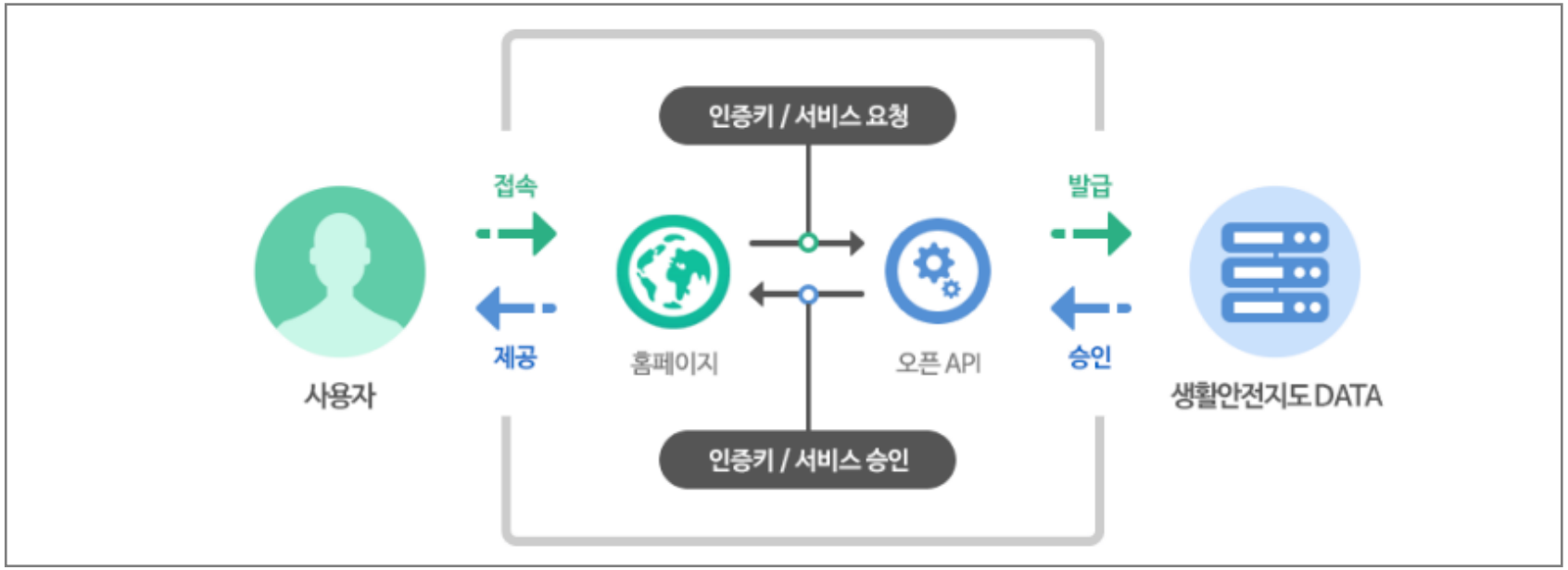
데이터 수집

데이터 수집

데이터 수집 방법

다운로드	공개 데이터베이스 사이트에서 관련 파일 다운로드
오픈 API 서비스	정보 제공자가 공개한 API를 호출해 데이터 다운로드
크롤링	웹페이지 내부에 기록되어 있는 정보를 선별적으로 수집



데이터 수집

데이터 형식

대여소 그룹	대여소 명	대여 일자 / 월	대여 건수
그룹명 없음	대여소명 없음	2017-01-01	0
광진구	500. 어린이대공원역 3번출구 앞	2017-01-01	20
광진구	501. 광진구의회 앞	2017-01-01	10
광진구	502. 뚝섬유원지역 1번출구 앞	2017-01-01	9
광진구	503. 더샵스타시티 C동 앞	2017-01-01	9
광진구	504. 신자초교입구교차로	2017-01-01	6
광진구	505. 자양사거리 광진아크로텔 앞	2017-01-01	11
광진구	515. 광양중학교 앞	2017-01-01	0
광진구	516. 광진메디칼 앞	2017-01-01	3
광진구	539. 군자교교차로	2017-01-01	5
광진구	540. 군자역 7번출구 베스트샵 앞	2017-01-01	11
광진구	542. 강변역 4번출구 뒤	2017-01-01	1
광진구	543. 구의공원(테크노마트 앞)	2017-01-01	6
광진구	544. 광남중학교	2017-01-01	2
광진구	546. 잠실대교북단 교차로	2017-01-01	3
광진구	548. 자양나들목	2017-01-01	11
광진구	549. 아차산역 3번출구	2017-01-01	5
광진구	551. 구의삼성쉐르빌 앞	2017-01-01	1

데이터 수집

데이터 형식

01 데이터 형태에 따른 분류

- 개별 데이터 : 조사 대상으로부터 직접 얻은 정보를 수치화한 자료
- 집계 데이터 : 원자료에 가공을 한 데이터(통계값)
- 데이터 조사 = 데이터 전처리

데이터 수집

데이터 형식

02 데이터 성질에 따른 분류

- 양적 자료 : 수치(이산/연속), 크기
  - 등간 척도 : 수치 사이의 간격에 의미가 있으며 원점이 없음

    예 온도
  - 비율 척도 : 수치 그 자체에 의미가 있으며 원점을 가짐

    예 신장, 체중, 연령
- 질적 자료 : 범주(명목), 순서를 가짐(순서)

데이터 수집

데이터 형식

03 분석 기준에 따른 분류

- 정적 자료 : 데이터 그 자체의 의미를 분석해도 되는 경우

  예 녹지 크기 및 유형과 미세먼지 수치와의 관련성 분석

- 시계열 자료 : 시간 흐름에 따른 값의 변화를 분석해야 하는 경우

  예 월별 미세먼지 수치 변화와 영향을 미치는 요인 분석

데이터 수집

데이터 형식

- 데이터 분석을 위한 데이터 형식

4개의 열 = 4개의 특징(속성)

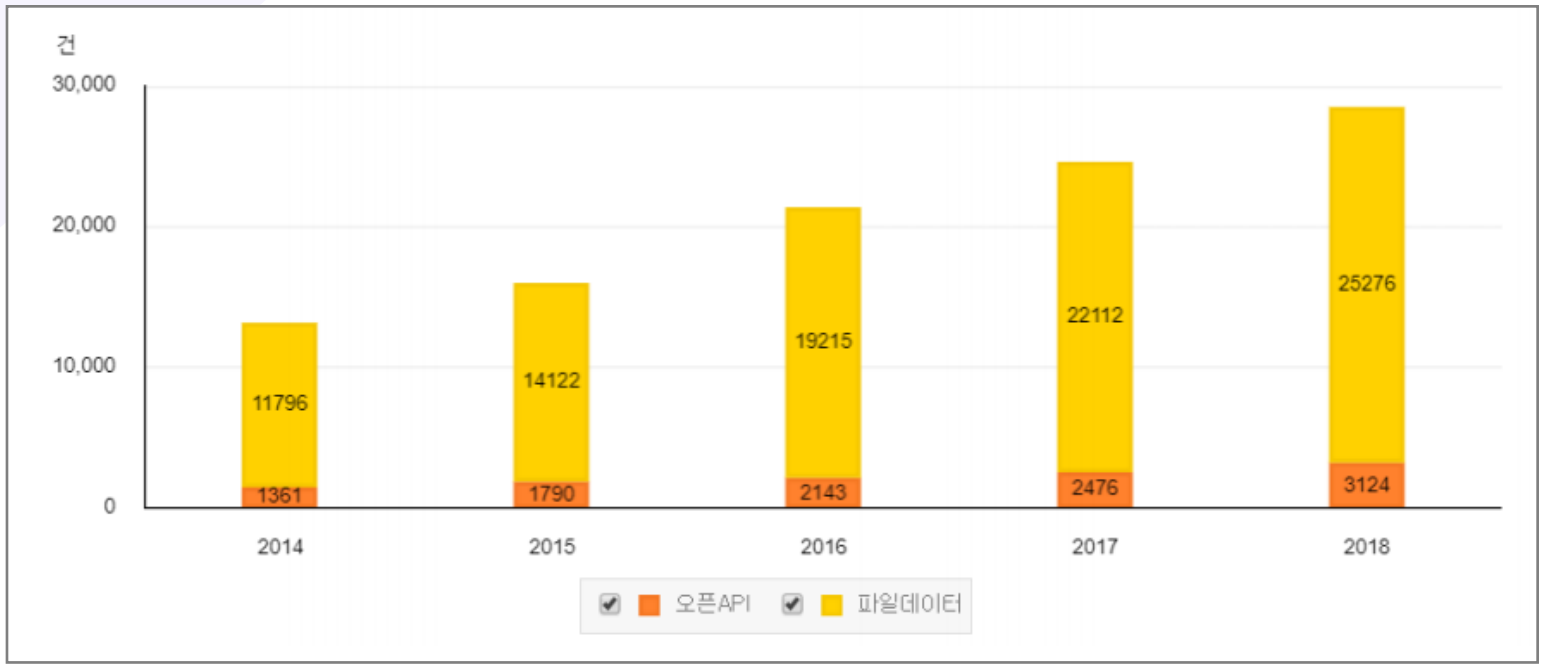
n개의 행
=
n개의
데이터

특징A	특징B	특징C	특징D
25.03	0.04	A	1030
16.67	0.23	B	5092
43.03	0.18	B	4096
...	...	...	...
10.2	0.08	A	6223
33.51	0.16	C	1028
12.23	0.32	B	6839

데이터 수집

데이터 형식

- 공공데이터 제공 현황



데이터 수집

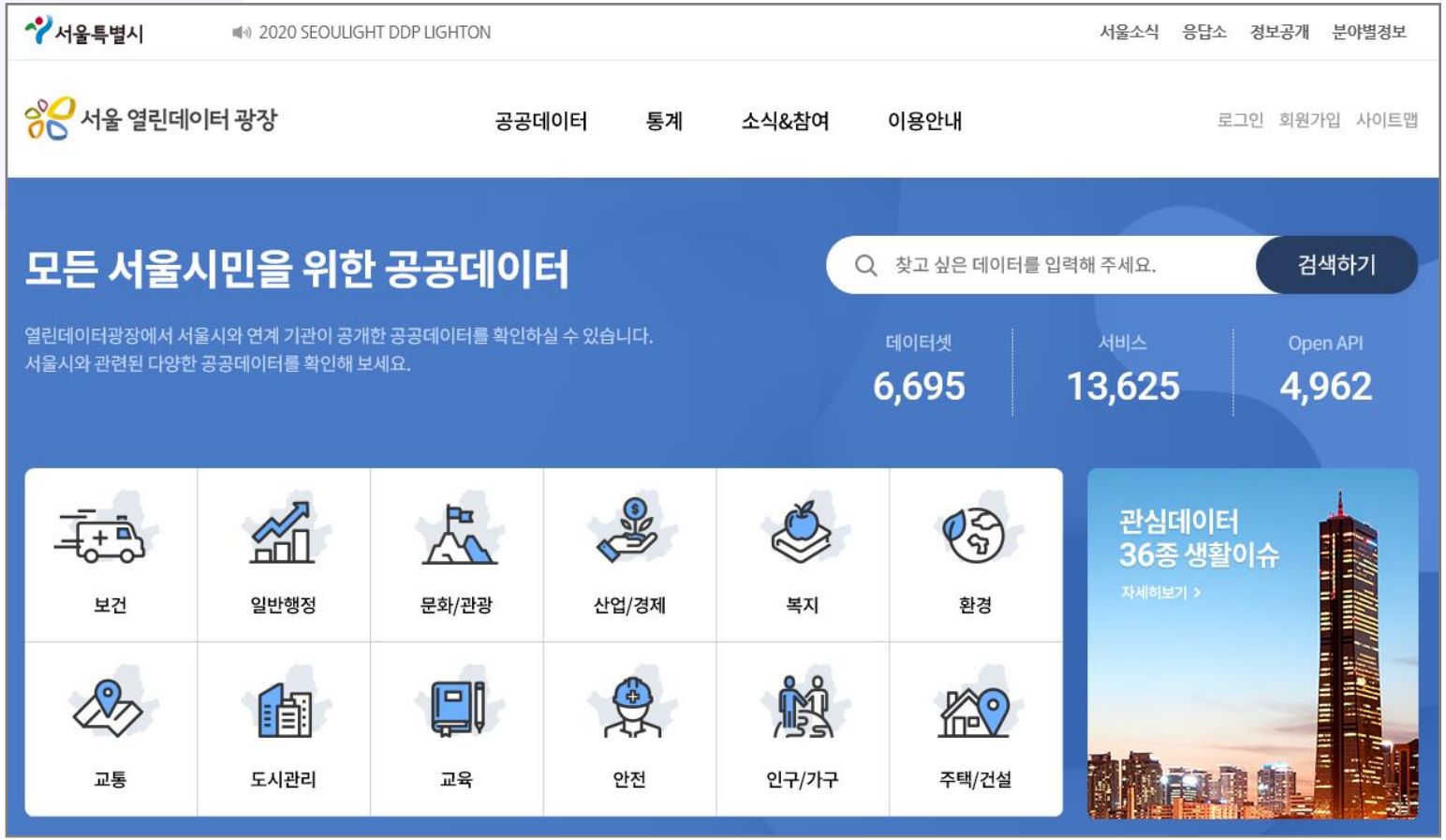
공공 데이터와 오픈 API 활용 데이터 수집

- 공공데이터 포털

데이터 수집

공공 데이터와 오픈 API 활용 데이터 수집

- 공공데이터 – 열린 데이터 광장



데이터 수집

공공 데이터와 오픈 API 활용 데이터 수집

공공데이터 – 열린 데이터 광장

01 12개 분야의 데이터셋을 제공

- 보건, 일반행정, 문화관광, 산업/경제, 복지, 환경, 교통, 도시 관리, 교육, 안전, 인구/가구, 주택/건설

02 데이터는 크게 두 가지로 나누어 제공

- 원천 데이터
- 통계(가공된) 데이터

데이터 수집

공공 데이터와 오픈 API 활용 데이터 수집

공공데이터 – 열린 데이터 광장

03 7가지 서비스 유형으로 데이터 제공

- = Sheet : 온라인 데이터시트(간단한 검색 및 분석 가능)
- = Open API : 개발자용 데이터베이스 제공 서비스
- = Chart : 데이터에 대한 간단한 시각화 결과 제공
- = Map : 지도상에 데이터 정보 표시(지리정보 시각화)
- = File : 데이터셋이 많은 경우 기간별/유형별 데이터를 별도의 파일로 제공
- = Link : 외부 데이터 주소 링크
- = LOD : 연결된 공개 데이터, Open API와 유사하나 여러 데이터를 연동해 구성이 가능

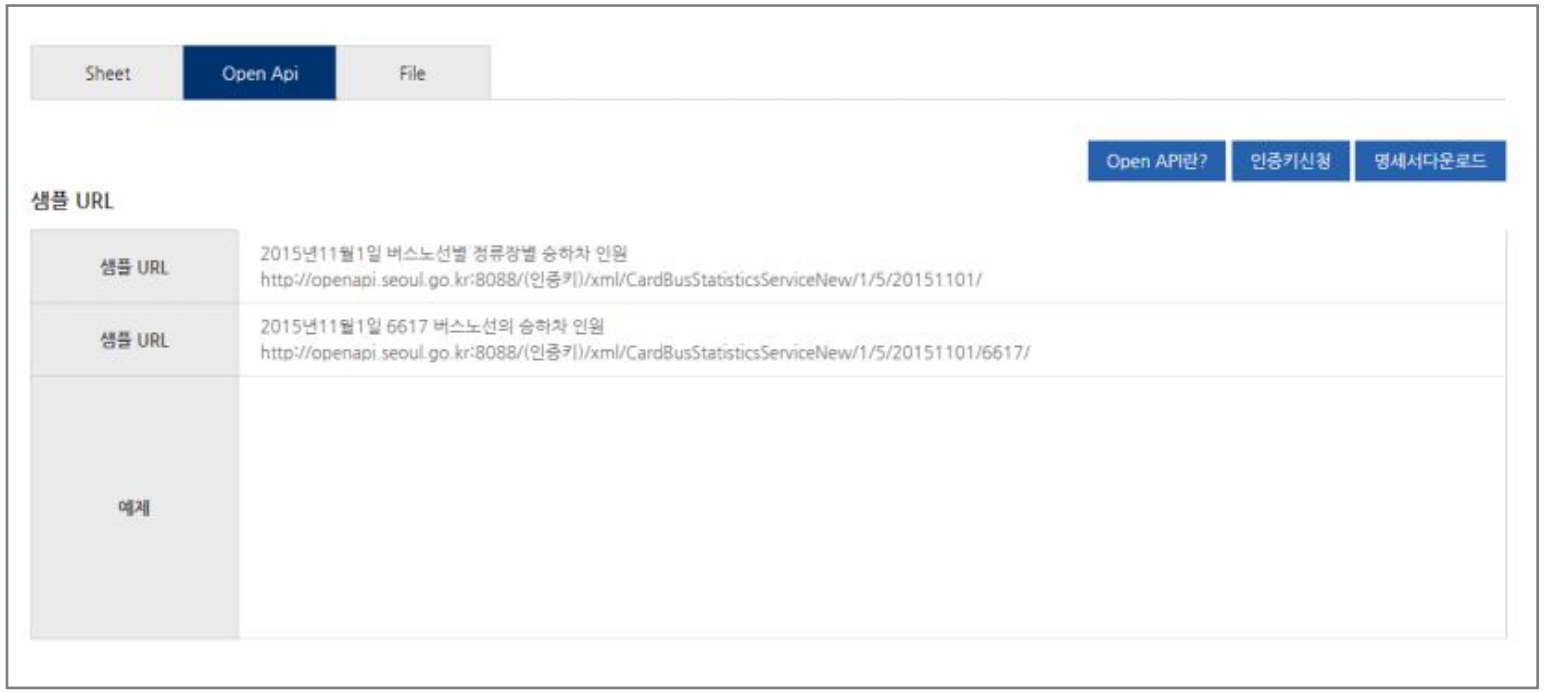
데이터 수집

공공 데이터와 오픈 API 활용 데이터 수집

- 오픈 API

개인 인증키와 데이터 제공 경로를 함께 보내 데이터를 받는 방식

서비스 제공 사이트에서 반드시 API 키를 신청해야 함(무료/유료)



데이터 수집

공공 데이터와 오픈 API 활용 데이터 수집

- 오픈 API

개인 인증키와 데이터 제공 경로를 함께 보내 데이터를 받는 방식

서비스 제공 사이트에서 반드시 API 키를 신청해야 함(무료/유료)

데이터 수집

공공 데이터와 오픈 API 활용 데이터 수집

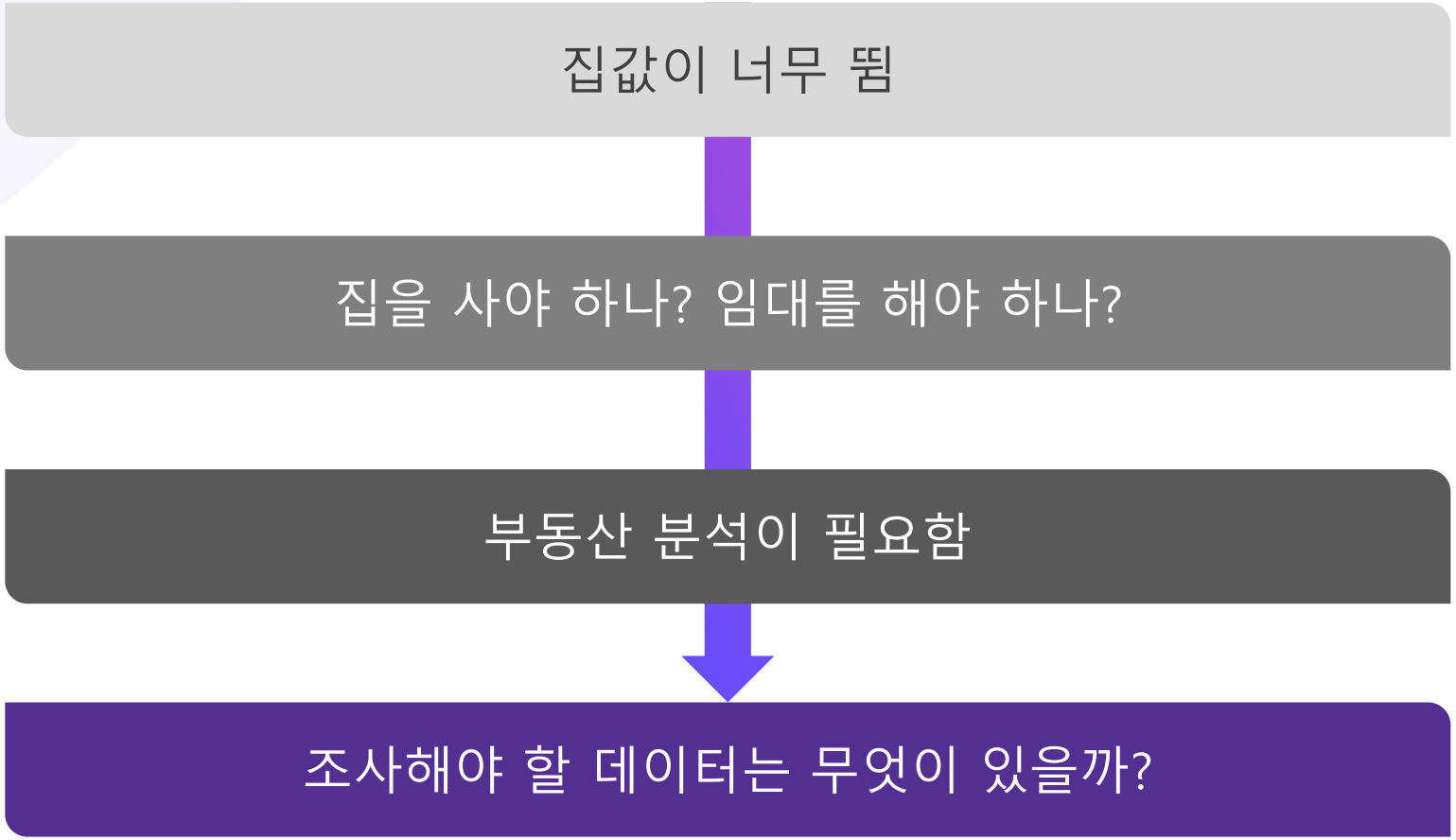
- 오픈 API

This XML file does not appear to have any style information associated with it. The document tree is shown below.


```
<CardBusStatisticsServiceNew>
  <list_total_count>55</list_total_count>
  <RESULT>
    <CODE>INFO-000</CODE>
    <MESSAGE>정상 처리되었습니다</MESSAGE>
  </RESULT>
  <row>
    <USE_DT>20151101</USE_DT>
    <BUS_ROUTE_ID>11110276</BUS_ROUTE_ID>
    <BUS_ROUTE_NO>6617</BUS_ROUTE_NO>
    <BUS_ROUTE_NM>6617번(양천차고지-목동우성아파트)</BUS_ROUTE_NM>
    <STND_BSST_ID>114000395</STND_BSST_ID>
    <BSST_ARS_NO>15718</BSST_ARS_NO>
    <BUS_STA_ID>9036852</BUS_STA_ID>
    <BUS_STA_NM>양천공영차고지</BUS_STA_NM>
    <RIDE_PASGR_NUM>41</RIDE_PASGR_NUM>
    <ALIGHT_PASGR_NUM>3</ALIGHT_PASGR_NUM>
    <WORK_DT>20151203</WORK_DT>
  </row>
  <row>
    <USE_DT>20151101</USE_DT>
    <BUS_ROUTE_ID>11110276</BUS_ROUTE_ID>
    <BUS_ROUTE_NO>6617</BUS_ROUTE_NO>
    <BUS_ROUTE_NM>6617번(양천차고지~목동우성아파트)</BUS_ROUTE_NM>
    <STND_BSST_ID>114000258</STND_BSST_ID>
    <BSST_ARS_NO>15363</BSST_ARS_NO>
    <BUS_STA_ID>0076525</BUS_STA_ID>
    <BUS_STA_NM>푸른마을3단지앞</BUS_STA_NM>
    <RIDE_PASGR_NUM>14</RIDE_PASGR_NUM>
    <ALIGHT_PASGR_NUM>7</ALIGHT_PASGR_NUM>
    <WORK_DT>20151203</WORK_DT>
  </row>
```

문제해결을 위한 시나리오 설정 및 데이터 검색

- 문제해결을 위한 시나리오
  - 문제제기



문제해결을 위한 시나리오 설정 및 데이터 검색

문제해결을 위한 시나리오

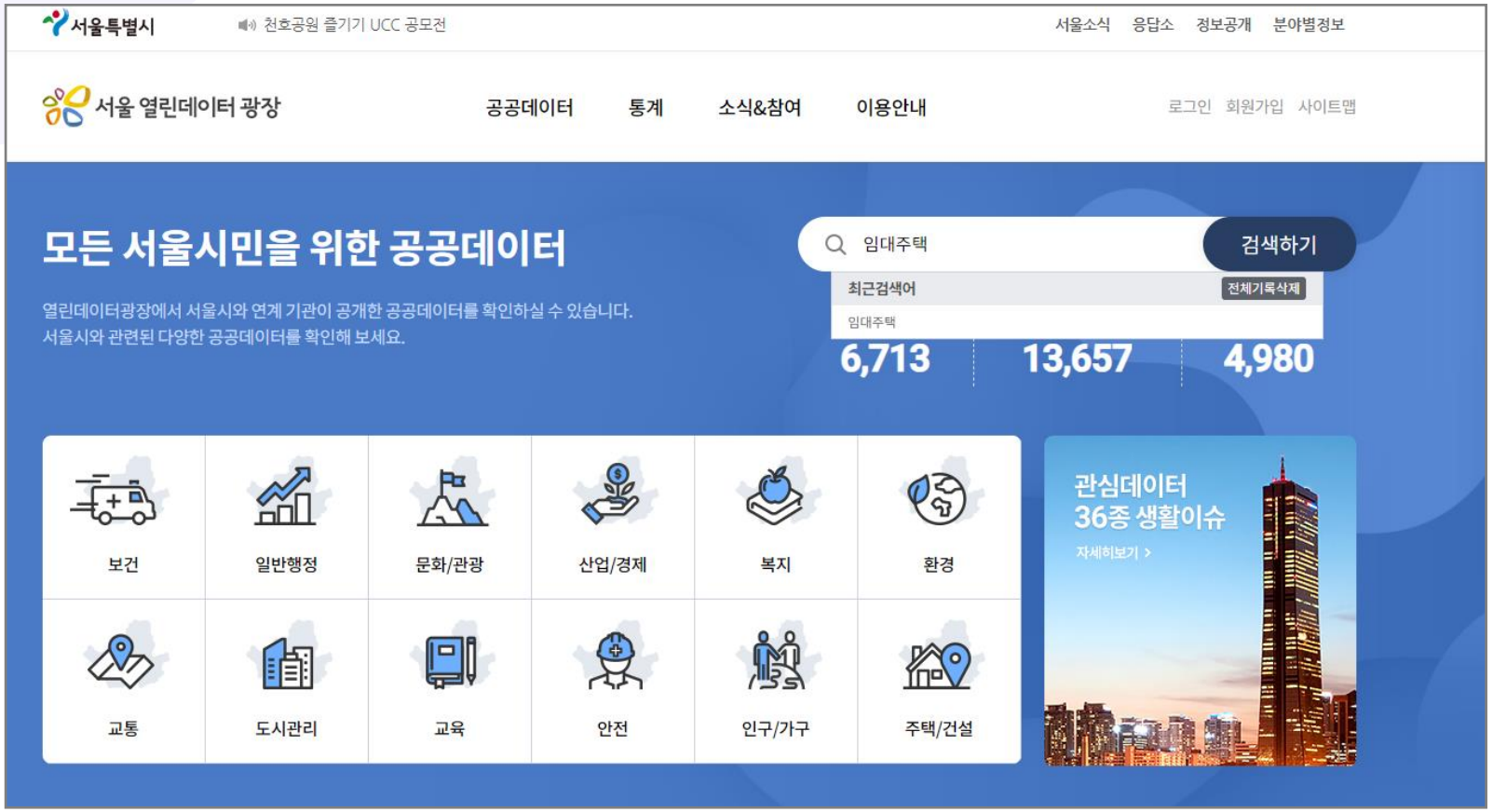
조사내역

- 해당 지역 집값 데이터 수집
  - 주택 유형별
- 해당 지역에 거주하고 있는 사람들에 대한 정보
  - 연령층, 교육수준
- 해당 지역 교통편의시설 분포/이용량
- 해당 지역 편의시설, 회사, 학교분포
- 해당 지역 전입/전출 인구
- 해당 지역 취약계층 분포

문제해결을 위한 시나리오 설정 및 데이터 검색

데이터 검색

검색어 : 임대주택



문제해결을 위한 시나리오 설정 및 데이터 검색

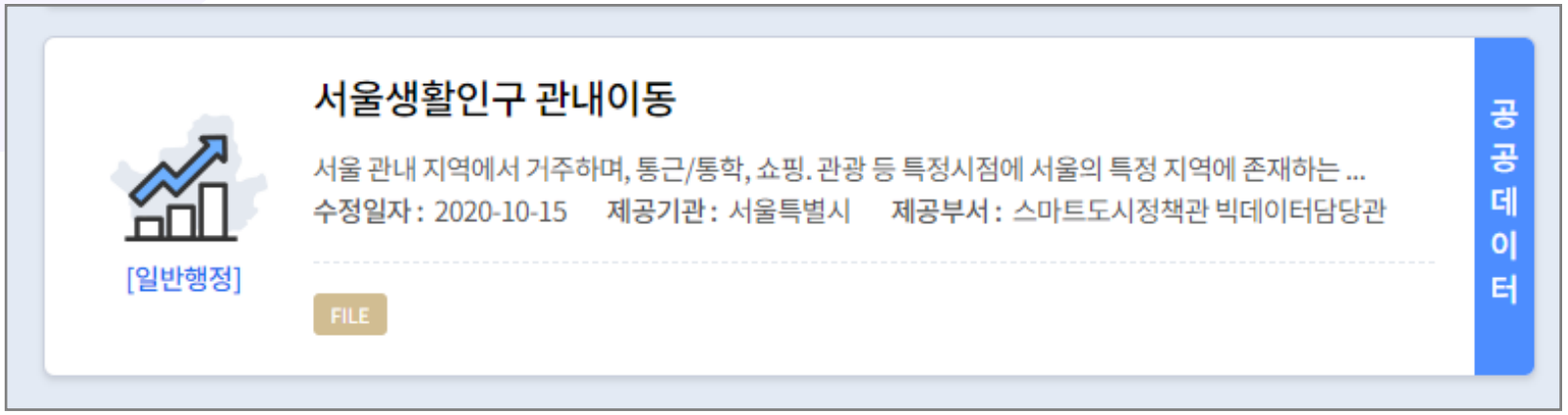
데이터 검색

검색어 : 임대주택

문제해결을 위한 시나리오 설정 및 데이터 검색

데이터 검색

검색어 : 생활인구



문제해결을 위한 시나리오 설정 및 데이터 검색

데이터 검색

검색어 : 교통

문제해결을 위한 시나리오 설정 및 데이터 검색

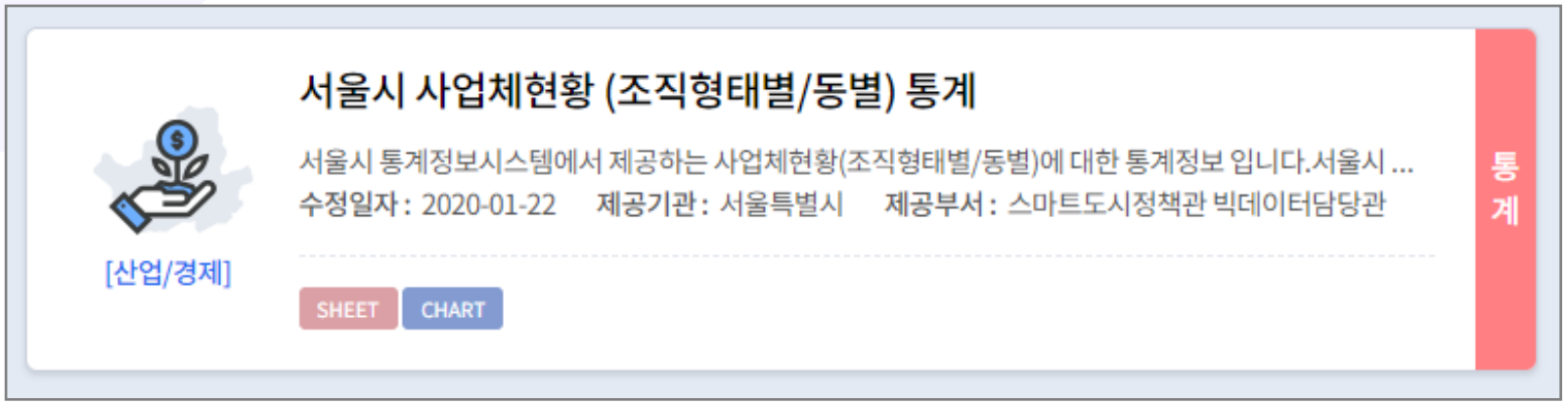
데이터 검색

검색어 : 부동산

문제해결을 위한 시나리오 설정 및 데이터 검색

데이터 검색

검색어 : 회사



문제해결을 위한 시나리오 설정 및 데이터 검색

이론 영상 후 실습 영상 제시
영상 2개로 나눠서 촬영

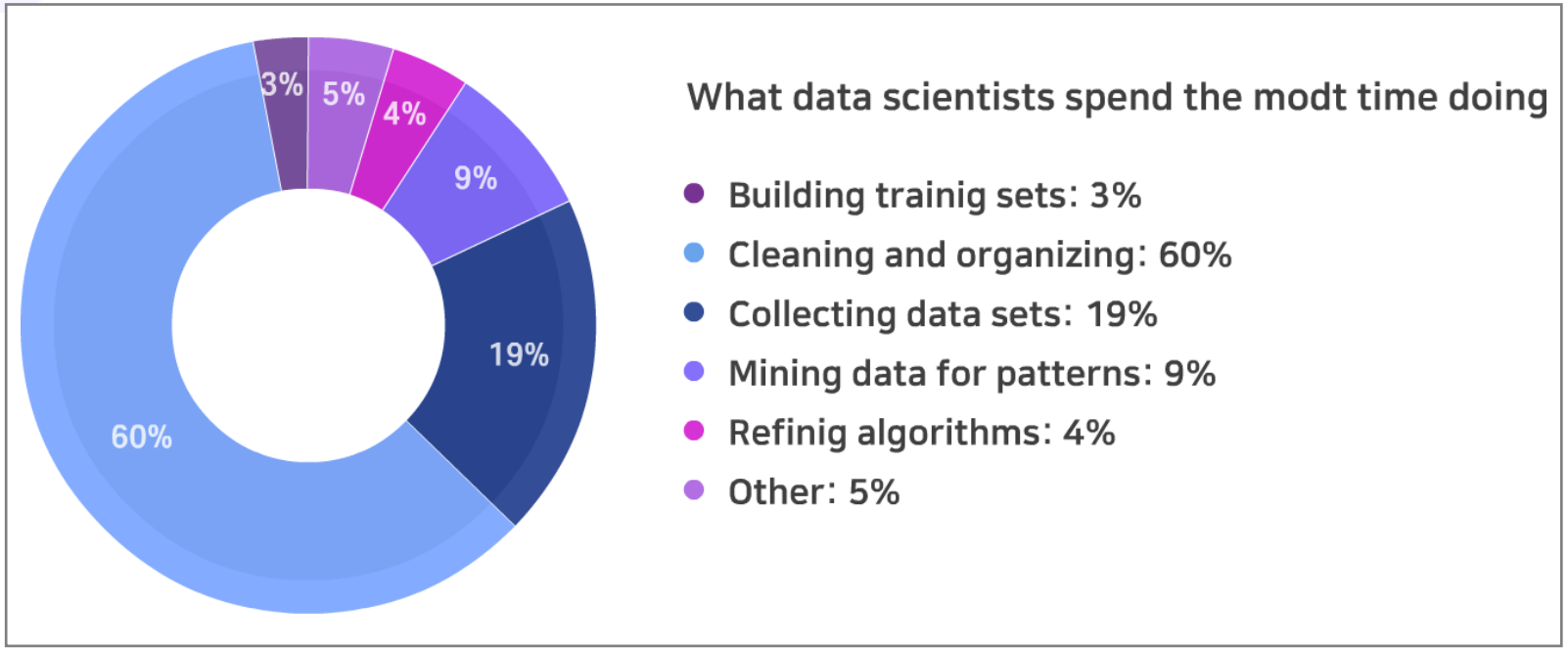
데이터 전처리, 모델링, 시각화

데이터 전처리, 모델링, 시각화

데이터 전처리

데이터를 확인하고, 정제하고, 변환하고, 다듬는 일련의 과정

데이터 분석 단계에서 가장 많은 시간과 노력이 필요한 단계

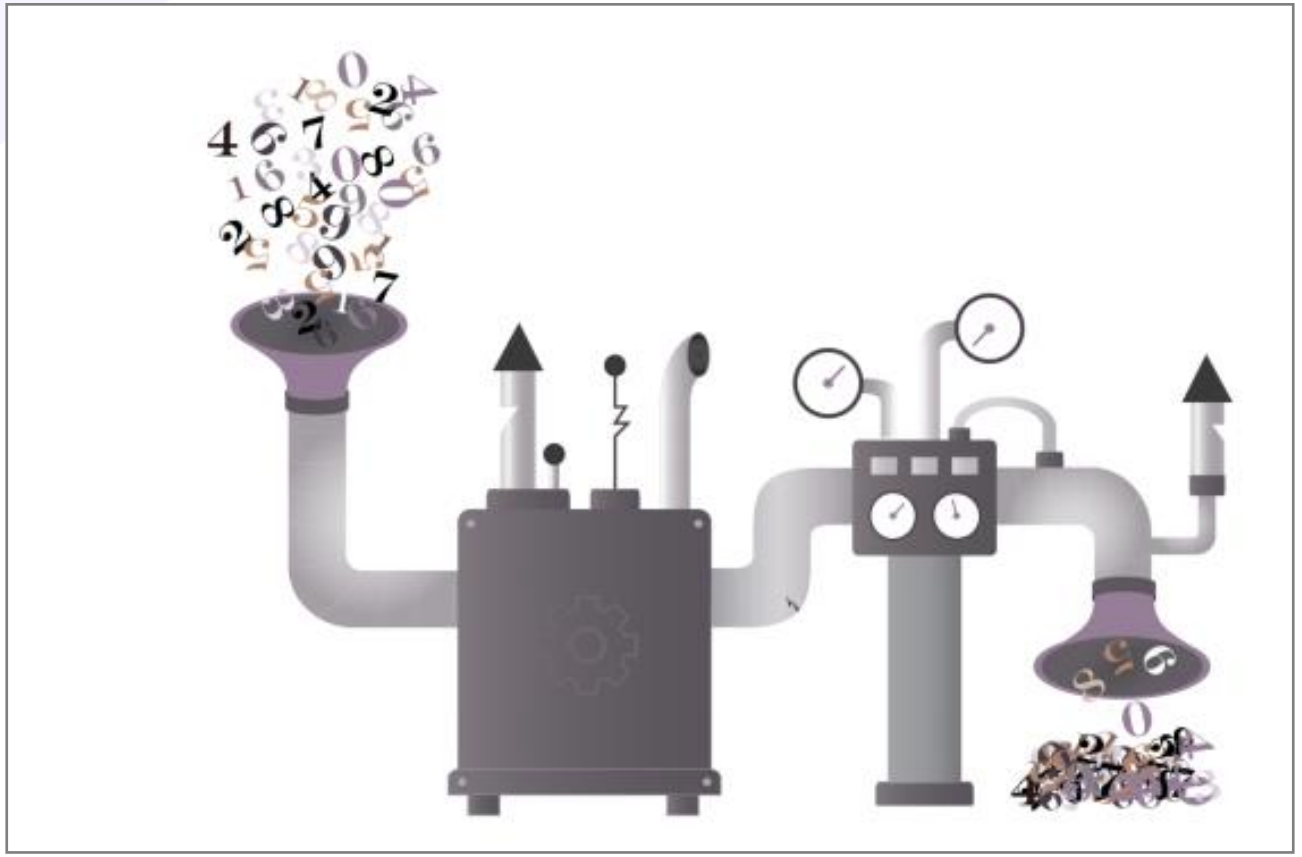


<출처 :https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-mosttime-consuming-least-enjoyable-data-science-task-survey-says/#551830c86f63

데이터 전처리, 모델링, 시각화

데이터 전처리

- GIGO(Garbage In Garbage Out)



데이터 전처리, 모델링, 시각화

데이터 전처리

- 데이터를 그대로 특징으로 사용할 수 있는 경우

형식화	▪ 분석 목적에 맞는 적절한 데이터 형식으로 변환 • 범주형/수치형
정제	▪ 문제 해결에 도움이 되지 않는 데이터는 제거, 누락된 (결측) 데이터 처리
정규화	▪ 개별 특징이 동일한 숫자 범위 안에 들어가도록 변환
분해	▪ 하나의 특징이 복잡한 여러 개념을 포함하는 경우 • 분리 or 선택
결합	▪ 두 가지를 모두 고려했을 때 더 큰 의미를 가지는 특징은 결합

데이터 전처리, 모델링, 시각화

데이터 전처리 방법

- 전처리에는 정답이 없으므로 경험이 중요함
- 가장 중요한 것은 데이터의 신뢰성 검증임
- 실무 – 분석 가능한 형태로 데이터를 변환하는 작업은 괴로움
- 나머지 데이터 형식화, 분해, 결합은 선택적으로 적용해야 함

다음과 같이 전처리를 하고 난 뒤, 파이썬/R을 분석과정에 동시 사용하고, 시각화는 파워 BI로 마무리하여 보고서를 작성해야 함

데이터 전처리, 모델링, 시각화

데이터 모델링

문제에 대한 깊은 이해를 위해선
서로 다른 데이터들을 관계를 지어 분석할 필요가 있음

데이터 모델링의 두 가지 의미

01 서로 다른 데이터들을 관계 맺음

02 데이터의 패턴을 찾아 수학적 모델로 만듦

- 보통 이를 데이터 모델링이라고 부름

데이터 전처리, 모델링, 시각화

데이터 모델링

파워 BI의 모델링 = 테이블 연결

- 카디널리티
- 1:1 관계
  - 테이블 A의 한 레코드가 테이블 B의 한 레코드에만 연결된 상태 (직원 - 성과)
- 1:N 관계
  - 테이블 A의 한 레코드가 테이블 B의 여러 레코드와 연결된 상태 (도서 회원 - 도서 대여 목록)
- N:N 관계
  - 테이블 A의 한 레코드가 테이블 B의 여러 레코드와 연결되고,
    테이블 B의 한 레코드가 테이블 A의 여러 레코드와 연결된 상태 (학생 - 수업)

데이터 전처리, 모델링, 시각화

데이터 시각화

반드시, 데이터 성질에 따른 분류에 의해
파워BI 의 축과 값, 측정값 등을 설정하라

양적자료

- 수치(이산/연속), 크기
- 등간 척도
  - 수치 사이의 간격에 의미가 있으며 원점을 가지지 않음

    예 온도
- 비율 척도
  - 수치 그 자체에 의미가 있으며 원점을 가짐

    예 신장, 체중, 연령

데이터 전처리, 모델링, 시각화

데이터 시각화

반드시, 데이터 성질에 따른 분류에 의해
파워BI 의 축과 값, 측정값 등을 설정하라

질적자료

- 범주(명목), 순서를 가짐(순서)

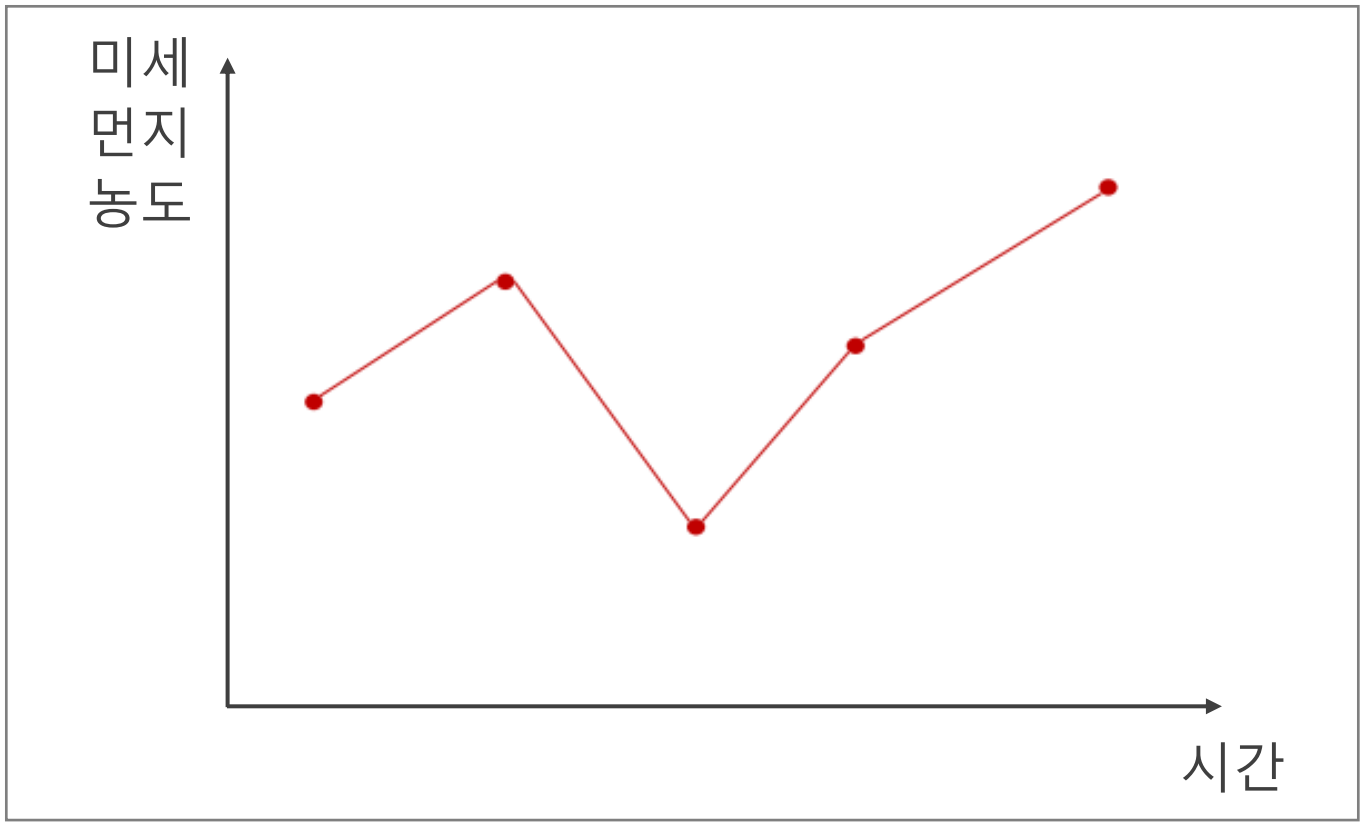
질적 자료는 그림을 그리는 기준,
양적 자료는 보고 싶은 값/기준임을 기억하자

데이터 전처리, 모델링, 시각화

데이터 시각화

- 데이터 성질에 따른 시각화

수치 데이터 + 수치 데이터

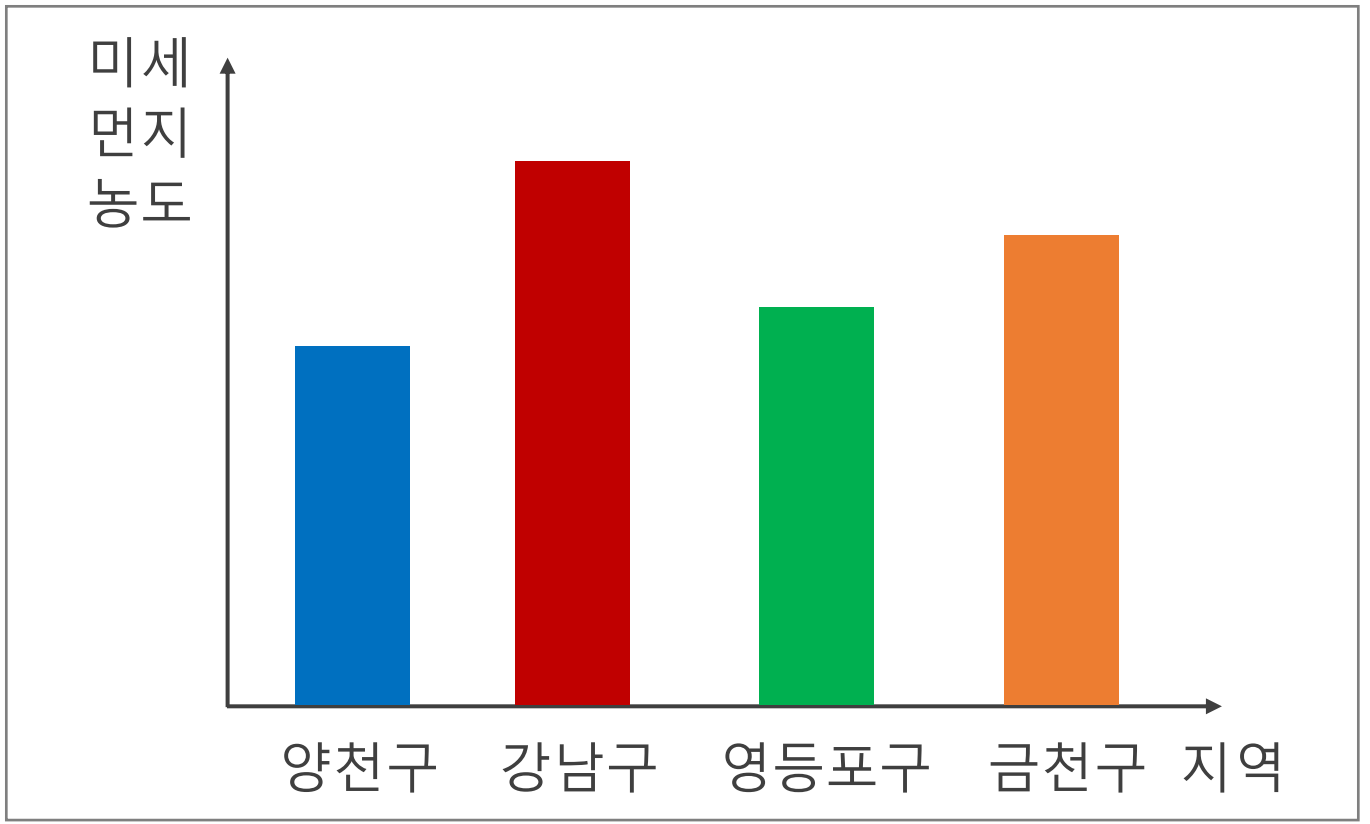


데이터 전처리, 모델링, 시각화

데이터 시각화

- 데이터 성질에 따른 시각화

범주 데이터 + 수치 데이터



데이터 전처리, 모델링, 시각화

데이터 시각화

01 선 그래프

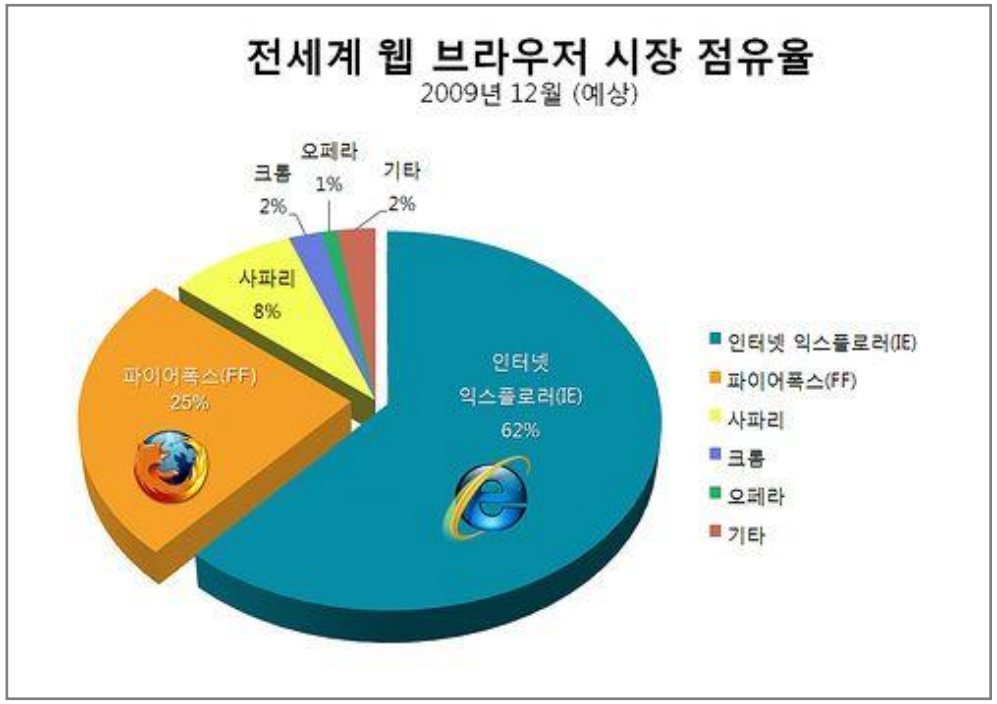
- X축 값의 변화에 따른 Y축 값의 변화
- 두 데이터 또는 한 데이터 내의 두 속성 사이의 관계

데이터 전처리, 모델링, 시각화

데이터 시각화

02 파이 차트(원 그래프)

- 전체에 대한 각 항목의 비율을 원 모양으로 나타낸 그래프
- 한 데이터 – 속성 내 여러 범주들 사이의 비율 확인

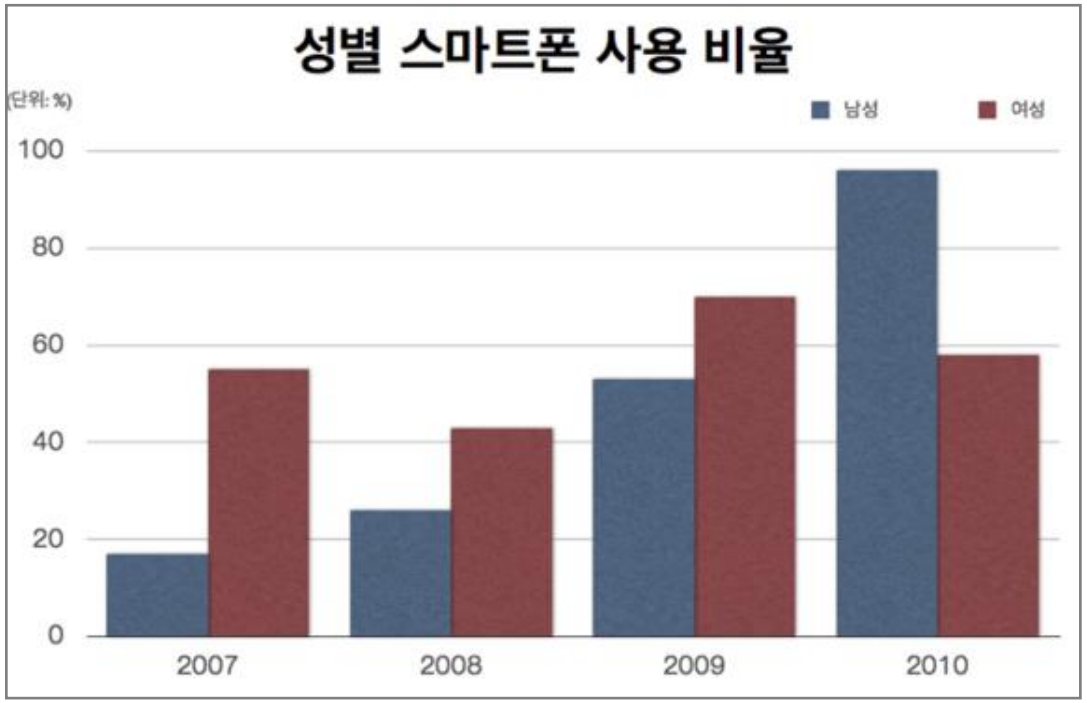


데이터 전처리, 모델링, 시각화

데이터 시각화

03 막대 그래프

- 자료의 양을 막대 모양의 길이로 나타낸 그래프
- 서로 다른 범주에 속하는 자료의 양을 비교

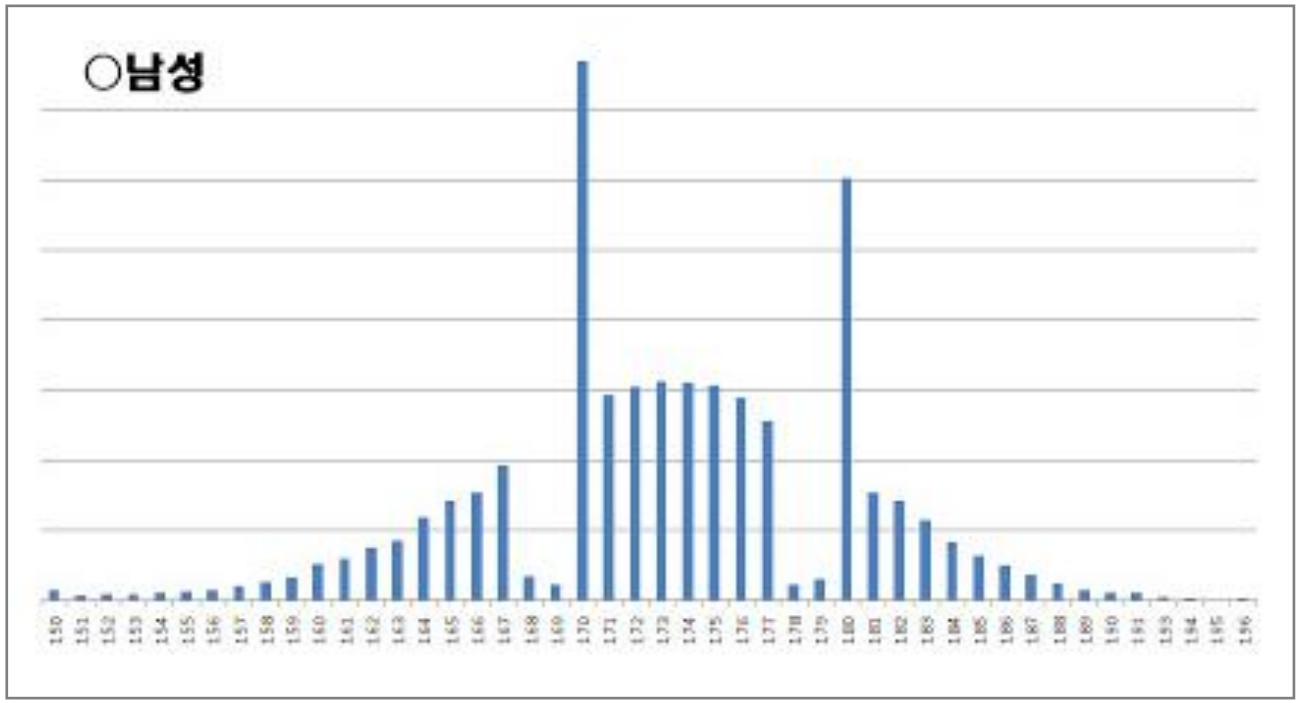


데이터 전처리, 모델링, 시각화

데이터 시각화

04 히스토그램

- 도수분포를 나타낸 막대그래프 모양의 그래프
- 데이터 내 값들의 빈도를 확인하고 싶을 때 사용

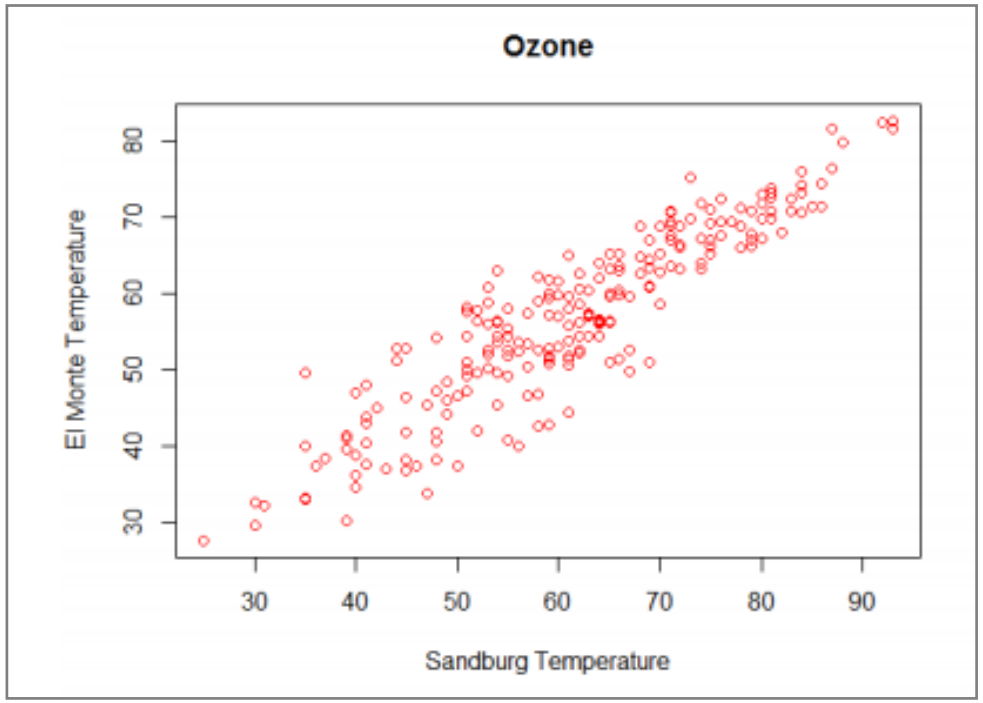


데이터 전처리, 모델링, 시각화

데이터 시각화

05 산점도 그래프

- 두 변수 간의 영향력을 보여주기 위해 가로 축과 세로 축에 데이터 포인트를 그리는 데 사용
  - 두 속성 사이의 관계(패턴)을 확인

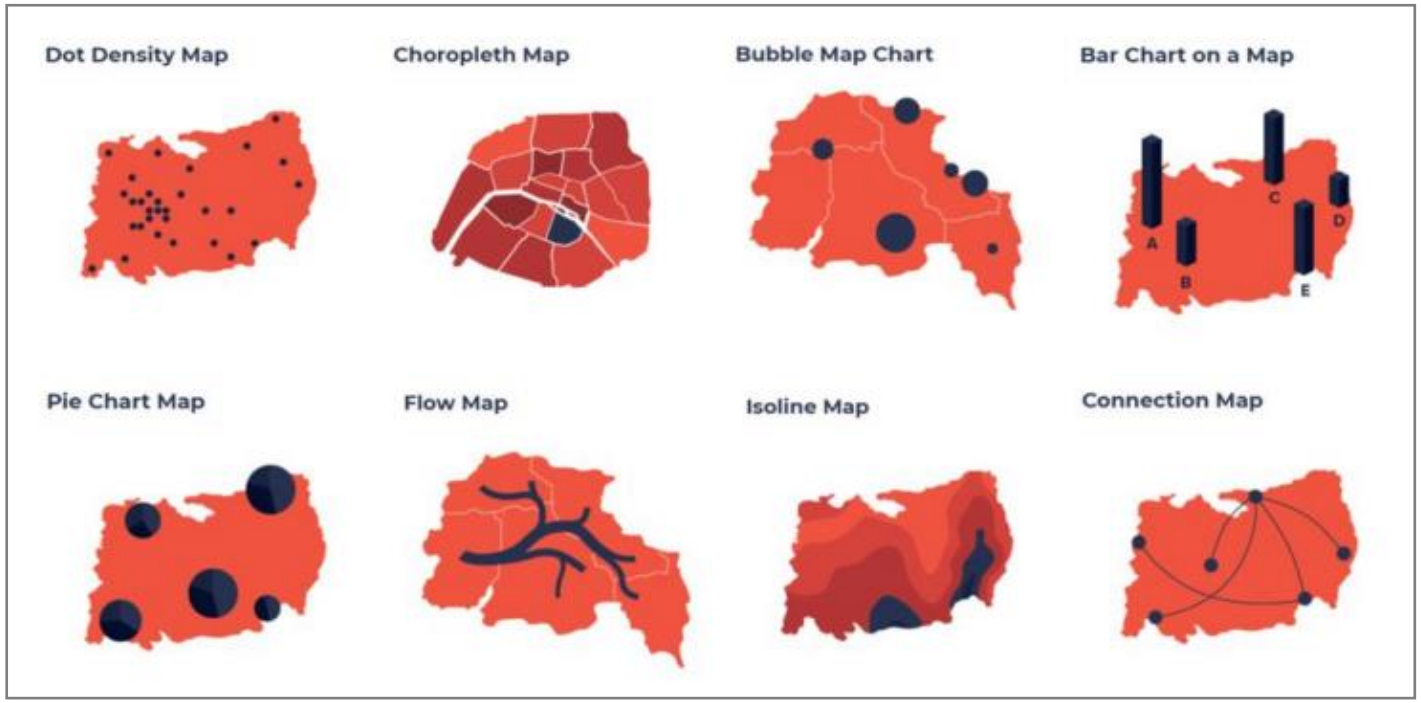


데이터 전처리, 모델링, 시각화

데이터 시각화

06 공간 그래프

- 데이터와 관련된 지리상의 위치 정보를 시각화
- 위치에 따른 빈도 분석이 가능



데이터 전처리, 모델링, 시각화

데이터 시각화

06 공간 그래프

- 국내 지진 발생 지점 공간 표시, 지진의 규모를 가중치로 표현

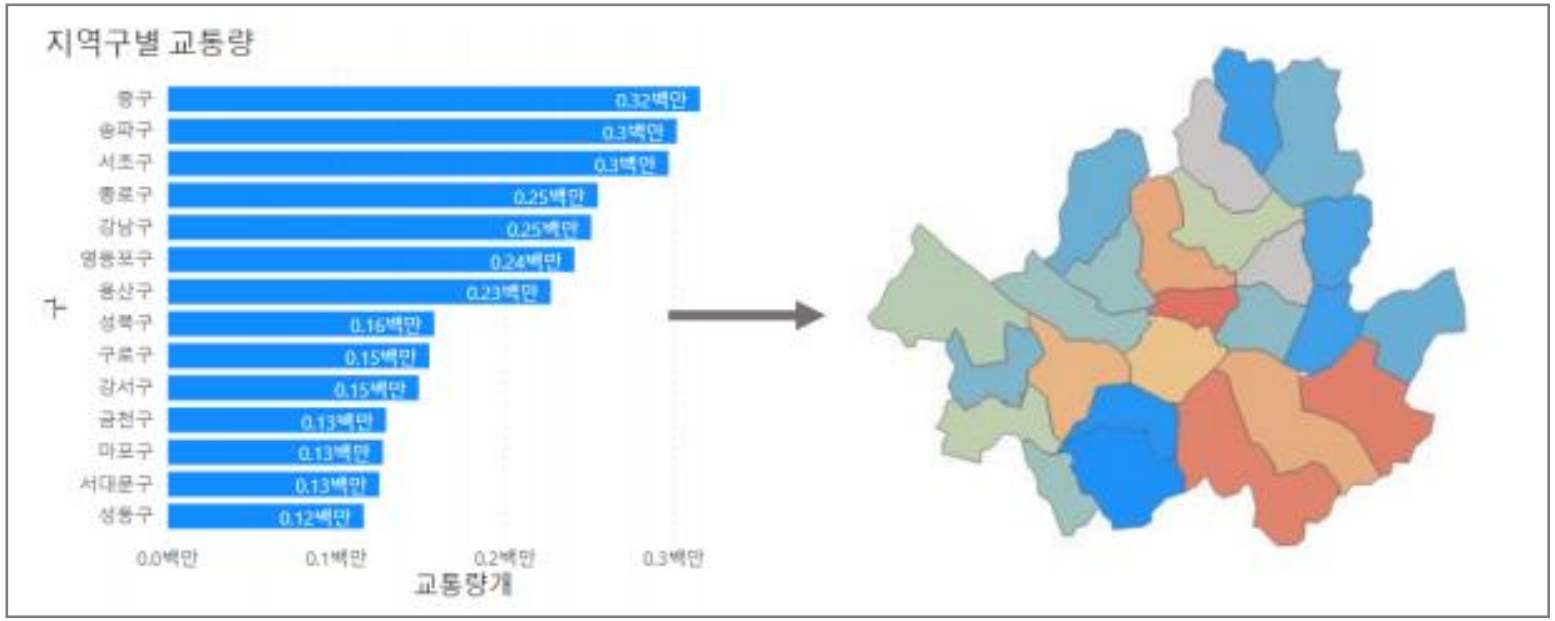
공간 데이터 분석

공간 데이터 분석

공간 데이터란?

지도 및 지도 위에 표현이 가능하도록 위치, 분포 등에서 알 수 있는
모든 정보로 일상생활이나 특정한 상황에서 행동이나 태도를 결정하는 중요한 기초 정보와 기준을 제시하는 데이터

- 단위에 따라 국토공간정보 또는 도시정보로 구분 가능



공간 데이터 분석

공간 데이터의 필요성

01 거시적으로 볼 때 안보, 관리, 안전에 중요한 역할을 수행함

02 국가별, 지역별 문제의 원인 등을 정확히 파악하고 분석함

02 단순 자료 비교만으로 패턴을 찾기 어렵기 때문에 필요함

- 지도상에 매핑이 필요함

  예 서울시 지역별 교통량, 지역별 매출 분석

공간 데이터 분석

지도(Map)

측량 결과에 따라 공간상의 위치와 지형 및 지명 등 여러 공간정보를 일정한 축척에 따라 기호나 문자 등으로 표시한 것

시공간에 존재하고 있는 여러 가지 상황을 일정한 약속에 따라 2차원 또는 3차원에 나타낸 것

공간 데이터를 표현한 결과물

공간 데이터 분석

지도(Map)

- 지도로 표현 가능한 데이터는?

01 주소(지번/도로명 주소)

  - 서울시 마포구 상암산로1길

02 위도 + 경도

02 행정동 코드 + 집계구 코드

공간 데이터 분석

지도(Map)

- 공간 정보를 담고 있는 파일 유형

공간 정보를 사용하는 목적, 용도에 따라 다양한 형식을 사용함

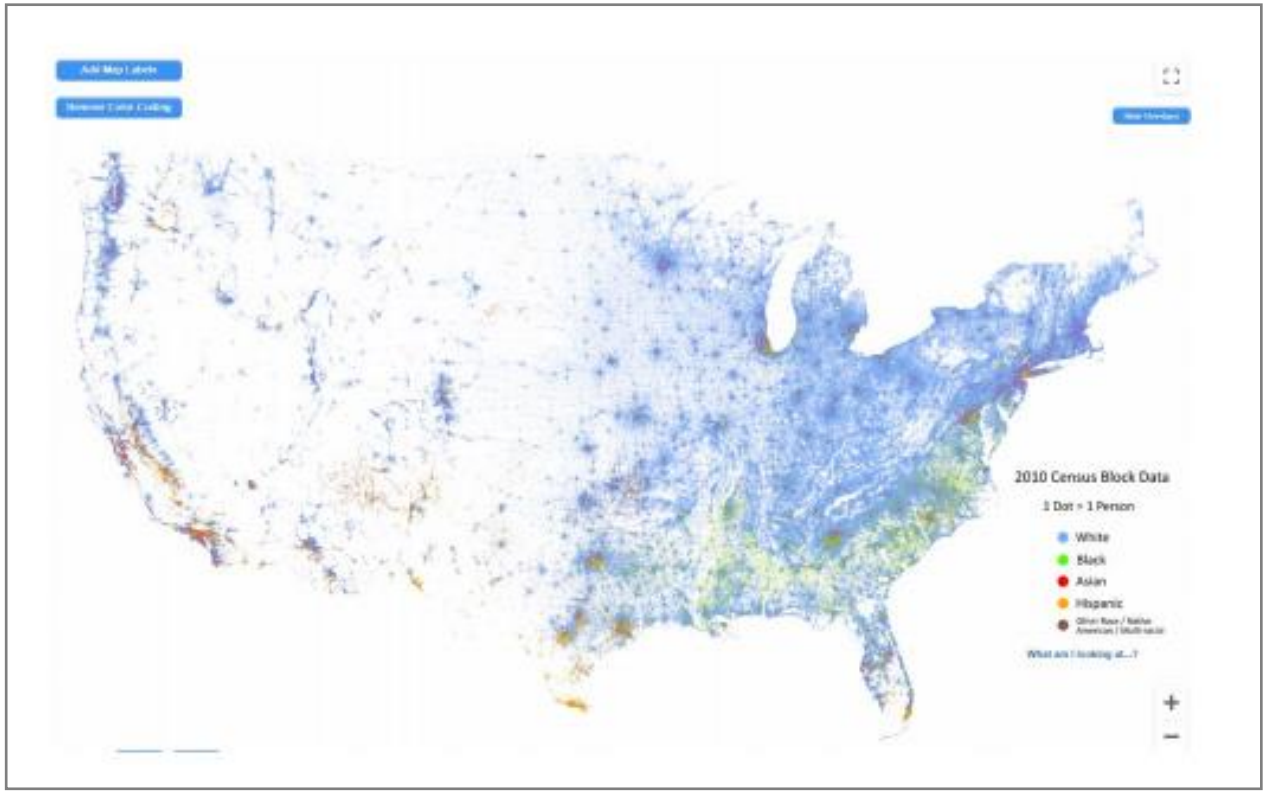
- .shp .dbf, .shx / .json (topo) / .geojson / .gml / .kml

공간 데이터 분석

공간 데이터 시각화 유형

01 Dot Density Map

- 지도 위에 데이터의 분포를 표현한 그림

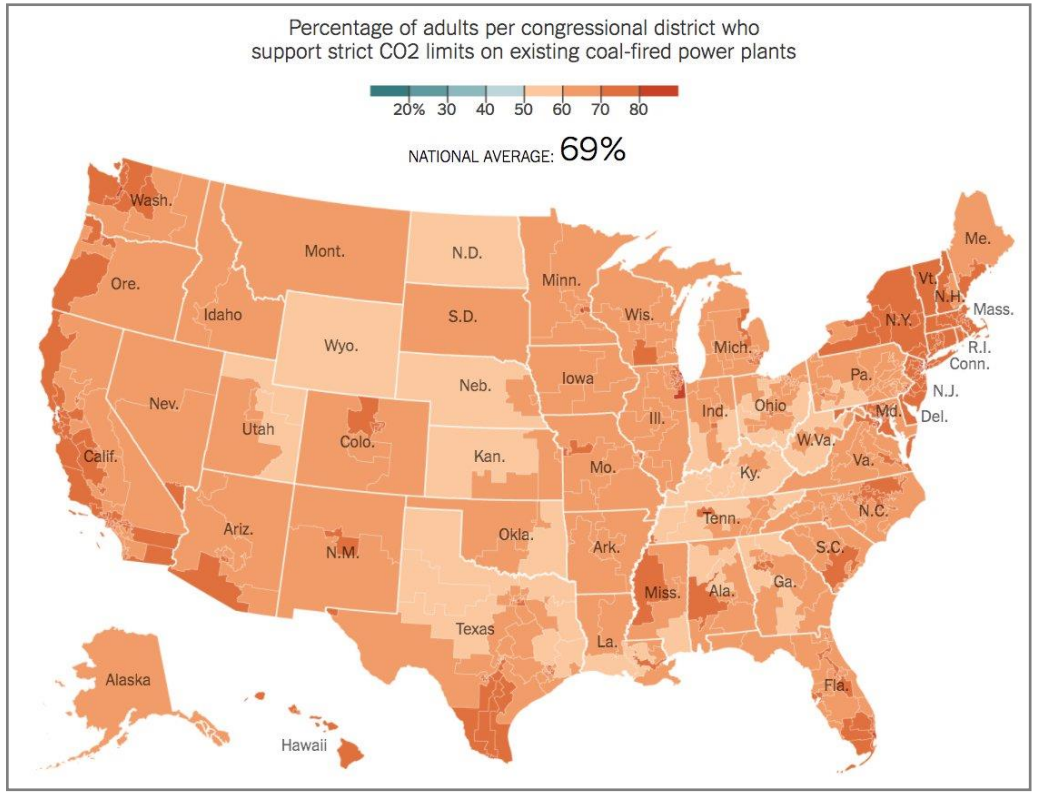


공간 데이터 분석

공간 데이터 시각화 유형

02 Choropleth Map

- 지리적 영역 범위 별 수치 데이터 값을 색으로 표현

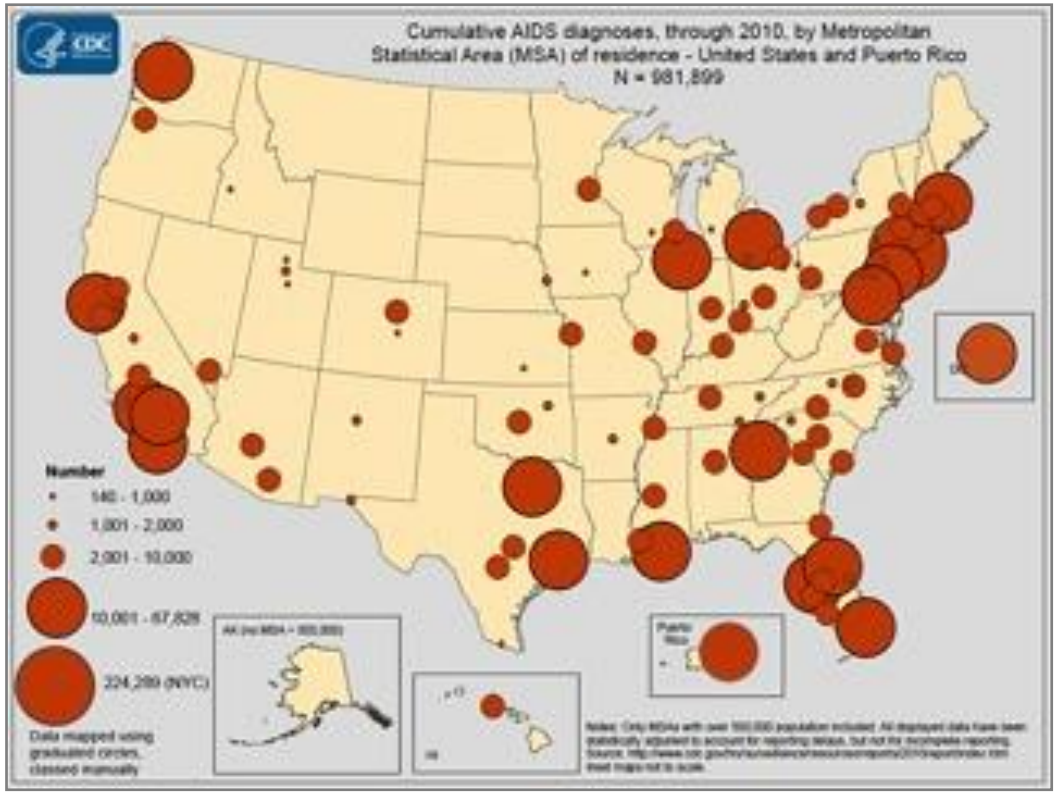


공간 데이터 분석

공간 데이터 시각화 유형

03 Symbol Map

- 지도의 특정 지점에 해당하는 수치 값을 심볼의 크기로 표현

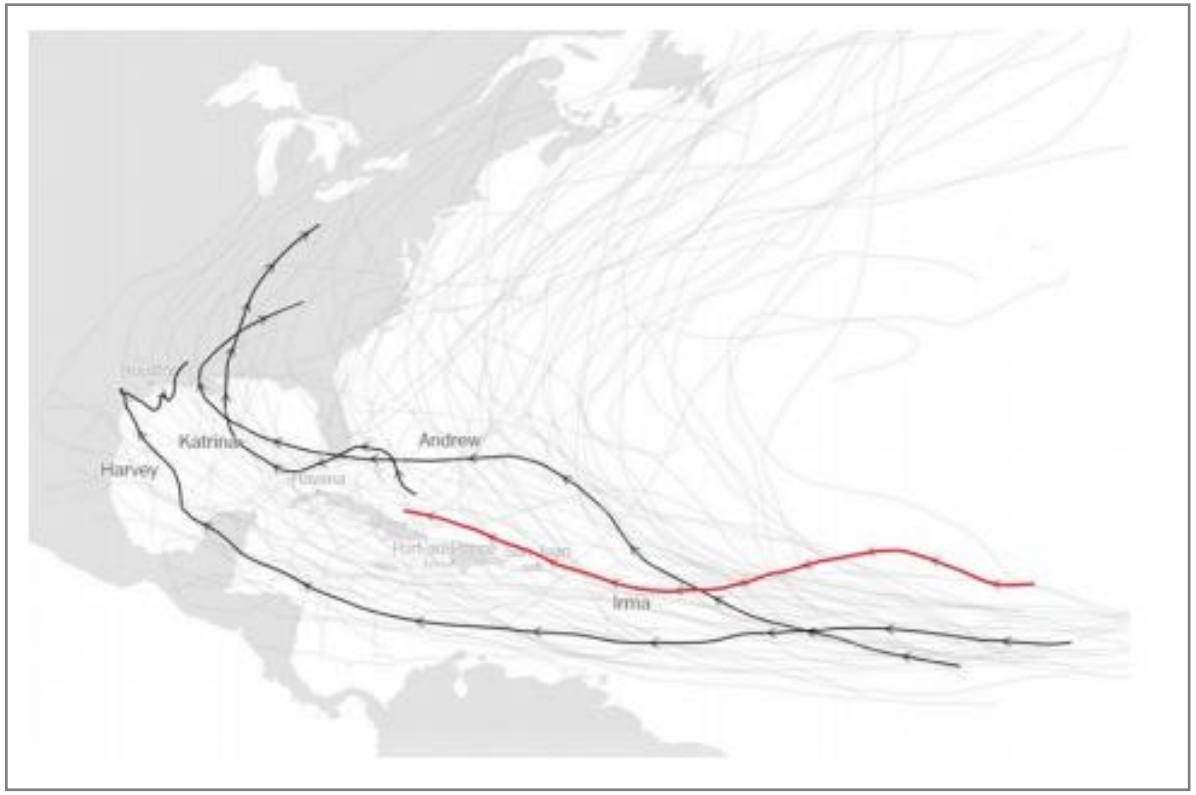


공간 데이터 분석

공간 데이터 시각화 유형

03 Connection, Flow Map

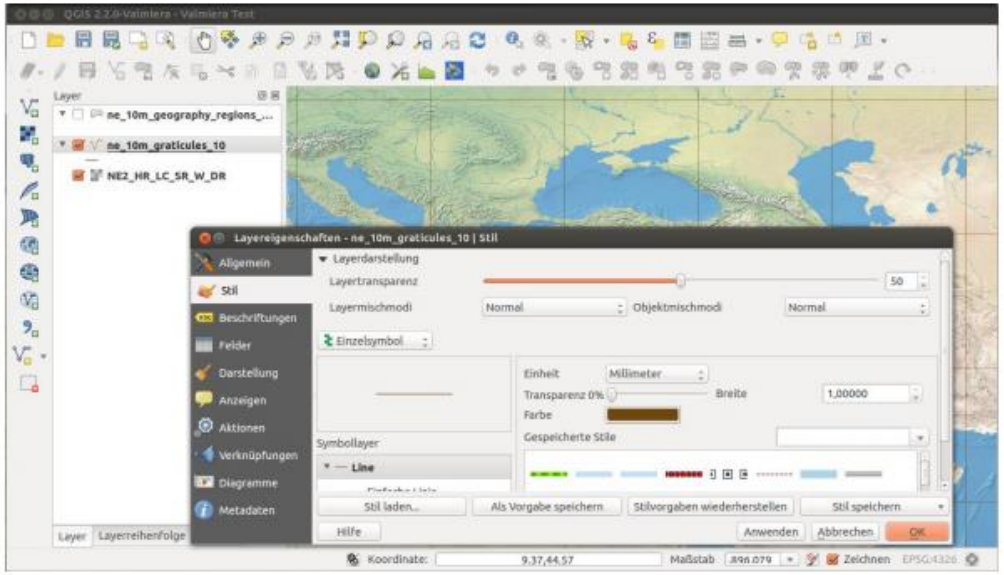
- 정보의 지리적인 이동 경로를 표현



공간 데이터 분석

공간 데이터 분석 도구

- 대표적 공간데이터 분석 도구 qGIS



- 무료 오픈소스 지리 정보 시스템
- 데이터 탐색, 공간 정보 분석, 확장 기능 등을 제공
- 공간 데이터 조회, 편집, 분석 기능을 제공하는 대표적인 오픈소스 데스크톱
  - 지리정보시스템 소프트웨어

공간 데이터 분석

이론 영상 후 실습 영상 제시