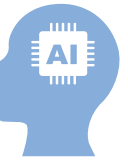
## CONTACT

010 - 2073 - 9671

[martusbang@naver.com](mailto:martusbang@naver.com)



# About Me



방은호 / Bang Eunho

생년월일: 1996.12.22

주소: 경기도 수원시 장안구 율전동

Tel. 010 - 2073 - 9671

Email. martusbang@naver.com

Github. [github.com/bbang-bbang](https://github.com/bbang-bbang)

Tistory. [bangeunho.tistory.com](https://bangeunho.tistory.com)

'유의미한 인사이트를 도출하는 **데이터 분석가** 방은호'

- 프로젝트와 자격증에 꾸준히 도전하며 자기 개발
- 데이터 분석을 위한 도메인 지식 학습
- 유의미한 인사이트 도출을 위해 노력

## 학력

2015 동원고등학교 졸업  
2021 성결대학교 졸업  
전공: 행정학과

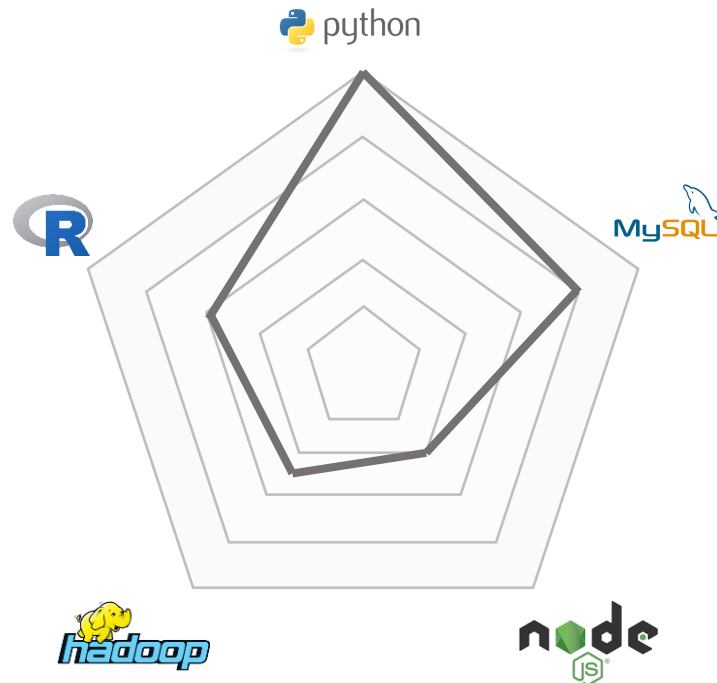
## 자격증

한국사 1급  
SQLD  
ADsP  
빅데이터 분석기사(필기)

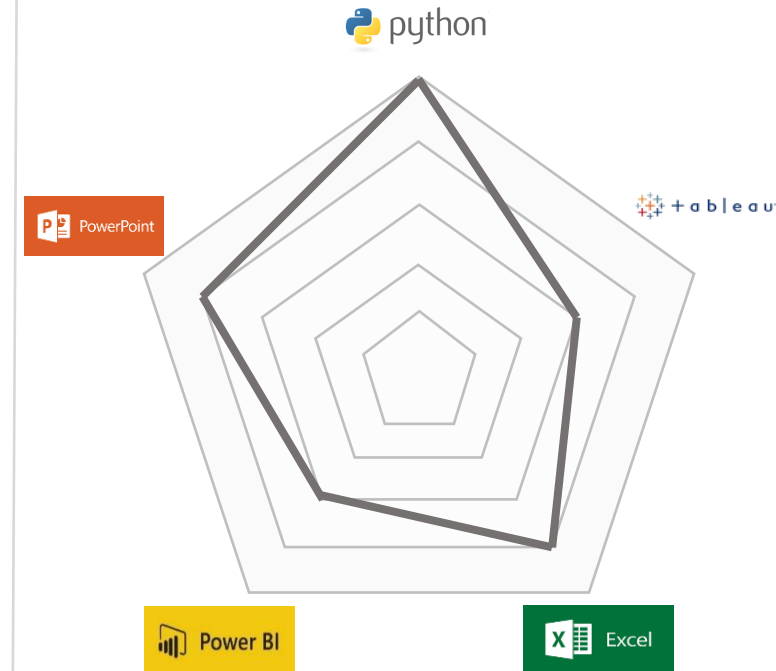
## 교육

2021.03. - 2021.08. 파이썬 기반 빅데이터 플랫폼&분석 전문가 양성과정[고용노동부 주관]  
2021.10. - 2022.04. 파이썬 기반 AI를 활용한 빅데이터 분석가 양성과정[과학기술정보통신부 주관]

## Programing Skill



## Visualize & OA



# Index :

---

## Data Analysis Projects

### 01 집단면역 위한 코로나 백신 접종

: 종류별 백신 접종률 파악 및 백신 접종의 효과성 제공

### 02 아동학대의 원인 추론

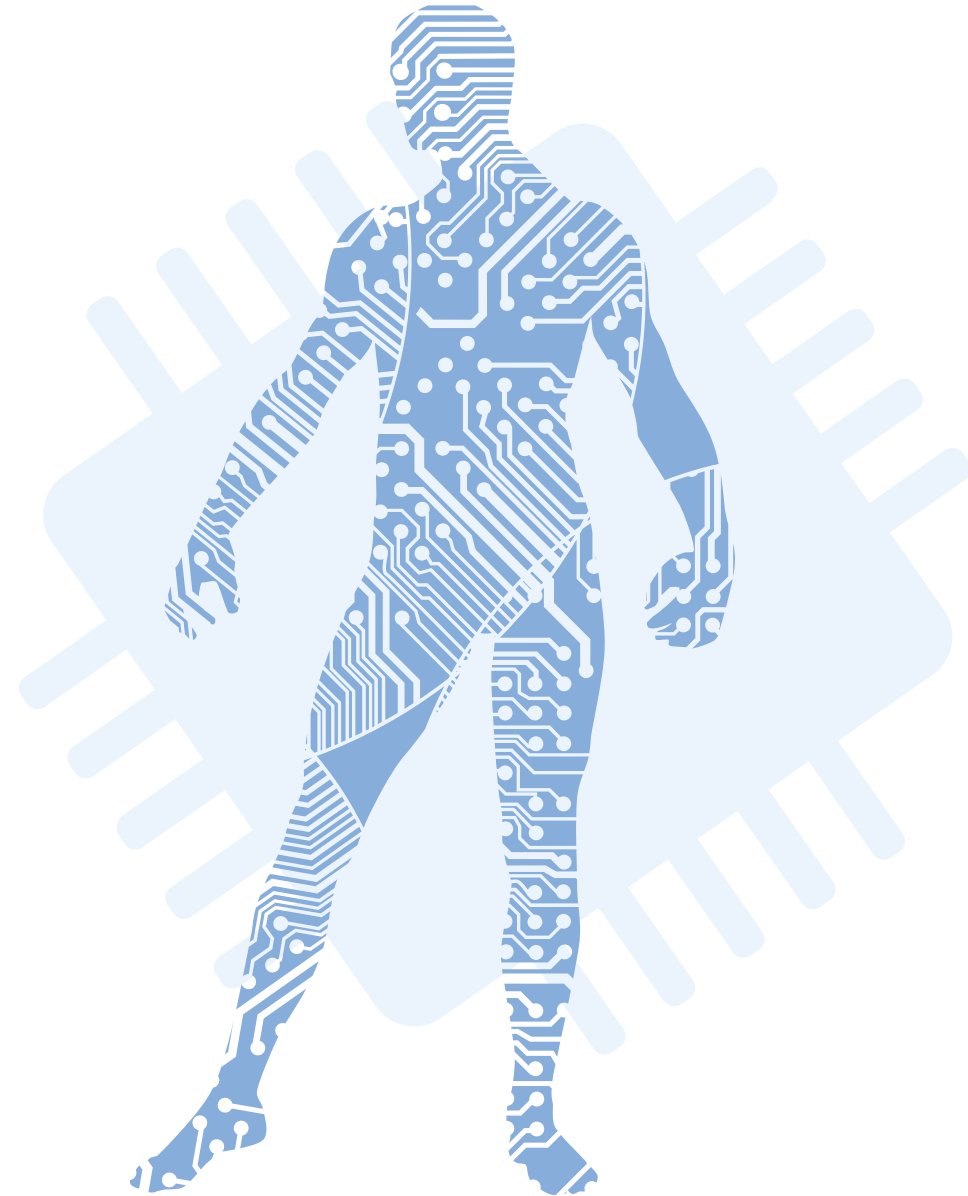
: 아동폭력의 원인에 대한 가설 수립 및 증명, 해결방안 제시

### 03 폐암의 원인 추론

: 폐암의 원인을 추론해보고, 합병증 간의 상관관계를 분석

### 04 데이터를 활용한 주식 투자 전략

: 주식 데이터를 기술적 지표를 활용하여 분석





*PROJECT. 1 (2021.06.~2021.07.)*

## 집단면역을 위한 코로나 백신 접종 권유

\* 주제: 영국의 백신 접종률 데이터 기반, 시각화를 통한 접종 중요성 권고

\* 가설: 해외사례 중 백신 접종 이후 코로나 감염율의 감소량 검증

\* 데이터 링크

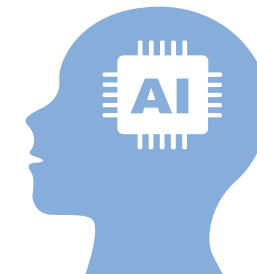
- 코로나 이상반응 데이터: <https://ncv.kdca.go.kr/menu.es?mid=a10118030000>
- 영국 백신 접종률 데이터: <https://coronavirus.data.gov.uk/details/vaccinations>
- 공공데이터 활용 지원 센터 데이터: <https://www.data.go.kr/tcs/dss/selectFile>

\* Used tool

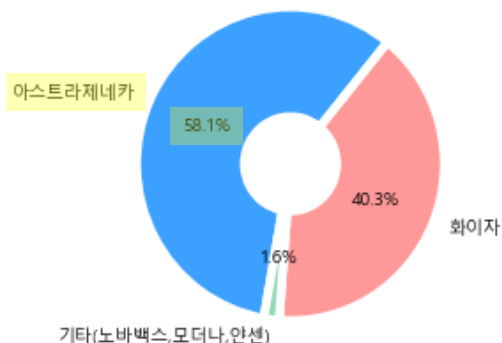


# Project 1

## EDA & 시각화



한국 백신 종류별 접종 비율(1-2분기)



Out[54]:

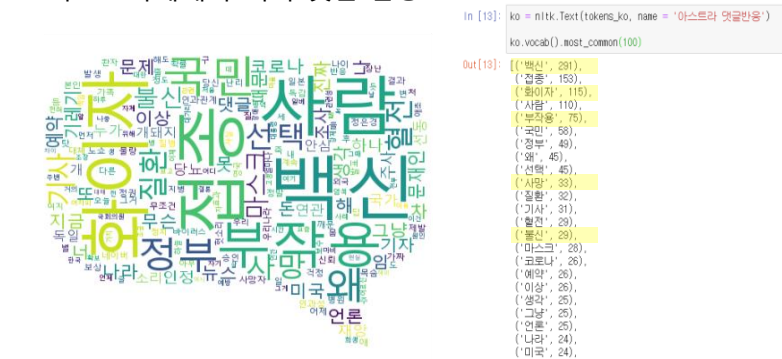
	구분	1분기	2분기	합계
0	아스트라제네카	200	866	1066
1	화이자	186	554	740
2	기타(노바백스, 모더나, 안센)	0	29	29

- 2021년 1, 2분기 백신 접종률 분석
- 아스트라제네카, 화이자, 기타 순으로 접종

< 아스트라제네카 기사반응 >

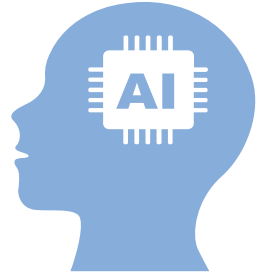


< 아스트라제네카 기사 댓글 반응 >



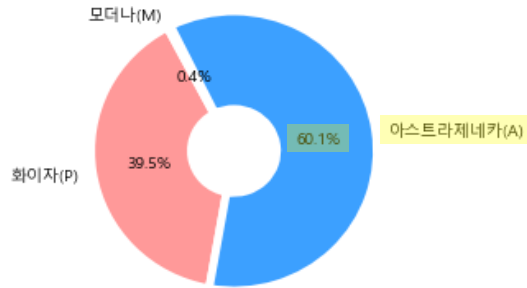
- ✓ 국민의 반 이상이 **아스트라제네카**를 접종했고, 추가적으로 접종 예정
- ✓ 아스트라제네카에 대한 네이버 기사와 댓글을 **Crawling + Word Cloud** 시각화
- ✓ 부정적인 단어가 상위 랭크 차지 ➡ **아스트라제네카에 대한 부정적인 여론**

# Project 1 : EDA & 시각화



## ■ 한국과 종류별 백신 접종률이 비슷한 영국의 데이터 분석

영국 백신 종류별 접종 비율



	백신	접종자수	이상반응	사망	이상반응 비율
0	화이자(P)	21600000.0	58065.0	374.0	0.268819
1	아스트라 제네카(A)	32900000.0	175057.0	786.0	0.532088
2	모더나(M)	200000.0	1462.0	2.0	0.731000

질병관리청 국외 코로나19 예방접종 후 이상반응 보고 현황  
2021. 5. 20

- ✓ 영국은 코로나 확진자가 가장 적고 백신 선진국으로 비교 대상 선정
- ✓ 영국의 2021년 1 - 2분기 백신 이상 반응 및 사망자의 비율 분석
- ✓ 하지만 이상반응이 조금이라도 있다면 심리적인 불안감을 느낌

➢ 백신접종에 따른 감염자 수 그래프 시각화  
→ **가설 검증**



	date	daily_corona	vaccinatedSum
0	2021-05-30	3240.0	39379411
1	2021-05-29	3398.0	39259168
2	2021-05-28	4182.0	39075642
3	2021-05-27	3542.0	38871200
4	2021-05-26	3180.0	38614683
...	...	...	...
135	2021-01-15	55761.0	3576263
136	2021-01-14	48682.0	3234946
137	2021-01-13	47525.0	2918252
138	2021-01-12	45533.0	2639309
139	2021-01-11	46169.0	2431648

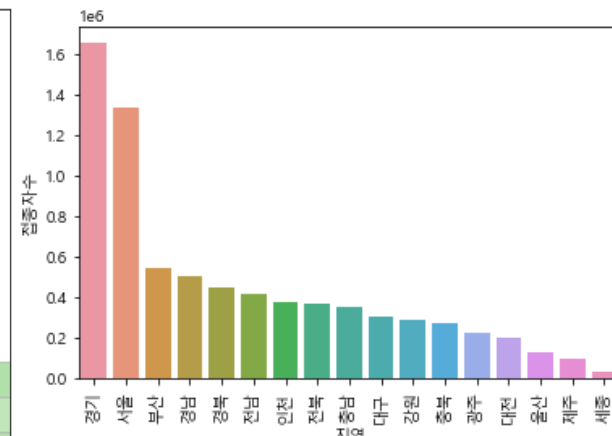
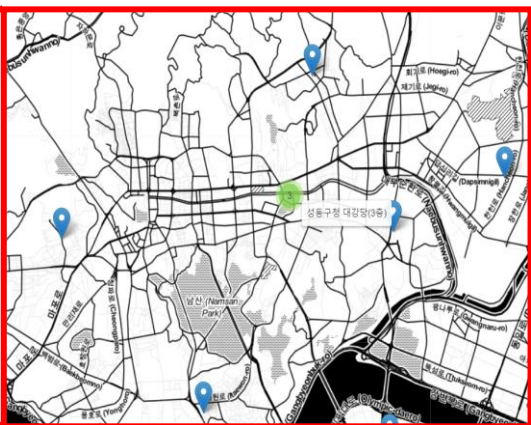
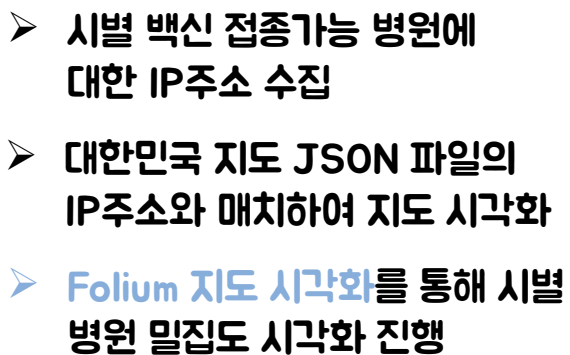
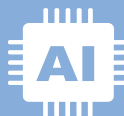
GOV.UK Coronavirus (COVID-19) in the UK  
2021. 5. 30

- 영국 정부의 2021.01 - 05 기간 데이터 사용
- 시각화 결과 백신을 맞는 인구가 증가할수록 일별 확진자 감소



즉, 이상반응 수치만큼 백신접종을 통한  
**코로나 예방 긍정적인 기대효과** 추론

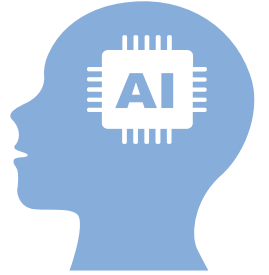
●



- ✓ 지도 시각화 결과 인구가 많이 거주하고 있는 **수도권, 특별시**에 집중된 것으로 파악
- ✓ 백신 접종 인원을 늘리기 위해서는 **접종 가능 병원 인프라 범위를 확대**시킬 필요가 있음
- ✓ 빈 칸으로 존재하는 곳은 부산 **중구** / 서울시 **중구**와 같이 동명이의 지역의 존재로 IP오류 발생



# Project 1 : EDA & 시각화



## ■ 현 시점 백신 접종자에 대한 혜택 내용

- ✓ 주요 공공시설의 입장료, 이용료 할인 혜택 제공
- ✓ 백신을 접종한 인원 대상으로 사적인 모임의 허용
- ✓ 요양병원 및 시설에 대한 면회 허용
- ✓ 영화 할인 쿠폰 제공

## ■ 수정 방안

- ✓ 해외에서는 예금 증서, 상품권, 현금카드와 같은 금전적인 혜택 진행
- ✓ 정부에서 이상 효과에 대한 원인 분석 및 피해보상을 철저히 진행하고, 백신 접종에 대한 적극적인 홍보 활동과 접종 가능 병원의 인프라를 확대 필요





*PROJECT. 2 (2021.11. ~ 2021.12.)*

## 아동학대의 원인 추론(공모전)

\* 주제: 데이터 기반 아동 학대 원인 분석 및 해결방안 제시

\* 가설: 1. 아동학대 증가 추세에 대한 진위 검증 & 2. 수입과 아동학대건수의 상관관계 검증

\* 제한 사항: **[통합 데이터 지도]**에 등록된 데이터만 사용 가능

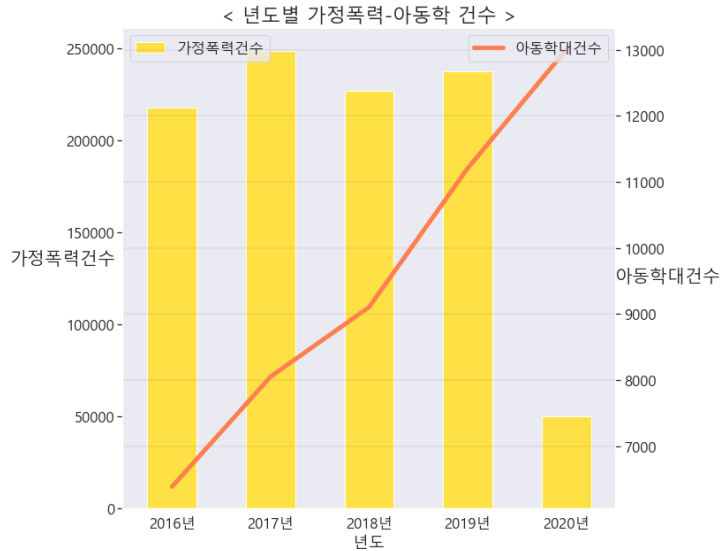
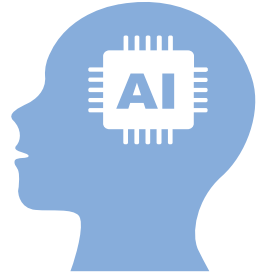
\* 데이터 링크

- 가정폭력 영향요소 융합데이터: [데이터 검색 | 통합 데이터 지도 \(bigdata-map.kr\)](#)
- 아동학대건수(13 - 20년) 데이터: [데이터 검색 | 통합 데이터 지도 \(bigdata-map.kr\)](#)
- 공공데이터 활용 지원 센터: [데이터 검색 | 통합 데이터 지도 \(bigdata-map.kr\)](#)
- 청소년 플랫폼 이용 데이터: [데이터 검색 | 통합 데이터 지도 \(bigdata-map.kr\)](#)

\* Used tool

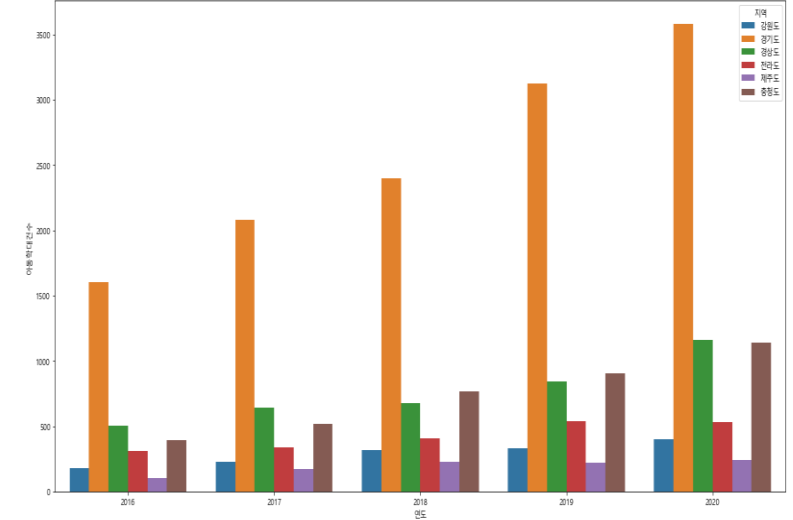
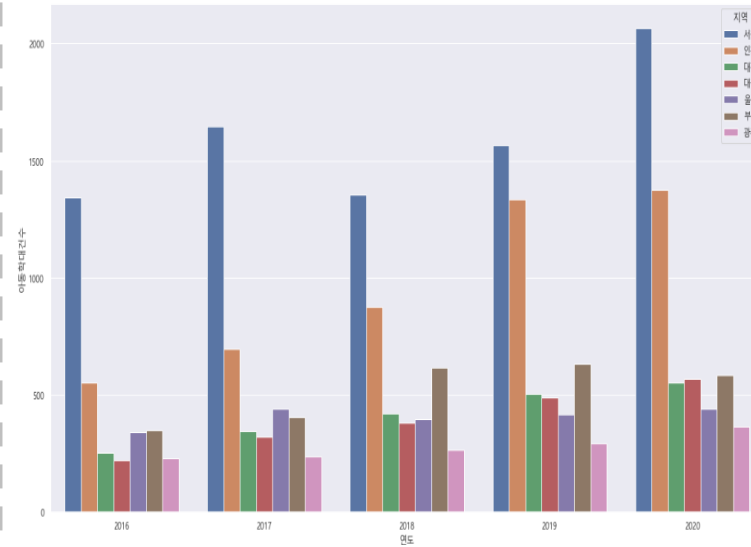


# Project 2 : EDA & 시각화



❖ 2020년 데이터는 1분기 데이터만 존재해 시각화의 한계 발생

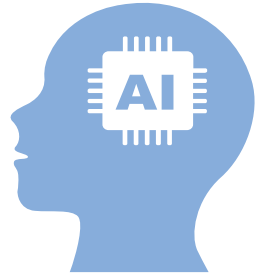
- 2016 - 20년도 가정폭력건수와 아동학대건수 비교 그래프 시각화
- 가정폭력 건수는 큰 변동이 없지만 **아동학대건수는 지속적으로 증가 추세**
- 아동학대 증가 추세 가설에 대한 **가설 증명**



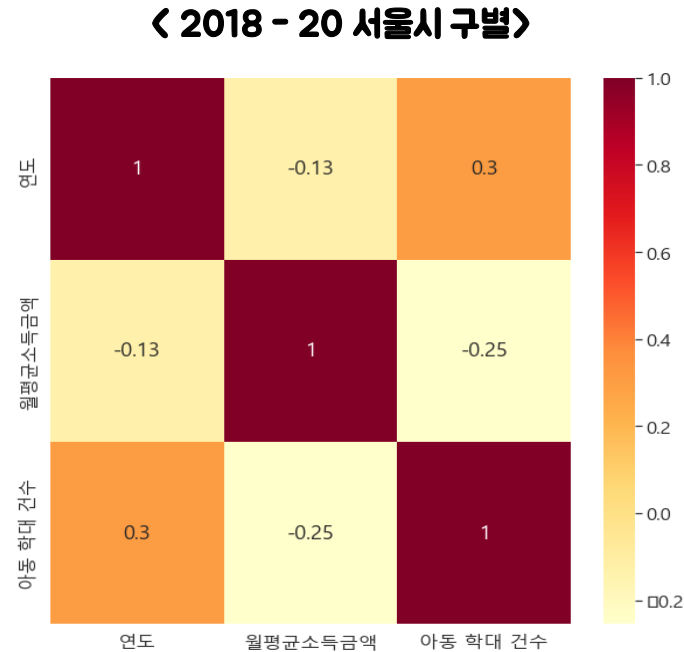
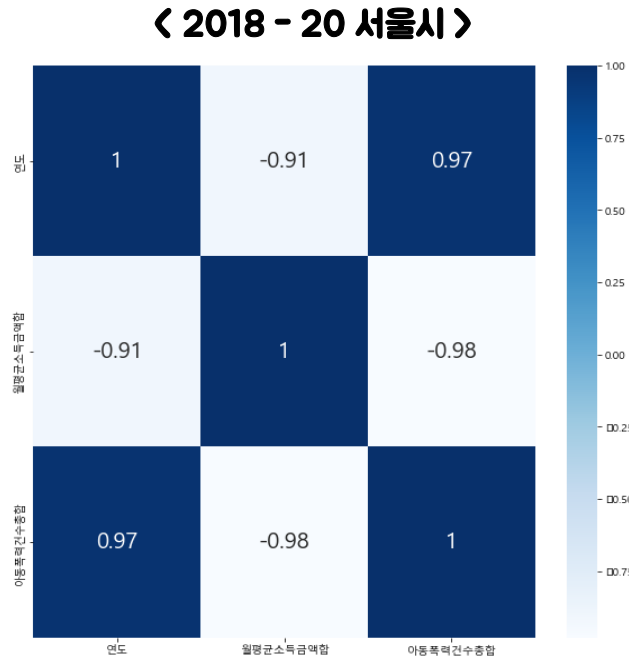
- 특별/광역시별, 도별 아동학대건수에 대한 시각화 결과 **서울, 경기도**가 상위에 위치
- 인구가 많은 지역일수록 아동학대건수도 많은 **비례관계**인 것으로 추론
- 아동학대건수가 많은 서울을 기준으로 아동학대건수와 가정소득의 상관분석 진행

➡ **가설 검증 진행**

# Project 2 : EDA & 시각화



## 18 - 20년도 서울시 아동학대건수와 월평균소득의 상관분석 진행



```
import pandas as pd
data = pd.read_csv('2019_1.csv')
data
data = data.loc[data['행정동코드'] == 11110]
jongro = data['월평균소득액'].sum()
jongro/16
```

```
import pandas as pd
data = pd.read_csv('2019_1.csv')
data
data = data.loc[data['행정동코드'] == 11740]
gangdong = data['월평균소득액'].sum()
gangdong/17
```

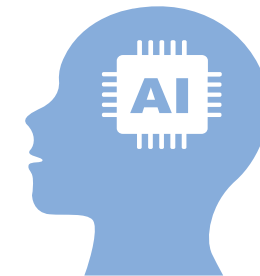
총 18구

- 서울시 월평균 소득 금액과 아동폭력 건수는 높은 상관관계
- 서울시 구별 월평균 소득 금액과 아동폭력 건수는 독립적인 관계

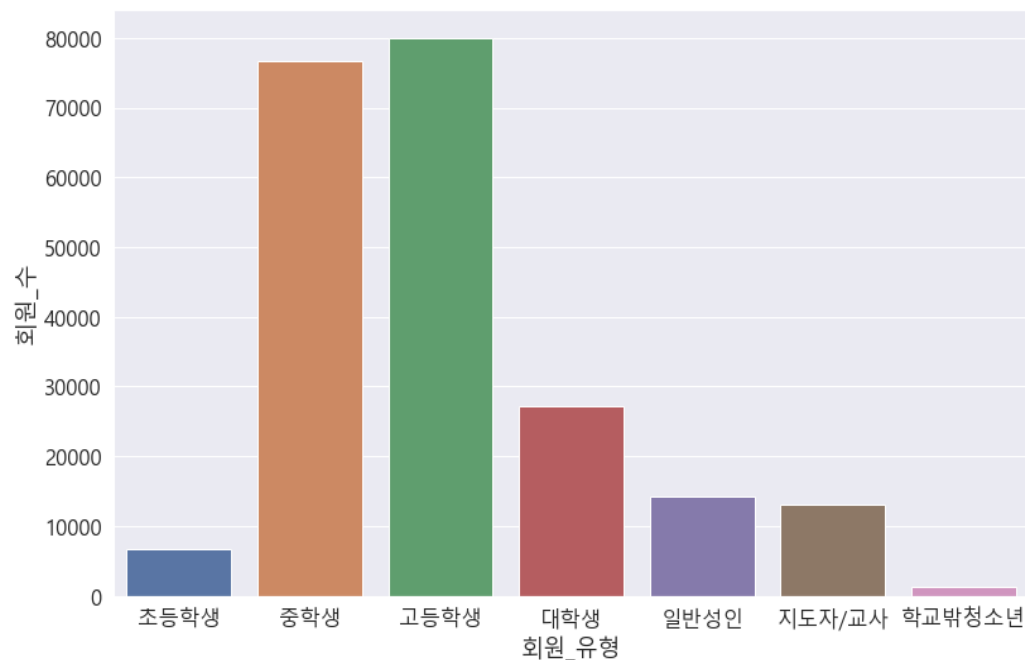


- ✓ 서울시 전체 인구를 기준일 경우 평균소득금액이 아동학대의 원인으로 추론할 수 있지만, 구별로 세분화할 경우 절대적인 원인으로 정의할 수 없음  
+) 구별로 소득의 정도와 동의 개수가 동일하지 않기에 오류가 존재
- ✓ 가정의 소득이 아동학대의 원인이라는 가설에 대한 가설 증명

# Project 2: EDA & 시각화



## ■ 결론



- ✓ 플랫폼 이용회원이 대부분 청소년(만 18세 미만)이므로 아동학대의 당사자를 만날 수 없는 경우, 플랫폼을 활용하여 아동학대 방지 및 대처방안에 대한 홍보 방법 제안
- ✓ 아동학대에 대해서 국가적으로 경각심을 가지고 정부는 아동학대 관련부서와 그에 관련된 모든 부서에 전문가를 양성하는 인적 지원과 경제적 지원 필요
- ✓ 서울 지역 외 타지역에 대한 데이터들이 충분하지 않아 지역별 비교가 불가능하게 되어 더 많은 결과를 도출하지 못해 아쉬움



*PROJECT. 3 (2021.12.~2022.01.)*

## 폐암의 원인 추론

\* 주제: 폐암 발병 원인 분석 및 합병증과의 상관분석

\* 가설: 1. 폐암의 절대적인 원인 흡연에 대한 검증 & 2. 폐암과 결핵의 관계성 검증

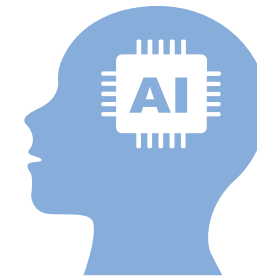
\* 데이터 링크([공공데이터포털](#))

- 국립암센터\_폐암 환자 수, 직업 데이터
- 국립암센터\_폐암 환자 음주, 흡연 여부 데이터
- 국립암센터\_종류별 세부진단, 암 병력 여부 데이터
- 국립암센터\_폐암 환자 고혈압, 결핵, 간질환, 당뇨병, 심장질환 병력 여부 데이터

\* Used tool



# Project 3 : EDA & 시각화



## 데이터 수집

- 모든 데이터의 형태가 API(JSON + HTML)구조
- REQUESTS 라이브러리를 활용해 API 데이터를 JSON 구조로 변환, 데이터프레임화

```
import requests

my_url = 'http://apis.data.go.kr/B551172/Lung09/luTubercholsisByType?'
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)
```

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

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

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

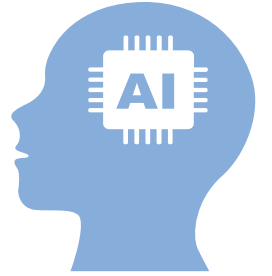
```
jsdf = pd.DataFrame(js_file['response']['items']['item'])
```

```
jsdf
```



	centerNm	critYr	ptAge	ptSexCd	statsTrgtNm	ncsNmvl	whoINcsDnmvl	ptCntNmvl	whoIPtCntDnmvl
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

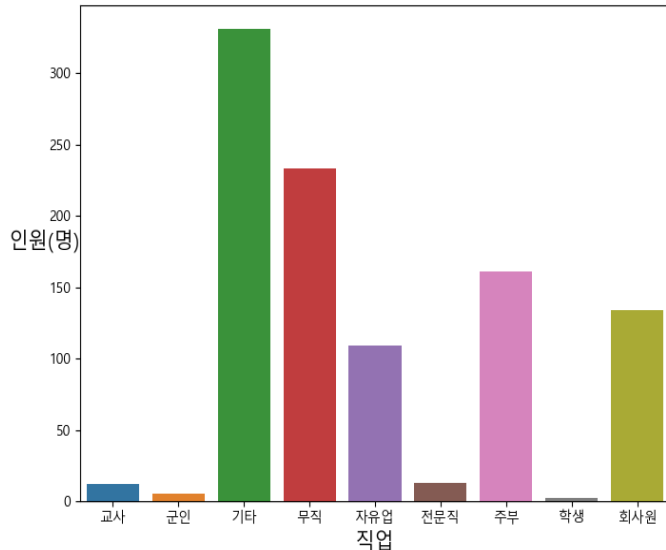
# Project 3 : EDA & 시각화



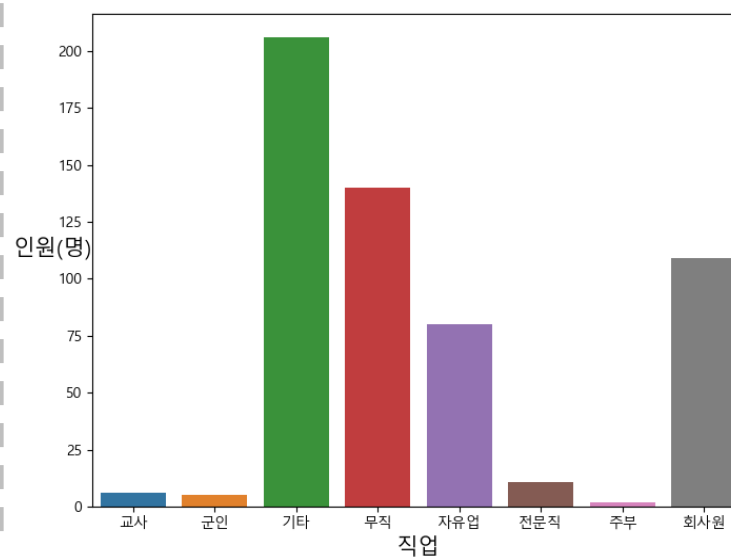
## ■ 직업별 폐암 환자 시각화

- 직업별 폐암환자 비율은 기타 > 무직 > 주부 순으로 상위를 차지
- 남성, 여성으로 나눈 결과, 남성은 기타 / 무직이 상위를 차지하고, 여성은 주부가 상위를 차지
- 남성 그래프를 통해 기타, 무직에 **대학생 또는 건설현장과 같은 4대 보험 미적용 직업들이** 속해 있을 것으로 추론
- 실제로 **[주부]** 직업이 폐암 발병에 원인을 제공하는지 조사 진행 ➡ **2021년 조리사 폐암 산재 인정**

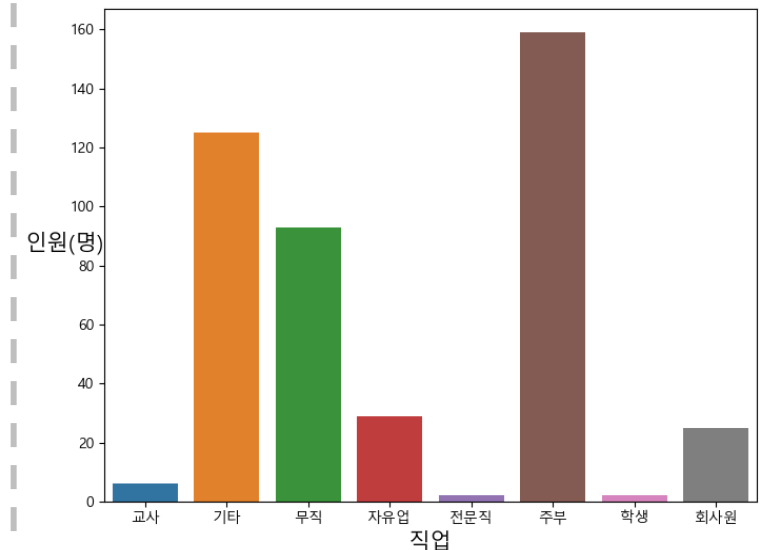
〈전체 폐암 환자〉



〈남성 폐암 환자〉

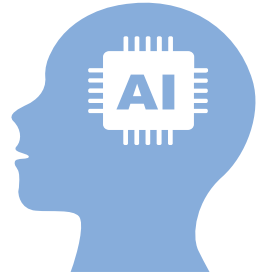


〈여성 폐암 환자〉





# Project 3 : EDA & 시각화



## ■ 흡연, 음주 여부와 폐암과의 관계성 분석(선형회귀, OLS 라이브러리 사용) → 가설 검정

➢ OLS : 최소제곱법, 선형모델의 파라미터를 추정하는 가장 기본적인 방법

### < 흡연 여부 >

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

0.11158296326569894

```
1 from sklearn.metrics import mean_squared_error
2 a_pred = lr.predict(X_test_val)
3
4 a = mean_squared_error(y_test_val, a_pred)**0.5
5 a
```

0.5702095684301874

### OLS

```
1 import statsmodels.api as sm
2
3 model = sm.OLS(y, X)
4 result = model.fit()
5 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

### < 음주 여부 >

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

0.029424695060832562

```
1 from sklearn.metrics import mean_squared_error
2 a_pred = lr.predict(X_test_val)
3
4 a = mean_squared_error(y_test_val, a_pred)**0.5
5 a
```

0.5115734565248449

### OLS

```
1 import statsmodels.api as sm
2
3 model = sm.OLS(y, X)
4 result = model.fit()
5 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

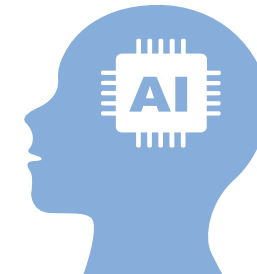
Omnibus:	4402.855	Durbin-Watson:	1.981
Prob(Omnibus):	0.000	Jarque-Bera (JB):	135.401
Skew:	0.112	Prob(JB):	3.96e-30
Kurtosis:	1.211	Cond. No.	1.00



• 음주가 흡연보다 더 높은 폐암 발병에 대한 관계성을 보임 → 가설 검증

• 폐암은 꼭 흡연이 아니더라도 직업, 환경적인 원인이 있기에, 접근성이 높은 음주 여부가 더 높게 나왔을 것으로 추론

# Project 3 : EDA & 시각화



- 폐암환자의 합병증 관계성 분석(결핵, 당뇨병, 고혈압, 간질환, 심장질환) → 가설 검증

OLS Regression Results

< 결핵 >

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.

OLS Regression Results

< 간질환 >

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

OLS Regression Results

< 당뇨병 >

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.

OLS Regression Results

< 심장질환 >

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.

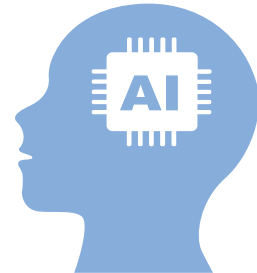
OLS Regression Results

< 고혈압 >

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.

- 가설에서 세운 결핵보다 **고혈압**이 폐암과 더 높은 관계성을 지님 → 가설 검증
- 고혈압약 **ACE 억제제**가 폐암 위험을 증가시키며, 특히 **10년 이상 복용한 환자는 폐암 위험이 31% 높다**는 연구 결과가 나옴

# Project 3 : EDA & 시각화



## ■ 흡연, 음주 여부와 고혈압과의 관계성 분석

- 흡연보다 음주가 고혈압 발병과 관련성이 높음( $0.360 > 0.178$ )

### < 음주 + 고혈압 >

```
1 lr.score(scaled_minmax_test, y_test_val)
0.029424695060832562
```

```
1 from sklearn.metrics import mean_squared_error
2 a_pred = lr.predict(X_test_val)
3
4 a = mean_squared_error(y_test_val, a_pred)**0.5
5 a
0.5115734565248449
```

#### OLS

```
1 import statsmodels.api as sm
2
3 model = sm.OLS(y, X)
4 result = model.fit()
5 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

### < 흡연 + 고혈압 >

```
1 lr.score(X_test_val, y_test_val)
-0.01169134478390288
```

```
1 from sklearn.metrics import mean_squared_error
2 a_pred = lr.predict(X_test_val)
3
4 a = mean_squared_error(y_test_val, a_pred)**0.5
5 a
0.49632968663479937
```

#### OLS

```
1 import statsmodels.api as sm
2
3 model = sm.OLS(y, X)
4 result = model.fit()
5 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



*PROJECT. 4 (2022.01.~2022.03.)*

## 데이터를 활용한 주식 투자 전략

\* 주제: 기술적 지표를 활용한 주식 데이터 분석

\* 가설: 1. 기술적 분석에 대한 성능 검증 & 2. 추세별 기술적 지표의 성능 검증

\* 데이터 링크

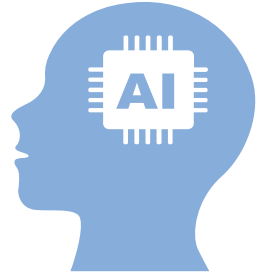
- 한국거래소 KOSPI 200 기업 주식 데이터
- Yahoo Finance KOSPI 200 기업 주식 데이터
- 네이버 증권 재무제표 데이터

\* Used tool



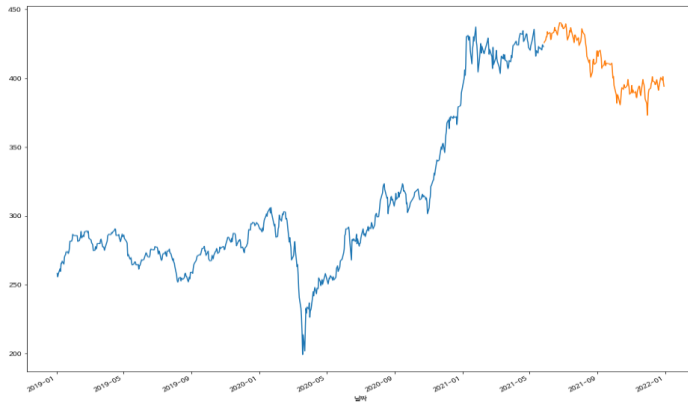


# Project 4 : EDA & 시각화



## 기술적 분석(ARIMA) - KOSPI 200 종목 데이터

### < 종가 데이터 >



### < 최적 ARIMA 모형 >

```
Performing stepwise search to minimize aic
ARIMA(1,1,1)(0,0,0)[0] intercept : AIC=3361.466, Time=0.24 sec
ARIMA(0,1,0)(0,0,0)[0] intercept : AIC=3360.247, Time=0.03 sec
ARIMA(1,1,0)(0,0,0)[0] intercept : AIC=3361.760, Time=0.10 sec
ARIMA(0,1,1)(0,0,0)[0] intercept : AIC=3361.849, Time=0.12 sec
ARIMA(0,1,0)(0,0,0)[0] : AIC=3360.943, Time=0.03 sec

Best model: ARIMA(0,1,0) intercept
Total fit time: 0.515 seconds
```

### < 잔차 검정 >

C- SARIMAX Results

Dep. Variable: y No. Observations: 593

Model: SARIMAX(0, 1, 0) Log Likelihood -1678.123

Date: Wed, 30 Mar 2022 AIC 3360.247

Time: 09:12:00 BIC 3369.014

Sample: 0 HQIC 3363.662

- 593

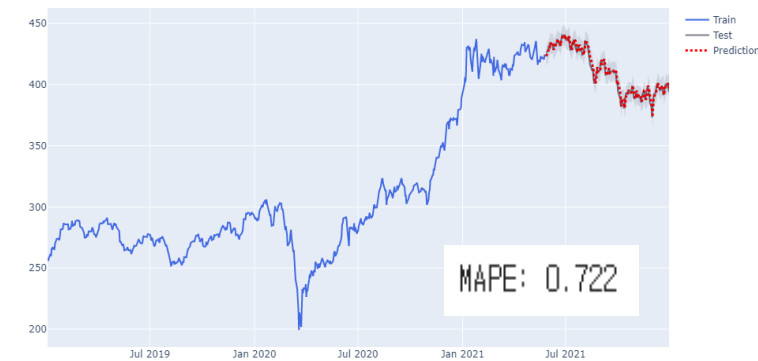
Covariance Type: opg

<

- ✓ **Jarque-Bera**
  - 잔차가 정규성을 따름
- ✓ **Ljung-Box**
  - 잔차는 자기상관성이 존재하지 않음(귀무 가설)
  - 귀무 가설 채택

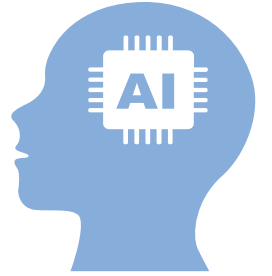
잔차는 자기상관성이  
존재 하지 않으며 비정규성 지남

### < 예측 시각화 >

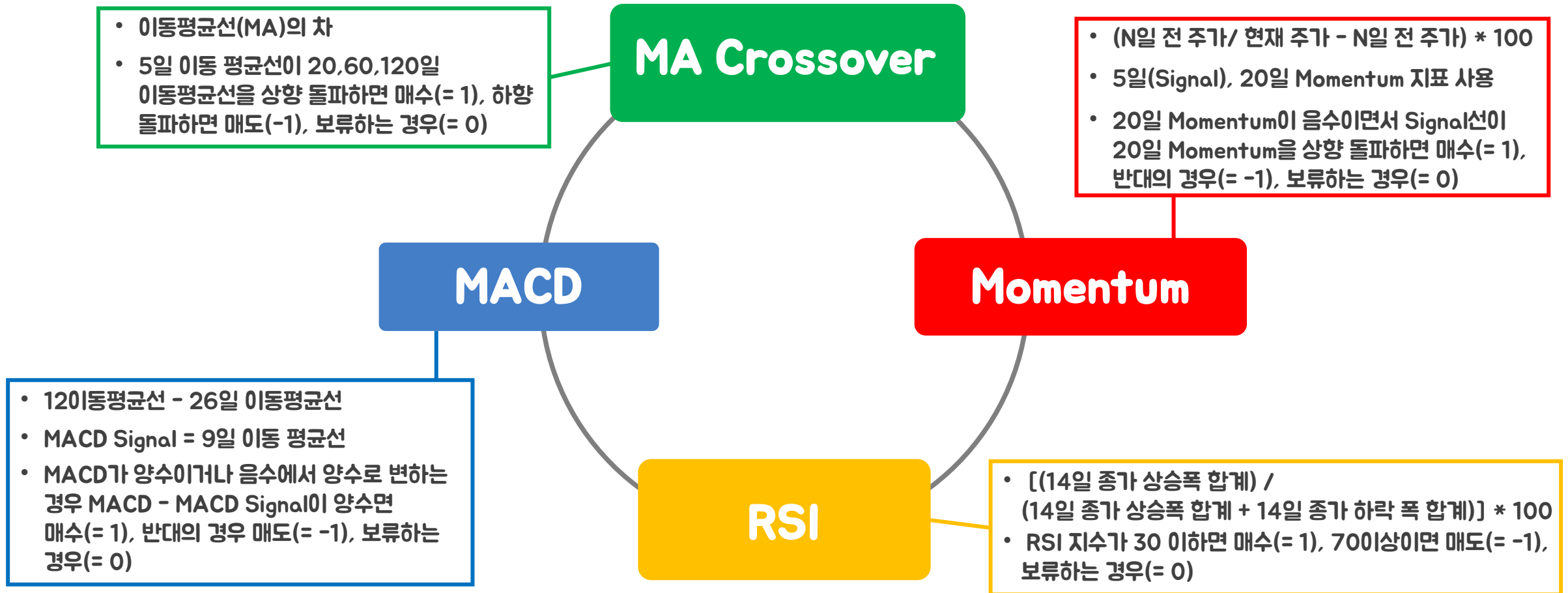


- ✓ 한 번에 예측하는 것이 아니라, 일자를 기준으로 하나씩 예측하고, 관측치(Close)를 업데이트
- ✓ MAPE(Mean Absolute Percentage Error)
- ✓  $[ |(\text{예측값} - \text{실제값})| / \text{실제값} ] * 100$
- 잔차가 실제값의 0.722%

# Project 4 : EDA & 시각화

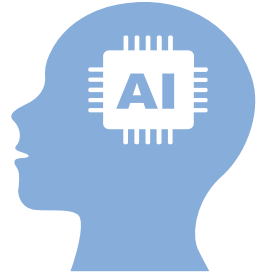


## 기술적 분석(MA Crossover, Momentum, RSI, MACD)





# Project 4 : EDA & 시각화



## ■ 앙상블(Classification) - 삼성전자

- 각 지표의 매수(1), 매도(0) 기준을 설정하고 Classification 앙상블 진행
- 과적합으로 성능이 좋게 나온 것일 수도 있지만 성능이 좋은 것으로 추론 → **가설 검증**

### < RSI >

```
SVC() = 0.5100671140939598
GaussianNB() = 0.6510067114093959
RandomForestClassifier() = 0.5973154362416108
AdaBoostClassifier() = 0.5838926174496645
SGDClassifier() = 0.47651006711409394
LogisticRegression() = 0.6174496644295302
KNeighborsClassifier(n_neighbors=9) = 0.5302013422818792
LGBMClassifier() = 0.6308724832214765
```

### < MACD >

```
SVC() = 0.6174496644295302
GaussianNB() = 0.5704697986577181
RandomForestClassifier() = 0.8322147651006712
AdaBoostClassifier() = 0.7986577181208053
SGDClassifier() = 0.6174496644295302
LogisticRegression() = 0.7919463087248322
KNeighborsClassifier(n_neighbors=9) = 0.5973154362416108
LGBMClassifier() = 0.8389261744966443
```

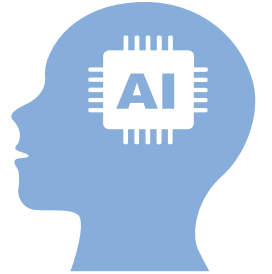
### < Momentum >

```
SVC() = 0.42953020134228187
GaussianNB() = 0.5838926174496645
RandomForestClassifier() = 0.5369127516778524
AdaBoostClassifier() = 0.47651006711409394
SGDClassifier() = 0.47651006711409394
LogisticRegression() = 0.5570469798657718
KNeighborsClassifier(n_neighbors=9) = 0.5033557046979866
LGBMClassifier() = 0.5369127516778524
```

### < MA Crossover >

```
SVC() = 0.5704697986577181
GaussianNB() = 0.6845637583892618
RandomForestClassifier() = 0.9261744966442953
AdaBoostClassifier() = 0.8926174496644296
SGDClassifier() = 0.5637583892617449
LogisticRegression() = 0.9328859060402684
KNeighborsClassifier(n_neighbors=9) = 0.5302013422818792
LGBMClassifier() = 0.912751677852349
```

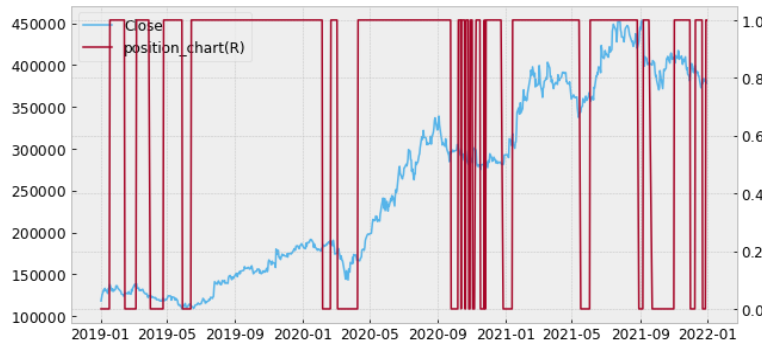
# Project 4 : EDA & 시각화



## 백테스팅 - KOSPI 200 기업 데이터(200 기업)

- 백테스팅 : 과거의 데이터를 이용해 투자전략을 시험해 보는 과정(MACD지표를 기준 비교)
- 벤치마크 : 특별한 기법 적용 없이 시작일에 매수하여 마지막날 매도하는 방식(수익률 평가 기준)

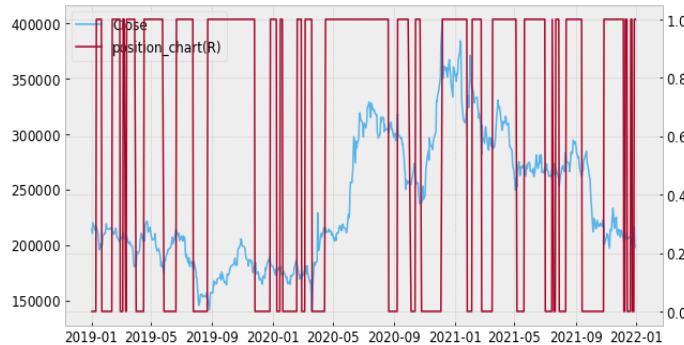
〈 NAVER(상승 추세) 〉



수익률(%) : 71.56	벤치마크	
	수익률	MDD
	220.76	-25.33

상승 추세에서는 **기술적 지표**가  
벤치마크보다 수익률이 낮음 → **가설 검증**

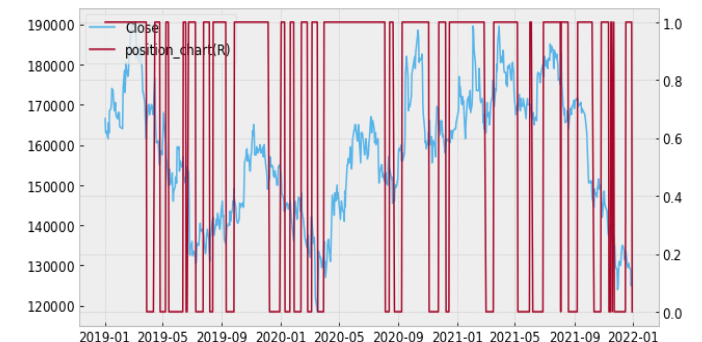
〈 POSCO(횡보 추세) 〉



수익률(%) : 372.88	벤치마크	
	수익률	MDD
	-7.69	-51.18

횡보 추세에서는 **기술적 지표**가  
벤치마크보다 수익률 높음

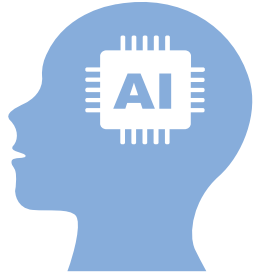
〈 CJ대한통운(하락 추세) 〉



수익률(%) : 372.88	벤치마크	
	수익률	MDD
	-24.32	-37.8

하락 추세에서는 **기술적 지표**가  
벤치마크보다 손실률이 낮음

# Project 4 : 분석 결과




## ✓ 결론

- 최대 수익률 : **저점에서 매수 / 고점에서 매도** 하지만 상승, 하락을 **예측하고 투자하는 것은 불가능**
- 현업에서는 기술적 지표를 활용한 수익률 보다 MDD(최대 손실율) 지표 활용
- 기술적 지표를 활용한 개인의 스탠스 성립할 경우 **최악의 결과** 예방 가능

감사합니다.  
잘 부탁드립니다.



CONTACT

 010 - 2073 - 9671

 [martusbang@naver.com](mailto:martusbang@naver.com)