# 0. 강사 및 강의 소개

## 강사 소개

### 기본 사항

김민준 | 1993년 11월 22일 | 대구 출생 | 이대역 근처 거주

서울대학교 자유전공학부 오랜 기간 동안 휴학 중 | 사회복무요원 복무 중

010-5511-4898 문자 전화 카톡 환영 | lakiu@naver.com

### 이력 및 수상

대구 대륜고등학교 졸업 | 수리 가형 + 사회탐구 응시

EBS 공부의왕도 출연 | 청소년을 위한 만만한 경제학 저술

TESAT 최우수상 | 매경TEST 대상 | 경제학원론(조순 外) 10판 개정작업 참여

삼성전자 C-Lab 인턴십 - VR 관련 기술 프로젝트 참여

인디게임 Bytes of Hexagon 기획 & 개발

색 조합 추천 딥러닝 "Daltonism" 개발

영어교육 웹 솔루션 스타트업 "Flowenglish" CTO

용산구 전공연구반 (2013년~) | 성심여고 방과후학교 (2012년~)

세종시 전공연구반 (2017년~)

### 관심과 취향

딥러닝, 데이터과학, 게임제작, 경제학, 창업, 특이점, 인공지능, VR, 3D 프린터

책 모으기, 게임하기, 오늘 할 일 미루기, 저축 안 하고 이거저거 막 지르기

호: 고래, 모찌, 부대찌개 | 불호: 카카오프렌즈

## 강의 목표

데이터과학과 머신러닝을 이해하고, Azure 클라우드 서비스를 통해 실습한다.

경영 혁신 및 사회적 가치 창출을 위한 데이터와 인공지능의 활용을 분석한다.

데이터 분석에 필요한 통계학 이론과 개념을 이해하고 실습을 통해 적용한다.

## 강의 구성

### 통계학 & 데이터과학 & 머신러닝

기본적인 통계 이론과 개념을 공부한 뒤, 이를 데이터에 적용해보는 것을 통해 데이터에서 의미와 가치를 만드는 과정을 실습하는 것을 목표로 한다. 나아가 이러한 데이터를 딥러닝에 활용하는 방안에 대해 탐구한다.

통계학은 고등학교 수준의 내용부터 대학교 1학년 과정의 내용을 다루며, 이는 『그림으로 설명하는 개념 쏙쏙 통계학』과 edX의 "DAT101x: Data Science Orientation" 강의를 참조하였다.

데이터마이닝과 머신러닝은 Azure 클라우드 서비스를 이용할 것이며, 어려운 코드를 작성하지 않고 배운 개념과 기법을 빠르고 강력하게 구현할 수 있다.

### 기업과 사회에서의 활용

데이터와 통계를 활용한 사례를 살펴보는 것을 통해 상황에 따른 적절한 데이터 활용 시나리오를 생각하고 설계한다. 다양한 실제 사례와 실습을 위한 모의 데이터를 함께 사용할 것이며, 수업 중 배운 내용을 직접 적용하는 것을 통해 특성과 장단점, 한계를 알 수 있도록 하였다

수강생들은 강의 진행 기간 동안 자신만의 데이터 활용 시나리오를 기획하고, 데이터를 수집한 뒤 분석 결과를 발표해야 한다.

## 일정

데이터, 통계학, 데이터 과학 / 머신러닝

데이터를 활용하는 방법

---

통계학 기초 이론

데이터의 수집과 가공

---

정규분포 | 통계적 검정

회귀분석

---

확률 | 베이즈 정리

분류 | ROC Curve

---

군집분석

데이터 시각화

---

머신러닝 서비스 배포

딥러닝의 구조와 학습과정

---

데이터 활용 시나리오 프로젝트 발표

## 유의사항

### 강의의 목적 및 한계

본 강의는 짧은 일정으로 인해, 데이터과학 및 통계학의 이론적 측면을 세세하게 탐구하는 대신 실습과 활용에 필요한 부분만을 다룰 것이다. 이에 대해 추가적으로 심화학습이 필요한 부분은 앞서 언급한 『그림으로 설명하는 개념 쏙쏙 통계학』과 edX의 "DAT101x: Data Science Orientation" 강의와 함께 『Head First Statistics』와 edX의 다른 강의들을 통해 보충할 수 있을 것이다.

### 사례

최대한 많은 분야에 적용될 수 있고, 주변에서 접할 수 있는 사례를 활용할 것이나, 만약 학생이 관심있는 사례나 분야가 있다면 반영할 것이다.

### 실습

매 강의 이후 집에서 Office Mix를 통해 Lab을 수행하고 퀴즈를 풀 것이다. 핵심 개념을 복습하고, 엑셀 실습 영상을 따라한 뒤 그 결과를 묻는 퀴즈를 해결한다.

### 데이터 분석 시나리오

관심있는 분야에서 데이터를 활용하고 분석하는 시나리오를 기획하고, 실제로 데이터를 수집한 뒤 강의에서 다룬 기법을 적용한 뒤 발표한다. **시나리오 발표와 Lab을 모두 수행해야만 강의를 수료할 수 있다.**

### Azure Machine Learning Studio

Microsoft가 제공하는 클라우드 기반 머신러닝 서비스로, 데이터마이닝 및 머신러닝을 코드를 작성하거나 복잡한 설치과정 없이 간단하게 웹에서 무료로 수행할 수 있다.

https://studio.azureml.net/

### 오픈 채팅방

https://open.kakao.com/o/g5KKEZv

강의자료 및 공지사항, 질문답변 등

# 1. 데이터, 통계학, 데이터과학

#데이터 #데이터의종류 #빅데이터 #3V
#통계학 #데이터과학 #데이터활용시나리오

## 데이터란?

Data라는 단어는 라틴어 Datum의 복수형인 Data에서 유래했으며, 그 뜻은 재료, 자료, 논거 등이 있다. 우리 말로 하자면 자료라고 할 수 있겠으나 그 미묘한 어감이 사라지는 관계로 그냥 다들 데이터라고 한다. 발음은 사실 데이ㅌㅏ 정도가 맞지만 이미 표준어로 굳어졌다. 최근 빅데이터 열풍을 타고 마케팅에서 마법의 단어로 사용되고 있으며, 뭔가 체계적이고, 객관적이고, 신뢰할 수 있고, 비싼 값을 받아도 될 거 같은 이미지를 형성하기 위해 쓰인다. 때로는 스마트폰, 인터넷 요금제에서 데이터 사용량 내지는 제공량을 데이터라고도 한다.

데이터를 정확하게 정의하기는 어렵지만, 어떤 정보를 가진 값, 숫자, 문자, 영상, 그림, 소리 등을 모두 데이터라고 할 수 있다. 21세기에는 이 모든 것을 결국 컴퓨터 파일로 저장하고 처리하기 때문에, 데이터는 곧 파일이고 파일이 곧 데이터가 된다고 봐도 큰 문제는 없을 것이다.

## 데이터의 종류

데이터는 다양한 기준을 통해 종류를 구분할 수 있으며, 다른 종류의 데이터는 다른 방식으로 처리되어야 한다. 따라서 이를 파악하는 것은 매우 중요하다.

하나의 데이터는 다양한 기준들에 대해 여러 종류로 구분될 수 있다. 가령 나이는 양적 | 이산 | 정형 데이터에 속한다.

### 질적 데이터 VS 양적 데이터

질적 데이터는 남자, 여자와 같은 범주(Category) 자료와 등급, 순위와 같은 순서(Ordinal) 자료가 포함된다. 이들에게 임의의 숫자를 부여할 수도 있으나 (남자는 0, 여자는 1 등) 이 숫자를 평균을 내거나 더하는 등의 연산은 무의미하다.

양적 데이터에는 온도와 같이 수치의 간격(Interval)이 의미를 가지는 자료와 간격과 함께 비율 (Ratio)도 의미를 가지는 무게, 시간 등의 자료가 있다.

### 연속 데이터 VS 이산 데이터

길이, 면적, 시간 등 연속성이 있는 자료를 연속 데이터라 하고, 주사위 눈, 등급, 나이 등 연속성이 없는 자료를 이산 데이터라고 한다. 간단히 말해 실수로 표현되어야 하는 것은 연속, 정수로 표현되는 것은 이산이라고 볼 수 있다.

### 정형 데이터 VS 비정형 데이터

형식이 정해진 데이터, 즉 신체검사 기록, 연락처, 주문 기록 등 표로 나타낼 수 있는 데이터를 정형 데이터라고 한다. 전통적으로 데이터의 분석과 수집 모두 정형 데이터 위주로 이루어졌으나, 최근에는 이미지, 비디오 등 다양한 비정형 데이터의 중요성이 높아지고 있다.

## 빅데이터

최근 자주 사용되는 빅데이터라는 단어는 그만큼 자주 오용되기도 한다. 대개의 경우 마케팅 용어로 사용하는 빅데이터는 빅데이터가 아닌 경우가 많은데, 이는 빅데이터를 정의하는 3V를 만족하지 못하기 때문이다.

### Volume

빅데이터는 이름에서 알 수 있듯 양이 많다. 이 양의 기준은 절대적인 것이 아니라, 데이터를 분석하고 활용하는 목적에 따라 달라진다.

### Velocity

데이터 입출력에 필요한 속도를 의미한다. 양이 많더라도 기존의 기술로 필요한 입출력을 감당할 수 있다면 빅데이터로 보기 힘들다.

### Variety

다루는 데이터의 종류가 다양함을 의미한다. 비정형 데이터의 처리가 빅데이터의 핵심인 이유기도 하다.

## 통계학

통계학은 데이터를 관찰하고 처리하는 것을 연구하는 학문이라고 할 수 있다. 데이터의 수집, 가공, 분석, 처리, 의미 추출 등이 통계학의 범위에 포함된다. 그 특성상 굉장히 실증적인 성향을 보이며, 동시에 이론적 토대도 굳건한 학문이다. 다시 말해 어렵다. 문제는 이 통계학이 쓰이지 않는 곳이 없다는 것이다.

현 시대는 정보가 모든 것을 지배하고 있다. 모든 기업은 나름의 방식으로 정보를 수집하고 가공하여 의사결정을 한다는 점에서 IT 기업이다. 개개인의 의사결정에도 정보가 중요한 것은 말할 것도 없다. 이 정보를 처리하는 방법이 바로 통계라는 점에서 통계학은 이 시대를 살아가는 모든 이들의 필수 학문이 되었다.

순수 인문대를 제외한 다른 모든 전공에는 통계학 과목이 존재한다. 대학 문턱을 밟는 순간 여러분은 통계학의 세계에 강제로 입장하게 되는 셈이다. 순수 인문대의 현실을 생각해보면 결국 먹고 살려면 통계를 해야한다는 결론이 나온다.

한편 통계는 이 정보를 정확하게 처리하여 올바른 선택을 내리는 것이 목적이지만, 반대로 정보를 왜곡하여 잘못된 선택을 유도하도록 악용되기도 한다. 이러한 사기에 당하지 않기 위해서도 통계학을 배울 필요가 있다.

## 데이터과학

데이터과학은 근래에 급부상한 단어로, 통일된 정의는 없으나 프로그래밍, 통계학, 수학, 기계학습, 그리고 각 분야에 대한 지식(Domain knowledge)가 데이터과학자가 갖춰야 할 소양이라고 볼 수 있다. 본래 통계학이 데이터를 분석하는 학문이란 점에서 겹치기도 하나, 여기에 프로그래밍과 배경지식이 추가되어 실제 기업과 사회 현장에서 데이터를 활용하는 시나리오를 계획하고, 이에 필요한 솔루션을 만들고 실행해 의사결정에 영향을 주는 일련의 과정이 중심이라는 점에서 구분할 수 있다.

데이터과학자는 현재 인기도 1위의 직업으로, 평균 연봉은 $116,840 약 1.35억이며, 데이터과학자에 대한 수요는 날이 갈수록 급증하고 있다.

## 데이터의 수집과 활용 시나리오

### 목표 설정

데이터를 수집하고 활용하려는 목적을 명확하게 설정한다. 사용 가능한 데이터의 종류와 양을 기반으로 당면한 문제에 도움을 줄 수 있는 솔루션을 찾는다.

### 수집

데이터를 수집한다. 내부 데이터, 웹 크롤링, API, 센서, 공공 데이터 등 이미 존재하는 데이터를 이용하거나 새로운 데이터를 수집한다.

### 저장

데이터를 저장한다. 데이터의 유형과 양에 따라 저장 자체가 개인 수준에서는 어려울 수 있으며, 이 경우에는 전용의 스토리지 서버를 임대하거나 구축할 필요가 있다. 또한 데이터의 백업과 보안도 신경 써야 할 요소다. 최근에는 클라우드 서비스를 통해 저비용으로 고성능의 서비스를 이용할 수 있다.

### 가공

데이터를 목적에 맞게 가공한다. 여러 출처에서 데이터를 수집했을 경우, 양식 등을 통일 시킬 필요가 있다. 또한 누락, 중복 등의 문제를 해결해야 한다. 필요한 경우 이 과정에서 기계학습을 통해 데이터를 1차적으로 가공할 수도 있다.

### 쿼리 및 마이닝

특정 조건을 만족하는 데이터를 찾거나, 데이터의 평균값이나 분산 등의 변수를 계산하는 등의 작업을 수행한다. 각종 통계적 기법과 기계학습을 통해 데이터에서 의미를 추출한다.

### 실제 활용

데이터에서 추출한 의미나 정보를 시각화 혹은 보고서 등으로 작성하고, 팀 및 경영진에게 데이터가 주는 메시지를 효과적으로 전달한다.

# 2. 머신러닝, 딥러닝

#머신러닝 #데이터마이닝 #딥러닝
#ANN #CNN #RNN #강화학습 #GPGPU

## 머신러닝이란?

기존의 프로그램과 달리, 컴퓨터가 스스로 주어진 문제로부터 학습하고 주어진 데이터에 대한 평가와 새롭게 들어올 데이터를 처리할 수 있는 능력을 갖추게 하는 것을 의미한다.

데이터에서 의미 있는 규칙과 패턴을 찾는 데이터마이닝과 비슷해 보일 수도 있고, 실제로 기계학습의 일부 과정에 데이터마이닝이 쓰이기도 하나 기계학습은 데이터의 검증과 예측이 목표란 점이 다르다.

### 머신러닝이 각광받는 이유

데이터는 머신러닝을 위한 자원이다. 기계학습의 본질은 결국 데이터의 분류와 검증, 예측이기 때문이다. 좋은 데이터가 많을수록 심화된 학습이 가능하며, 이로부터 더 정확하고 세밀한 처리 능력이 개발된다.

한편, IT혁명 이후로 웹, 스마트폰, SNS, 사물인터넷의 발달과 확산은 무수한 데이터의 생산과 수집이 가능해졌다. 본래 이러한 데이터는 정리나 분류가 되어있지 않아 쓰레기에 가까웠으나 데이터마이닝, 빅데이터 기술의 발달로 데이터에서 의미 있는 정보를 빠르게 추출할 수 있게 됐다. 이용할 수 있는 데이터의 양과 질의 획기적 개선과 GPGPU를 비롯한 컴퓨터 성능의 발전은 머신러닝의 잠재력을 폭발시켰으며, 수많은 기업과 사회가 머신러닝에 주목하는 이유가 됐다.

## 딥러닝이란?

기계학습의 한 분야로, 사람의 뇌와 같은 뉴런 신경망을 인공적으로 구현한 ANN(Artificial Neural Network)를 기반으로 한 알고리즘들을 통칭한다. 다른 방법과 달리 높은 추상화가 특징이며, 사람이 직접 규칙을 지정하지 않아도 네트워크 스스로 문제에 대한 해결법을 구축하는 듯한 모습이 장점이자, 그 내부 구조를 개발자도 정확히 파악하기 어렵다는 점에서 단점이 되기도 한다.

### 딥러닝이 각광받는 이유

딥러닝의 원류인 인공신경망은 학계에서 사장됐던 분야이다. 학습이 매우 느리고 부정확했으며, 사전에 규칙을 넣어주지 않아도 되는 만큼 더 많은 데이터가 필요했기 때문이다. 그러나 학습 알고리즘의 개선과 함께 컴퓨터의 발달, 특히 GPU의 발달로 인해 학습에 걸리는 시간이 획기적으로 줄어들었으며, 빅데이터의 등장은 바로 이러한 딥러닝에 필요한 무지막지한 양의 데이터를 마련할 수 있게 하였다. 내외적으로 모든 조건이 딥러닝에게 최고의 환경을 조성한 것이다.

또한 딥러닝을 위한 여러 기법이 개발되고 연구되며 딥러닝의 잠재력 또한 더 커졌다. 초기 딥러닝은 숫자 몇 개를 입력으로 받아 역시 숫자로 된 출력을 내놓는 정도였으나, CNN(Convolution Neural Network)를 활용한 이미지 분석, RNN(Recurrent Neural Network)를 활용한 시계열, 자연어 분석 등을 통해 더 많은 데이터와 문제에 딥러닝을 적용할 수 있게 됐다. 나아가 알파고와 같은 강화학습(Reinforcement Learning)은 게임이나 운전과 같은 분야에 딥러닝을 사용할 수 있는 가능성을 열었다.

## GPGPU

General Purpose computing on Graphics Processing Units의 약자로, 일반적으로 2D, 3D 이미지를 처리하기 위한 장치인 GPU를 수치연산 등의 다른 프로그램의 계산에 활용하는 것을 말한다. CPU는 소수의 천재가 있는 방이라면, GPU는 수천의 평범이들이 있는 방이라 볼 수 있으며, 단순하고 병렬화하기 쉬운 연산에 대해선 CPU 대비 수 백, 수 천 배의 속도를 낼 수 있다.

간단한 머신러닝의 경우 일반적인 CPU로도 충분하나, 딥러닝의 경우 매우 간단한 경우를 제외하면 GPU의 사용이 필수적이다. 대부분 NVIDA의 그래픽카드를 사용하며, Tesla, Titan X(p), GTX 1080 (Ti) 등 최고사양의 제품이 딥러닝 연구, 개발을 위한 컴퓨터에 장착되고 있다. 또한 AWS, Azure 등 클라우드 서비스에서도 GPU가 장착된 서버를 합리적인 가격에 대여할 수 있다.

## Race against the Machine

데이터과학과 머신러닝, 딥러닝은 막대한 가치를 창출하며, 과거에는 수많은 인력과 비용이 들었을 일을 빠르게 자동으로 처리하게 한다. 이 말을 뒤집으면 어떻게 되는가? 기업은 비용절감을 위해 기계가 데이터를 수집하고 처리하는 것보다 느린 노동자를 해고할 것이다. 실제로 최근 기술 발달로 인해 사라질 일자리들에 대한 예측이 쏟아져 나오고 있다.

과거에는 타이피스트란 직업이 있었다. 타자가 일인 직업이다. 각 부서에서 펜으로 쓴 메모나 회의록을 가져오면 타자기로 쳐서 문서로 만드는 일을 했다. 지금 타이피스트는 사라졌다. 타자를 대신 쳐준다는 것도 웃기지만 애당초 타자를 칠 필요가 사라지고 있다. 고객상담센터에 전화를 했는데, ARS가 응대를 하고 음성인식을 통해 담당자를 연결해 준 경험이 있는가? 담당자는 살아남았지만 응대를 하던 사람의 일자리는 사라진 것이다.

한편 오토 파일럿이 등장했지만 조종사는 사라지지 않았다. 사라진 것은 항법사, 항공기관사다. 로봇 변호사 ROSS가 미국의 대형 법무법인에 채용됐다. ROSS는 변호사들과 함께 일한다. 변호사들과 함께 판례를 검색하고 분석하던 사무원들은 사라졌다. 이제 우리는 기계가 대체할 직업과, 대체하지 못할 직업에 대해 생각할 필요가 있다. 그 차이는 무엇인가?

감성이나 감정이 필요한 일자리는 살아남을 것이다. 의사결정권을 가진 직업도 살아남을 것이다. 그러나 그 일을 보조하는 일자리는 크게 줄어들 것이다. 다시 말해 데이터를 처리하고 분석하는 직업은 전망이 좋지 않다. 업무의 매뉴얼이 구체적이고 반복적일수록 기계에 쉽게 대체될 것이다. 텔레마케터, 상담원, 운전기사가 그 예가 될 것이다.

Q. **기계와의 경주에서 유리한 / 불리한 직업**?

## Race with the Machine

기계가 인간보다 나은 분야는 분명 많으나, 인간을 완전히 대체할 수 있는 분야는 한정되어 있다. 반면 인간이 기계를 통해 생산성을 역사상 전례없이 끌어올릴 수 있는 기회는 많다. 인간이 수행하기에 복잡하거나, 귀찮거나 시간이 오래 걸릴 작업은 기계에게 맡기고 중요한 작업에 집중할 수 있는 것이다.

명함을 예로 들어보자, 직장인들에게 명함은 중요한 비즈니스 요소다. 상대의 소속, 직급, 이름, 연락처를 파악해야만 결례를 범하지 않으며, 필요한 일이 있을 때 즉각 연락할 수 있다. 그러나 명함을 받고 일일이 기록하는 것은 번거롭다. 명함 인식 앱은 그저 명함을 카메라로 찍기만 하면 이러한 정보를 자동으로 인식하고, 연락처에 추가해준다.

또한 기계를 이용하여 새로운 가치를 창출할 수도 있다. SNS는 끝없이 수많은 사람들의 의견이 자유롭게 게시된다는 점에서 빅데이터와 기계학습의 주목을 받고 있다. 특히, 트렌드 분석에서 그 잠재력과 성과를 입증하고 있다. 특정인, 특정 단체에 대한 글에서 호와 불호를 판별하고, 이들의 평소 행동과 정보를 결합해 분석하는 것이다.

## Deus ex machina

딥러닝의 특징은 우리가 원리나 규칙, 알고리즘을 모르는 일에 대해서도 기계가 스스로 학습하여 처리할 수 있다는 것이다. 지금까지의 기계는 설계자의 능력에 제한되어 왔으나, 이제는 그 제한이 사라지는 것이다. 이는 극단적으로 말해, 충분히 많은 양의 데이터와 충분히 빠른 컴퓨터만 있으면 인간보다 뛰어난 기계를 만들 수 있다는 것이다.

실제로 사진에 있는 항목을 분류하는 ImageNet 경진대회에서의 우승 프로그램들은 이미 인간보다 더 우수한 성적을 보이고 있다. 알파고는 이미 인간과 비교할 수 없는 수준의 실력을 갖췄으며, 포커대회도 인공지능이 승리하였다. 문장의 구문을 분석하는 구글의 클라우드 서비스는 숙련된 모국어 화자의 정확도보다 더 높은 정확도를 보이고 있다.

# 3. 데이터를 활용하는 방법

#존스노우 #심야버스 #TARGET #구글어스
#데이터분석도구 #데이터리터러시

## 데이터 분석 사례

### 존 스노우

왕좌의 게임 등장인물이 아니라, 빅토리아 시대의 의사로 본래는 마취방법 등을 연구하며 빅토리아 여왕을 직접 치료하기도 하였다. 아이러니하게도 후대에는 마취가 아닌 역학조사의 선구자로서 유명하다. 현대의 역학은 최첨단 통계기법과 빅데이터를 통해 발전했지만, 그 근본은 존 스노우의 콜레라 조사에 있다.

당시 런던의 상하수도는 엉망진창에 가까웠으며, 따라서 콜레라와 같은 수인성 전염병이 창궐하였다. 스노우는 콜레라 환자의 집을 지도에 표시해보았고, 거리의 특정한 지점을 중심으로 콜레라가 점점 퍼진다는 사실을 찾아낸다. 그 지점은 바로 펌프였다.

허나 당대 의학계에서는 전염병의 원인이 물이 아닌 미아즈마라는 공기 중의 독기라는 의견이 절대적 지지를 받고있었으며, 펌프의 물을 퍼올려 현미경으로 보아도 특별한 것을 찾을 수 없었다. 스노우는 이에 좌절하지 않고 환자들이 실제로 펌프의 물을 마셨는지 집요하게 확인하고, 먼 지역에서 발병한 환자가 사실은 그 펌프물을 가져와 마셨다는 사실 등을 밝혀내며 기어이 지역사회를 설득해 펌프를 폐쇄하고 콜레라를 종식시킨다. 이후 해당 펌프 근처에서 원인요소를 찾은 결과 첫 발병자의 집 정화조가 펌프와 지나치게 가까이 있었고, 부식된 벽으로 인해 오염이 발생한 것이다.

### 심야버스

서울시가 처음 심야버스를 계획하면서 한정된 버스 노선으로 최대한 많은 지역과 이용자를 커버해야 한다는 문제에 직면했다. 또한 심야 시간대에는 낮과는 다른 이용 패턴이 나타나기에, 문제는 더욱 복잡해졌다.

이는 서울시가 가진 택시 승하차 데이터를 분석하는 것을 통해 해결됐다. 심야에 가장 승하차가 많은 구간을 중심으로 노선을 계획한 것이다.

### TARGET

타겟은 미국내 업계 2위의 할인매장으로, 2002년 고객 데이터 분석 서비스 부서를 신설한 이후로 빅데이터 분석을 통해 고객들에게 맞춤 쿠폰을 보내고 있다. 과거의 구매 내역을 추적해 자주 사는 품목의 신제품 및 경쟁제품에 대한 쿠폰을 보내거나 내역에 변동이 있을 경우 이를 예측해 곧 구매하게 될 물건에 대한 할인을 제공하는 것이다.

타겟의 맞춤 쿠폰은 2012년 10대 여고생에게 유아용품 할인 쿠폰을 보낸 사건이 알려지면서 유명세를 탔다. 쿠폰을 본 아버지는 딸이 임신이라도 한 줄 아냐며 타겟에 강력하게 항의했으나, 상황은 며칠만에 반전된다. 그 학생이 정말로 임신을 한 것이었기 때문이다.

타겟은 임산부의 일반적인 구매 패턴을 통해 특정 고객의 임신 여부를 판단한다. 가령 임신 초기에는 칼슘, 마그네슘 등이 포함된 영양제를 구매하고, 20주를 넘으면 튼살 방지를 위한 로션을, 출산이 임박하면 유아용품을 구매한다는 것이다. 학생이 영양제와 로션을 일정 기간을 두고 구매한 것이 포착되었고, 이를 역산하여 출산시기를 계산해 때에 맞춰 쿠폰을 보낸 것이다.

### 동물들은 지구 자기장을 느낀다

사빈 베갈이 이끄는 연구팀은 동물들의 자기 감각(Magnetic sense), 즉 지구 자기장을 감지할 수 있는 능력에 대해 연구하고 있었다. 새나 물고기, 쥐 등이 이 감각을 가지고 있다는 것은 많이 알려졌으나, 문제는 더 큰 동물에게도 같은 능력이 있냐는 것이었다. 이들은 구글 어스를 이용해 위성사진에 찍힌 소와 사슴들의 자세를 조사하였고, 머리가 북쪽을 향하려는 경향이 있음을 밝혀냈다.

Q. **사례들의 공통점과 배울 점은 무엇인가?**

## 데이터를 처리하기 위한 도구

### EXCEL

종종 과소평가 당하기도 하나, 일반적인 개인용 컴퓨터로는 감당할 수 없는 양의 데이터를 처리하거나 매우 복잡한 연산을 하는 것이 아니라면 엑셀의 기능적 한계로 분석을 못할 경우는 거의 없다. 또한 필요한 경우 VBA나 각종 추가기능을 통해 전문 솔루션 못지않는 효율성을 보일 수도 있다.

### Python

파이썬은 배우기 쉽지만 강력한 프로그래밍 언어로, 일반적인 코딩부터 웹, 서버, 데이터 마이닝, 인공지능 등에 이르기까지 매우 다양한 영역에서 사용되고 있다. 파이썬은 웹 크롤링이나 데이터 전처리 등을 적은 비용을 통해 할 수 있다는 점에서 그 자체로도 뛰어나지만, 언어와 솔루션을 가리지 않고 궁합이 잘 맞는다는 점에서 이들 사이의 다리 역할을 한다는 것도 빼놓을 수 없는 장점이다.

“Life is too short. You need python.”

### R, Matlab

파이썬을 비롯한 범용 프로그래밍 언어와 달리, 이들은 태생이 수치 연산과 통계 처리에 있다. 특히 기업들이 주로 파이썬을 이용하는 경우가 많은 반면, 학계에서는 R과 매트랩을 주로 사용하는 편이다. 많은 연구, 활용 사례와 확장기능 등이 있으며 기본적으로 데이터에 최적화되어 있는 것이 장점이다.

### 데이터베이스 관리 시스템 - SQL, 하둡 등

SQL은 데이터베이스를 관리하고 질의하기 위한 프로그래밍 언어로, 웹페이지, 재고관리, 각종 프로그램 등 다양한 곳에서 표준으로 사용되고 있다. 따라서 SQL은 데이터 처리의 표준어라고도 볼 수 있다.

반면 빅데이터의 경우는 SQL로는 한계가 있기에, 클라우드나 그리드 컴퓨팅을 통한 분산처리를 지원하는 하둡 등의 빅데이터 프레임워크가 이용되고 있으며, 구글, 페이스북을 비롯한 IT 공룡들은 자체적으로 솔루션을 구축하고 있다.

## 데이터 분석을 위해서는

기업, 조직에서 데이터 분석을 의사결정에 활용하는 것은 단순한 문제가 아니다. 데이터 분석 담당자만이 아니라 구성원 모두의 협조와 결정권자의 열린 태도가 필요하며, 데이터의 품질 유지와 관리를 위한 밑작업과 업무 프로세스 개선도 동반되어야 한다. 그러나 데이터 분석에 대한 수요는 모든 산업에서 급속도로 늘어나고 있으며, 보유한 데이터에서 가치를 창출하는 것이 경쟁에서의 핵심이 되면서 이를 위한 준비가 되지 않은 기업은 곧 도태될 전망이다.

또한 최근에는 Quantified self, 즉 개인의 일상생활에서 자신의 신체 상태를 각종 스마트, 웨어러블 기기로 측정하고 관리하는 것을 통해 삶의 질을 높이려는 움직임이 확산되고 있다. 체중, 심박수, 걸음, 운동 기록, 수면, 영양 정보 등의 수많은 생활 데이터가 다양한 센서를 통해 수집되고 있으며, 한 연구결과에 따르면 미국 성인의 69%가 적어도 하나의 생활 데이터를 추적하고 있다.

이렇듯 데이터 분석은 모두에게나 중요하지만, 그만큼 준비해야할 것도 많다.

### 기록하기

분석을 하기위해선 당연히 데이터가 존재해야 한다. 데이터 자체가 파일 등으로 잘 관리되고, 백업되고 있을 뿐만 아니라 데이터의 생성 및 수집 과정에 관한 메타데이터도 착실히 있어야 한다. 공장이나 사무실 등의 시설, 차량 등에 센서를 부착하고 네트워크에 연결하는 것도 떠오르는 경영 혁신 사례다.

### Data literacy

글을 읽고 해석하며 맥락을 파악하는 것처럼, 데이터 리터러시는 데이터를 읽고 해석하며 의미를 파악하는 능력을 말한다.

### 열린 마음

모든 변화와 혁신이 요구하는 것이지만, 기존의 프로세스를 고수하는 대신 열린 마음으로 새로운 아이디어와 방법을 받아들일 필요가 있다. 당장은 조금 귀찮더라도, 데이터를 착실히 기록한다면 이는 곧 커다란 자산이 되어 돌아올 것이다.

# 4. 데이터의 수집과 가공

#데이터베이스 #크롤링 #API #전처리
#회귀분석 #군집분석

## 데이터 수집

### 데이터베이스

리빙포인트: 데이터를 수집하려면 데이터가 저장되어 있는 데이터베이스를 찾아 연결하면 된다. 문제는 세상 어디에도 "내가 필요한 데이터가 저장된 DB" 같은 건 없다는 것이다. 설사 존재한다 하더라도 외부에 공개되어 있지 않거나, 불완전한 경우가 많다. 이를 극복하기 위해서 문제에 관련된 데이터를 담고 있는 DB가 무엇이 있는지 탐색하고, 때로는 심야버스 노선을 위해서 택시 데이터를 이용한 것처럼 창의력과 상상력을 발휘해야 할 수도 있다.

### 문헌자료

논문이나 보고서의 표, 차트, 각종 문서도 전통적이지만 여전히 유용한 데이터 소스다. 디지털화가 되어 있지 않다면 파일로 만드는 과정이 수반되어야 한다는 단점이 있으나, 때로는 유일한 대안인 경우도 있다.

### 웹 크롤링

웹 페이지와 게시판 등의 자료를 문자 그대로 긁어온다. 해당 사이트의 구조와 HTML에 대한 이해가 있어야 원하는 데이터를 추출할 수 있다. 주로 파이썬과 크롬 개발자 도구를 이용해 웹 크롤러를 제작한다.

### API (Application Programming Interface)

API란 특정 서비스나 소프트웨어에 대해 외부에서 접근할 수 있도록 만들어 둔 기능이라고 할 수 있다. 가령 페이스북 API는 다른 프로그램이 페이스북 계정에 대한 정보를 알아내고 연동이 가능하게 한다.

### Digital Exhaust

페이지 방문기록, 쿠키, 검색, 머무른 시간 등 인터넷을 사용하면서 자연스럽게 생기는 정보들을 의미한다. 구글 맞춤 광고 등이 이를 사용하는 대표적 사례다.

## 데이터 전처리

### 변수 확인 및 형태 변형

데이터 처리에 앞서 해당 데이터의 메타데이터를 파악한다. 즉 필드와 레코드의 의미, 출처 및 수집 과정에 대해 확인한다. 몇몇 데이터는 분석 기법을 적용하기 어려운 형태로 저장되어 있을 수 있는데, 이 또한 적절하게 변형한다.

### 단위 및 양식 통일

다양한 소스에서 데이터를 수집한 경우, 같은 변량에 대해서도 다른 단위나 양식을 사용할 수 있다. 섭씨, 화씨, 미터, 야드 등이 대표적이며 이외에도 주소, 전화번호 등도 형식을 통일시킬 필요가 있다.

### 중복 및 누락, 이상값

동일한 레코드가 여러 개 있다면 하나만 남기고 제거한다. 단, 이 때 일치성의 판단 기준이 문제가 될 수 있다. 가령 이름은 동명이인이 흔히 존재할 수 있어 적합하지 않다. 누락의 경우 데이터의 특성에 따라 적절한 처리방법을 결정해야 한다. 일반적으로 0으로 놔둘 경우 분석에 심각한 왜곡을 주기 때문에, 집단의 평균을 넣어주거나 앞뒤 변수의 평균을 이용한다. 혹은 참고할 수 있는 다른 값을 이용해도 좋다.

### 통합

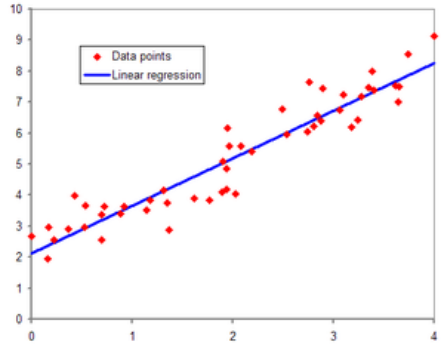
데이터 사이의 관계를 이용해 데이터들을 통합할 수 있다. 대표적으로 주문내역으로, 주문내역에 있는 상품 필드를 재고관리내역의 상품과 연결할 수 있다.

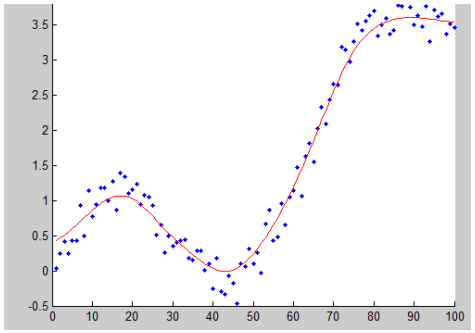
### 변환

지수나 제곱, 로그적 분포를 보이는 데이터는 선형으로 변환하면 더 좋은 결과를 얻을 수 있다. 또한 텍스트나 음성, 영상 등의 비정형 데이터는 문장 부호나 노이즈를 제거하고, Word embedding이나 FFT를 통해 변환하는 과정이 필요하다.

## 회귀분석(Regression Analysis)

회귀분석을 통해 연속 데이터들 사이의 관계를 설명하는 모형을 구할 수 있다. 아래와 같은 선형회귀가 대표적이며, 간단하게는 변수들의 분포를 가장 그럴듯하게 보여주는 선이나 함수를 찾아내는 과정이라고 이해할 수 있다.



회귀분석을 통해 경향성을 찾아낼 수 있으며, 이러한 경향이 데이터를 얼마나 설명할 수 있는지를 적합도라는 지표를 통해 알아낼 수도 있다. 또한 회귀분석을 통해 독립변수와 종속변수의 관계를 수치적으로 알아낼 수 있다. 전단지를 돌리는 것이 매출에 어떤 영향을 미치는가? 100장을 돌리면 얼마를 더 버는가? 와 같은 질문에 대해 답할 수 있는 것이다. 또한 데이터에 없는 경우에 대해서도 예측을 할 수 있다.

선형회귀의 경우, 데이터의 관계를 아래와 같은 식으로 나타낸다.

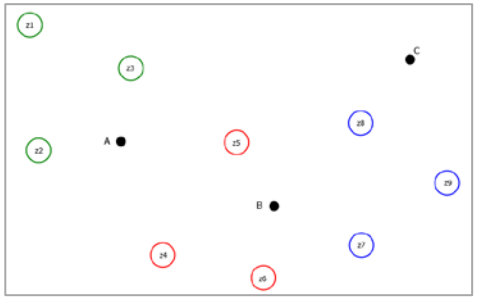
$$Y = aX_1 + bX_2 + cX_3 + \cdots + z$$

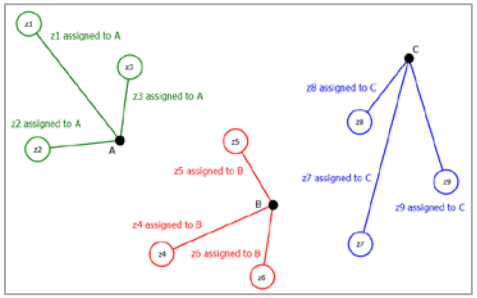
각 X들 앞에 있는 계수는 곧 그 변수가 Y에 미치는 영향이며, z는 절편이 된다. 가령 손님 수 = 0.1 * 전단지 수 + 10이라면 전단지를 10장 더 돌리면 손님이 1명 더 올 것이라는 예측을 할 수 있는 것이다. 이렇듯 선형회귀는 해석이 간단하기에 유용하게 사용되나, 변수들 사이에 선형 관계가 없으면 이용할 수 없다. 이 경우 변수에 로그나 제곱, 루트 등을 취해 선형으로 변환하는 방법이 있으며 이를 통해 다양한 상황에서 선형회귀를 적용할 수 있다.

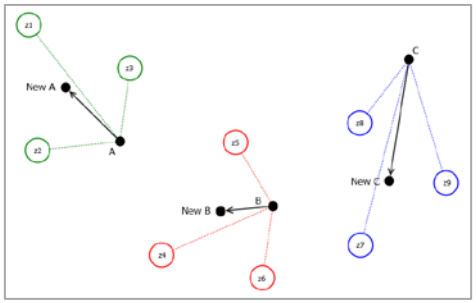
## 군집 분석(Cluster Analysis)

데이터의 특성을 더 잘 이해하고 구분하기 위해 데이터를 비슷한 것끼리 그룹화하는 것을 군집화라고 한다. 군집 분석은 데이터를 잘 설명할 수 있는 그룹을 정하고, 각 그룹의 대표점을 찾는 데이터 마이닝 기법이다. 마케팅 및 고객관리, 제품분석, 실험 결과 분석 등 다양한 목적을 위해 유용하게 쓰인다.

군집화를 위한 알고리즘은 매우 다양하나, 가장 널리 알려진 것은 k-means 알고리즘이다. 이 알고리즘은 주어진 데이터를 k개의 군집으로 묶으며, 각 데이터와 군집의 중심 사이의 거리를 최소화하는 방식으로 작동한다.






Q. **회귀 및 군집 분석을 적용할 수 있는 사례?**

# 5. 데이터를 설명하는 통계 - 기술통계학

#통계량 #평균 #중앙값 #편차 #분산 #상관계수 #표 #그래프

## 통계량

데이터의 표본 하나하나를 살펴보며 특성을 알아볼 수도 있지만, 시간과 비용이 많이 들고 알아낼 수 있는 것도 제한적이다. 기술통계학(Descriptive statistics)은 이처럼 방대한 데이터를 표나 그래프로 알아보기 쉽게 나타내고, 각종 통계량을 계산해 데이터의 특성을 요약하는 학문이다.

고등학교 한 반의 내신 성적표를 데이터라고 할 때, 이 표에 학생 한 줄을 레코드라고 한다. 반이 30명이면 이 표에는 30개 레코드가 있는 것이다. 각 레코드에는 국어, 수학, 영어 등의 항목이 있고 해당하는 점수가 있는데, 이 개개의 점수를 변량이라고 한다.

### 평균(Mean)

$$\text{평균 } \overline{x} = \sum x_i = \frac{x_1 + x_2 + x_3 + x_4 + \cdots + x_n}{N}$$

N개의 변량에 대해 평균은 변량을 모두 더한 뒤 그 개수인 N으로 나눈 것으로 구한다. 쉽게 구할 수 있으면서도 데이터의 가장 중요한 특성을 나타낸다는 점에서 널리 쓰이나 이상값(Outlier)에 영향을 많이 받는다는 점에서 한계가 있다.

### 중앙값(Median)

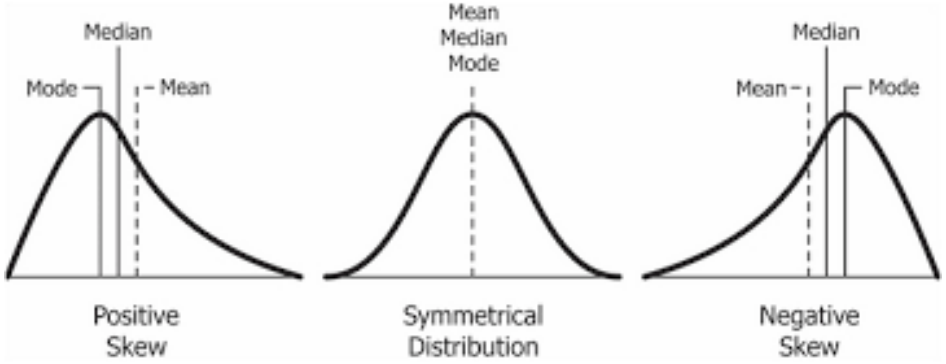
변량을 순서대로 놓았을 때 딱 가운데에 있는 변량의 값을 말한다. 변량이 짝수개이면 가운데의 2개 변량의 값을 평균한 것으로 구한다. 가령 1, 2, 2, 3, 3의 경우 2가 중앙값이며, 1, 2, 2, 3, 3, 3은 2와 3의 평균인 2.5가 중앙값이다.

### 최빈값(Mode)

변량 중 가장 많이 나타나는 값을 말한다. 1, 1, 1, 1, 2, 2, 2, 3, 3의 경우 1이 가장 많이 나타났으므로 1이 최빈값이 된다.

데이터가 정규분포에 가까울수록 이들 3가지 통계량의 값이 수렴한다.

### 왜도(Skewness)




데이터의 분포가 치우친 정도를 말한다. 위의 그림에서 볼 수 있듯, 꼬리가 오른쪽으로 길어진 경우 Right tailed라 하며 양의 왜도 값을, 왼쪽으로 길어진 경우 Left tailed라 하며 음의 왜도 값을 가진다. 왜도 값에 따라 평균, 중앙값, 최빈값의 상대적 분포가 변하는 것도 주목하자.

### 백분위수(Percentile)

변량들을 순서대로 나열했을 때, 백분율로 나타낸 위치에 해당하는 변량의 값을 의미한다. 100개의 변량이 있을 때, 25백분위수의 값은 25번째 변량의 값에 해당한다. 만약 200개의 변량이라면 50번째 변량의 값이 될 것이다. 보통 25, 50, 75 백분위수를 각각 제1, 제2, 제3사분위수라고 한다.

### 편차(Deviation)와 분산(Variation)

편차는 각 변량에서 평균값을 뺀 것을 의미한다. 이 때, 각 변량들의 편차를 제곱한 뒤 더한 것을 편차제곱합이라고 한다. 편차제곱합을 변량의 수로 나눈 것을 분산이라고 한다. 수식으로는 다음과 같다.

$$V = s^2 = \frac{1}{\mathrm{N}}\{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2\}$$

이 때 분산의 제곱근인 s를 표준편차(Standard deviation)이라고 한다.

분산은 데이터가 평균, 즉 중심으로부터 얼마나 퍼져 있는지를 나타낸다. 이 때 편차를 그냥 더해도 되는데 굳이 저렇게 제곱을 해서 더하는지 의문이 들 수도 있다. 이는 양의 편차와 음의 편차가 서로 상쇄되는 것을 막기 위한 것이다.

가령 2, 2, 2와 1, 2, 3을 생각해보자. 둘다 평균은 2이나 분산은 다르다. 만약 편차를 그냥 더할 경우 전자는 0 + 0 + 0 = 0, 후자는 −1 + 0 + 1 = 0으로 분산이 같은 것으로 나오게 된다. 따라서 이를 막기 위해 편차제곱합을 개수로 나누어 분산을 구한다. 그러나 이렇게 하면서 단위가 바뀌고 (미터에서 제곱미터로 바뀐다거나) 또한 수치가 뻥튀기되기 때문에, 루트를 취해서 다시 원래 단위로 바꾸는 것이다.

어떤 변량의 편차와 표준편차를 비교하는 것을 통해 해당 변량이 평균으로부터 얼마나 떨어져 있는지를 알아볼 수도 있다. IQ가 대표적이다.

### 공분산(Covariance)

지금까지의 통계량이 한 변수의 특징을 요약했다면, 공분산은 두 변수의 관계를 나타낸다. x와 y의 공분산은 다음과 같이 구한다.

$$S_{xy} = \frac{(x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + \cdots + (x_n - \overline{x})(y_n - \overline{y})}{N}$$

이 값이 양수이면 x와 y가 양의 상관, 즉 x와 y가 같이 증가하거나 감소하고, 음수이면 음의 상관, 즉 x가 증가할 때 y가 감소하거나 그 반대인 관계가 있다. 만약 0인 경우, x와 y 사이에 상관관계가 없다고 본다.

### 상관계수(Pearson correlation coefficient)

분산과 마찬가지로 공분산도 수치가 단위에 따라 뻥튀기 되는 문제가 있다. 따라서 공분산 값만을 보고 상관관계의 여부나 강도를 알기는 힘들다. 이를 해결하기 위한 것이 아래의 상관계수다.

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

이 값은 −1 < r < 1의 범위를 가지며, 1에 가까울수록 양의 상관, −1에 가까울수록 음의 상관을 나타내며 값이 극단적일수록 상관관계가 크다고 볼 수 있다.

## 표와 그래프

데이터를 분석하고 시각화할 때 표와 그래프를 활용한다. 엑셀을 활용하면 손쉽게 만들 수 있지만, 필요한 정보를 효과적으로 나타내는 것은 또다른 문제다.

### 표(Table)

표는 행과 열로 구성되어 있다. 엑셀 기준으로 숫자로 표시되는 게 행이며 문자로 표시되는 게 열이다. 하나의 행은 하나의 레코드를 나타낸다. 열은 필드라고도 하며, 각 레코드에 포함된 변량들의 종류를 구분한다. 행과 열이 교차하는 각각의 칸을 셀이라 하며 개별 변량은 이 셀에 들어있다.

표를 활용할 때 가장 중요한 것은 적절한 열을 정하는 것이다. 열은 표의 데이터가 가진 정보의 종류를 정하는 것이기에, 적절한 열을 구성하면 곧 적절한 표가 만들어진다. 표를 막 만드는 대신, 지금 필요한 정보가 무엇인지 고심하고, 필요한 경우 평균, 합계, 순위 등 다양한 통계량을 계산하여 열을 추가한다.

### 그래프

나이팅게일이 야전에서 병사들의 사망원인을 나타내기 위해 사용한 Rose graph 이후로, 지금까지 수많은 종류의 그래프가 만들어지고 사용되고 있다. 그래프를 사용할 때 가장 중요한 것은 단순히 모양이 이쁜 그래프를 골라서는 안 된다는 것이다. 각 데이터는 그 특성에 적합한 종류의 그래프가 정해져 있으며, 또한 목적에 따라 다른 그래프를 써야할 수도 있다. 가령 비율 데이터는 일반적으로 원이나 도넛 그래프를 많이 쓰지만, 시간에 따른 비율의 변화를 보여주기 위해서는 꺾은선 그래프가 더 적합하다.

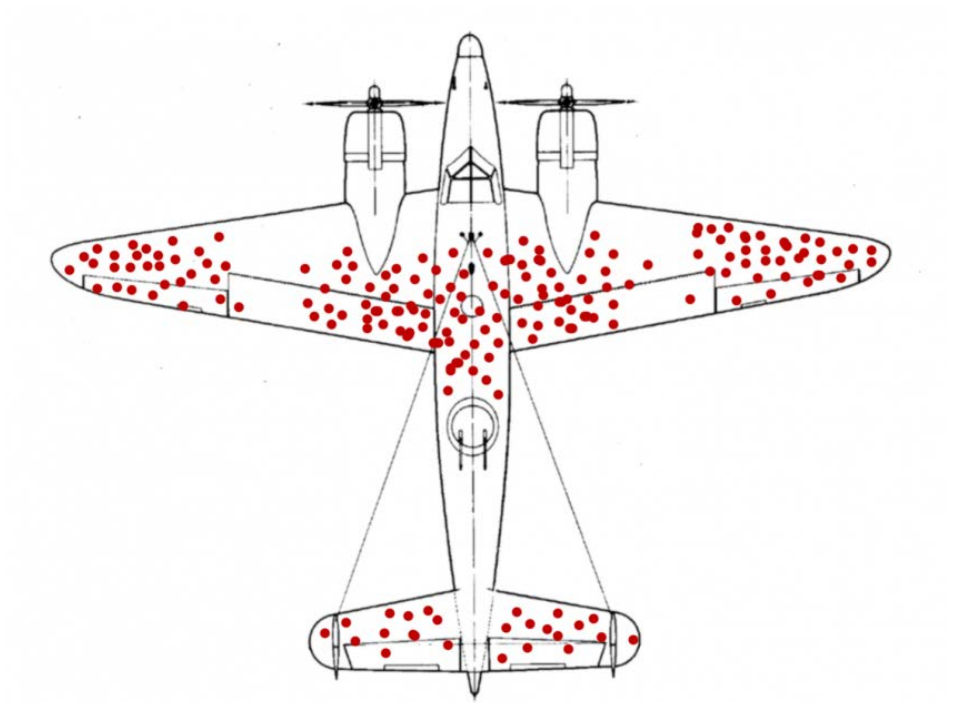
또한 축과 단위를 정하는 것도 중요하다. 그래프의 각 축에 어떤 것을 나타낼지, 그리고 그 단위를 어떻게 할지는 그래프의 명확성을 결정하는 핵심요소다. 때로는 원래 값 대신 로그를 씌우거나 백분율값을 이용하는 것도 답이 될 수 있다.

그래프는 편향된 정보를 전달할 수도 있다. 특히, 길이와 면적이 문제가 될 수 있는데 이는 길이가 2, 3배 늘어날 때 면적은 4, 9배 증가하기 때문으로 이를 고려하지 않으면 잘못된 비례의 그래프를 그리게 된다.

# 6. 데이터와 오류

#편향 #오류 #왜곡 #상관관계
#인과관계 #시각화오류

## 전투기의 피탄 흔적



2차대전 중 미 해군에서는 전투기의 생존성을 높이기 위해 작전 중 총알에 맞고 살아 돌아온 비행기들의 총알자국의 분포를 조사하였다. 위 그림의 빨간 점이 바로 그 분포로, 이에 따라 군 당국은 빨간 점이 밀집된 부분에 장갑을 더 강화해야한다고 생각하였다.

그러나 연구를 이끈 아브라함 발트란 통계학자는 이와 반대로 빨간 점이 없는 곳에 장갑을 둘러야 한다고 주장하였다.

이는 조사의 대상이 “총알에 맞고 살아 돌아온 비행기”들이므로, 저 점들은 다시 말해 맞아도 살 수 있는 위치들을 의미하며 도리어 점이 없는 곳일수록 맞으면 치명적인 부위라는 것을 의미하기 때문이다.

## 데이터 분석에서의 오류

데이터 분석 과정에는 항상 수많은 오류를 범할 가능성이 있다. 완벽한 분석을 하는 것은 불가능할지라도, 잘못된 결론을 내릴 수 있는 중대한 오류들에 대해 미리 파악하고, 대처하는 것은 데이터과학의 필수요소다.

### 데이터 자체의 오류와 편향

“인터넷 및 컴퓨터 보급 실태”에 관한 설문조사를 한다고 하자, 이 설문조사를 구글 폼으로 수행한다면 아주 편향된 결과가 나타날 것이다. 열악한 IT 환경에 있는 사람들은 해당 설문에 참여할 수 없기 때문에, 보급률은 과대평가된다.

이처럼 수집 방법, 조사 대상의 선정에 따라 데이터 자체가 이미 오류와 편향을 내재하고 있을 수 있다. 따라서 메타데이터를 분석해 이러한 잠재적 오류의 가능성을 탐색하고, 이를 제거하기 위한 방법에 대해 생각할 필요가 있다. 나아가 표본조사를 진행할 경우, 탐구하고자 하는 전체 집단에 대해 충분한 대표성을 가질 수 있는 표본을 선정해야 한다.

### 분석 과정에서의 잘못

데이터 자체는 문제가 없더라도, 통계 분석 기법을 잘못 적용하여 오류가 발생할 수 있다. 잘못된 식을 사용하거나, 해당 변수의 특성과 맞지 않는 기법을 사용하는 것이다. 또한 계산 및 분석 과정에서 실수로 인해 잘못된 통계량을 산출하거나, 데이터 전처리 과정에서 큰 왜곡이 발생할 가능성도 있다. 따라서 잘 알려진 샘플 데이터를 활용해 분석을 테스트하고, 제대로 된 결과가 나오는지 확인 후 실제 분석을 진행하는 방법이 유용하다.

### 상관관계와 인과관계의 혼동

회귀분석 및 상관계수 분석 등으로 나오는 변수들의 관계는 단순히 상관관계고 이것이 인과를 보장하지는 않으나 이들을 구분하지 못한 오류가 빈번하다.

# 인과의 판단

## 상관관계가 말해주는 것

A와 B 두 변수 사이에 양의 상관관계가 관측됐다면, 이는 A가 증가할 때 B도 증가하는 패턴이 데이터에서 나타났다는 것이며, 추가적인 조사 없이 이로부터 어떤 인과적 판단을 내릴 수는 없다.

에어컨 판매량과 선풍기 판매량에 대해 상관분석을 진행하면 분명히 강한 양의 상관관계가 있다고 나올 것이다. 그러나 에어컨 판매를 촉진시키기 위해 선풍기 할인판매를 시작해야한다고 주장하는 사람은 없을 것이다. 이들의 관계를 좀 더 정확히 파악하기 위해선 두 변수와 관련이 있는 다른 숨은 변수의 존재를 생각해야한다. 이 경우 그 변수는 기온이 될 것이다. 기온이 높아지면 에어컨의 판매가 활발해지고, 선풍기 또한 판매량이 늘어나기에 기온이 올라가고 내려가면서 에어컨과 선풍기 판매량 변수의 움직임이 같은 방향으로 나타날 것이다.

상관관계는 심증이 될 수는 있으나 확증이 될 수는 없다. 상관관계는 매우 다양한 원인에 의해 나타날 수 있다. 실제 인과관계에 의해서도, 위처럼 제3의 숨은 변수에 의해서도, 데이터 자체의 편향에 의해서도, 혹은 단순히 우연에 의해서도 나타날 수 있다. 심지어는 실제로 인과관계가 있는 두 변수가, 다른 요인에 의해 영향을 받아 통계상으로는 상관관계가 없는 것으로 나타날 수도 있다.

## 상관에서 인과로

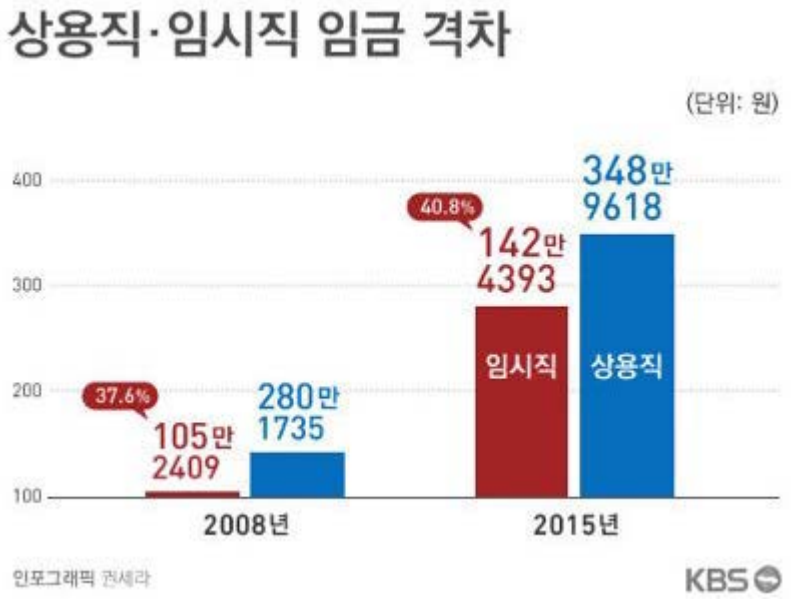
인과관계를 밝히기 위해선 나타난 상관관계에 다른 변수가 영향을 미치지 않았다는 것을 입증해야 한다. 실험실 환경에서는 실제로 다른 변인들을 통제하고 독립변수만을 변화시킬 수 있으나, 대부분의 데이터는 그런 통제가 불가능하다.

또 하나의 방법은 해당 변수를 제외한 다른 변수들을 모두 임의로 할당하거나 임의적으로 분포된 것을 밝히는 것이다. 가령 게임 플레이 시간과 점수의 관계를 알기 위해 대상 집단에게 게이밍 경험이나 성별, IQ 등의 변수와 관련없이 임의로 플레이할 시간을 할당하고 이후 점수를 분석해 나온 상관관계는 인과관계로 해석할 수 있다.

# Area Principle

데이터와 분석 결과가 옳더라도 데이터를 시각화하는 과정에서 오류나 왜곡이 나타날 수 있다. 대표적인 것으로 Area Principle을 위배하는 것이다. 이 원칙은 데이터를 나타내는 도표나 차트에서 도형의 면적이 데이터의 크기와 대응해야 한다는 것으로 이를 지키지 않을 경우 오해나 혼동을 일으킬 수 있다. 그런데 역으로 보는 사람이 잘못된 해석을 내리게 하기 위해 이를 악용하기도 한다.

"매직 그래프"라는 표현이 바로 그러한 종류의 그래프를 말하는 것이다. 주로 막대 그래프에서 Y축이 0부터 시작하는 대신, 특정 값부터 시작해 전체 값에 비해서는 미미한 차이가 그래프상으로는 크게 나타나는 수법을 사용한다.




Q. 데이터 분석 과정에서의 오류, 왜곡 사례?