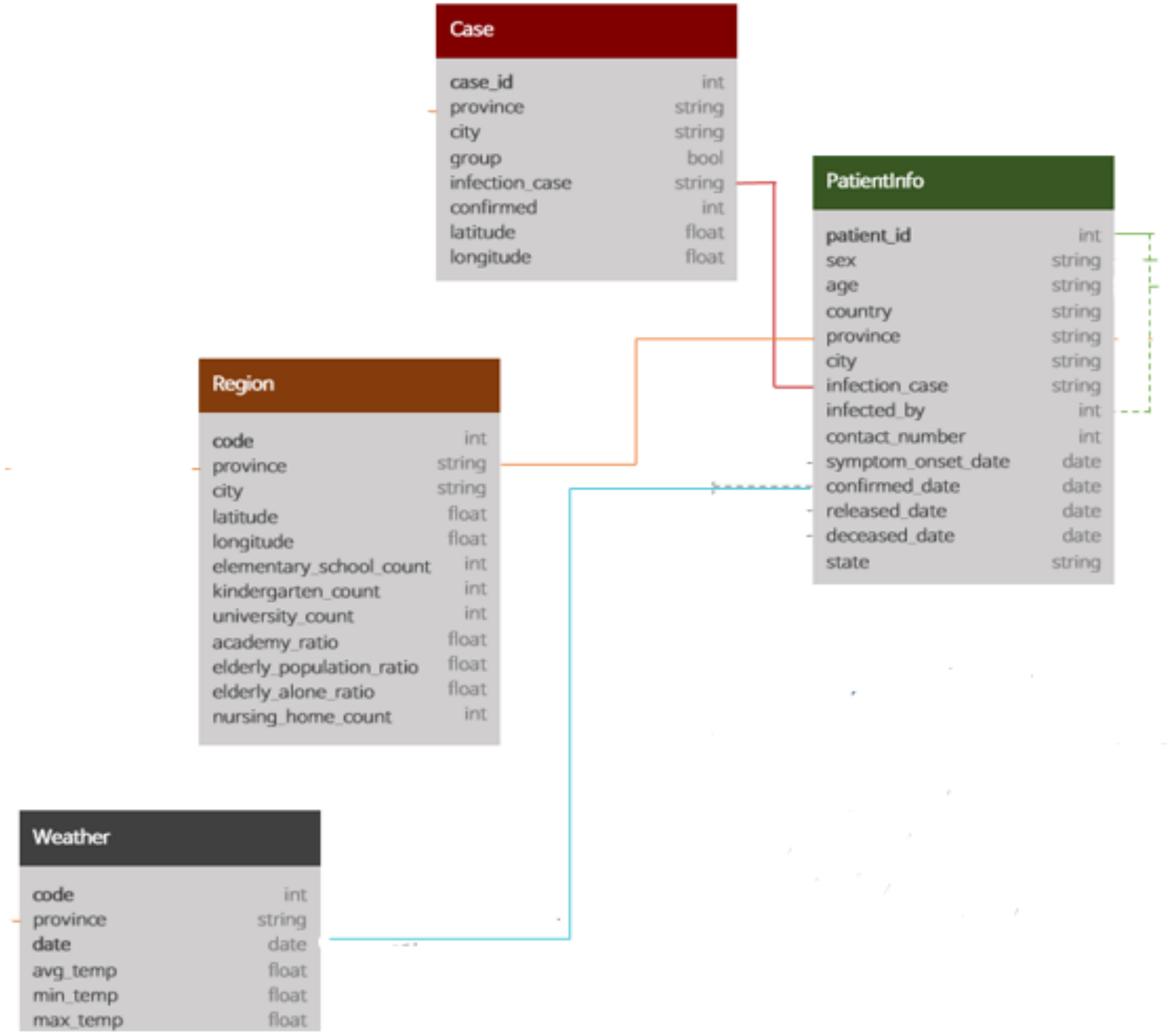
Team Project. K_COVID19

Donghyeon Kim, Donghee Kim, Inuk Jung

College of IT Engineering, School of Computer Science and Engineering

2020-2

Entire Database



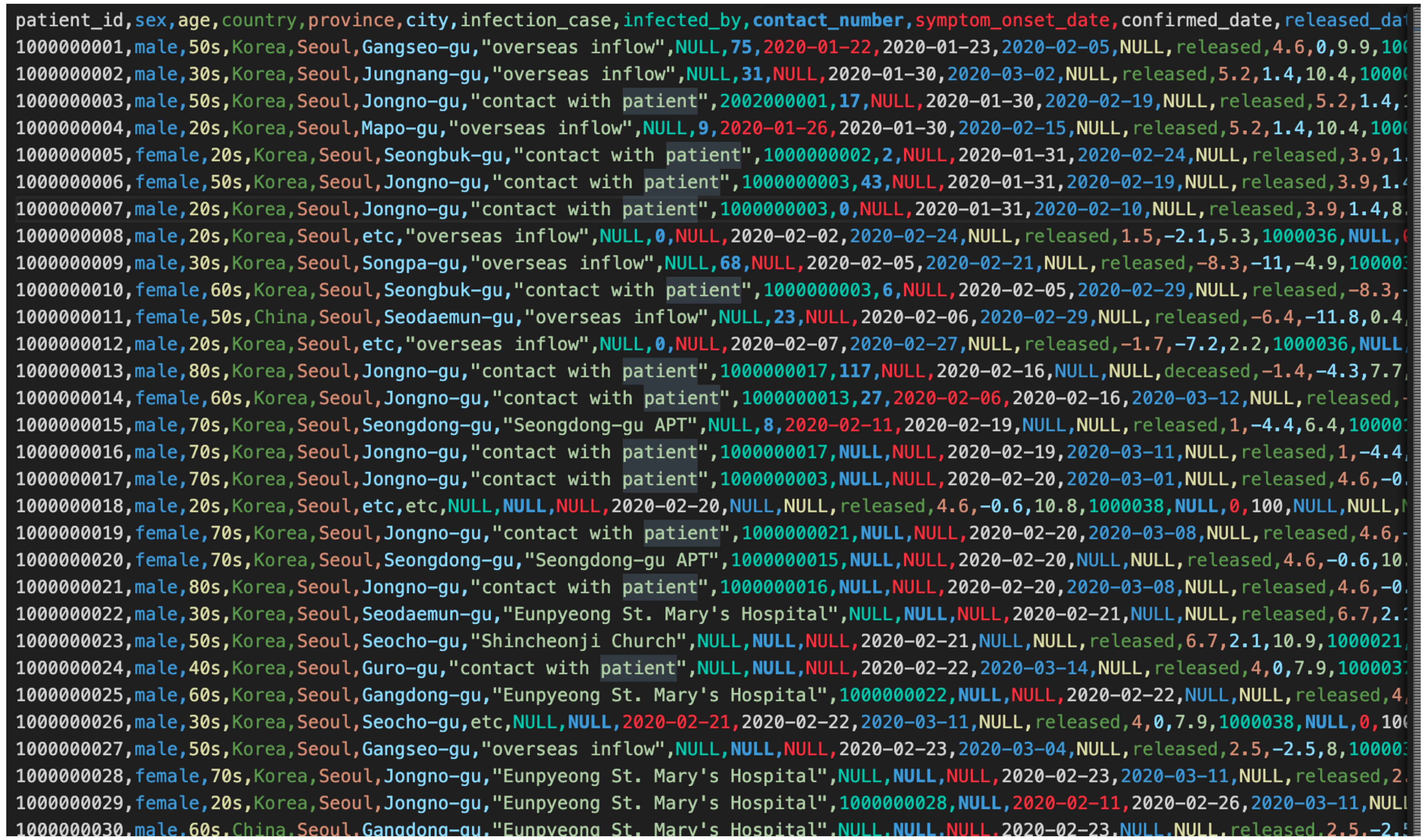
Case	
case_id	int
province	string
city	string
group	bool
infection_case	string
confirmed	int
latitude	float
longitude	float

PatientInfo	
patient_id	int
sex	string
age	string
country	string
province	string
city	string
infection_case	string
infected_by	int
contact_number	int
symptom_onset_date	date
confirmed_date	date
released_date	date
deceased_date	date
state	string

Region	
code	int
province	string
city	string
latitude	float
longitude	float
elementary_school_count	int
kindergarten_count	int
university_count	int
academy_ratio	float
elderly_population_ratio	float
elderly_alone_ratio	float
nursing_home_count	int

Weather	
code	int
province	string
date	date
avg_temp	float
min_temp	float
max_temp	float

제공 되는 데이터 : K_COVID19.csv



```
patient_id,sex,age,country,province,city,infection_case,infected_by,contact_number,symptom_onset_date,confirmed_date,released_dat
1000000001,male,50s,Korea,Seoul,Gangseo-gu,"overseas inflow",NULL,75,2020-01-22,2020-01-23,2020-02-05,NULL,released,4.6,0,9.9,10
1000000002,male,30s,Korea,Seoul,Jungnang-gu,"overseas inflow",NULL,31,NULL,2020-01-30,2020-03-02,NULL,released,5.2,1.4,10.4,1000
1000000003,male,50s,Korea,Seoul,Jongno-gu,"contact with patient",2002000001,17,NULL,2020-01-30,2020-02-19,NULL,released,5.2,1.4,
1000000004,male,20s,Korea,Seoul,Mapo-gu,"overseas inflow",NULL,9,2020-01-26,2020-01-30,2020-02-15,NULL,released,5.2,1.4,10.4,100
1000000005,female,20s,Korea,Seoul,Seongbuk-gu,"contact with patient",1000000002,2,NULL,2020-01-31,2020-02-24,NULL,released,3.9,1.
1000000006,female,50s,Korea,Seoul,Jongno-gu,"contact with patient",1000000003,43,NULL,2020-01-31,2020-02-19,NULL,released,3.9,1.
1000000007,male,20s,Korea,Seoul,Jongno-gu,"contact with patient",1000000003,0,NULL,2020-01-31,2020-02-10,NULL,released,3.9,1.4,8.
1000000008,male,20s,Korea,Seoul,etc,"overseas inflow",NULL,0,NULL,2020-02-02,2020-02-24,NULL,released,1.5,-2.1,5.3,1000036,NULL,
1000000009,male,30s,Korea,Seoul,Songpa-gu,"overseas inflow",NULL,68,NULL,2020-02-05,2020-02-21,NULL,released,-8.3,-11,-4.9,10000
1000000010,female,60s,Korea,Seoul,Seongbuk-gu,"contact with patient",1000000003,6,NULL,2020-02-05,2020-02-29,NULL,released,-8.3,
1000000011,female,50s,China,Seoul,Seodaemun-gu,"overseas inflow",NULL,23,NULL,2020-02-06,2020-02-29,NULL,released,-6.4,-11.8,0.4,
1000000012,male,20s,Korea,Seoul,etc,"overseas inflow",NULL,0,NULL,2020-02-07,2020-02-27,NULL,released,-1.7,-7.2,2.2,1000036,NULL
1000000013,male,80s,Korea,Seoul,Jongno-gu,"contact with patient",1000000017,117,NULL,2020-02-16,NULL,NULL,deceased,-1.4,-4.3,7.7,
1000000014,female,60s,Korea,Seoul,Jongno-gu,"contact with patient",1000000013,27,2020-02-06,2020-02-16,2020-03-12,NULL,released,
1000000015,male,70s,Korea,Seoul,Seongdong-gu,"Seongdong-gu APT",NULL,8,2020-02-11,2020-02-19,NULL,NULL,released,1,-4.4,6.4,10000
1000000016,male,70s,Korea,Seoul,Jongno-gu,"contact with patient",1000000017,NULL,NULL,2020-02-19,2020-03-11,NULL,released,1,-4.4,
1000000017,male,70s,Korea,Seoul,Jongno-gu,"contact with patient",1000000003,NULL,NULL,2020-02-20,2020-03-01,NULL,released,4.6,-0.
1000000018,male,20s,Korea,Seoul,etc,etc,NULL,NULL,NULL,2020-02-20,NULL,NULL,released,4.6,-0.6,10.8,1000038,NULL,0,100,NULL,NULL,
1000000019,female,70s,Korea,Seoul,Jongno-gu,"contact with patient",1000000021,NULL,NULL,2020-02-20,2020-03-08,NULL,released,4.6,
1000000020,female,70s,Korea,Seoul,Seongdong-gu,"Seongdong-gu APT",1000000015,NULL,NULL,2020-02-20,NULL,NULL,released,4.6,-0.6,10.
1000000021,male,80s,Korea,Seoul,Jongno-gu,"contact with patient",1000000016,NULL,NULL,2020-02-20,2020-03-08,NULL,released,4.6,-0.
1000000022,male,30s,Korea,Seoul,Seodaemun-gu,"Eunpyeong St. Mary's Hospital",NULL,NULL,NULL,2020-02-21,NULL,NULL,released,6.7,2.
1000000023,male,50s,Korea,Seoul,Seocho-gu,"Shincheonji Church",NULL,NULL,NULL,2020-02-21,NULL,NULL,released,6.7,2.1,10.9,1000021
1000000024,male,40s,Korea,Seoul,Guro-gu,"contact with patient",NULL,NULL,NULL,2020-02-22,2020-03-14,NULL,released,4,0,7.9,100003
1000000025,male,60s,Korea,Seoul,Gangdong-gu,"Eunpyeong St. Mary's Hospital",1000000022,NULL,NULL,2020-02-22,NULL,NULL,released,4
1000000026,male,30s,Korea,Seoul,Seocho-gu,etc,NULL,NULL,2020-02-21,2020-02-22,2020-03-11,NULL,released,4,0,7.9,1000038,NULL,0,10
1000000027,male,50s,Korea,Seoul,Gangseo-gu,"overseas inflow",NULL,NULL,NULL,2020-02-23,2020-03-04,NULL,released,2.5,-2.5,8,10000
1000000028,female,70s,Korea,Seoul,Jongno-gu,"Eunpyeong St. Mary's Hospital",NULL,NULL,NULL,2020-02-23,2020-03-11,NULL,released,2.
1000000029,female,20s,Korea,Seoul,Jongno-gu,"Eunpyeong St. Mary's Hospital",1000000028,NULL,2020-02-11,2020-02-26,2020-03-11,NUL
1000000030,male,60s,China,Seoul,Gangdong-gu,"Eunpyeong St. Mary's Hospital",NULL,NULL,NULL,2020-02-23,NULL,NULL,released,2.5,-2.
```

주의 !
PatientInfo, Case, Region, Weather
테이블의 row의 갯수는 다릅니다!

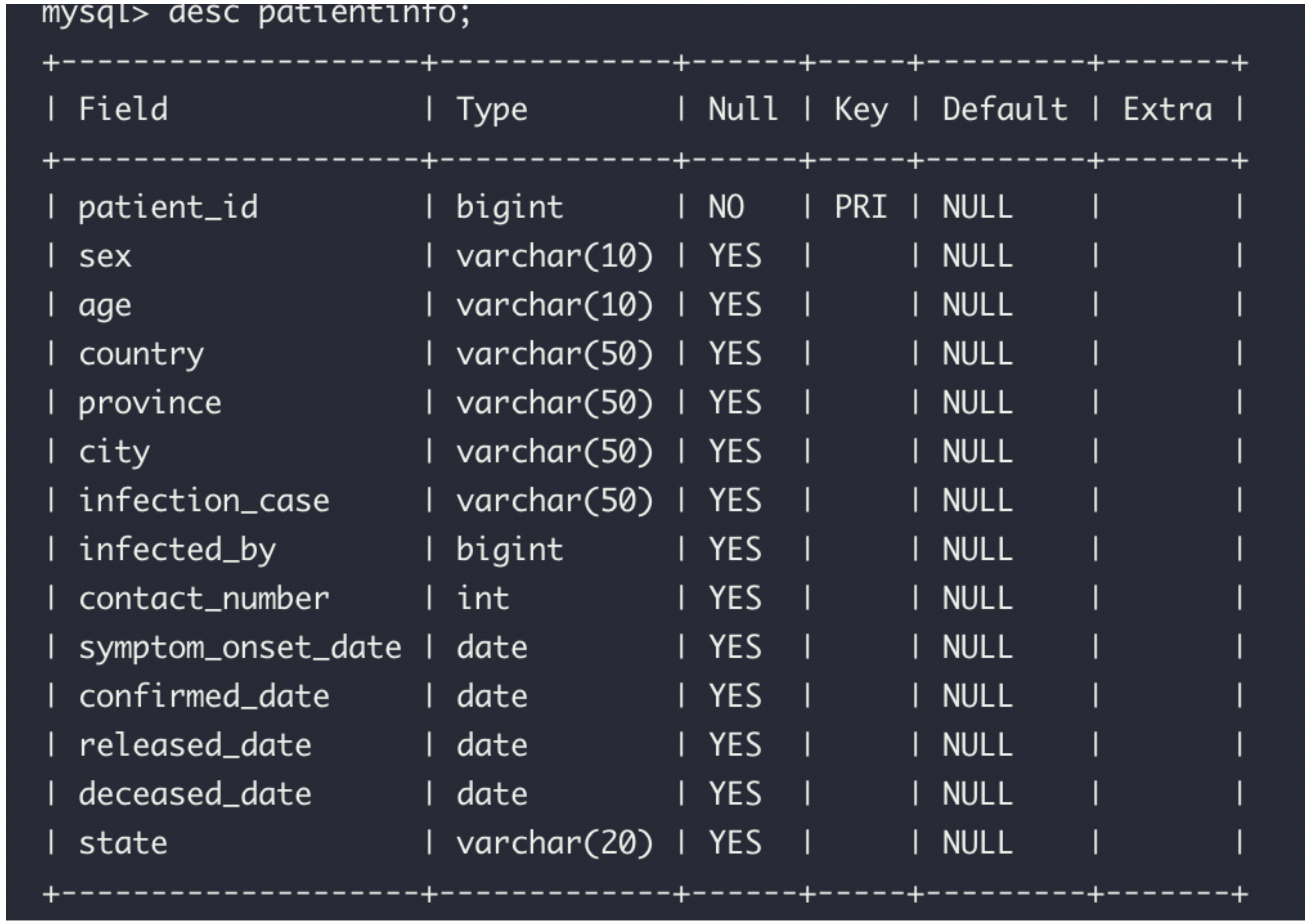
- 4개의 테이블을 만들기 위한 데이터들이 하나의 csv 파일에 들어 있습니다.

- 해당 csv 파일은 총 33개의 column 으로 구성되어 있습니다. 1행에 각 열이 어떤 속성값인지 명시되어 있습니다.

- 각 row는 환자 한명에 대한 row 입니다.

- 각 row로 부터 각 4개의 테이블에 알맞게 파싱을 하신후 insert를 하시면 됩니다.

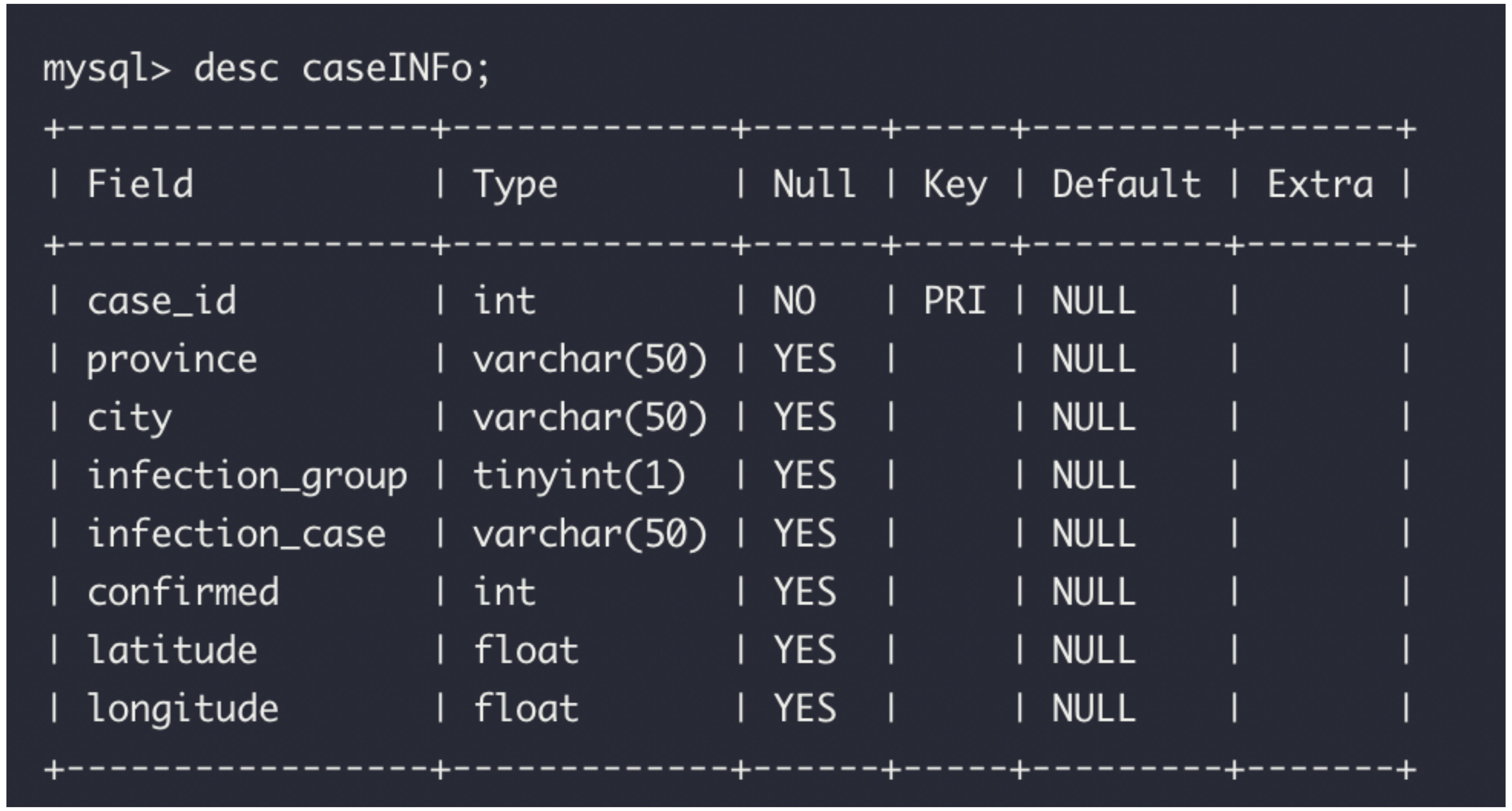
Patientinfo 테이블 : Epidemiological data of COVID-19 patients in South Korea



```
mysql> desc patientinfo;
+-------------------+-------------+------+-----+---------+-------+
| Field             | Type        | Null | Key | Default | Extra |
+-------------------+-------------+------+-----+---------+-------+
| patient_id        | bigint      | NO   | PRI | NULL    |       |
| sex               | varchar(10) | YES  |     | NULL    |       |
| age               | varchar(10) | YES  |     | NULL    |       |
| country           | varchar(50) | YES  |     | NULL    |       |
| province          | varchar(50) | YES  |     | NULL    |       |
| city              | varchar(50) | YES  |     | NULL    |       |
| infection_case    | varchar(50) | YES  |     | NULL    |       |
| infected_by       | bigint      | YES  |     | NULL    |       |
| contact_number    | int         | YES  |     | NULL    |       |
| symptom_onset_date | date       | YES  |     | NULL    |       |
| confirmed_date    | date        | YES  |     | NULL    |       |
| released_date     | date        | YES  |     | NULL    |       |
| deceased_date     | date        | YES  |     | NULL    |       |
| state             | varchar(20) | YES  |     | NULL    |       |
+-------------------+-------------+------+-----+---------+-------+
```

- Patient_id : region_code(5) + patient_number(5)
- Provice : 서울, 부산 같은 특별시 및 광역시 또는 경기도 강원도 와 같은 도
- City :
  1) province가 서울 부산 같은 특별시, 광역시인 경우 City는
  강남구, 서초구, 해운대구
  2) province가 경상북도 경기도 같은 경우에는 City가
  구미시, 안동시
- Infection_case : 감염 원인
  ex) overseas inflow, contact with patient, Eunpyeong St. Mary's Hospital
- Infected_by : the ID of who infected this patient
  cf) this column refers to the 'patient_id' column.
- Contact_number : 접촉한 사람들 수
- Symptom_onset_date : 증상발생 날짜
- Confirmed_date : 확진(양성 판정) 일
- Released_date : 완치(퇴원)날짜
- Deceased_date : 사망일
- State : isolated / released / deceased

Case 테이블 : Data of COVID-19 infection cases in South Korea



```
mysql> desc caseINFo;
+-----------------+-------------+------+-----+---------+-------+
| Field           | Type        | Null | Key | Default | Extra |
+-----------------+-------------+------+-----+---------+-------+
| case_id         | int         | NO   | PRI | NULL    |       |
| province        | varchar(50) | YES  |     | NULL    |       |
| city            | varchar(50) | YES  |     | NULL    |       |
| infection_group | tinyint(1)  | YES  |     | NULL    |       |
| infection_case  | varchar(50) | YES  |     | NULL    |       |
| confirmed       | int         | YES  |     | NULL    |       |
| latitude        | float       | YES  |     | NULL    |       |
| longitude       | float       | YES  |     | NULL    |       |
+-----------------+-------------+------+-----+---------+-------+
```

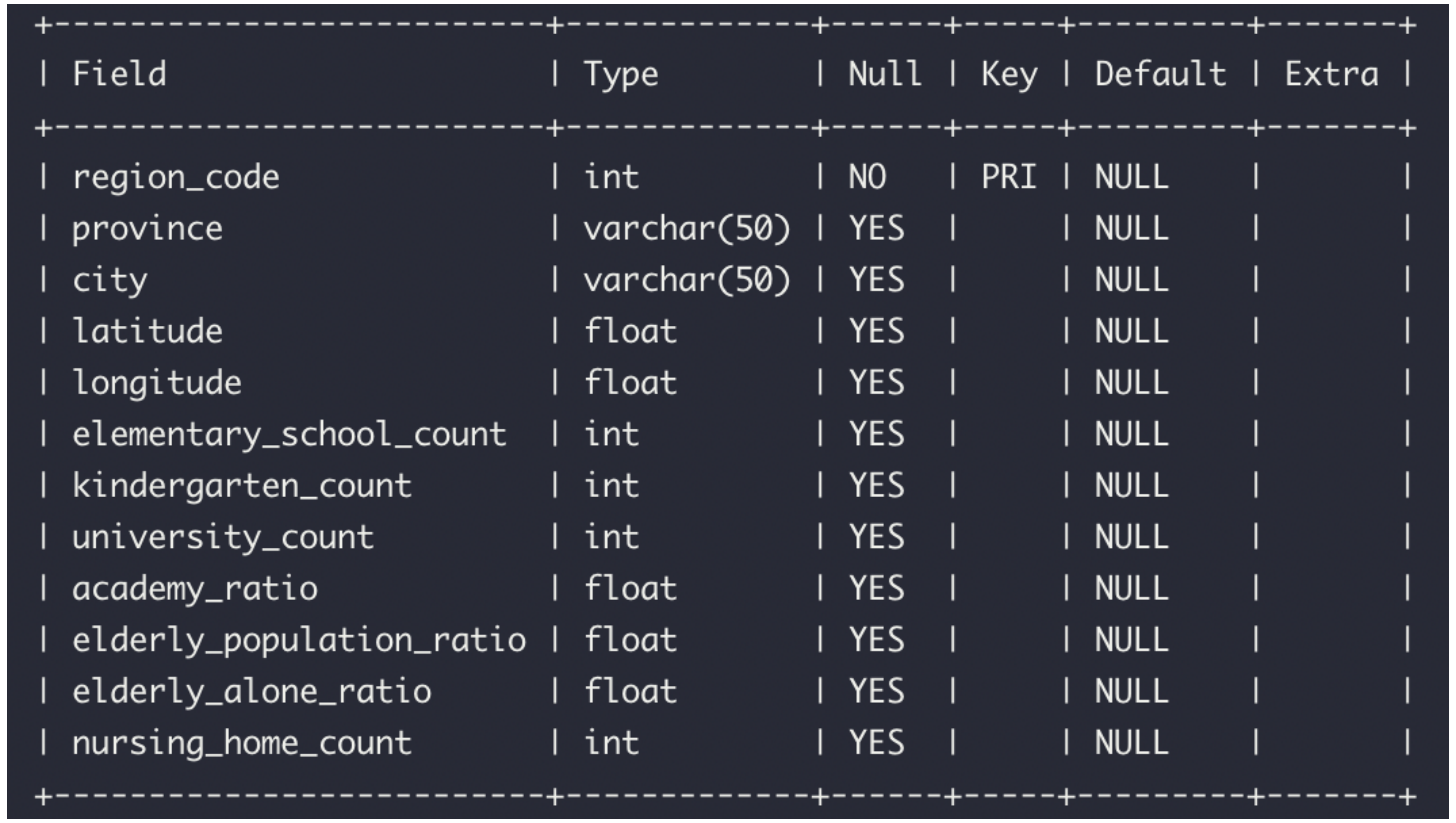
- Case_id : The ID of the infection case
  case_id(7) = region_code(5)+case_number(2)

- Infection_group : 집단감염 여부
  TRUE = Group infection
  FALSE =not group

- infection_case : the infection case (the name of group or other cases)
  ex) Itaewon Clubs, Guro-gu Call Center

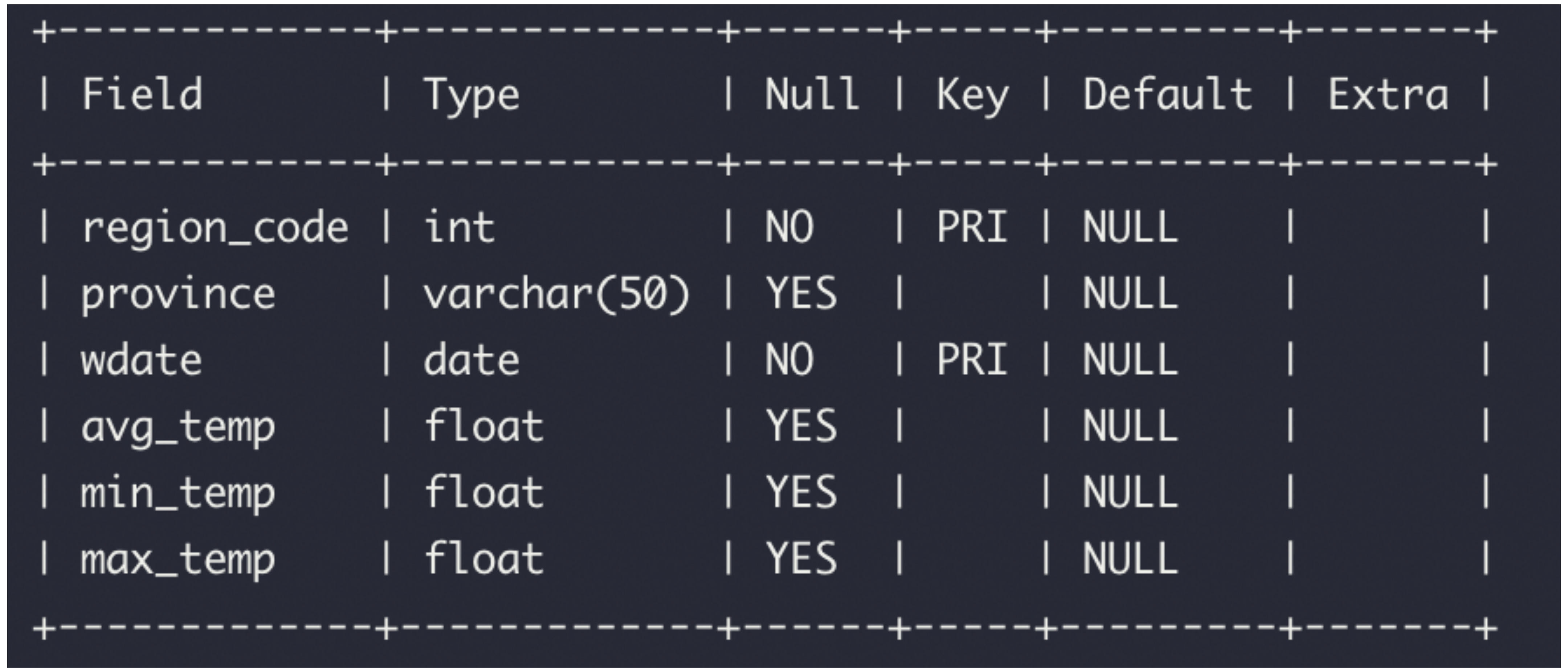
- Confirmed : 확진자 수

Region 테이블 : Location and statistical data of the regions in South Korea



```
+--------------------------+-------------+------+-----+---------+-------+
| Field                    | Type        | Null | Key | Default | Extra |
+--------------------------+-------------+------+-----+---------+-------+
| region_code              | int         | NO   | PRI | NULL    |       |
| province                 | varchar(50) | YES  |     | NULL    |       |
| city                     | varchar(50) | YES  |     | NULL    |       |
| latitude                 | float       | YES  |     | NULL    |       |
| longitude                | float       | YES  |     | NULL    |       |
| elementary_school_count  | int         | YES  |     | NULL    |       |
| kindergarten_count       | int         | YES  |     | NULL    |       |
| university_count         | int         | YES  |     | NULL    |       |
| academy_ratio            | float       | YES  |     | NULL    |       |
| elderly_population_ratio | float       | YES  |     | NULL    |       |
| elderly_alone_ratio      | float       | YES  |     | NULL    |       |
| nursing_home_count       | int         | YES  |     | NULL    |       |
+--------------------------+-------------+------+-----+---------+-------+
```

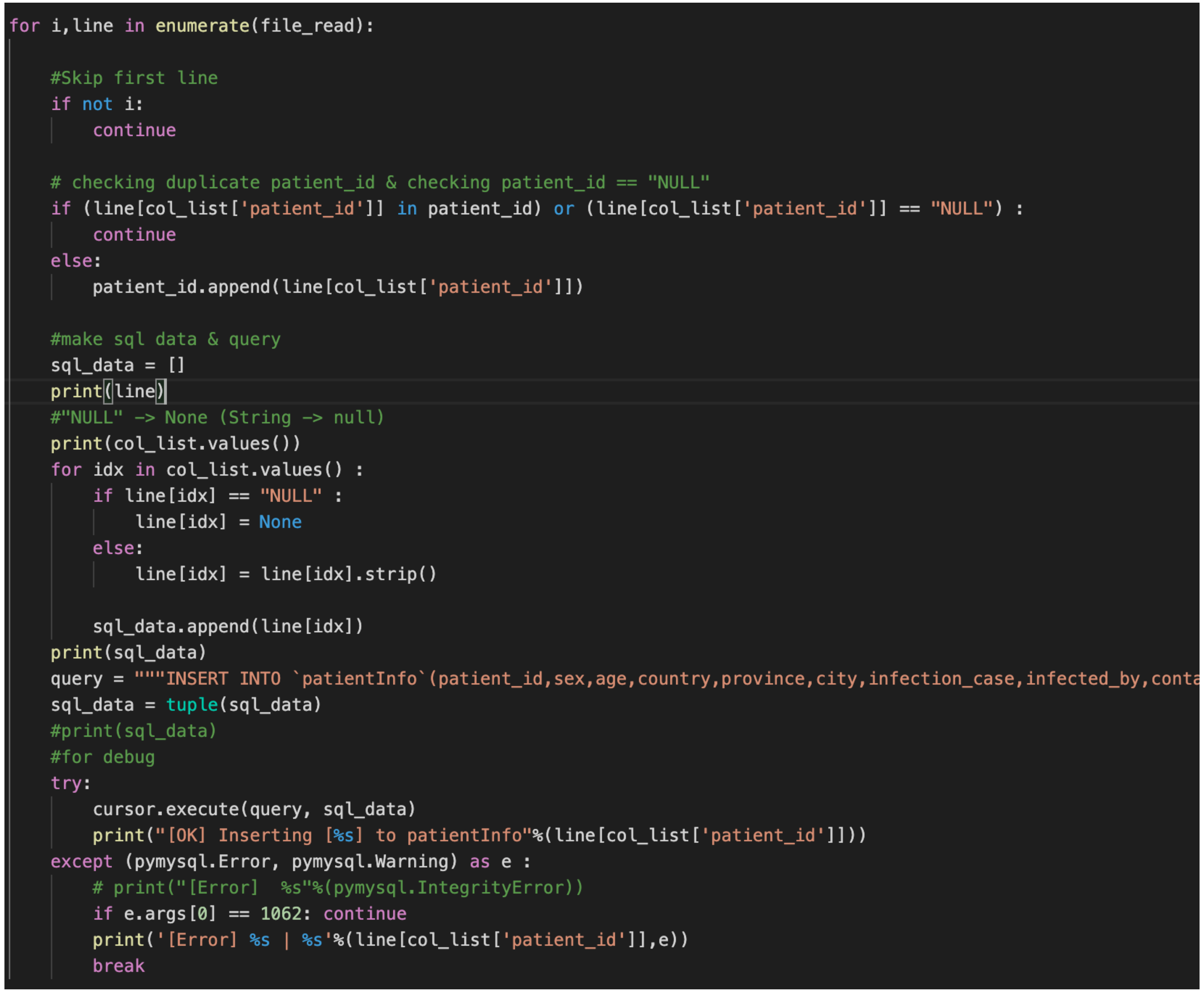
Weather 테이블 : Data of the weather in the regions of South Korea



```
+-------------+-------------+------+-----+---------+-------+
| Field       | Type        | Null | Key | Default | Extra |
+-------------+-------------+------+-----+---------+-------+
| region_code | int         | NO   | PRI | NULL    |       |
| province    | varchar(50) | YES  |     | NULL    |       |
| wdate       | date        | NO   | PRI | NULL    |       |
| avg_temp    | float       | YES  |     | NULL    |       |
| min_temp    | float       | YES  |     | NULL    |       |
| max_temp    | float       | YES  |     | NULL    |       |
+-------------+-------------+------+-----+---------+-------+
```

- Region_code: the code of the region
- Wdate = Date

CSV파일 → 데이터 베이스 예시 : parsing_patient.py (lms 업로드)



```
for i,line in enumerate(file_read):

    #Skip first line
    if not i:
        continue

    # checking duplicate patient_id & checking patient_id == "NULL"
    if (line[col_list['patient_id']] in patient_id) or (line[col_list['patient_id']] == "NULL") :
        continue
    else:
        patient_id.append(line[col_list['patient_id']])

    #make sql data & query
    sql_data = []
    print(line)
    #"NULL" -> None (String -> null)
    print(col_list.values())
    for idx in col_list.values() :
        if line[idx] == "NULL" :
            line[idx] = None
        else:
            line[idx] = line[idx].strip()

        sql_data.append(line[idx])
    print(sql_data)
    query = """INSERT INTO `patientInfo`(patient_id,sex,age,country,province,city,infection_case,infected_by,conta
    sql_data = tuple(sql_data)
    #print(sql_data)
    #for debug
    try:
        cursor.execute(query, sql_data)
        print("[OK] Inserting [%s] to patientInfo"%(line[col_list['patient_id']]))
    except (pymysql.Error, pymysql.Warning) as e :
        # print("[Error]  %s"%(pymysql.IntegrityError))
        if e.args[0] == 1062: continue
        print('[Error] %s | %s'%(line[col_list['patient_id']],e))
        break
```

K_COVID19.csv파일에서 테이블당
유효한 attribute만 뽑아내어 insert!

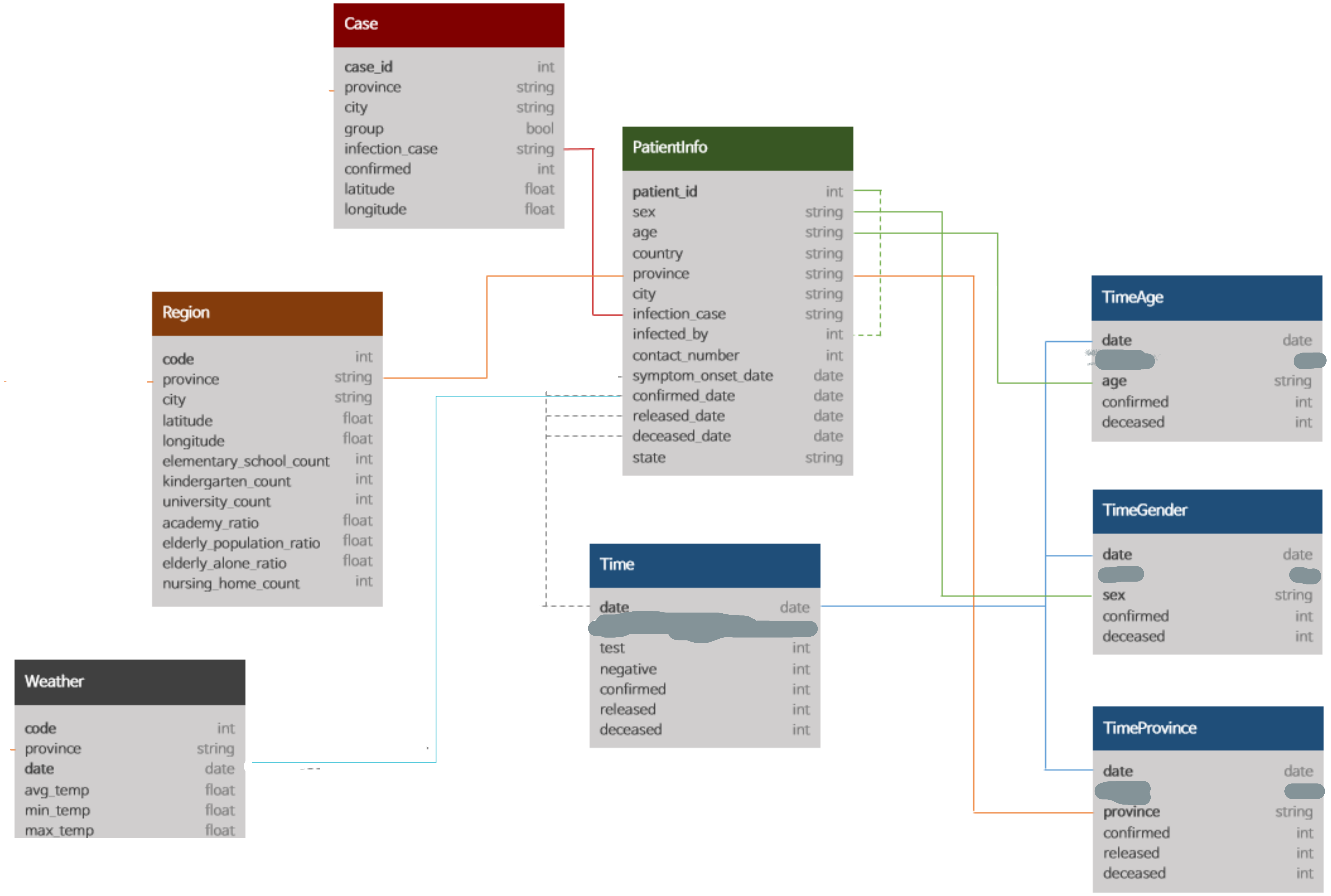
Pasring_case.py
Parsing_region.py
Parsing_weather.py를
팀원당 하나씩 만들어서 제출

힌트)
case 테이블 약 : 120 여개의 row
region 테이블 약 : 170 여개의 row
weather 테이블 약 : 2500여개의 row

Next Week

PatientInfo 테이블을 활용하여
새로운 Time테이블을 만들고(코드 제공)

Time테이블로부터
TimeAge, TimeGender, TimeProvice
테이블을 각 팀원당 한 개씩 생성



Case	
case_id	int
province	string
city	string
group	bool
infection_case	string
confirmed	int
latitude	float
longitude	float

PatientInfo	
patient_id	int
sex	string
age	string
country	string
province	string
city	string
infection_case	string
infected_by	int
contact_number	int
symptom_onset_date	date
confirmed_date	date
released_date	date
deceased_date	date
state	string

Region	
code	int
province	string
city	string
latitude	float
longitude	float
elementary_school_count	int
kindergarten_count	int
university_count	int
academy_ratio	float
elderly_population_ratio	float
elderly_alone_ratio	float
nursing_home_count	int

Weather	
code	int
province	string
date	date
avg_temp	float
min_temp	float
max_temp	float

Time	
date	date
[redacted]	
test	int
negative	int
confirmed	int
released	int
deceased	int

TimeAge	
date	date
[redacted]	[redacted]
age	string
confirmed	int
deceased	int

TimeGender	
date	date
[redacted]	[redacted]
sex	string
confirmed	int
deceased	int

TimeProvince	
date	date
[redacted]	[redacted]
province	string
confirmed	int
released	int
deceased	int

DO YOUR BEST!