

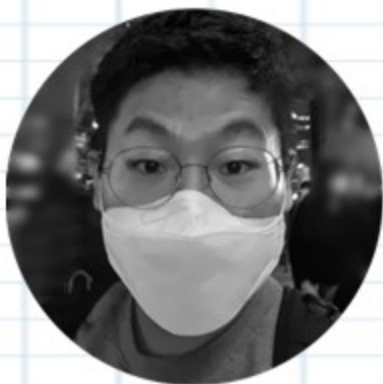
PART 3. 데이터 분석 - 1장.데이터 분석 개요

데이터 준전문가

ADSP, Advanced Data Analytics semi-Professional

류영표 강사

ryp1662@gmail.com



류영표

Youngpyo Ryu

동국대학교 수학과/응용수학 석사수료

現 Upstage AI X 네이버 부스트 캠프 AI tech 1~4기 멘토

前 Innovation on Quantum & CT(IQCT) 이사

前 한국파스퇴르연구소 Image Mining 인턴(Deep learning)

前 (주)셈웨어(수학컨텐츠, 데이터 분석 개발 및 연구인턴)

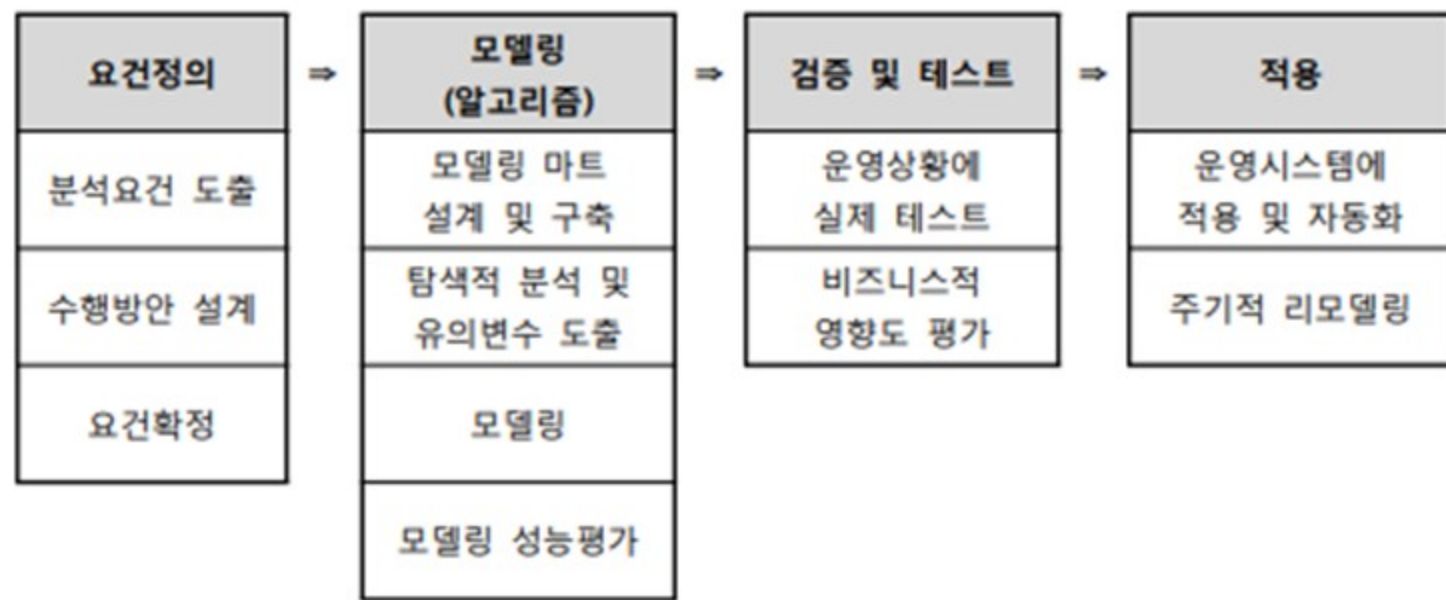
강의 경력

- 현대자동차 연구원 강의 (인공지능/머신러닝/딥러닝/강화학습)
- (주)모두의연구소 Aiffel 1기 퍼실리테이터(인공지능 교육)
- 인공지능 자연어처리(NLP) 기업데이터 분석 전문가 양성과정 멘토
- 공공데이터 청년 인턴 / SW공개개발자대회 멘토
- 고려대학교 선도대학 소속 30명 딥러닝 집중 강의
- 이젠 종로 아카데미(파이썬, ADSP 강사) / 강남 : ADSP
- 최적화된 도구(R/파이썬)을 활용한 애널리스트 양성과정(국비과정) 강사
- 한화, 하나금융사 교육
- 인공지능 신뢰성 확보를 위한 실무 전문가 자문위원
- 보건 · 바이오 AI활용 S/W개발 및 응용전문가 양성과정 강사
- Upstage AI X KT 융합기술원 기업교육 모델최적화 담당 조교

주요 프로젝트 및 기타사항

- 개인 맞춤형 당뇨병 예방·관리 인공지능 시스템 개발 및 고도화(안정화)
- 페플라스틱 이미지 객체 검출 경진대회 3위
- 인공지능(AI)기반 데이터 사이언티스트 전문가 양성과정 1기 수료
- 제 1회 산업 수학 스터디 그룹 (질병에 영향을 미치는 유전자 정보 분석)
- 제 4,5회 산업 수학 스터디 그룹 (피부암, 유방암 분류)
- 빅데이터 여름학교 참석 (혼잡도를 최소화하는 새로운 노선 건설 위치의 최적화 문제)

데이터 분석 기법의 이해



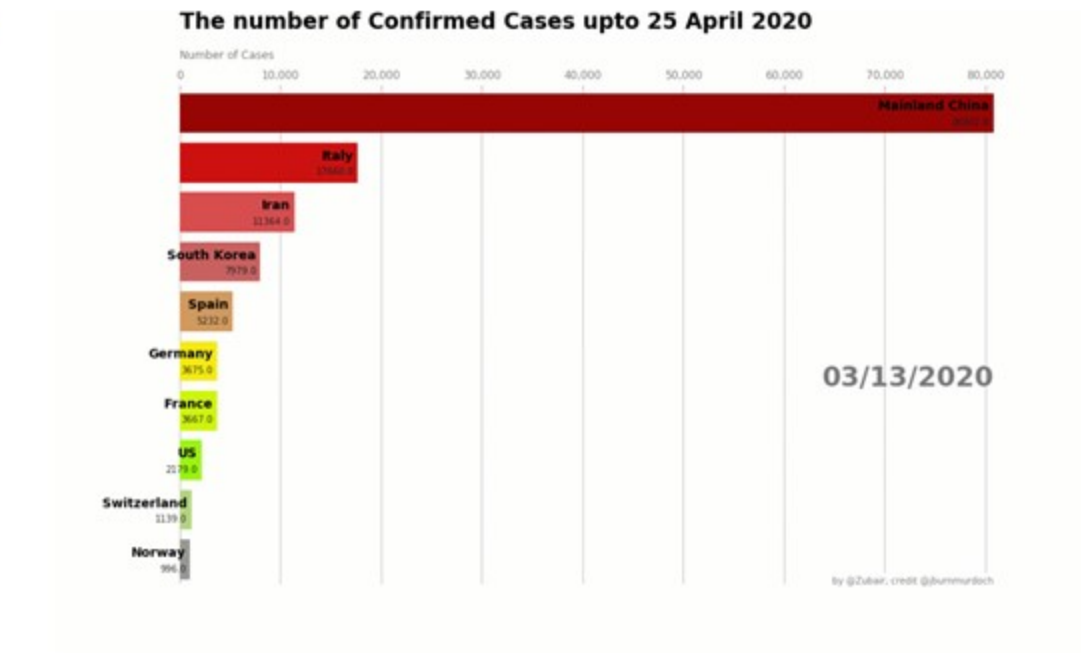
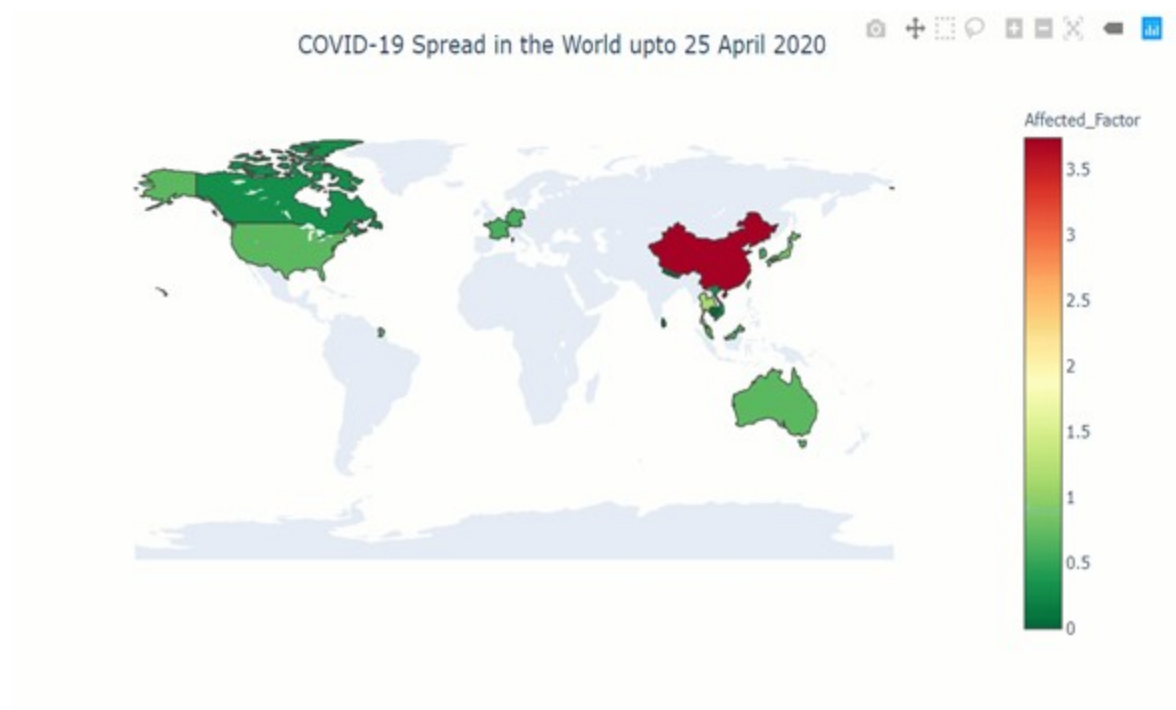
- 데이터 분석에 대한 정의는 매우 다양하고 수준과 복잡성, 목적도 다르다.
- 분석은 많은 데이터를 이용해 **인사이트를 얻거나 의사결정을 하는데 사용**됨.
- 다양한 산업에 대한 이해가 중요하다.
->상식수준에서 벗어난 해당 업계 신입사원 수준의 **산업 분야 이해가 필요**하다.
- 평상시 관심을 갖고 업무와 관련되어 조금씩 늘 학습할 것 추천한다.

데이터 분석 기법의 이해

- 데이터 분석은 통계에 기반을 두고 있지만, 통계지식과 복잡한 가정이 상대적으로 적은 실용적인 분야.
- 신규 데이터나 DW에 없는 데이터는 기존 운영시스템(legacy)에서 직접 가져오거나 운영데이터저장소(ODS)에서 정제된 데이터를 가져와서 DW의 데이터와 결합하여 활용
- 분류값과 입력변수들을 연관시켜 인구통계, 요약변수, 파생변수 등을 산출한다.
- 파생변수 : 기존의 변수를 조합하여 새로운 변수를 만들어 내는 것을 의미.
- 요약변수 : 수집된 정보를 분석에 맞게 종합(aggregate)한 변수
- 독립변수 : 종속변수의 변화를 가져오거나 영향을 미치는 원인 변수
- 종속변수 : 독립변수의 영향으로 나타나는 결과가 되는 결과 변수

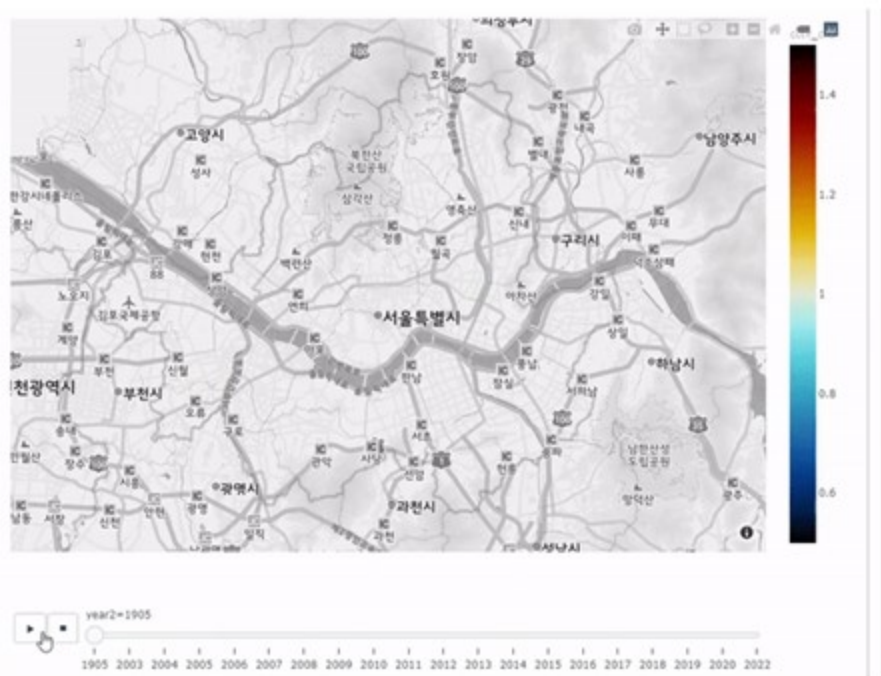
시각화(시각화 그래프)

- 시각화는 가장 낮은 수준의 분석이지만, 잘 사용하면 복잡한 분석보다도 더 효율적이다.
- 대용량 데이터를 다루는 빅데이터 분석에서는 시각화는 필수
- 탐색적 분석을 할 때 시각화는 필수



공간분석(Spatial Analysis)

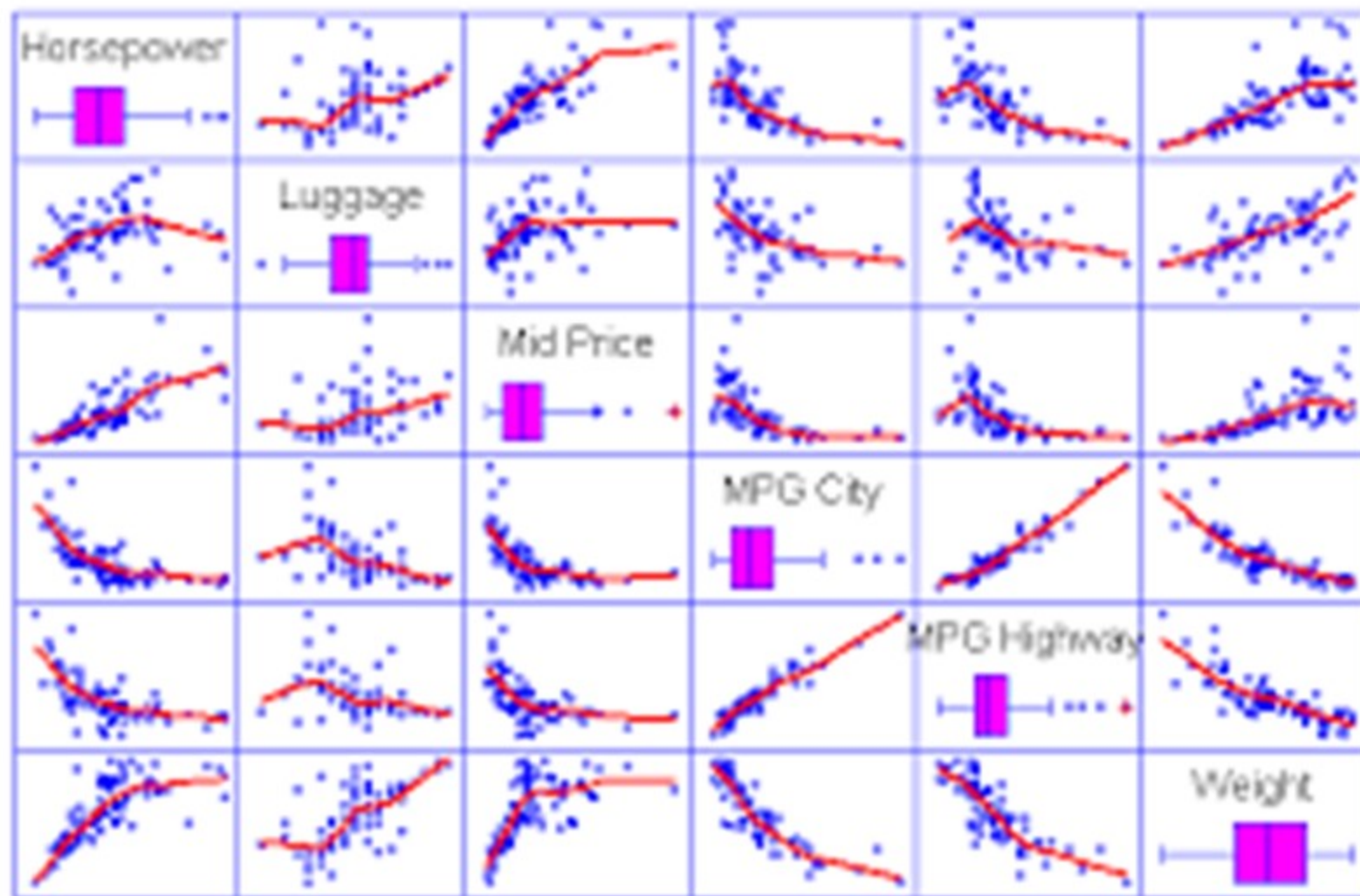
- 공간적 차원과 관련된 속성들을 시각화하는 분석
- 지도 위에 관련 속성들을 생성하고 크기, 모양, 선 굵기 등으로 구분하여 인사이트를 얻는다.



탐색적 자료 분석(Exploratory Data Analysis)

- 다양한 차원과 값을 조합해가며 특이한 점이나 의미 있는 사실을 도출하고 분석의 최종 목적을 달성해가는 과정으로 데이터의 특징과 내재하는 구조적 관계를 알아내기 위한 기법들의 통칭
- 프린스턴 대학의 듀키교수가 1977년 저서를 발표함으로써 EDA가 등장.
- 4가지 주제
 - 저항성의 강조(Resistance) : 데이터 파손에 대한 저항성
 - 잔차(Residual) 계산 : 개별 데이터가 주경향성에 얼마나 벗어났는지 확인
 - 자료 변수의 재 표현(re-expression) : 분포의 선형성, 안정성, 대칭성
 - 시각화(Graphical Representation) : 그래프

탐색적 자료 분석(Exploratory Data Analysis)



통계 분석(Statistics analysis)

- 통계 : 어떤 현상을 종합적으로 한눈에 알아보기 쉽게 일정한 체계에 따라 숫자와 표, 그림의 형태로 나타나는 것.
- 기술통계(Descriptive statistics) : 모집단으로부터 표본을 추출하고, **표본이 가지고 있는 정보를 쉽게 파악**할 수 있도록, 데이터를 정리하거나 요약하기 위해 하나의 숫자 또는 그래프의 형태로 표현하는 절차.
- 추측(토론)통계 (inferential statistics) : 모집단으로부터 추출된 **표본의 표본 통계량으로 부터 모집단의 특성인 모수**에 **관해 통계적으로 추론**하는 절차.

데이터 마이닝

- 대용량의 자료로부터 정보를 요약하고, 미래에 대한 예측을 목표로 자료에 존재하는 **관계, 패턴, 규칙 등을 탐색**하고, 이를 모형화 함으로써 이전에 알려지지 않은 **유용한 지식을 추출**하는 분석 방법.



인공지능

인간의 지적 능력을 컴퓨터를 통해 구현하는 기술



머신러닝

컴퓨터가 데이터를 통해 스스로 학습하여 예측이나 판단을 제공하는 기술

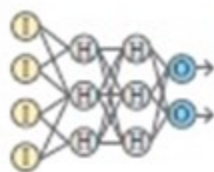
전문가
시스템

규칙기반
시스템

결정
트리

선형
회귀

퍼셉트론



딥러닝

깊은 인공신경망 알고리즘을 활용하는 머신러닝 기술

합성곱 신경망
(CNN)

심층
강화학습

순환 신경망
(RNN)

연습문제

1. 데이터 마이닝의 모델링에 대한 설명이다. 설명이 가장 잘못된 것은?

- ① 데이터마이닝 모델링은 통계적 모델링이 아니므로 지나치게 통계적 가설이나 유의성에 집착하지 말아야 한다.
- ② 모델링 방법은 여러 가지가 있으므로 모델링 시 반드시 다양한 옵션을 줘서 모델링을 수행하여 최고의 성과를 도출하여야 한다.
- ③ 분석데이터를 학습 및 테스트 데이터로 6:4, 7:3, 8:2 비율로 상황에 맞게 실시한다.
- ④ 성능에 집착하면 분석 모델링의 주목적인 실무 적용에 반하여 시간을 낭비할 수 있으므로 훈련 및 테스트 성능에 큰 편차가 없고 예상 성능을 만족하면 중단한다.



연습문제

2. 탐색적 데이터 분석의 목적은 데이터를 이해하는 것이다. 다음 중 이에 대한 설명으로 가장 부적절한 것은?

- ① 데이터에 대한 전반적인 이해를 통해 분석 가능한 데이터인지 확인하는 단계이다.
- ② 탐색적 데이터 분석 과정은 데이터에 포함된 변수의 유형이 어떻게 되는지를 찾아가는 과정이다.
- ③ 데이터를 시각화하는 것만으로는 이상점(Outlier) 식별이 잘 되지 않는다.
- ④ 알고리즘이 학습을 얼마나 잘 하느냐 하는 것은 전적으로 데이터 품질과 데이터에 담긴 정보량에 달려있다.





Thank you.

ADSP / 류영표 강사
ryp1662@gmail.com