

데이터 분석 프로세스

데이터 분석 프로세스

데이터 분석

- 여러 데이터를 수집, 정리, 변환하여 유용한 정보를 발견하여 결론을 내리고 의사결정을 지원하는 과정
- 우리는 평상시에도 데이터를 바탕으로 의사결정을 하고 있다.
 - ex) 가게부를 보고 ‘~ 지출을 줄여야지.’
- data is everything, and everything is data
- 데이터 분석가가 아니라 다양한 영역에서 데이터 분석에 대한 니즈는 커지고 있다.

데이터 분석 프로세스

접근 방법

- 생산성, 효율성 향상 : 데이터 시각화를 활용해서 비효율 공정을 확인하여 낭비 되는 부분을 식별하고 불필요하게 낭비 되는 부분을 찾는다.
- 의사결정 : 데이터를 활용하여 현 상황에 대해서 이해하고 미래를 예측하여 의사결정에 도움을 준다.
- 설득 : 데이터를 활용하여 주장을 뒷바침할 근거를 제시 한다.

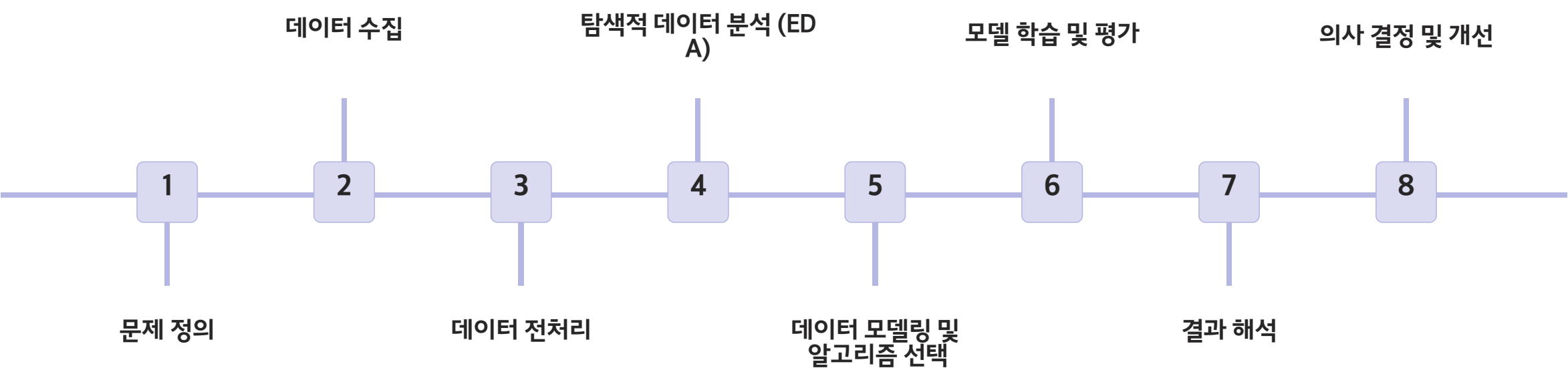
데이터 분석 프로세스

마음 가짐

- 서로 다른 경험과 배경지식, 프로젝트를 통해 얻고자 하는 목적이 달라 같은 문제도 서로 다르게 이해할 수 있다.
-> 원활한 의사소통 필요
- 시각화로 표현 된 데이터를 수용하는 자세
- 데이터를 어디서 어떻게 활용할 것인지 정의
-> 가지고 있는 역량, 장비에 따라 달라질 수 있음
- 의미있는 가치 창출 -> 결국 돈이 되는가?

데이터 분석 프로세스

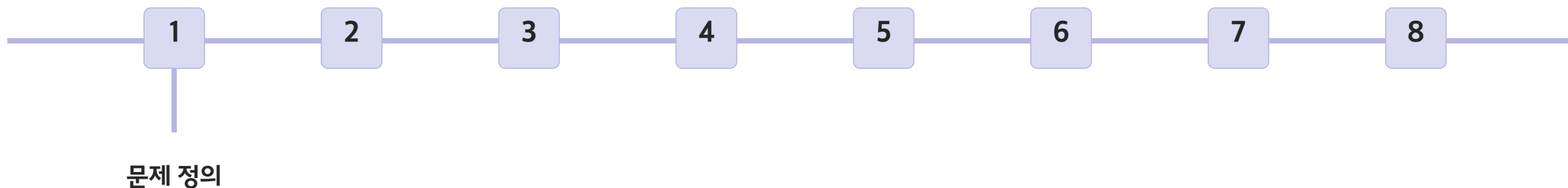
프로세스



데이터 분석 프로세스 (문제정의)

데이터 분석 프로세스

문제정의



데이터를 기반으로 해결하고자 하는 **문제**가 무엇인가?

문제를 제대로 이해하기 위해서 **무엇**이 필요할까?

데이터 분석 프로세스

문제정의

- 데이터의 분석과 활용의 목적 분명하게 정의
- 해결하기 위해 필요한 데이터는 무엇인지?
- 현업 내부에서 수집할 수 있는 데이터인지?
- 외부 협력이 필요한 데이터인지?
- 언제 발생한 데이터를 수집해야 하는지?



Why?



What?



Where



Where



When

데이터 분석 프로세스

문제정의

1. 문제 정의 : 데이터 분석의 목적과 해결해야 할 문제를 명확하게 정의합니다. 이는 분석의 방향성을 제시하고 필요한 데이터와 분석 기법을 결정하는 데 도움을 준다.
2. 절차
 1. 목표 이해 (과제 정의): 분석을 해야하는 이유, 현 상태를 이해, 필요 정보 수집
 2. 분석계획 수립 : 필요한 자원을 바탕으로 분석 내용을 정의하고 스케줄 작성

데이터 분석 프로세스

목표 이해

- 데이터 분석을 통해 얻고자 하는 것 & 우리가 해결하려고 하는 문제는 무엇인가? (프로젝트의 목표 설정)
- 데이터 분석에 필요한 정보 (우리가 집착한 내부, 외부 상황)를 파악하여 식별하여 수집한다.
 - 우리가 프로젝트를 진행 해야 하는 이유와 배경
 - 우리가 가진 문제
- 분석에 들어 가기 전 데이터 분석을 활용해서 문제를 해결하거나 확인할 수 할 수 있는 것인지 할 수 없는 것인지 생각해보기

데이터 분석 프로세스

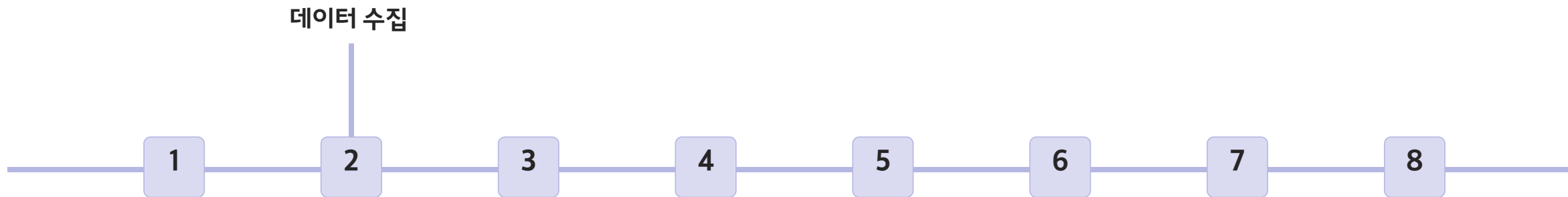
분석계획 수립

- 프로젝트 일정 수립 : 인력, 자원, 시간을 고려하여 상세 일정을 수립
- 하고자 하는 분석 & 모델링에서 필요한 데이터는 무엇인지?
- 주어진 자원, 데이터에서 분석이 가능한지 확인
- 데이터를 어떻게 정제할 것인지?
- 어떤 방식으로 시각화 하는 것이 좋은지?

데이터 분석 프로세스 (데이터 수집)

데이터 분석 프로세스

데이터 수집



목적에 맞는 데이터를 어떻게 수집할 것인지?

데이터 분석 프로세스

데이터 수집 방법

데이터베이스

데이터베이스는 구조화된 데이터를 저장하고 관리하는 시스템입니다. SQL 쿼리를 사용하여 데이터베이스에 접근하여 필요한 데이터를 추출합니다.

공공 데이터

공공 데이터는 정부 기관이나 공공 기관에서 제공하는 공개 데이터 세트입니다. 데이터 세트를 검색하고 다운로드하여 분석에 활용합니다.

Open API

Open API는 외부 애플리케이션이나 서비스에서 제공하는 데이터에 접근하기 위한 인터페이스입니다. 다양한 서비스들이 API를 통해 데이터를 제공하고 있으며, 이를 활용하여 데이터를 수집할 수 있습니다.

RPA(Robotic Process Automation)

RPA는 소프트웨어 로봇이 반복적이고 규칙적인 작업을 자동화하는 기술입니다. RPA 툴을 사용하여 작업 흐름을 구성하고, 필요한 작업 단계를 자동으로 수행하는 로봇을 생성합니다.

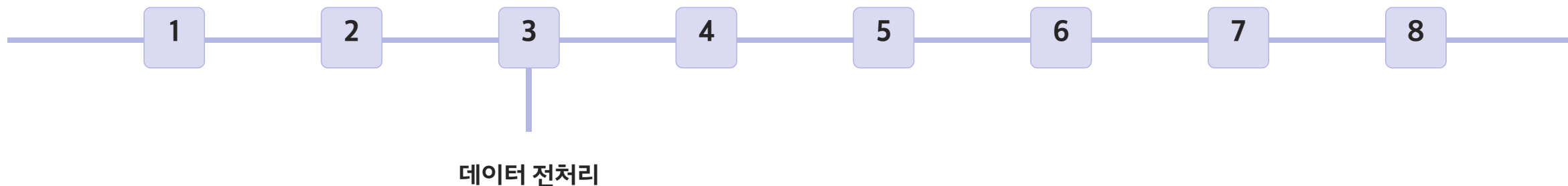
웹 스크래핑

웹 페이지에서 필요한 데이터를 자동으로 추출하는 기술입니다. 웹 스크래핑 라이브러리나 프레임워크를 사용하여 웹 페이지에 접속하고 필요한 데이터를 스크래핑합니다.

데이터 전처리

데이터 분석 프로세스

데이터 전처리



“수집된 데이터를 분석이 가능한 형태로 구조화하고 정제하는 단계”

“데이터의 품질을 유지하는 단계”

데이터 분석 프로세스

데이터 전처리

- 데이터 품질 판단 기준
 - 정확성 : 정확하고 가독성이 높은가?
 - 관련성 : 정의한 문제, 주제와 관련이 있는가?
 - 완전성 : 누락되거나 오류가 존재하지 않는가?
 - 적시성 : 분석결과에 영향을 줄 만큼의 시기 적절한 데이터인가?
 - 일관성 : 데이터의 형식(단위, 날짜, 숫자, 문자 등)은 일정한가?

데이터 분석 프로세스

데이터 전처리

- 데이터 전처리를 활용하여 분석 가능한 상태로 만든다.
- 전처리 과정은 전체 과정의 80%정도 시간이 들어 간다.
- 전처리를 할 때 데이터는 항상 문제가 있다고 생각하고 접근하는게 좋다.
- 데이터 무결성 확인
 - 데이터 타입이 일치하는가?
 - 데이터 단위가 일치하는가?
 - 중복, 누락, 이상값은 없는가?

데이터 분석 프로세스

데이터 전처리 사례

[예] 고양이와 개 사진 입력시, 고양이인지 개인지 맞추는 인공지능

Cats



Dogs



데이터 분석 프로세스

데이터 전처리

1

결측치 처리

결측치(누락된 값)를 확인하고 적절한 방법으로 처리합니다. 결측치를 제거하거나 대체하는 등의 방법을 사용할 수 있습니다.

2

이상치 처리

이상치(정상적인 범위를 벗어난 값)를 탐지하고 처리합니다. 이상치를 제거하거나 대체하는 등의 방법을 사용하여 데이터의 정확성과 신뢰성을 향상시킵니다.

3

데이터 타입 변환

데이터의 타입을 적절하게 변환합니다. 예를 들어, 문자열로 저장된 숫자 데이터를 숫자형으로 변환하거나, 날짜/시간 데이터를 적절한 형식으로 변환하는 작업을 수행합니다.

4

중복 데이터 처리

중복된 데이터를 확인하고 처리합니다. 중복된 데이터를 제거하거나 통합하는 등의 방법을 사용하여 데이터의 중복성을 제거합니다.

5

데이터 정규화

데이터를 일관된 형태로 조정합니다. 예를 들어, 텍스트 데이터의 대소문자를 통일하거나, 단위를 일관성 있게 조정하는 작업을 수행합니다.

6

피처 스케일링/정규화

피처(변수)들의 값 범위를 조정합니다. 일반적으로는 피처 스케일링(Feature Scaling) 또는 피처 정규화(Feature Normalization)를 수행하여 피처들 간에 상대적인 크기 차이를 줄이는 작업을 수행합니다.

7

피처 인코딩

범주형 데이터를 숫자형 또는 이진 형태로 변환합니다. 원-핫 인코딩, 레이블 인코딩 등의 방법을 사용하여 범주형 데이터를 처리합니다.

8

피처 생성

기존의 피처들을 기반으로 새로운 피처를 생성합니다. 예를 들어, 날짜 데이터에서 요일 정보를 추출하여 새로운 피처를 생성하는 작업을 수행할 수 있습니다.

9

피처 선택

분석에 필요한 피처를 선택합니다. 불필요한 피처를 제거하거나, 중요한 피처를 선택하는 작업을 수행합니다.

데이터 분석 프로세스

데이터 전처리 사례(공통)

결측치 처리 :

고객 만족도 조사 데이터에서 일부 응답자의 연락처 정보가 누락되어 있는 경우, 결측치를 확인하고 다른 응답자의 연락처 정보를 사용하여 누락된 값을 채워넣습니다.

이상치 처리 :

주식 거래 데이터에서 거래 금액이 음수인 이상치를 탐지하고 제거합니다. 음수 값은 잘못된 입력 또는 오류로 간주되므로, 이상치를 처리하여 분석 결과의 신뢰성을 높입니다.

데이터 타입 변환 :

날짜 정보가 문자열로 저장된 경우, 날짜/시간 형식으로 변환하여 분석에 활용합니다. 예를 들어, 판매 기록 데이터에서 "2021-07-15"와 같은 문자열을 날짜 형식으로 변환하여 월별 판매량을 분석할 수 있습니다.

중복 데이터 처리 :

고객 주문 데이터에서 중복된 주문을 확인하고 제거합니다. 동일한 고객이 중복으로 주문한 경우, 중복 데이터를 처리하여 정확한 고객 수와 판매량을 계산할 수 있습니다.

데이터 정규화:

텍스트 데이터에서 대소문자를 통일합니다. 예를 들어, 영화 리뷰 데이터에서 모든 텍스트를 소문자로 변환하여 단어 빈도수 분석이나 감성 분석에 활용할 수 있습니다.

피처 생성:

웹 로그 데이터에서 시간 정보를 기반으로 시간대 피처를 생성합니다. 예를 들어, 웹 사이트의 방문 기록에서 "오전"과 "오후"를 나타내는 시간대 피처를 생성하여 다른 시간대에 따른 사용자 행동 패턴을 분석할 수 있습니다.

데이터 통합:

고객 정보가 여러 개의 데이터베이스에 분산되어 있는 경우, 데이터베이스 조인을 사용하여 고객 정보를 통합합니다. 이를 통해 고객에 대한 종합적인 정보를 분석하고 개인화된 마케팅 또는 서비스를 제공할 수 있습니다.

데이터 분석 프로세스

데이터 전처리 사례(모델링)

피처 인코딩:

고객 분류 모델에서 성별 정보를 활용할 때, "남성"과 "여성"을 각각 0과 1로 인코딩하여 모델에 입력합니다. 이를 통해 성별에 따른 고객 특성의 영향력을 분석할 수 있습니다.

피처 선택:

고객 이탈 예측 모델에서 다양한 피처들 중 상관 관계 분석이나 변수 중요도를 활용하여 가장 영향력이 큰 피처들을 선택합니다. 이를 통해 모델의 복잡도를 줄이고 예측 성능을 향상시킬 수 있습니다.

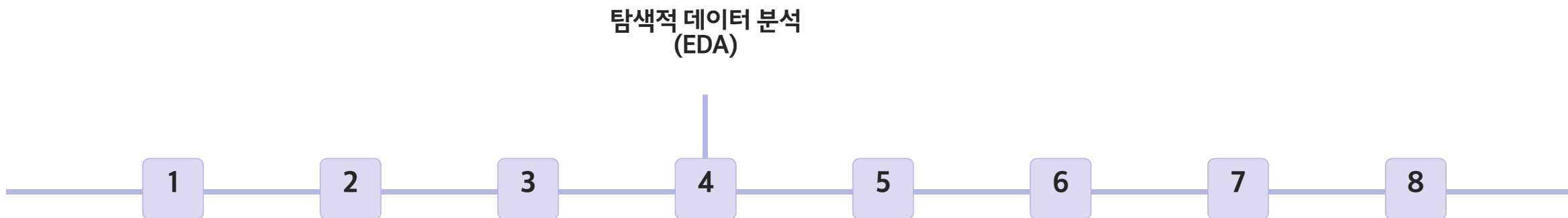
피처 스케일링/정규화:

주택 가격 예측 모델에서 면적과 가격 데이터를 사용할 때, 면적을 표준화 또는 정규화하여 가격과의 상대적인 크기 차이를 줄입니다. 이를 통해 모델이 면적과 가격 간의 관계를 더 잘 이해하고 예측할 수 있습니다.

탐색적 데이터 분석

데이터 분석 프로세스

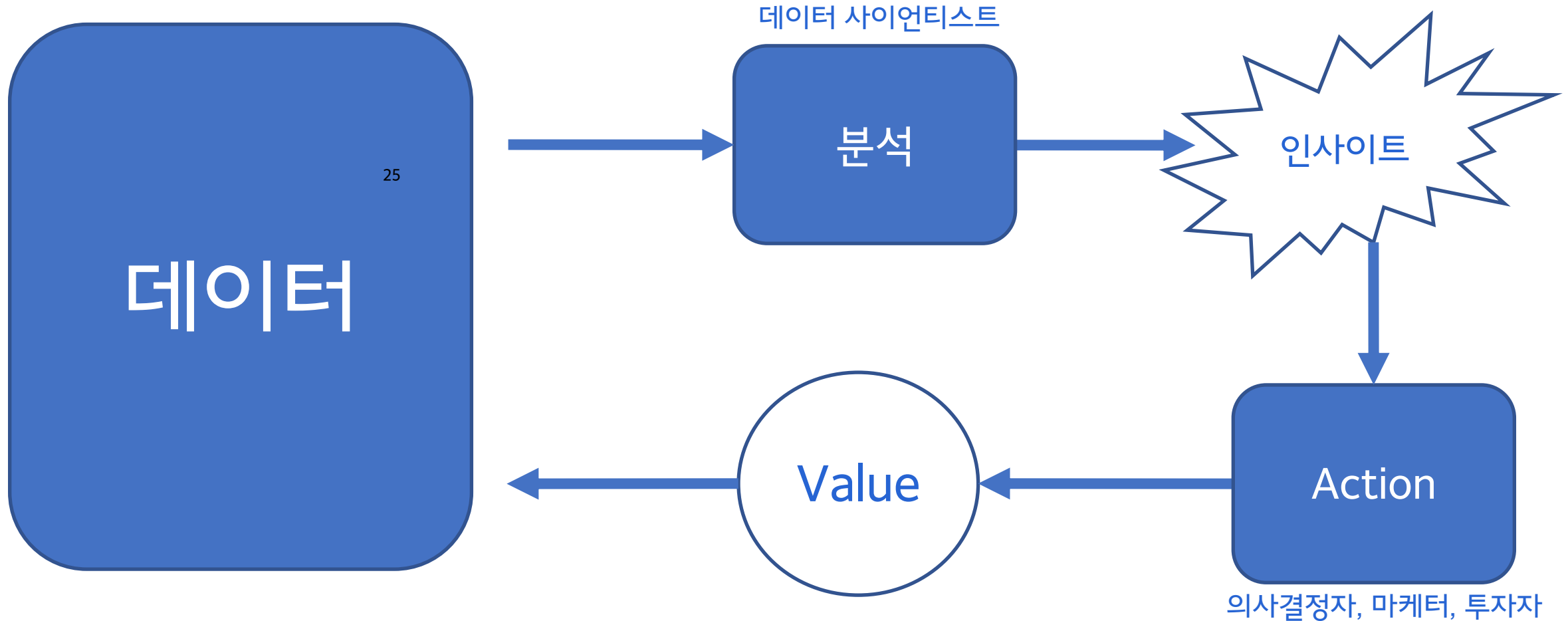
탐색적 데이터 분석 (EDA)



“데이터를 탐색하고 시각화하여 데이터의 패턴, 관계, 특징 등을 이해”
“데이터의 특성을 파악하고 잠재적인 문제나 가설을 발견”

데이터 분석 프로세스

탐색적 데이터 분석 (EDA)



데이터 분석 프로세스

탐색적 데이터 분석 (EDA)

1

데이터의 구조 파악

데이터의 크기, 변수(피처)의 개수, 데이터 형식 등을 확인하여 데이터의 전반적인 구조를 파악합니다.

2

기술 통계 분석

데이터의 통계적 특성을 요약하고 기술합니다. 평균, 중앙값, 표준편차, 최소/최대값 등을 계산하여 데이터의 분포와 중심 경향성을 파악합니다.

3

데이터 시각화

다양한 시각화 기법을 활용하여 데이터를 시각적으로 탐색합니다. 히스토그램, 박스 플롯, 산점도, 히트맵 등을 사용하여 데이터의 패턴과 상관 관계를 시각화합니다.

4

변수 간 상관 관계 분석

변수들 간의 상관 관계를 분석합니다. 상관 행렬, 히트맵, 산점도 행렬 등을 통해 변수들 간의 상관 관계를 시각화하고 이를 통해 변수들 사이의 의존성을 파악합니다.

5

데이터의 분포 탐색

데이터의 분포를 탐색하여 데이터가 정규 분포를 따르는지, 왜곡이 있는지 등을 평가합니다. 히스토그램, Q-Q 플롯, 밀도 추정 등을 사용하여 데이터의 분포를 시각화합니다.

6

클래스 불균형 처리

분류 작업에서 클래스 간의 불균형이 있는 경우, 클래스의 분포를 분석하고 적절한 처리 방법을 적용합니다. 언더샘플링, 오버샘플링, 가중치 부여 등의 기법을 활용하여 불균형을 해결합니다.

7

시계열 데이터 분석

시계열 데이터의 특성을 파악하고 추세, 계절성, 자기상관 등을 분석합니다. 시계열 그래프, 자기상관 플롯, 분해 등을 사용하여 시계열 데이터의 패턴을 이해합니다.

8

클러스터링

비슷한 특성을 가진 데이터들을 그룹화하는 클러스터링 분석을 수행합니다. K-평균 클러스터링, 계층적 클러스터링 등의 알고리즘을 활용하여 데이터를 클러스터로 분류합니다.

데이터 분석 프로세스

탐색적 데이터 분석 (EDA) 예시

데이터의 구조 파악:

- 데이터 세트의 변수(피처) 목록 확인
- 변수의 데이터 형식 확인
- 데이터의 행과 열의 개수 확인
- 데이터의 일부 샘플 확인

기술 통계 분석:

- 주택 가격 데이터에서 평균 가격, 중앙값, 표준편차, 최소/최대값 등을 계산하여 데이터의 분포와 중심 경향성 파악

데이터 시각화:

- 히스토그램을 사용하여 주택 가격 데이터의 분포 확인
- 박스 플롯을 사용하여 주택 가격의 이상치 여부 확인
- 산점도를 사용하여 주택 가격과 다른 변수(예: 면적) 사이의 관계 시각화

변수 간 상관 관계 분석:

- 상관 행렬을 생성하여 주택 가격과 다른 변수들 간의 상관 관계 확인
- 히트맵을 사용하여 상관 관계를 시각화하여 변수들 간의 의존성 파악

데이터의 분포 탐색:

- 주식 시장 데이터에서 일일 수익률의 분포를 탐색하여 정규 분포를 따르는지 확인
- Q-Q 플롯을 사용하여 데이터가 정규 분포에 근접하는지 시각화

클래스 불균형 처리:

- 고객 이탈 예측 모델에서 이탈 클래스의 비율이 매우 작은 경우, 언더샘플링 또는 오버샘플링을 사용하여 클래스 간의 균형 유지

시계열 데이터 분석:

- 판매 데이터에서 월별 판매량 추세 분석
- 자기상관 플롯을 사용하여 데이터의 자기상관 구조 파악

클러스터링:

- 고객 구매 데이터에서 고객 그룹을 클러스터링하여 세그먼트별 특성 파악

데이터 분석 프로세스


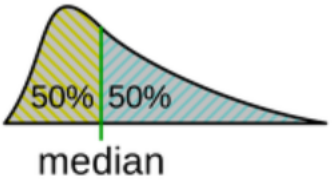
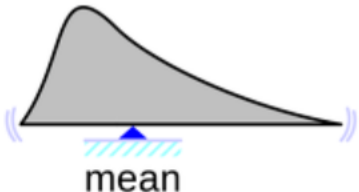
탐색적 데이터 분석 (EDA)

- 데이터에서 찾아내고자 하는 특성에 따라 2가지 분석 기법 존재
 - Central tendency(집중화 경향)
 - 데이터를 대표하는 값, 데이터가 어떤 값에 집중되어 있는지 분석
 - mean(평균), median(중앙값), 최빈값(mode), 최대값/최소값 범위(range)
 - Variation(분산도)
 - 데이터가 전반적으로 어떻게 퍼져 있는지를 분석
 - standard deviation(표준편차), quartile(사분위)

데이터 분석 프로세스

대표값

- 데이터 전체를 대표하는 통계 값
- 최빈값, 중앙값, 평균값, 최댓값, 최솟값 등이 있다.

	최빈치(mode)	중앙치(median)	산술평균(mean)
의미	<p>• 가장 빈번하게 나타나는 값</p> 	<p>• 자료를 크기 순으로 나열했을 때, 중앙에 위치하는 값</p> 	<p>• 자료를 모두 더해서 자료의 개수로 나눈 값</p> 
특징	<p>• 명목자료에서는 최빈치가 대푯값이다.</p>	<p>• 서열자료의 경우 평균을 사용할 수 없으므로 중앙치를 사용한다.</p>	<p>• 일부 극단적인 값들에 크게 영향을 받는다. • 수학적 연산에 의해 계산되므로 수리적인 조작이 용이하다.</p>
예	<p>유행하는 가방 인기 투표</p>	<p>학교 석차 100명 중 50 등</p>	<p>년간 평균 강우량 기말 고사 평균 점수</p>

데이터 분석 프로세스

분산도

- 데이터의 분포를 설명하는 통계 값
 - Range(범위)
 - Quatile deviation(사분 편차)
 - Variance(분산)
 - Standard deviation(표준편차)

데이터 분석 프로세스

Range(범위)

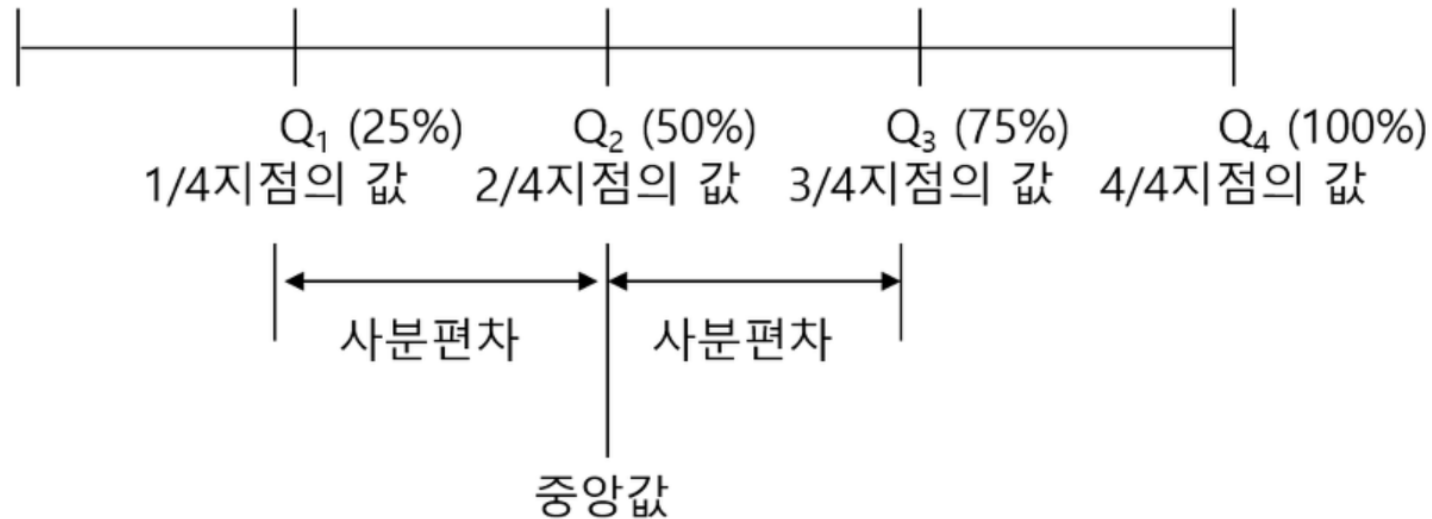
- 데이터 중 가장 큰 값과 가장 작은 값의 차이
- 최소값, 최대값 이라는 극값을 의미하는 통계치여서 데이터의 분포 분석 불가



데이터 분석 프로세스

Range(범위)

- 데이터를 크기순으로 정렬 하여 $\frac{1}{4}$, $\frac{3}{4}$ 지점의 값 차이의 반
- Quatile : $\frac{1}{4}$



데이터 분석 프로세스

탐색적 데이터 분석 (EDA)

- 분석이 막여할 때에는 아래의 내용을 함께 고민하고 정리해야 한다.

- 왜 이런 일이 발생했을까?
- 앞으로 무슨일이 일어 날 것인가?
- 우리가 무엇을 해야할까?

- 숫자에만 집중하지 말고 전체적인 맥락을 잘 파악해야 한다.

- 분석 툴

마케팅	통계 분석	BI	프로그래밍
Google Analytics	엑셀(Excel)	Power BI	Python
Amplitude	R	Tableau	MATLAB
	SAS		SQL
	SPSS		R

데이터 시각화

데이터 분석 프로세스

데이터 시각화

- 시각화된 자료는 결국 조직의 전략
- 비즈니스 의사결정을 위한 핵심 자료
- 사람이 이해할 수 있는 형태로 시각화 하는 단계
- 주의 사항
 - 데이터의 왜곡은 없는지?
 - 분석한 주제와 목적에 맞는 시각화 방식인지?

Revenue and Customer Overview - Q1 2016

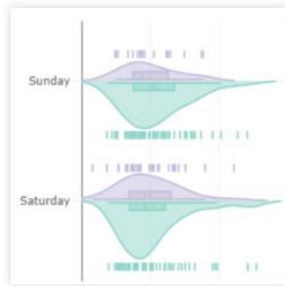


데이터 분석 프로세스

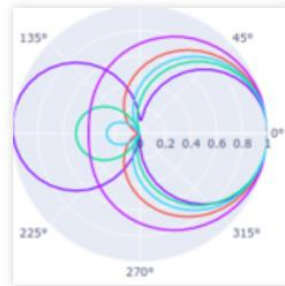
데이터 시각화

Fundamentals

[More Fundamentals »](#)



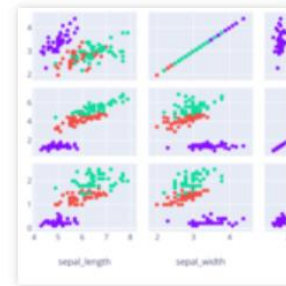
The Figure Data Structure



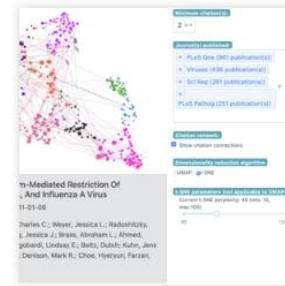
Creating and Updating Figures



Displaying Figures



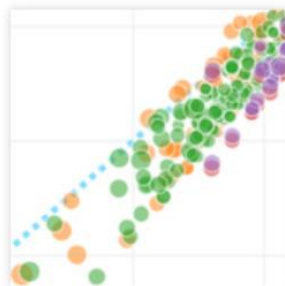
Plotly Express



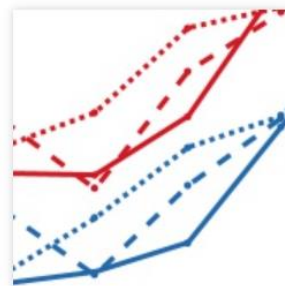
Analytical Apps with Dash

Basic Charts

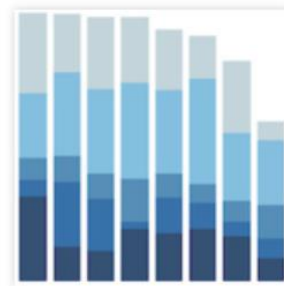
[More Basic Charts »](#)



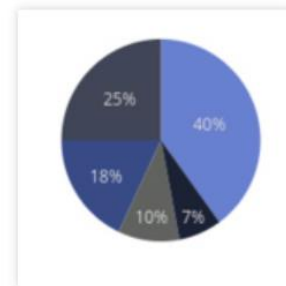
Scatter Plots



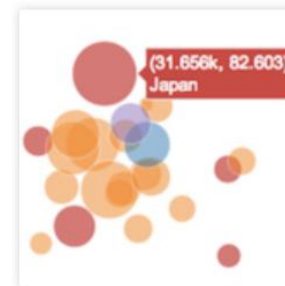
Line Charts



Bar Charts



Pie Charts



Bubble Charts

데이터 분석 프로세스

데이터 시각화

- 시각화는 데이터를 이해하기 쉬운 형태로 변환.
- 의사결정을 위한 스토리텔링
- 분석은 기술적으로 하더라도 자료에는 누구나 이해할 수 있는 언어를 활용
- 결과 자료의 종류
 - 의사결정 지원 : 필수적이고 간결한 정보를 포함 (간결하고 명료한 쉬운 차트 활용)
 - 실무자 지원 : 세분한 정보와 기술적인 차트와 그래프로 작성

데이터 분석 프로세스

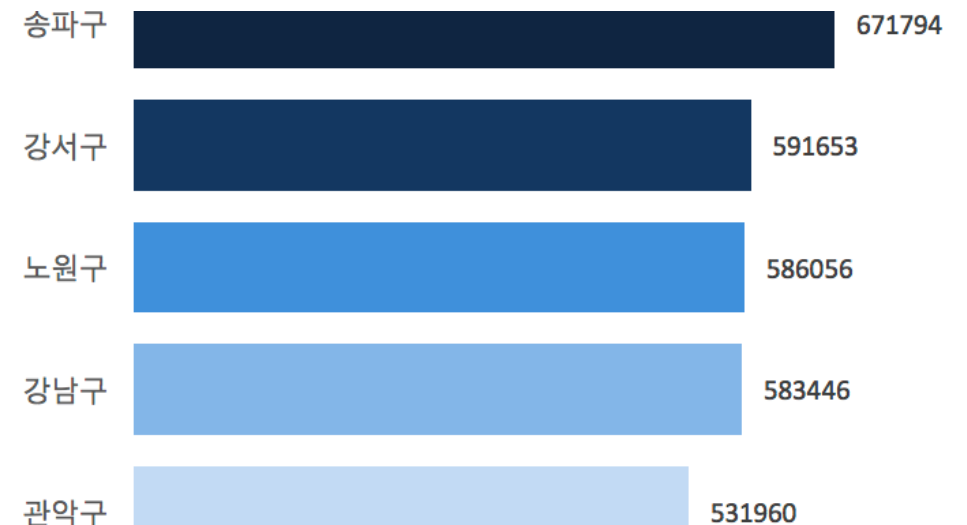
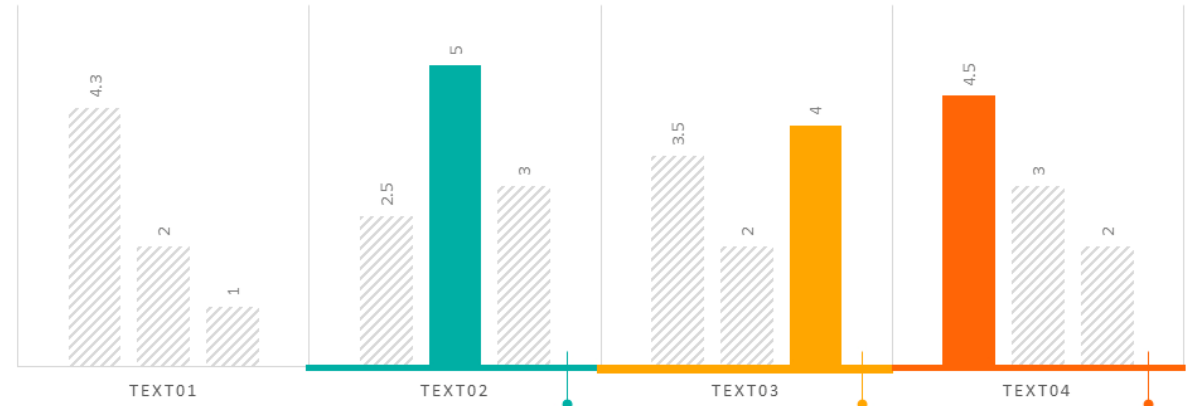
데이터 시각화

- 종류
 - 비교 분석 : 막대차트 / 라인차트 등
 - 구성 분석 : 파이차트 등
 - 관계 분석 : 스캐터 차트 등
 - 분포 분석 : 히스토그램 / 박스플롯 등

데이터 분석 프로세스

막대그래프

- 여러 항목 간 비교가 수월하며 이해하기 쉽다.
- 주의 사항
 - 중복된 내용은 피하고
그래프에서 강조할 부분은 색상을
통해 강조하자.
 - 명도 부분으로 데이터의 정도를
표현해 줄 수 있다.
 - 축의 범위로 데이터 해석의 과장,
축소 되지 않게 주의



데이터 분석 프로세스

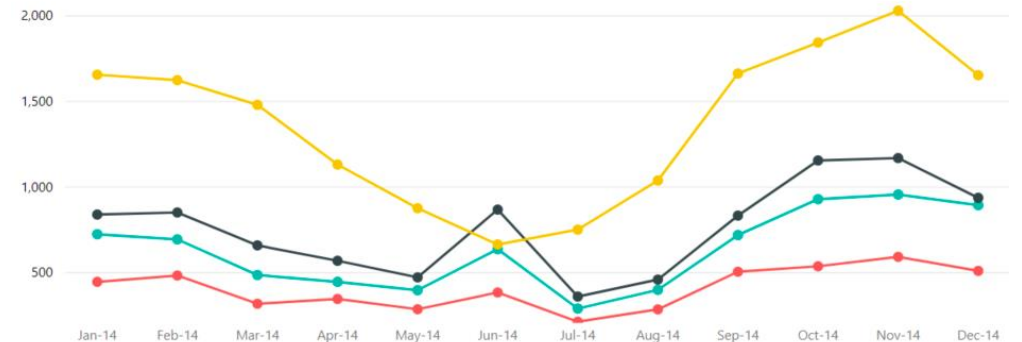
라인그래프

- 시간에 따른 데이터의 경향을 파악하기 용이하다.
- 항목간 비교가 쉬우며 편화를 파악하기 용이
- 주의사항
 - 여러 항목을 비교할 때에는 색상 대비를 확실히 하여 구분해주자.
 - 너무 많은 라인을 넣지 말 것



Total Units by Month and Manufacturer

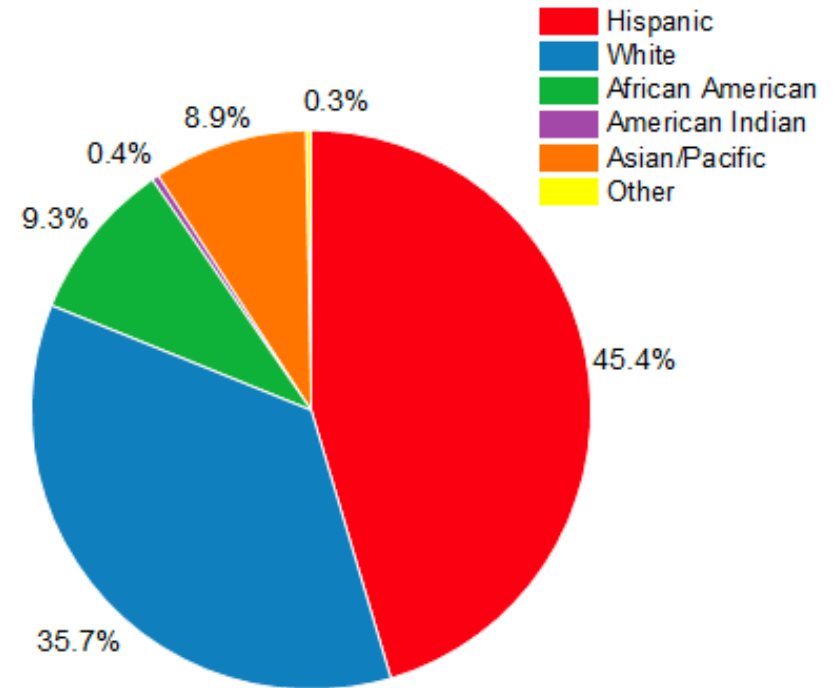
Manufacturer Aliqui Natura Pirum VanArsdel



데이터 분석 프로세스

파이차트

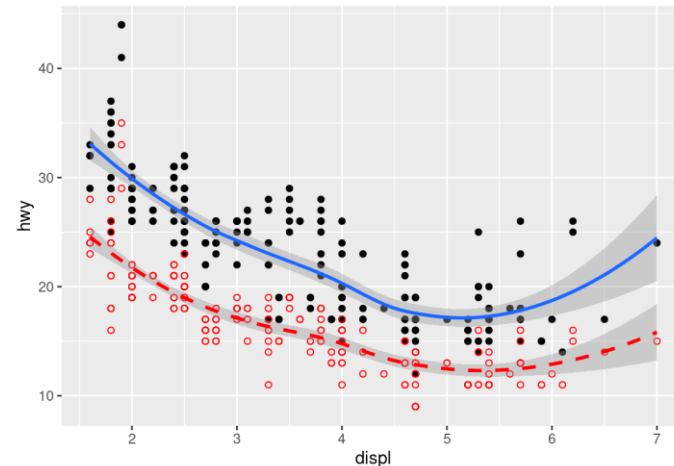
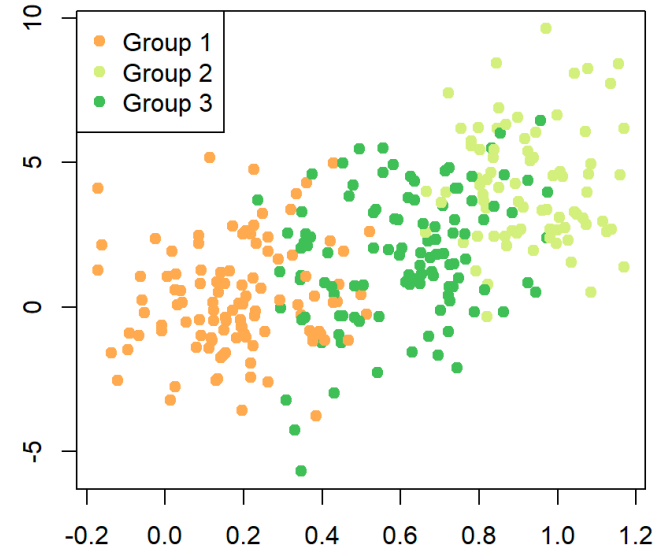
- 여러 항목들의 상대적인 비율을 이해하기 쉽다.
- 여러 개의 변수를 활용할 수 없다.
- 전체 데이터의 크기를 파악할 수 없다.
- 주의 사항
 - 5개의 넘는 항목을 비교하지는 않는다.
 - 만약 넘는다면 그외 or 기타로 표현한다.



데이터 분석 프로세스

스케터차트

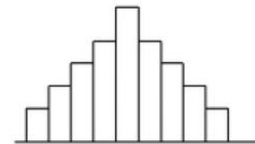
- 두 변수간의 관계를 파악하기 용이
- 데이터의 개수가 많은 경우 시각화하기 어렵다.
- 3개 이상의 변수를 시각화 할 수 없다.
- 주의 사항
 - 추세선을 활용하는게 좋다고 판단되면 추세선도 활용하자.
 - 색상으로 구분하여 데이터를 구분한다.



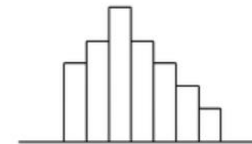
데이터 분석 프로세스

히스토그램

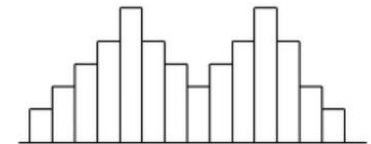
- 데이터의 분포를 파악하기 용이
- 대용량 데이터를 확인하기 용이
- 이상치 여부 판별하기 용이
- 정확한 값을 알기 어렵다.
- 다양한 변수를 한번에 확인하기 어렵다.
- 주의 사항
 - 시각적으로 보기 좋은 구간 수를 활용하자.
 - 추세선을 그리는게 좋다고 판단되면 추세선도 잘 활용하자.



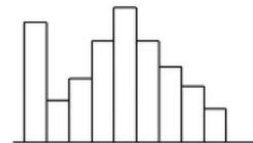
정규분포



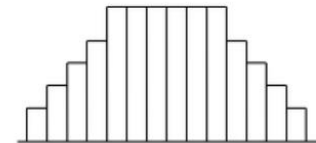
특정한 값보다 작은 값을
모집단(표본)으로부터
제거한 경우



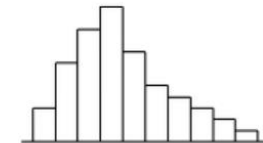
두 모집단이 혼합된 경우



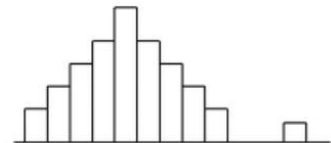
한계값에서 벗어난
값을 모두 한계값
으로 대신한 경우



여러개의 모집단이
혼합된 경우



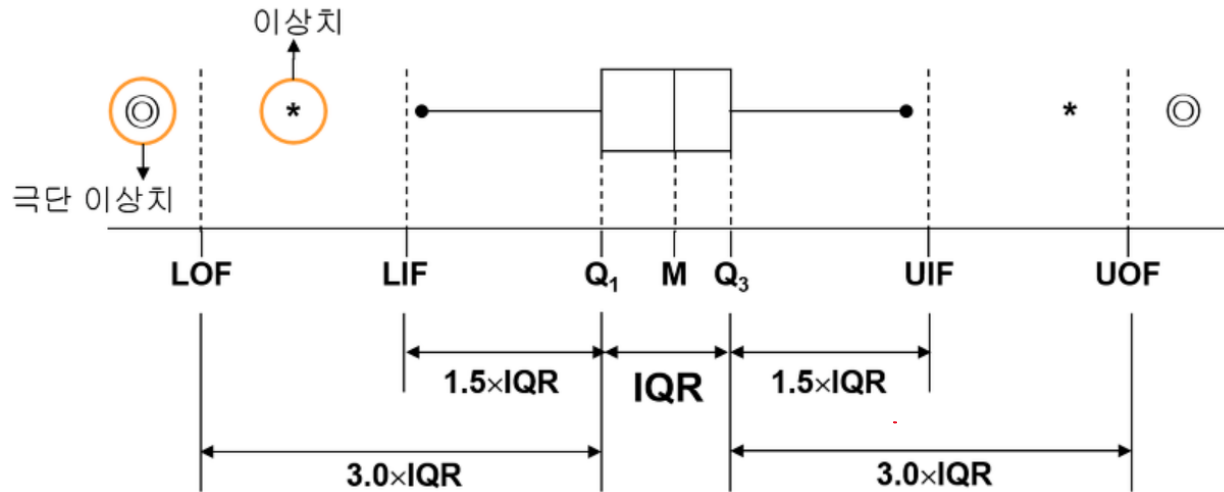
비대칭 분포



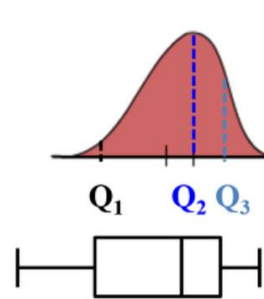
이상값이 존재한 경우

데이터 분석 프로세스

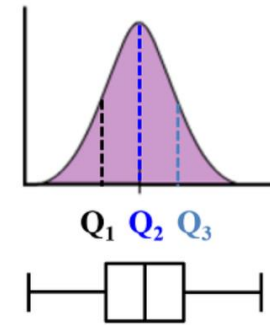
박스 플롯



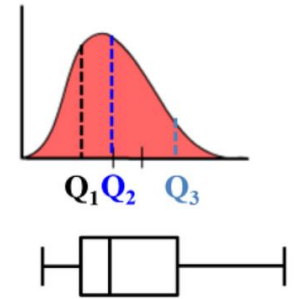
Negatively-Skewed



Symmetrical



Positively-Skewed



- 사분위 값을 차트를 이용해 시각화 하는 차트
- Box whisker(수염) 을 벗어난 값을 이상치라고 부름
- 이상치는 통계 분석 전 처리 방법을 고민해야 함
- 데이터의 분포를 쉽게 파악 가능

데이터 분석 적용 사례

적용 사례

#1 올빼미버스



<http://bus.go.kr/nBusMain.jsp>

서울시에서 운영하는 심야 버스 서비스
자정부터 오전 5시까지 운행

심야 버스 노선

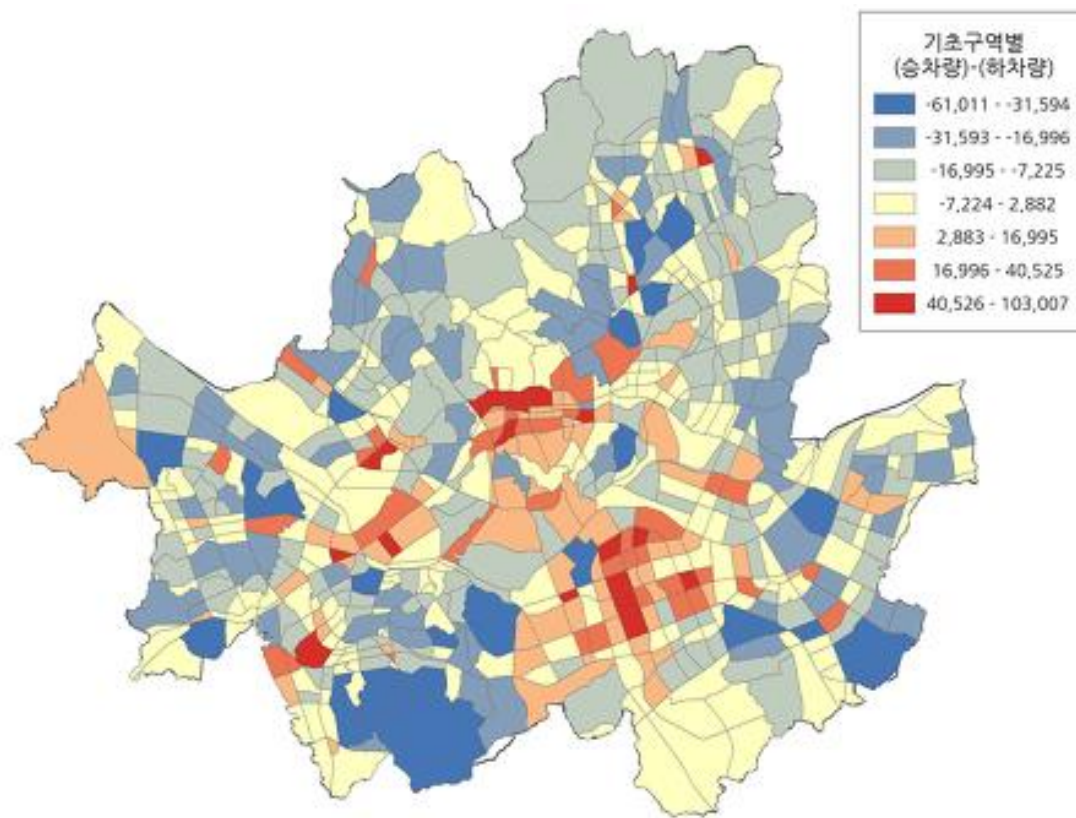
어떻게 정할 수 있을까요?

적용 사례

#1 올빼미버스



자정부터 오전 5시



자정부터 오

#1 올빼미버스

노선별 시뮬레이션으로 최적화



적용 사례

#2 경기도 CCTV 설치

경기도 31개 시군, 매년 CCTV 설치

위치 선정에 대한 객관적인 기준 부재

대부분 민원에 근거하여 특정 지역에 집중적으로 설치됨
최적화된 설치 위치는? 이를 뒷받침할 객관적인 데이터는?

적용 사례

#2 경기도 CCTV 설치

구분	데이터명	출처	사이즈(개수)
공공	CCTV 설치현황	지자체	6,578
	CCTV 설치예정현황		620
	CCTV 설치요청 민원데이터		914
	50 가로등초·보안등(방범등)설치현황	국립재난안전연구원	121,571
	개발정보	지자체	41
	어린이집현황	경기도 교육청	12,591
	학교기본현황(유치원, 초·중·고)		4,455
	용도지역지구 도시지역	국토교통부	106,262
	연속지적도	경기도	4,812,860
	하천용도지역	하천지리정보시스템	3,893
	공원	경기도청	6,288
	어린이놀이시설 현황	놀이시설안전관리시스템	14,199
	상가업소데이터	소상공인시장진흥공단	135,361
	가중치 항목의 중요도	지자체	77

적용 사례

#2 경기도 CCTV 설치

민간	경기도 내국인 유동인구	민간업체	4,129,517
	경기도 외국인 유동인구		4,129,517
	도로링크 ⁵¹		1,946,621
	아파트 단지 정보		24,713
	도로명주소 건물데이터		1,579,235
	인구정보		71,740
	블록 영역		83,392
	가구정보		71,110
	거처유형별 주택정보		785,038
	표준지 공시지가		59,843
	버스정류장		146,452
	건물유형별 가구 비율		533,666

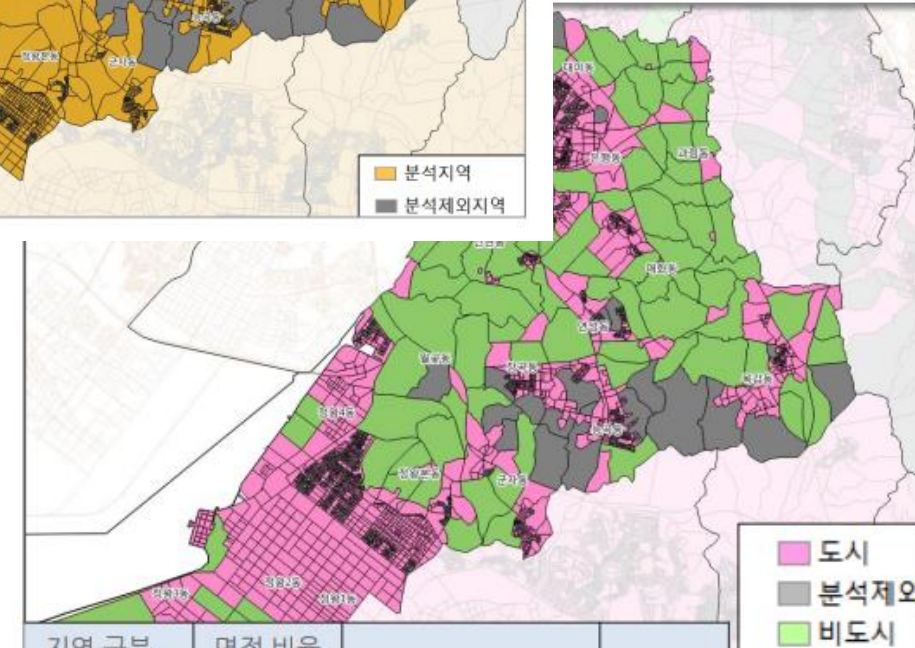
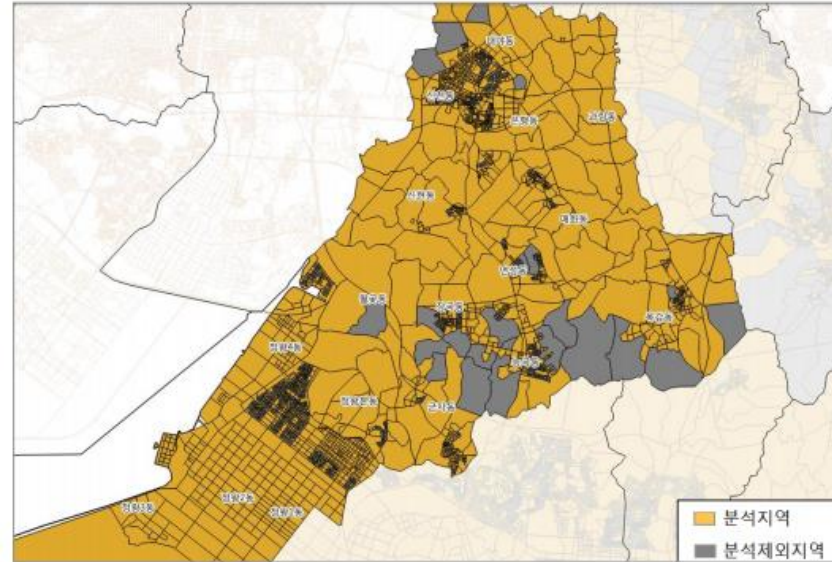
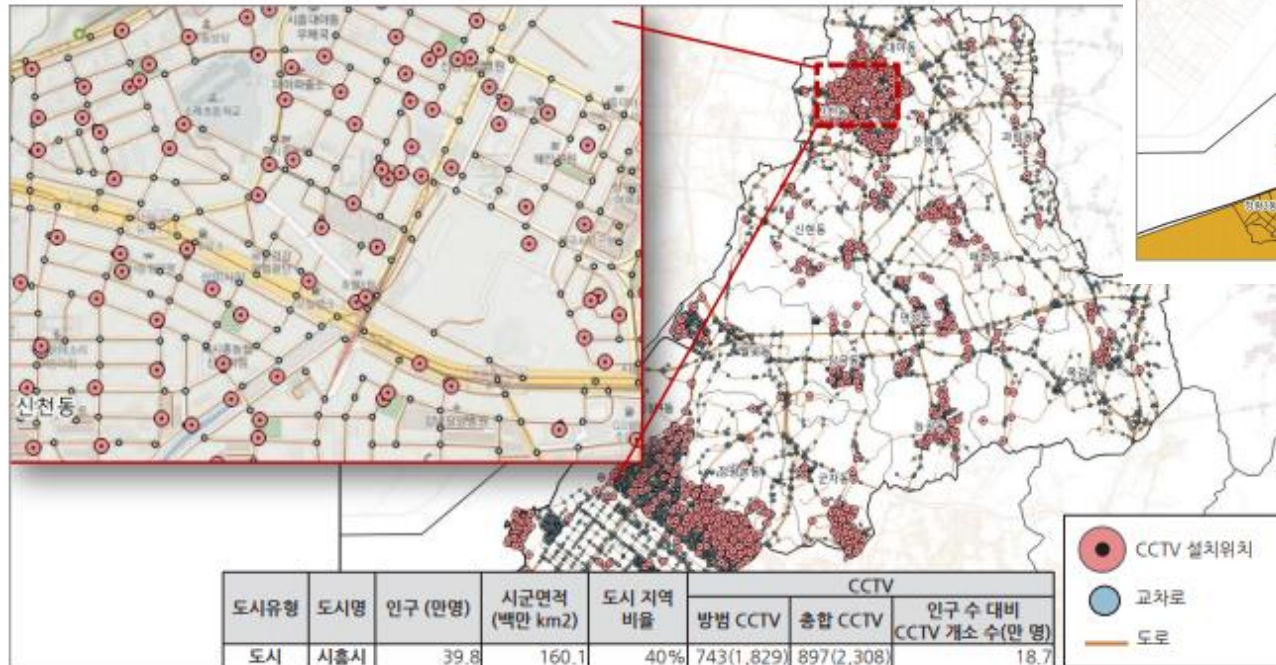
적용 사례

#2 경기도 CCTV 설치



시흥시 CCTV 설치지역 및 교차로 현황

52



적용 사례

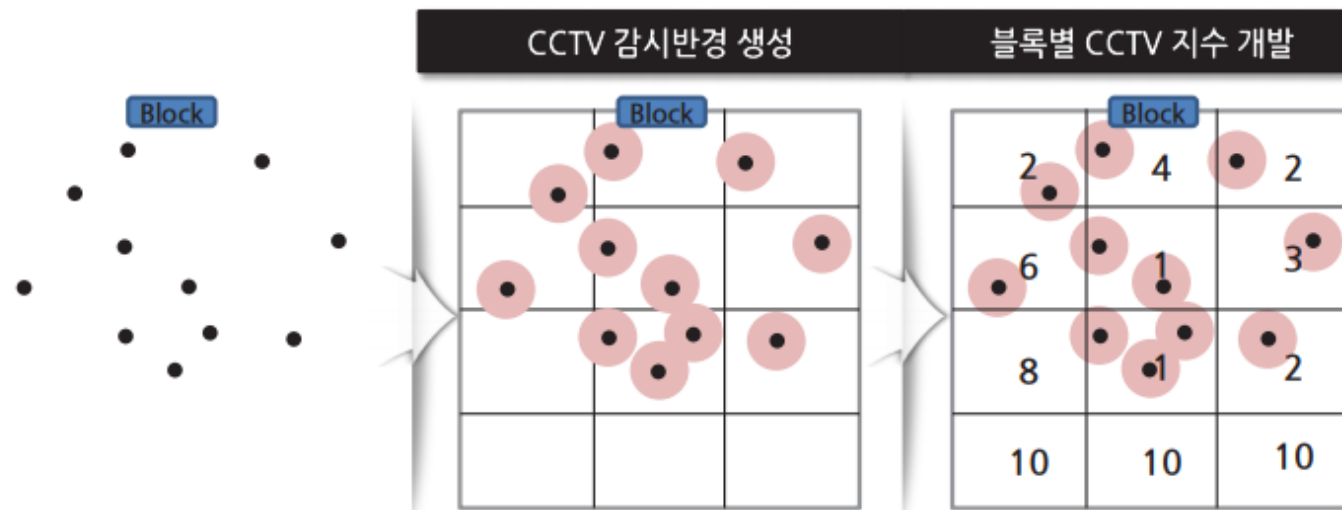
#2 경기도 CCTV 설치

범죄 발생 가능성이 높은 범죄 예측 지수

기존에 설치된 CCTV의 감시 취약성을 나타내는 CCTV취약지수

현업 부서들의 CCTV 설치 선호위치(학교 근처 등)에 대한 가중치를 반영

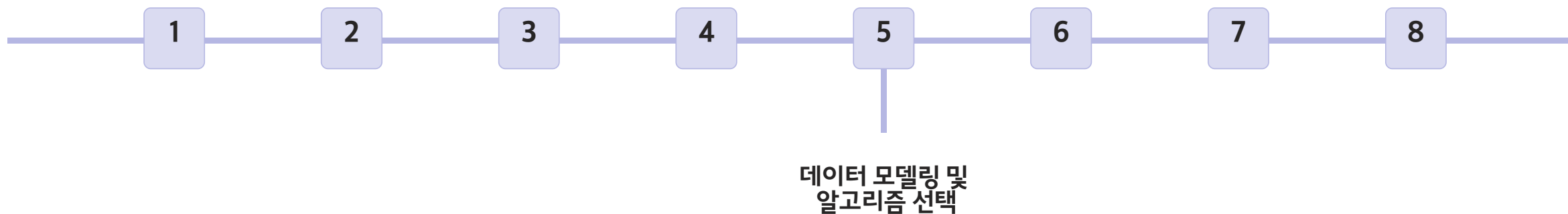
⁵³ 이를 점수화하여 최종지수가 높은 순으로 “CCTV우선 설치 지역”을 도출



데이터 모델링 및 알고리즘 선택

데이터 분석 프로세스

데이터 모델링 및 알고리즘 선택



“ 분석에 적합한 **모델이나 알고리즘을 선택하고 구축** ”
“ Ex) 회귀 분석, 분류, 군집화, 시계열 분석 등 ”

데이터 분석 프로세스

머신러닝 활용 사례

- 머신러닝 지도학습(Supervised Learning)의 활용 사례
 - 스팸 메일 필터링: 이메일 데이터를 기반으로 스팸과 정상 메일을 분류하는 모델을 개발
 - 주택 가격 예측: 주택의 특징과 가격 데이터를 활용하여 주택 가격을 예측하는 모델을 개발
 - 의료 진단: 환자의 증상과 진단 결과를 기반으로 질병을 예측하거나 진단하는 모델을 개발
- 머신러닝 비지도학습(Unsupervised Learning)의 활용 사례
 - 군집화(Clustering): 고객 세그멘테이션, 이미지 분류 등에서 유사한 특징을 가진 데이터들을 군집화하여 분류
 - 이상치 탐지(Anomaly Detection): 네트워크 트래픽 분석, 금융 거래 모니터링 등에서 이상한 행동을 감지하여 탐지
 - 차원 축소(Dimensionality Reduction): 고차원 데이터의 특징을 보존하면서 저차원으로 축소하여 시각화나 계산 효율성 향상

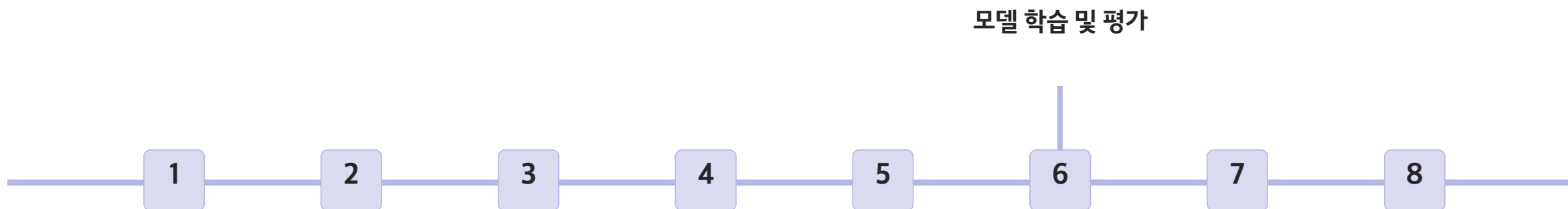
데이터 분석 프로세스

딥러닝 & 시계열 활용 사례

- 딥러닝의 활용 사례
 - 이미지 분류: 이미지 데이터를 분석하여 각각의 카테고리로 분류하는 모델을 개발
 - 자연어 처리: 텍스트 데이터를 처리하여 문장 분류, 감정 분석, 기계 번역 등을 수행하는 모델을 개발
 - 음성 인식: 음성 데이터를 처리하여 음성 명령 인식, 음성 텍스트 변환 등을 수행하는 모델을 개발
- 시계열 데이터 분석의 활용 사례
 - 주식 가격 예측: 과거 주식 가격 데이터를 분석하여 향후 주식 가격을 예측하는 모델을 개발
 - 수요 예측: 과거 판매 기록을 기반으로 제품 수요를 예측하여 재고 관리나 생산 계획을 수립하는 모델을 개발
 - 기상 데이터 분석: 기상 관측 데이터를 활용하여 기상 패턴 예측, 날씨 변동 예측 등을 수행하는 모델을 개발

데이터 분석 프로세스

모델 학습 및 평가



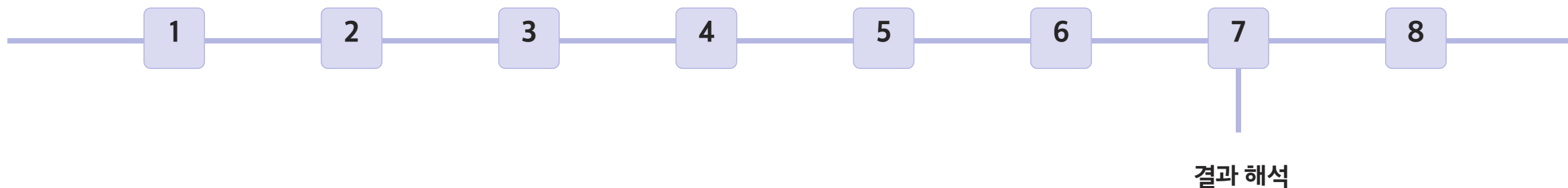
“선택한 모델에 데이터를 학습시키고 **모델의 성능을 평가**”

“학습 데이터와 평가 데이터를 분리하여 **과적합을 방지**”

“**모델의 일반화** 성능을 평가”

데이터 분석 프로세스

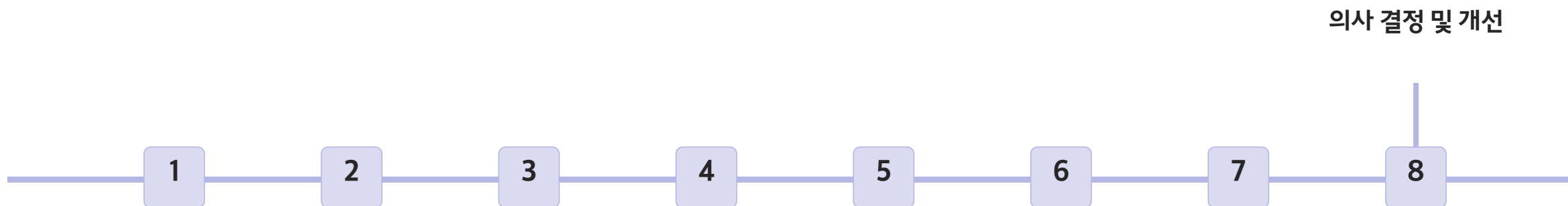
결과 해석



“모델의 결과를 해석하고 이해”
“문제에 대한 통찰력을 얻고 의사 결정에 활용”

데이터 분석 프로세스

의사 결정 및 개선



“분석 결과를 바탕으로 의사 결정을 내리고 필요에 따라 **개선 방안을 제시**”