한양대학교 언론정보대학 정보사회미디어학과
데이터사이언스 학회 DAYS

ORIENTATION

hy.days.official@gmail.com

ABOUT DAYS

- Data AnalYsis Society
  - Data Science
    - Data Analysis
    - Machine Learning
    - Deep Learning

hy.days.official@gmail.com

HOW?

- Statistics & Math
- Programming
- Domain Knowledge

hy.days.official@gmail.com

WHY?

- 의사결정을 위한 도구

hy.days.official@gmail.com

WHERE?

- 우리가 살아가고 있는 이 사회의 모든 곳

hy.days.official@gmail.com

IN DAYS

- 밑바닥을 탄탄하게 다지는 기초 공사
- 원하는 분야로 진출할 수 있는 밑거름
- 언젠가 반드시 겪어보아야 할 시행착오

hy.days.official@gmail.com

IN DAYS - 2020년 2학기

- 1. 데이터 분석의 기초 기술 -> Python/Pandas
- 2. 기초 통계학 -> Basic Statistics
- 3. 개인 프로젝트 경험 -> 데이터 분석 Flow

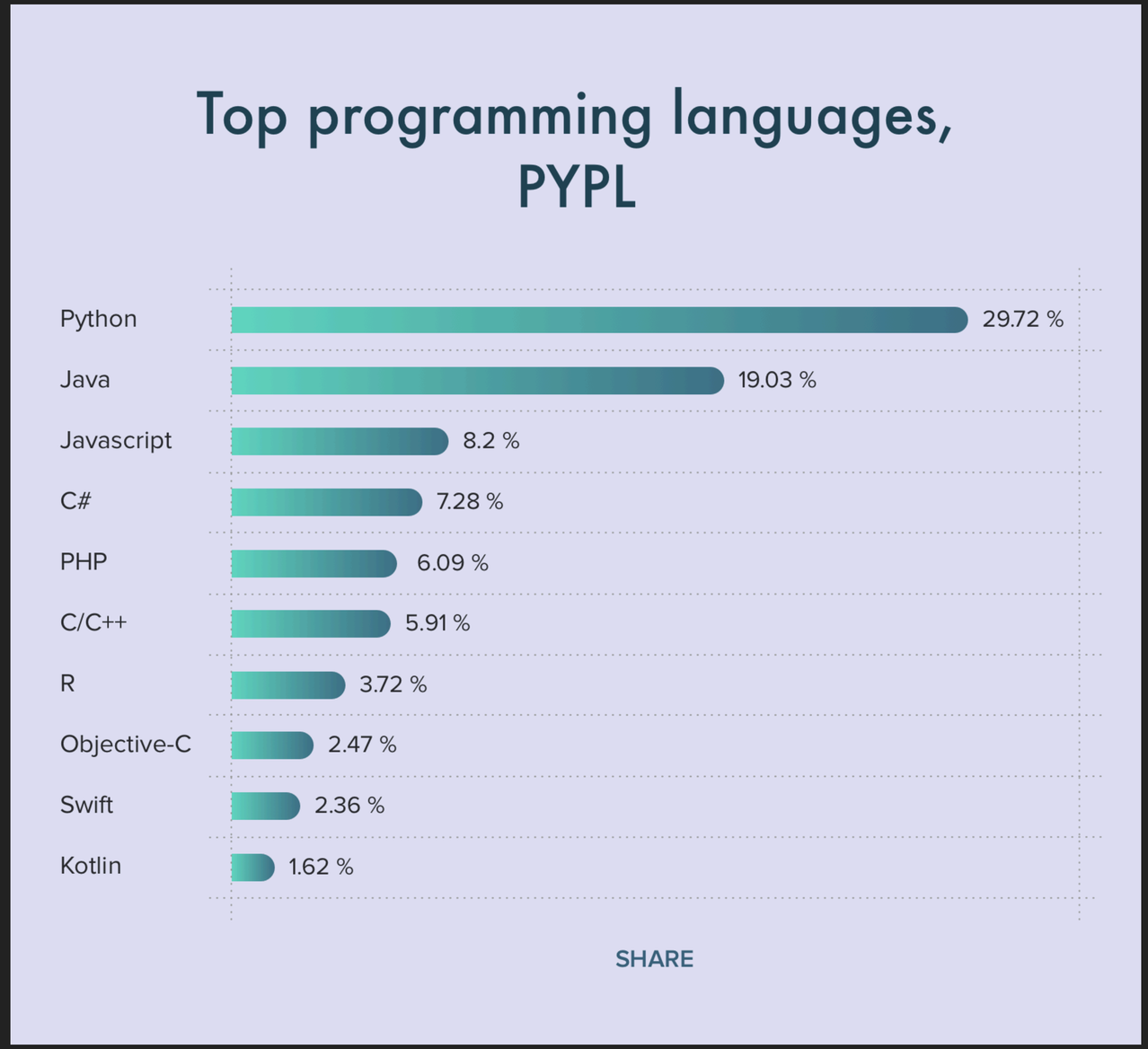
hy.days.official@gmail.com

PYTHON VS R

hy.days.official@gmail.com

PYTHON

- 더욱 다양한 분야에 활용 가능
- 쉽게 배울 수 있음



hy.days.official@gmail.com

PYTHON WITH JYPTER NOTEBOOK

- Python = 인터프리터 프로그래밍 언어
- Pandas = Python Library,
  Python에서 데이터를 다룰 수 있도록 도와주는 도구

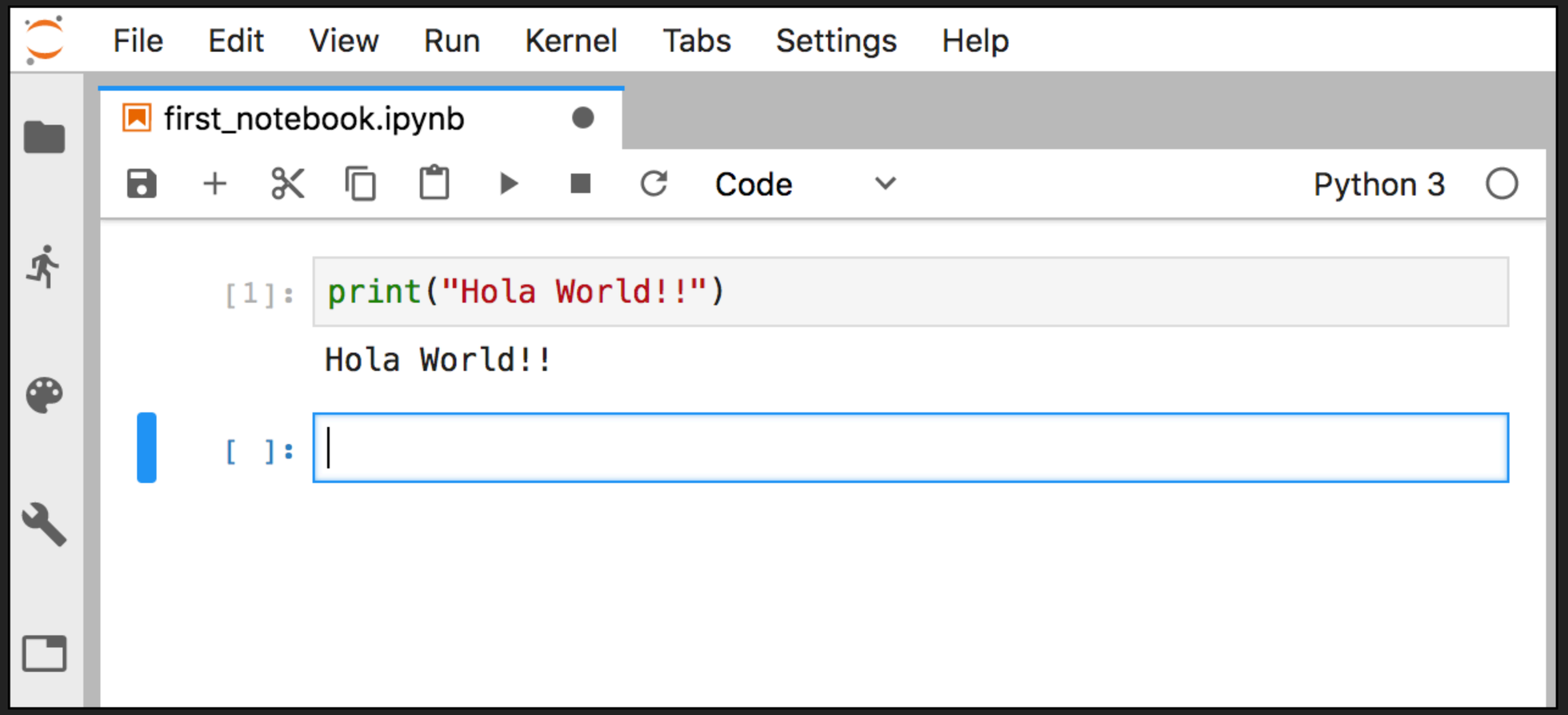
settings.py | Python - Get Started | crawler.py ● | views.py

myProject > likelion-crawling_bgm > heechan > testProj > testApp > crawling > cr

```python
from urllib.parse import quote_plus
from bs4 import BeautifulSoup
from selenium import webdriver


def crawler():
    url = 'https://vibe.naver.com/chart/total'

    driver = webdriver.Chrome(
        executable_path='C:/Users/Admin/Desktop/likelion-cr

    driver.get(url)

    html = driver.page_source
    soup = BeautifulSoup(html)
    title = soup.select('.tracklist .link_text')
    rank = soup.select('.rank .text')
    singer = soup.select('tbody .artist')
    resultArr = []

    for i in range(len(title)):
        tempObj = {}
        tempObj['rank'] = rank[i].get_text()
        tempObj['title'] = title[i].attrs['title']
        tempObj['singer'] = singer[i].attrs['title']
        resultArr.append(tempObj)

    driver.close()
    return resultArr

```

hy.days.official@gmail.com

PYTHON WITH JYPTER NOTEBOOK



hy.days.official@gmail.com

BASIC STATISTICS

- 산포도, 상관관계, 인과관계, 종속성, 독립성 ... ...
- 확률, 연속분포, 정규분포, 중심극한정리 ... ...
- 통계적 가설검정, P값, 신뢰구간 ... ...

hy.days.official@gmail.com

데이터 분석 FLOW

- 데이터 EDA (Exploratory Data Analysis)
- 데이터 전처리
- 분석 모델링
- 분석 모델 평가

- 최종 결과 도출

hy.days.official@gmail.com

데이터 분석 FLOW

In [1]:
```
import pandas as pd
```

In [5]:
```
df = pd.read_csv("./days_member.csv")
df
```

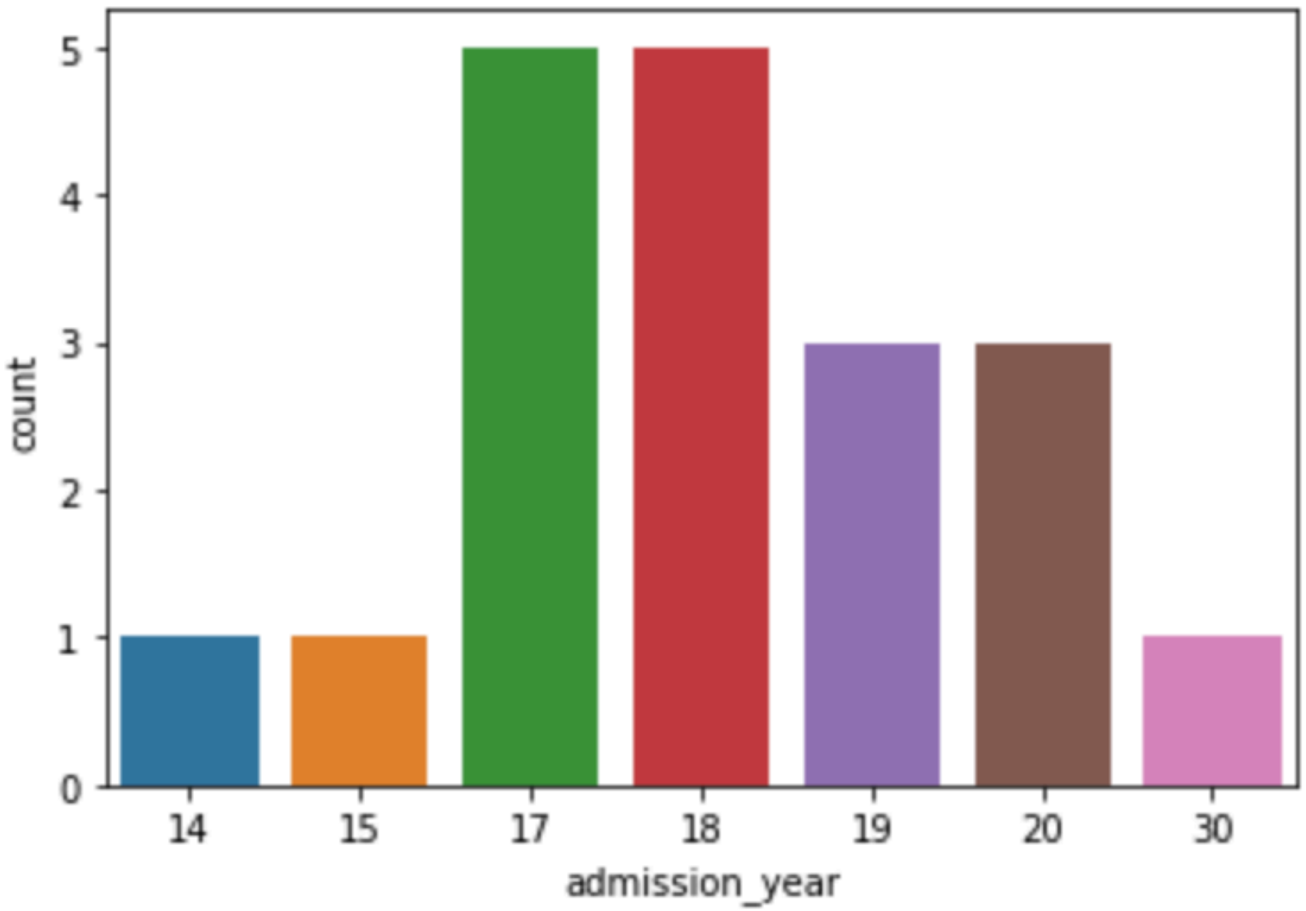
Out[5]:

	name	admission_year	position
0	김재훈	15	lead_member
1	차주희	18	lead_member
2	이정규	14	lead_member
3	서재현	17	lead_member
4	이수진	17	lead_member
5	박재현	17	lead_member
6	한예림	18	lead_member
7	배나영	19	lead_member
8	장예림	17	normal_member
9	임연수	18	normal_member
10	정은진	20	normal_member
11	정현수	17	normal_member
12	이유진	18	normal_member
13	장시은	19	normal_member
14	최동연	20	normal_member
15	이재성	18	normal_member
16	이재원	19	normal_member
17	최은선	20	normal_member
18	데이즈	30	member

hy.days.official@gmail.com

데이터 분석 FLOW

```
ax = plt.subplots()
ax = sns.countplot(df["admission_year"])
```

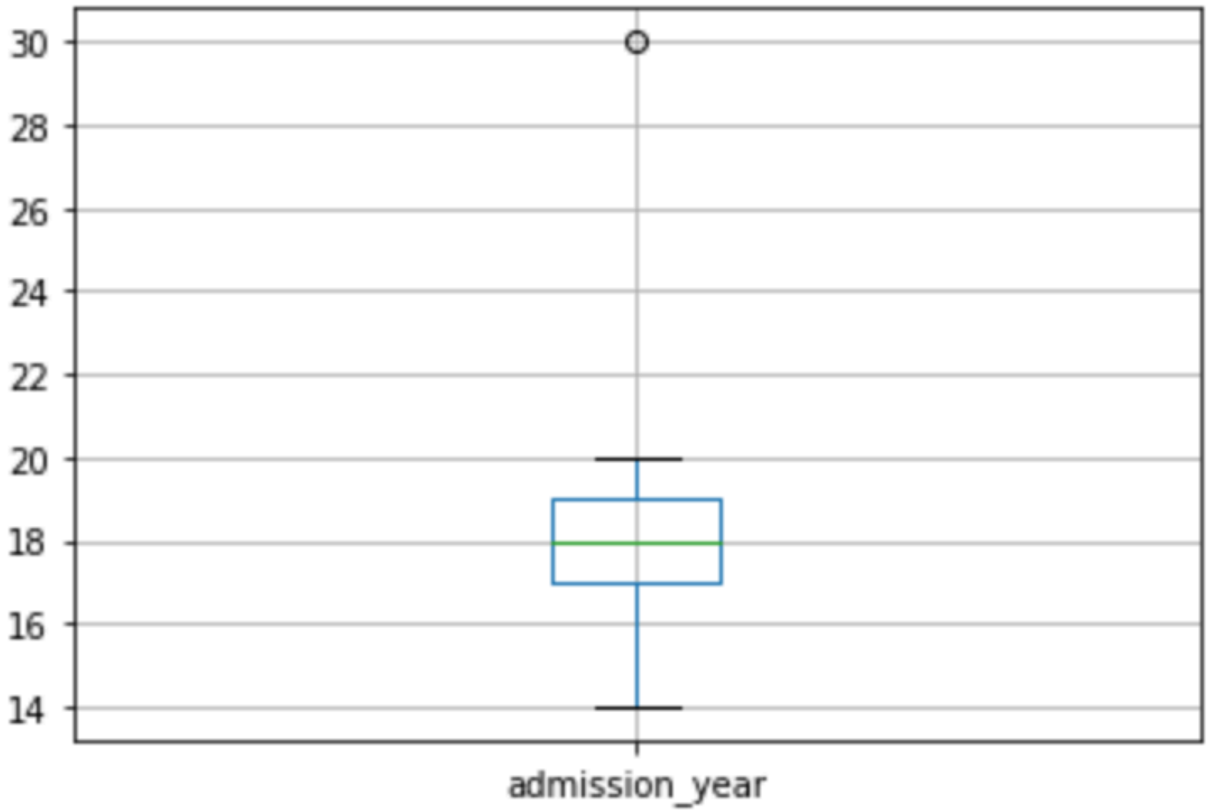


hy.days.official@gmail.com

데이터 분석 FLOW

```
df.boxplot()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fb8f9e1f610>
```



hy.days.official@gmail.com

개인 프로젝트

- 자율 주제 선정 (회귀분석 관련 프로젝트 추천)
  - Ex. 프로 야구선수(타자)의 시즌당 홈런수와 연봉 간의 상관관계 분석
  - Ex. 일정 범위 내의 카페 숫자와 주택 가격 간의 상관관계 분석
  - etc...

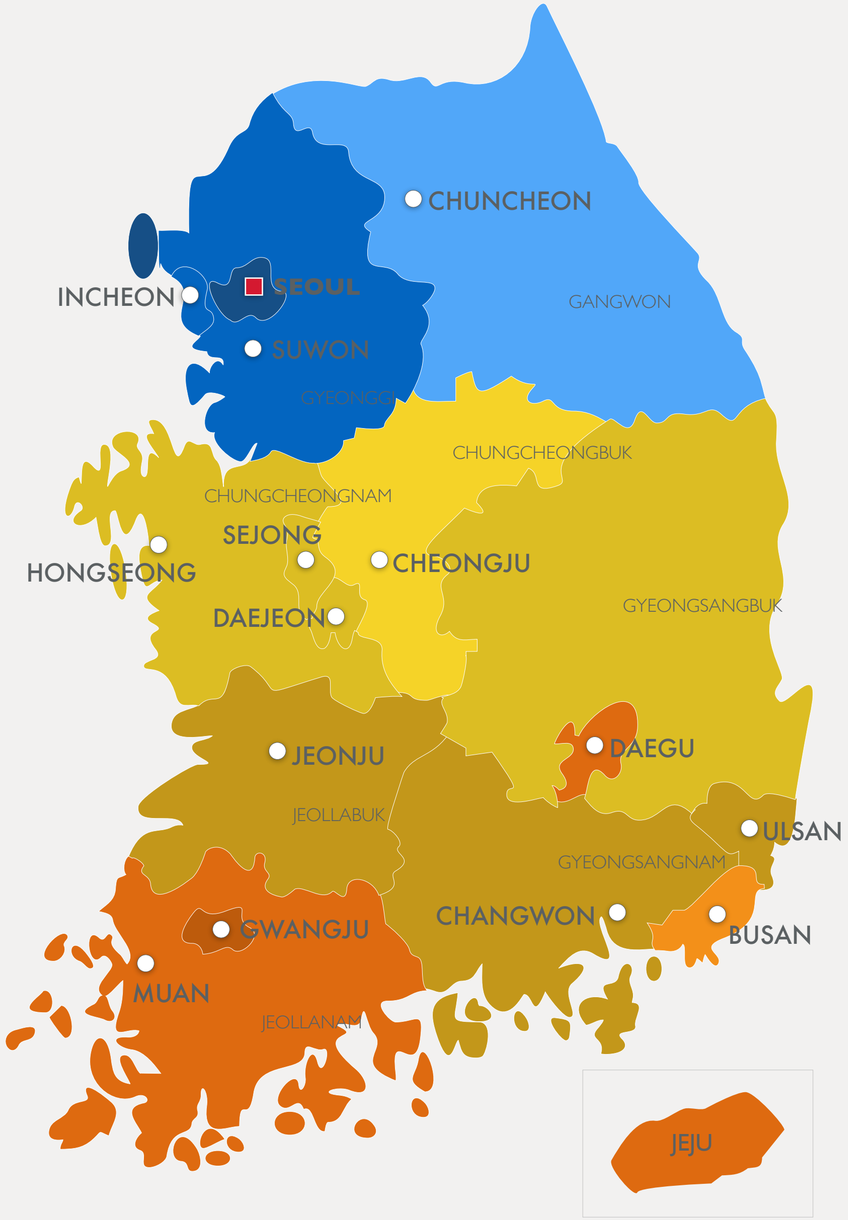
<프로젝트 예시>

Topic

국내 중증 심혈관질환과 계절적 특성간의 연관성 분석

South Korea Map

<프로젝트 예시>



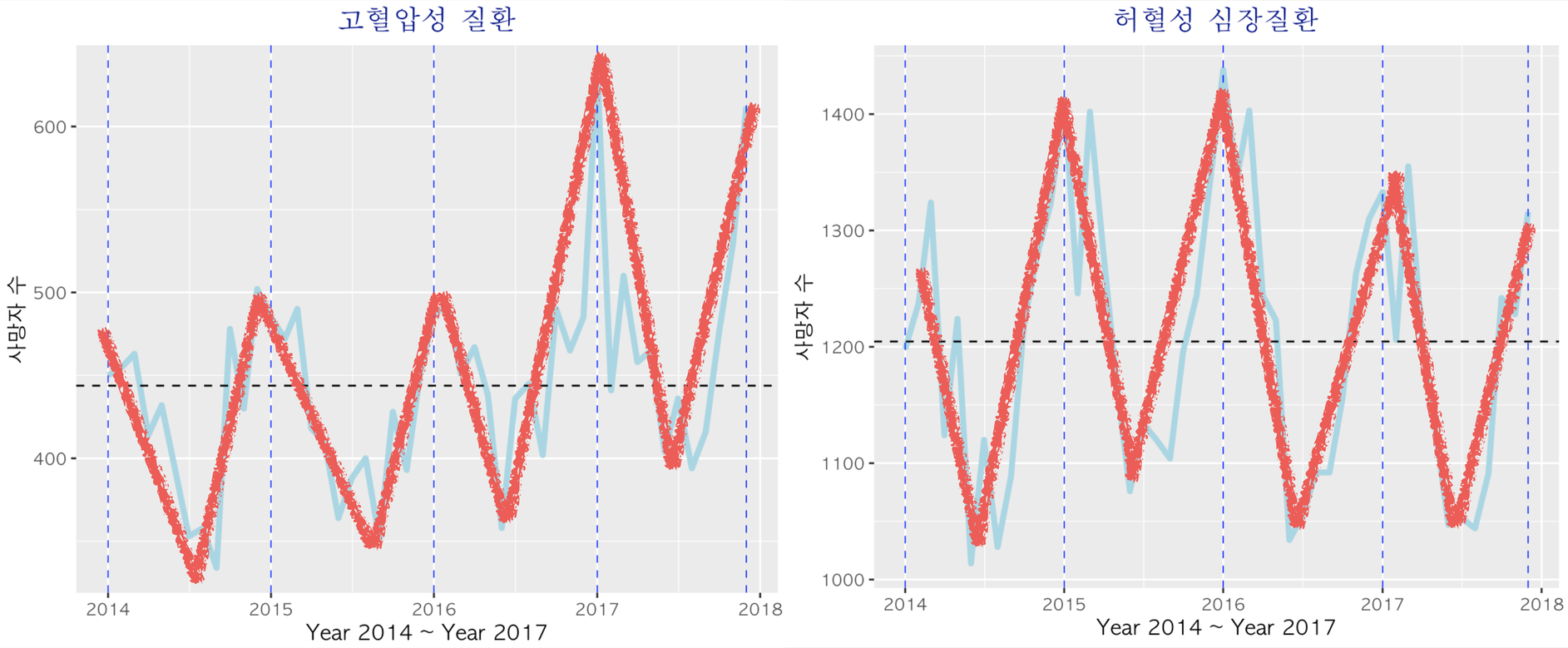
- 연구 필요성

- 한국인 사망원인 (국내 2017 사망원인통계)
  - 2위 : 심장 질환
  - 3위 : 뇌혈관 질환
- 세계 사망원인 (세계보건기구 2015 통계)
  - 1위 : 심혈관 질환

- 심혈관 질환의 특성
  - 특별한 전조증상 없이 급성 발발
  - 골든타임 놓칠 시 사망률

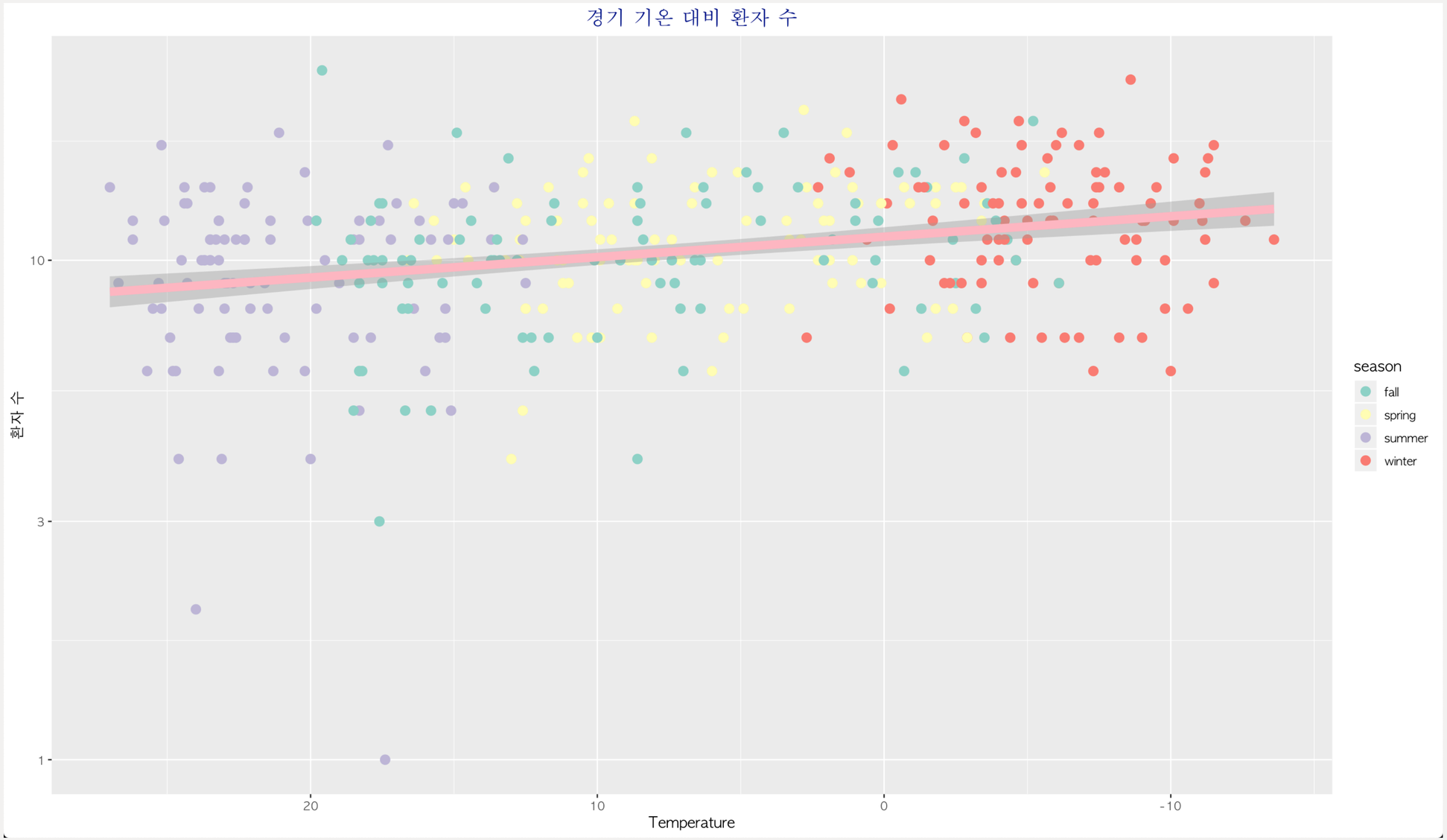
=> 국내 기후 데이터를 사용하여,
기온과 심혈관 질환 발병의 연관성 분석

<프로젝트 예시>



geom_vline을 통해 1년(12개월) 마다 세로줄로 그래프 상 년도 구분

<프로젝트 예시>



Column으로 "seasons"을 추가하여, 계절 구분

<프로젝트 예시>

*유의수준 0.001 =>
*서울, 부산, 경기, 인천 4개 지역
*p-값 < 유의수준
So, 귀무 가설을 기각

=> 신뢰수준 99.9%에서
일별 최저기온과 일별 중증 심혈관질환 사망자수는
상관 관계가 있다.

```
> #경기 summary(일 평균 환자수, 최소 최대 환자수 조회)
> summary(gyeonggi_df$gyeonggi_patient)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    9.00   11.00   10.95   13.00   24.00
> predict(gyeonggi_lm, gyeonggi_pred_temp_x)
        1         2         3         4         5
 8.922656 10.295865 11.669073 13.042282 13.957754
>
```

일별 최저 기온 [30, 15, 0, -15, -25] 도에

대한 일별 사망자 수 예측 모델

<프로젝트 예시>

[심혈관 질환 <-> 기온] 상관관계 분석

=> 일정 수준 이상 유의한 상관관계 관찰

"추운 날씨에 심혈관 질환 발병을 더욱 주의해야 한다"는 사회적 통념

=> 일정 수준 이상으로 유의한 주의 사항

추후, 대용량의 Data Set을 활용한 모델에 대해 Change Point를 감지하여, 지역별 일정 기온 이하로 내려가는 시점에서 유의미한 환자수 변화를 감지

=> 심혈관 질환 방지 캠페인, 심혈관 질환 유의 경보 등 국민 건강 개선에 기여 가능
=> 심혈관 질환 사망자 및 발병자 감소에 기여

CURRICULUM

- 1차시 - Orientation & 활동 안내 [학회장 김재훈]
- 2차시 - Python으로 데이터 다루기, 데이터 시각화하기 [학회 고문 이정규]
- 3차시 - 기초 통계학 1 [부학회장 차주희]
- 4차시 - 기초 통계학 2 [부학회장 차주희]
- 5차시 - 기초 통계학 3 [학회 멘토 서재현]
- 6차시 - 회귀 분석과 통계 응용, 개인 프로젝트 안내 [학회 고문 이정규]
- 7차시 - 머신러닝 기초 개념 1 (기계학습과 k-NN) [학회 멘토 박재현]
- 8차시 - 머신러닝 기초 개념 2 (로지스틱 회귀, 의사결정트리, 신경망과 딥러닝) [학회장 김재훈]
- 9차시 - 인공지능과 데이터 사이언스 [학회장 김재훈]
- 10차시 - 개인 프로젝트 발표

MENTORING TEAM

- Team_1 = [["김재훈", "서재현", "배나영"], ["장예림", "임연수", "정은진",]]
- Team_2 = [["차주희", "이수진"], ["정현수", "이유진", "장시은", "최동연"]]
- Team_3 = [["이정규", "박재현", "한예림"], ["이재성", "이재원", "최은선"]]

hy.days.official@gmail.com

Q&A