1) Case Data

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

2) Patient Data

* **PatientInfo**: Epidemiological data of COVID-19 patients in South Korea
* **PatientRoute**: Route data of COVID-19 patients in South Korea

3) Time Series Data

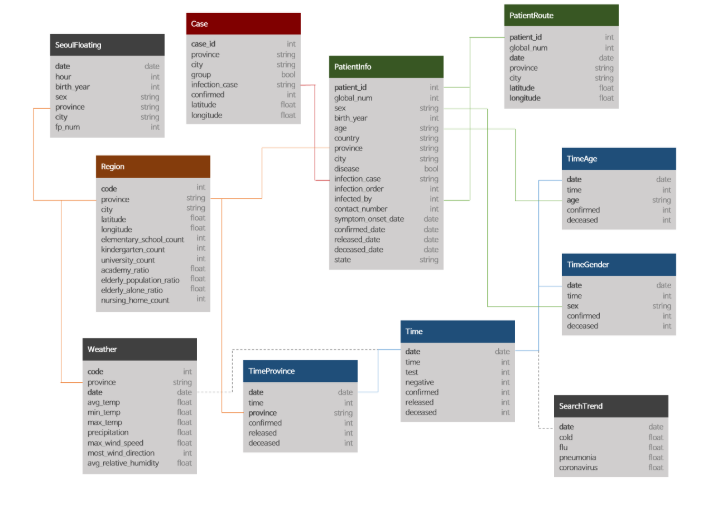
* **Time**: Time series data of COVID-19 status in South Korea
* **TimeAge**: Time series data of COVID-19 status in terms of the age in South Korea
* **TimeGender**: Time series data of COVID-19 status in terms of gender in South Korea
* **TimeProvince**: Time series data of COVID-19 status in terms of the Province in South Korea

4) Additional Data

* **Region**: Location and statistical data of the regions in South Korea
* **Weather**: Data of the weather in the regions of South Korea
* **SearchTrend**: Trend data of the keywords searched in NAVER which is one of the largest portals in South Korea
* **SeoulFloating**: Data of floating population in Seoul, South Korea (from SK Telecom Big Data Hub)
* **Policy**: Data of the government policy for COVID-19 in South Korea

# **2. The Structure of our Dataset**

* What color means is that they have similar properties.
* If a line is connected between columns, it means that the values of the columns are partially shared.
* The dotted lines mean weak relevance.



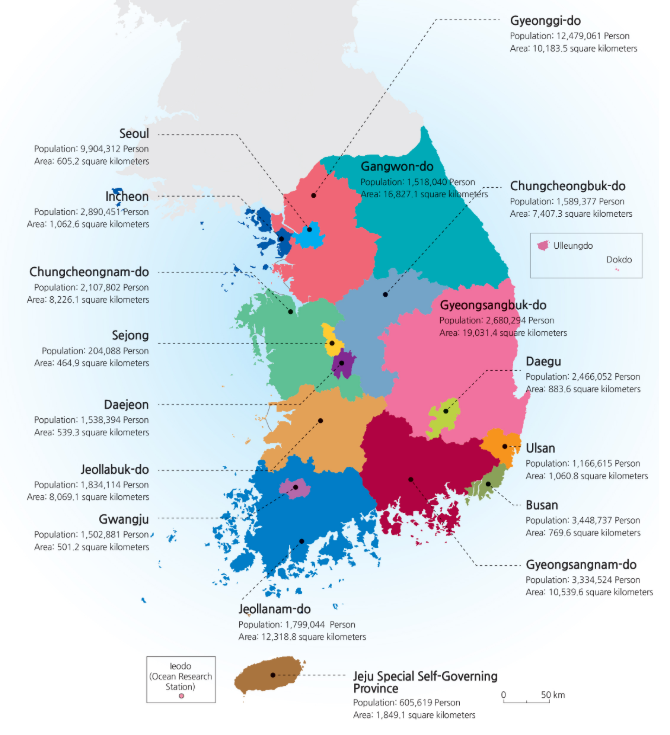
### Levels of administrative divisions in South Korea

#### Upper Level (Provincial-level divisions)

* **Special City**: Seoul
* **Metropolitan City**: Busan / Daegu / Daejeon / Gwangju / Incheon / Ulsan
* **Province(-do)**: Gyeonggi-do / Gangwon-do / Chungcheongbuk-do / Chungcheongnam-do / Jeollabuk-do / Jeollanam-do / Gyeongsangbuk-do / Gyeongsangnam-do

#### Lower Level (Municipal-level divisions)

* **City(-si)** [List of cities in South Korea](https://en.wikipedia.org/wiki/List_of_cities_in_South_Korea)
* **Country(-gun)** [List of counties of South Korea](https://en.wikipedia.org/wiki/List_of_counties_of_South_Korea)
* **District(-gu)** [List of districts in South Korea](https://en.wikipedia.org/wiki/List_of_districts_in_South_Korea)



### 1) Case

#### Data of COVID-19 infection cases in South Korea

1. case\_id: the ID of the infection case
   * case\_id(7) = region\_code(5) + case\_number(2)
   * You can check the region\_code in 'Region.csv'

* province: Special City / Metropolitan City / Province(-do)
* city: City(-si) / Country (-gun) / District (-gu)
  + The value 'from other city' means that where the group infection started is other city.
* group: TRUE: group infection / FALSE: not group
  + If the value is 'TRUE' in this column, the value of 'infection\_cases' means the name of group.
  + The values named 'contact with patient', 'overseas inflow' and 'etc' are not group infection.
* infection\_case: the infection case (the name of group or other cases)
  + The value 'overseas inflow' means that the infection is from other country.
  + The value 'etc' includes individual cases, cases where relevance classification is ongoing after investigation, and cases under investigation.
* confirmed: the accumulated number of the confirmed
* latitude: the latitude of the group (WGS84)
* longitude: the longitude of the group (WGS84)

### 3) PatientRoute

#### Route data of COVID-19 patients in South Korea

* patient\_id: the ID of the patient
* global\_num: the number given by KCDC
* date: YYYY-MM-DD
* province: Special City / Metropolitan City / Province(-do)
* city: City(-si) / Country (-gun) / District (-gu)
* latitude: the latitude of the visit (WGS84)
* longitude: the longitude of the visit (WGS84)

### 4) Time

#### Time series data of COVID-19 status in South Korea

* date: YYYY-MM-DD
* time: Time (0 = AM 12:00 / 16 = PM 04:00)
  + The time for KCDC to open the information has been changed from PM 04:00 to AM 12:00 since March 2nd.
* test: the accumulated number of tests
  + A test is a diagnosis of an infection.
* negative: the accumulated number of negative results
* confirmed: the accumulated number of positive results
* released: the accumulated number of releases
* deceased: the accumulated number of deceases

### 7) TimeProvince

#### Time series data of COVID-19 status in terms of the Province in South Korea

* date: YYYY-MM-DD
* time: Time
* province: the province of South Korea
* confirmed: the accumulated number of the confirmed in the province
  + The confirmed status in terms of the provinces has been presented since Feburary 21th.
  + The value before Feburary 21th can be different.
* released: the accumulated number of the released in the province
  + The confirmed status in terms of the provinces has been presented since March 5th.
  + The value before March 5th can be different.
* deceased: the accumulated number of the deceased in the province
  + The confirmed status in terms of the provinces has been presented since March 5th.
  + The value before March 5th can be different.

### 8) Region

#### Location and statistical data of the regions in South Korea

* code: the code of the region
* province: Special City / Metropolitan City / Province(-do)
* city: City(-si) / Country (-gun) / District (-gu)
* latitude: the latitude of the visit (WGS84)
* longitude: the longitude of the visit (WGS84)
* elementary\_school\_count: the number of elementary schools
* kindergarten\_count: the number of kindergartens
* university\_count: the number of universities
* academy\_ratio: the ratio of academies
* elderly\_population\_ratio: the ratio of the elderly population
* elderly\_alone\_ratio: the ratio of elderly households living alone
* nursing\_home\_count: the number of nursing homes

Source of the statistic: [KOSTAT (Statistics Korea)](http://kosis.kr/)

### 9) Weather

#### Data of the weather in the regions of South Korea

* code: the code of the region
* province: Special City / Metropolitan City / Province(-do)
* date: YYYY-MM-DD
* avg\_temp: the average temperature
* min\_temp: the lowest temperature
* max\_temp: the highest temperature
* precipitation: the daily precipitation
* max\_wind\_speed: the maximum wind speed
* most\_wind\_direction: the most frequent wind direction
* avg\_relative\_humidity: the average relative humidity

Source of the weather data: [KMA (Korea Meteorological Administration)](http://data.kma.go.kr/)

### 10) SearchTrend

#### Trend data of the keywords searched in NAVER which is one of the largest portal in South Korea

* date: YYYY-MM-DD
* cold: the search volume of 'cold' in Korean language
  + The unit means relative value by setting the highest search volume in the period to 100.
* flu: the search volume of 'flu' in Korean language
  + Same as above.
* pneumonia: the search volume of 'pneumonia' in Korean language
  + Same as above.
* coronavirus: the search volume of 'coronavirus' in Korean language
  + Same as above.

Source of the data: [NAVER DataLab](https://datalab.naver.com/)

### 12) Policy

#### Data of the government policy for COVID-19 in South Korea

* policy\_id: the ID of the policy
* country: the country that implemented the policy
* type: the type of the policy
* gov\_policy: the policy of the government
* detail: the detail of the policy
* start\_date: the start date of the policy
* end\_date: the end date of the policy

### 2) PatientInfo

#### Epidemiological data of COVID-19 patients in South Korea

1. patient\_id: the ID of the patient
   * patient\_id(10) = region\_code(5) + patient\_number(5)
   * You can check the region\_code in 'Region.csv'
   * There are two types of the patient\_number  
     1) local\_num: The number given by the local government.  
     2) global\_num: The number given by the KCDC

* global\_num: the number given by KCDC
  + There are some patients having no global\_num.
  + The paitents in Busan doesn't have the global\_num.
* sex: the sex of the patient
* birth\_year: the birth year of the patient
* age: the age of the patient
  + 0s: 0 ~ 9
  + 10s: 10 ~ 19  
    ...
  + 90s: 90 ~ 99
  + 100s: 100 ~ 109
* country: the country of the patient
* province: the province of the patient
* city: the city of the patient
* disease: TRUE: underlying disease / FALSE: no disease
* infection\_case: the case of infection
* infection\_order: the order of infection
* infected\_by: the ID of who infected the patient
  + This column refers to the 'patient\_id' column.
* contact\_number: the number of contacts with people
* symptom\_onset\_date: the date of symptom onset
* confirmed\_date: the date of being confirmed
* released\_date: the date of being released
* deceased\_date: the date of being deceased
* state: isolated / released / deceased
  + isolated: being isolated in the hospital
  + released: being released from the hospital
  + deceased: being deceased