17기 정규세션

ToBig's 16기 정수연

전처리 및 EDA

Contents

Introduction

인공지능?

머신러닝?

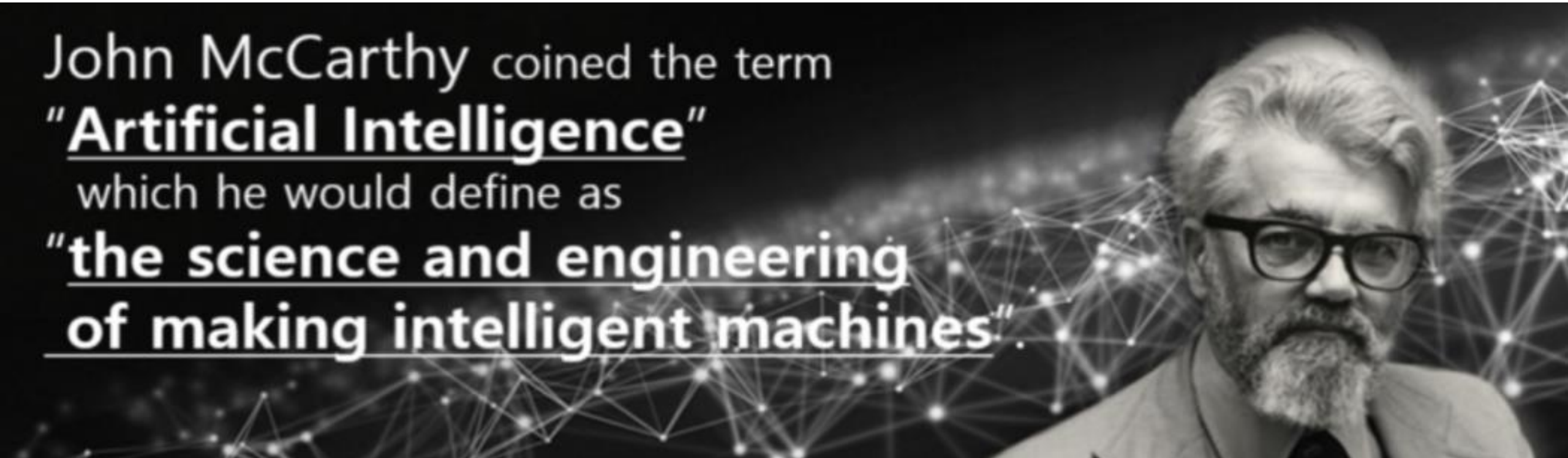
딥러닝?

쉽게 접할 수 있고
자주 사용하는 용어지만,
정확한 정의나 차이를
제대로 알지 못하는
경우가 많음

인공지능

Intelligent한 기계를 만드는 과학과 공학

기계를 인간 행동의 지식에서와 같이 행동하게 만드는 것을 의미



머신러닝

컴퓨터가 사전에 미리 프로그램 되어 있지 않고 데이터로부터 패턴을 학습하여 새로운 데이터에 대해 적절한 작업을 수행하는 일련의 알고리즘이나 처리 과정

"컴퓨터가 명시적으로 프로그램되지 않고도 학습할 수 있도록 하는 연구 분야"
Arthur Lee Samuel

별도의 프로그래밍 없이도 데이터를 통해 학습을 수행하고 시간이 지나면서 자체 정확도를 향상시키는 애플리케이션을 구축하는 데 집중하는 인공지능(AI)의 한 분야

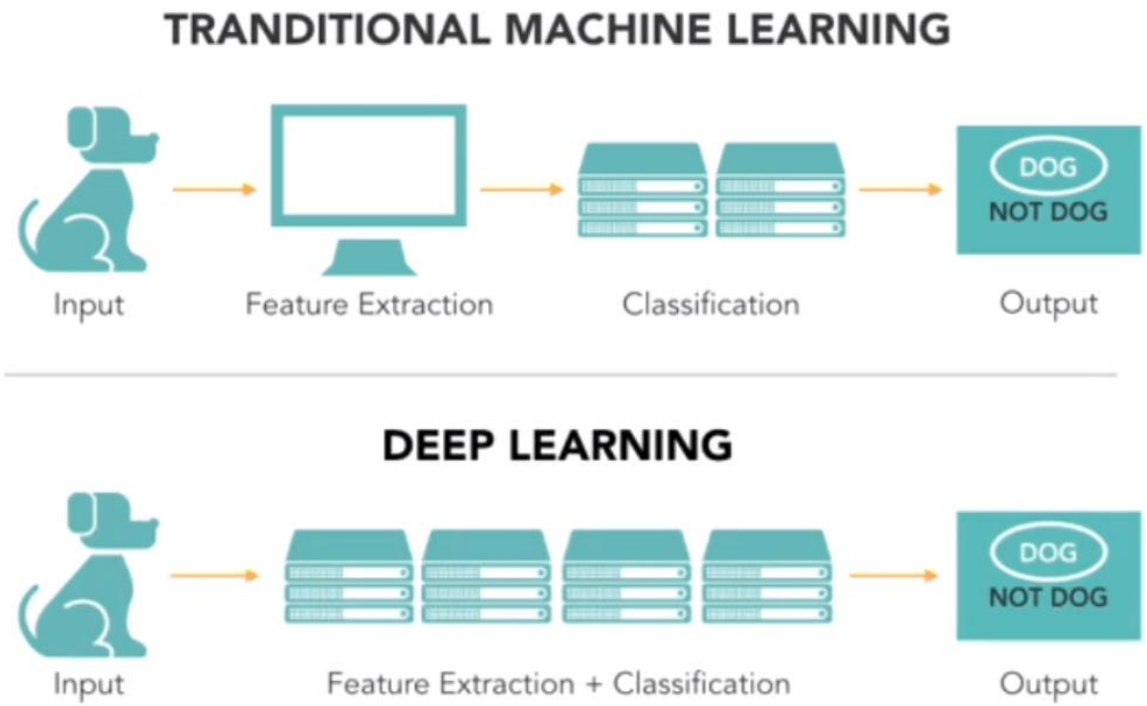
기록된 데이터에서 학습하고 이를 기반으로 예측하며, 불확실성 하에서 기본 유틸리티 기능을 최적화하고, 데이터에서 숨겨진 구조를 추출하고, 데이터를 간결한 설명으로 분류할 수 있는 알고리즘의 모음

머신러닝의 종류

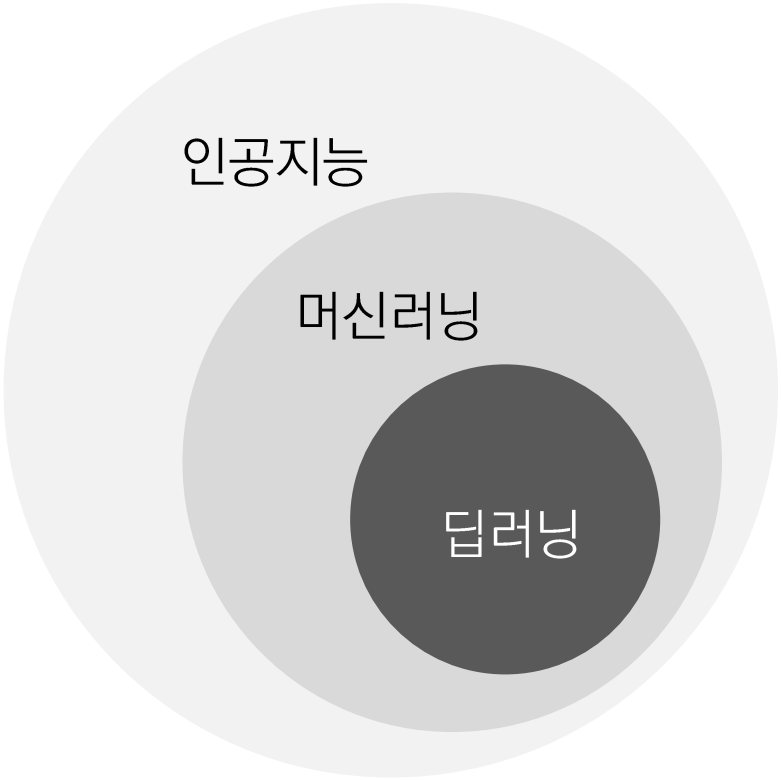

기계학습
(Machine Learning)
지도학습
(Supervised Learning)
• 레이블 된 데이터로 학습
• 미래 데이터 예측
• y = f(x)
분류
(Classification)
회귀
(Regression)
비지도학습
(Unsupervised Learning)
• 레이블 없이 학습
• 데이터의 숨겨진 구조/특징 발견
• x ~ p(x), x = f(x)
클러스터링
(Clustering)
강화학습
(Reinforcement Learning)
• 보상 시스템으로 학습
• 의사결정을 위한 최적의 액션 선택

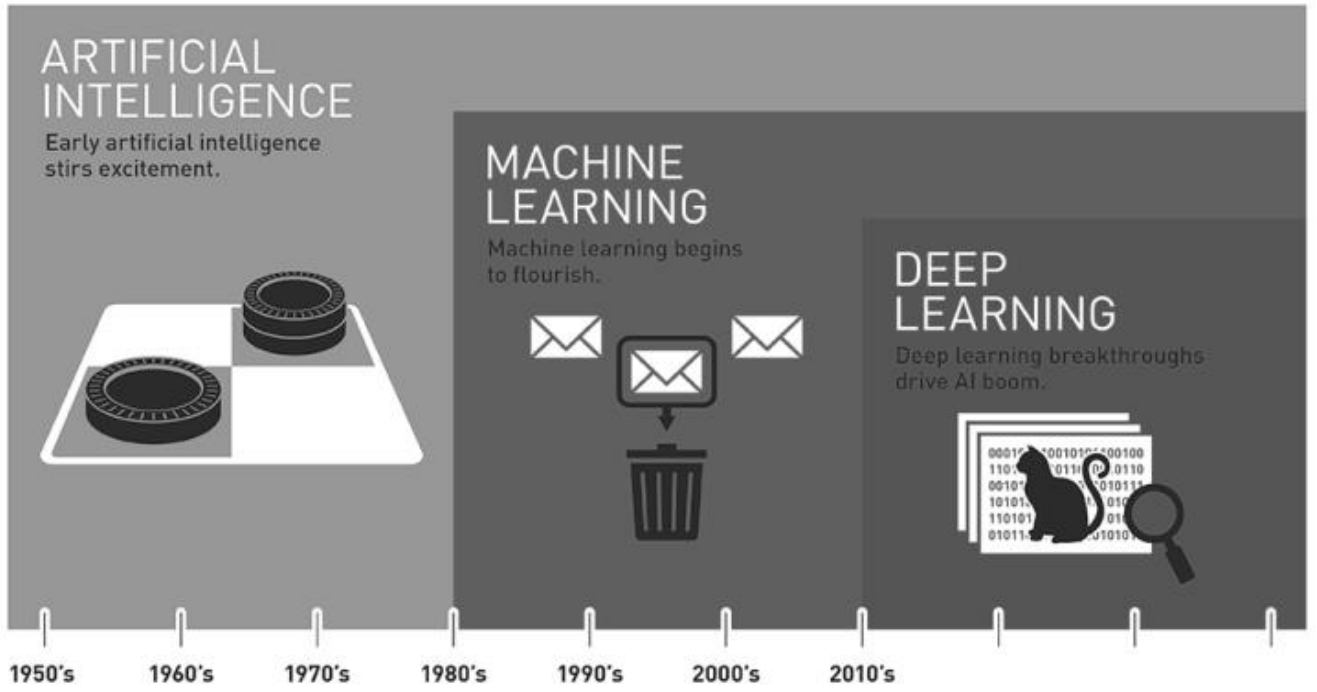
딥러닝

If machine learning is about mimicking how humans learn,
why not go all the way and try to mimic the human brain?



인공지능, 머신러닝, 딥러닝의 관계





Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

정규세션 일정

날짜	강의	담당자
1주차 (01/19)	EDA	정수연
	Git & Framework	김송민
2주차 (01/26)	Regression	이예림
	Optimization	김건우
3주차 (02/09)	SVM & SVR	김종우
	Naïve Bayes & Decision Tree	김송민
4주차 (02/16)	Dimensionality Reduction	김주호
	KNN & Clustering	박한나
5주차 (02/23)	Neural Net Basic	주지훈
	Ensemble	김건우

날짜	강의	담당자
6주차 (03/02)	Neural Net Advanced	김종우
	Time-Series	이예림
7주차 (03/09)	Vision Basic (CNN)	김윤혜
	NLP Basic (Word Embedding & LM)	전민진
8주차 (03/16)	Vision Advanced (Classification & Object Detection)	김경준
	NLP Advanced (Contextual Embedding : Transformer)	장준원
9주차 (03/23)	Recommender system Basic (Concept of Recommender system)	이승주
	Generative Basic (Vanilla GAN & VAE)	김권호
10주차 (03/30)	Recommender system Advanced (NCF & Deep FM)	이승주
	Generative Advanced (GAN Adv & WGAN-GP)	김권호

정규세션 일정

날짜	강의	담당자
[illegible]	EDA	정수영
[illegible]	Git & Framework	김송민
2주차 (01/26)	Regression	이예림
	Optimization	김건우
3주차 (02/09)	SVM & SVR	김종우
	Naïve Bayes & Decision Tree	[illegible]
4주차 (02/16)	Dimensionality Reduction	김주호
	KNN & Clustering	박한나
5주차 (02/23)	Neural Net Basic	주지훈
	Ensemble	김건우

날짜	강의	담당자
[illegible]	Neural Net Advanced	김종우
[illegible]	Time-Series	이예림
7주차 (03/09)	Vision Basic (CNN)	김윤혜
	NLP Basic (Word Embedding & LM)	전민진
8주차 [illegible]/16)	Vision Advanced (Classification & Object Detection)	김경준
	NLP Advanced (Contextual Embedding : Transformer)	장준원
9주차 (03/23)	Recommender system Basic (Concept of Recommender system)	이승주
	Generative Basic (Vanilla GAN & VAE)	김권호
10주차 (03/30)	Recommender system Advanced (NCF & Deep FM)	이승주
	Generative Advanced (GAN Adv & WGAN-GP)	김권호

앞으로 배우게 될 모든 과정에서 데이터에 대한 충분한 이해가 중요

EDA 필수!

정규세션 일정

앞으로 배우게 될 모든 과정에서 데이터에 대한 충분한 이해가 중요

날짜				담당자
1주차 (01/19)				김종우
				이예림
2주차 (01/26)				김윤해
				전민진
3주차 (02/09)				김경준
				장준원
4주차 (02/16)				이승주
	KNN & Clustering	박한나	(Vanilla GAN & VAE)	김권호
5주차 (02/23)	Neural Net Basic	주	Recommender system Advanced (NCF & Deep FM)	이승주
	Ensemble	김건우	Generative Advanced (GAN Adv & WGAN-GP)	김권호

EDA 필수!

Unit 01 | EDA란

Exploratory Data Analysis

탐색적 데이터 분석

데이터를 분석하고 결과를 내는 과정에 있어서
지속적으로 해당 **데이터에 대한 탐색과 이해**를 기본적으로 가져야 한다는 의미

Unit 01 | EDA란

Comfirmatory Data Analysis

- 목적을 가지고 데이터를 확보하여 분석하는 방법
- 관측된 형태나 효과의 재현성 평가, 유의성 검정, 신뢰구간 추정 등 통계적 추론을 하는 단계
- 가설검정, 보통은 설문조사, 논문에 대한 내용을 입증하는데 많이 사용

e.g) 사람은 죽는다(이론) → 소크라테스는 사람이다(조작화) → 따라서 소크라테스는 죽는다(관찰, 경험)

Exploratory Data Analysis

- 쌓여 있는 데이터를 기반으로 가설을 세워 데이터를 분석하는 방법
- 데이터의 구조와 특징을 파악하며 여기서 얻은 정보를 바탕으로 통계모형을 만드는 단계
- 빅데이터 분석에 사용됨

e.g) 뉴턴은 죽는다(관찰) → 칸트도 죽는다.(관찰) → 퓨리에도 죽는다.(관찰) → 모든 사람은 죽는다.(이론)

Unit 02 | 전처리 및 EDA

1) 데이터 출처와 주제에 대한 이해

- Uncover the factors that lead to employee attrition and explore important questions such as ‘show me a breakdown of distance from home by job role and attrition’ or ‘compare average monthly income by education and attrition’. This is a fictional data set created by IBM data scientists.
- IBM 데이터 과학자들이 만든 가상의 HR 데이터 set
- 1,470명에 대한 35개의 변수가 있으며, 종속변수는 Attrition, 즉 0 또는 1의 퇴사 여부

Unit 02 | 전처리 및 EDA

✓ 변수 종류 확인

- Classification이 목적일 경우
  - Target variable = Categorical 변수
  - 분류의 대상이 되는 feature를 target variable 로 설정

- Regression이 목적일 경우
  - Target variable = Numerical 변수
  - 다른 변수들에 의해 얼마나 영향을 받을 지에 대해 알고 싶은 feature를 target variable로 설정

Unit 02 | 전처리 및 EDA

2) 전체적인 데이터 살펴보기

- 분석의 목적과 해결해야할 문제에 집중하며 데이터가 전체적으로 어떤 모양인지 살펴보기
- 주로 head(), tail()을 통해 데이터를 전체적으로 살펴보기
- 컬럼명, 형태, 속성값 체크하기
- info() 통해 모든 기본 정보 체크

```
data.head()
```

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	Educatio
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sc
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sc
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sc

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 35 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       1470 non-null   int64
 1   Attrition                 1470 non-null   object
 2   BusinessTravel            1470 non-null   object
 3   DailyRate                 1470 non-null   int64
 4   Department                1470 non-null   object
 5   DistanceFromHome          1470 non-null   int64
```

Unit 02 | 전처리 및 EDA

3) 데이터의 개별 feature 값 살펴보기

범주형 자료 (Categorical Variable)	명목형 (Nominal)	– values represent discrete units – changing the order of units does not change their value.
	순서형 (Ordinal)	– values represent discrete and ordered units. – distance between units is not the same.
수치형 자료 (Quantitative)	연속형 (Continuous)	– variable whose value is obtained by measuring, i.e., one which can take on an uncountable set of values.
	이산형 (Discrette)	– can only take certain values – there is a positive minimum distance to the nearest other permissible value.

Unit 02 | 전처리 및 EDA

✓ 범주형 변수 처리

– 범주형 변수를 숫자형 벡터로 만들기 위해 다양한 인코딩 기법들 사용

['red' ,'blue',' purple']

Label Encoding	One-hot Encoding	Dummy variable Encoding
[1, 2, 3]	[1, 0, 0] [0, 1, 0] [0, 0, 1]	[0, 0] [1, 0] [0, 1]

Unit 02 | 전처리 및 EDA

✓ 결측치 처리

1. 결측치 확인

`isnull(), isna()`

: Generate a Boolean mask indicating missing values

2. 결측치 제거

- 결측치가 있는 행을 모두 제거
- 결측치가 있는 변수 자체를 제거
- 결측치가 있는 변수를 다른 값으로 대체
  - 수치형 변수 : 0, 평균, 최솟값, 중앙값 등으로 대체
  - 범주형 변수 : 별도의 범주를 생성하여 대체

Unit 02 | 전처리 및 EDA

✓ Nominal Data

- Frequencies
  - Count the number of events of interest
- Proportion(relative frequency)
  - Divide frequency by total number of events
- Percentage
  - Multiply proportion by 100
- Illustrate with bar chart or pie chart

✓ Ordinal Data

- Summarize
  - Frequencies, proportions, and percentages
  - Percentiles
  - Mode
  - Median
  - Interquartile range
- Illustrate
  - Bar chart and pie chart

Unit 02 | 전처리 및 EDA

✓ Continuous Data

- Summarize
  - Percentiles, median, interquartile range
  - Mean, median, or mode
  - Standard deviation, range or IQR
- Illustrate
  - Histogram, Boxplot

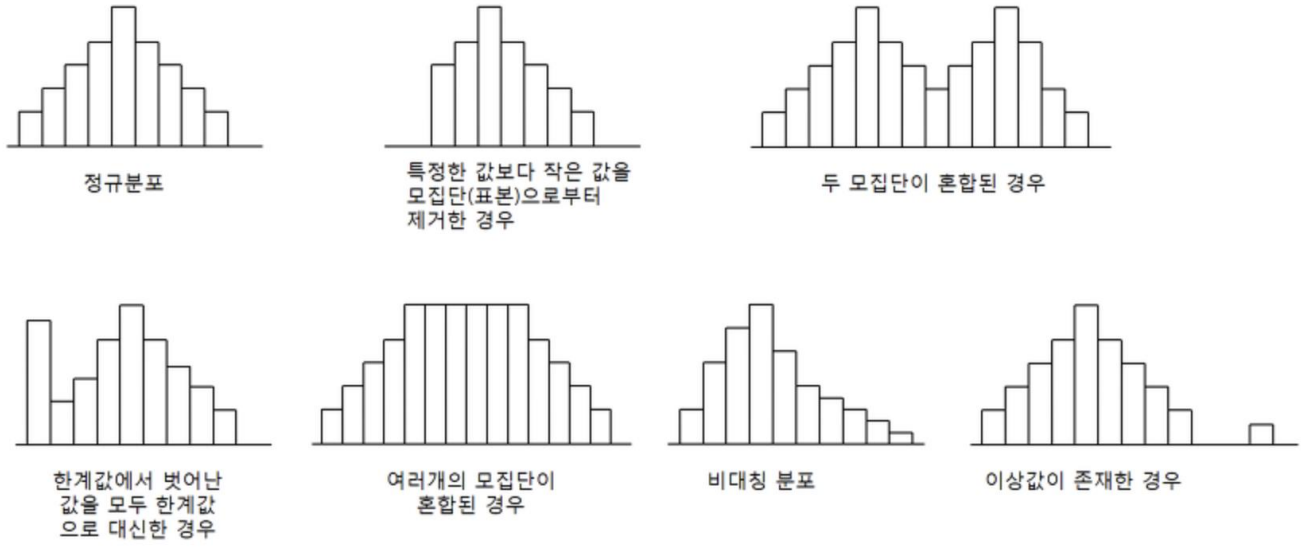
```
data.describe()
```

	Age	DistanceFromHome	JobLevel	MonthlyIncome	PercentSalaryHike	StockOptionLevel	TotalWorkingYears	TrainingTimesLastYear	YearsAtCompany	YearsInCurrentRole	YearsSinceLastPromotion	YearsWithCurrManager
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000
mean	36.923810	9.192517	2.063946	6502.931293	15.209524	0.793878	11.279592	2.799320	7.008163	4.229252	2.187755	4.123129
std	9.135373	8.106864	1.106940	4707.956783	3.659938	0.852077	7.780782	1.289271	6.126525	3.623137	3.222430	3.568136
min	18.000000	1.000000	1.000000	1009.000000	11.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	30.000000	2.000000	1.000000	2911.000000	12.000000	0.000000	6.000000	2.000000	3.000000	2.000000	0.000000	2.000000
50%	36.000000	7.000000	2.000000	4919.000000	14.000000	1.000000	10.000000	3.000000	5.000000	3.000000	1.000000	3.000000
75%	43.000000	14.000000	3.000000	8379.000000	18.000000	1.000000	15.000000	3.000000	9.000000	7.000000	3.000000	7.000000
max	60.000000	29.000000	5.000000	19999.000000	25.000000	3.000000	40.000000	6.000000	40.000000	18.000000	15.000000	17.000000

Unit 02 | 전처리 및 EDA

✓ Histogram

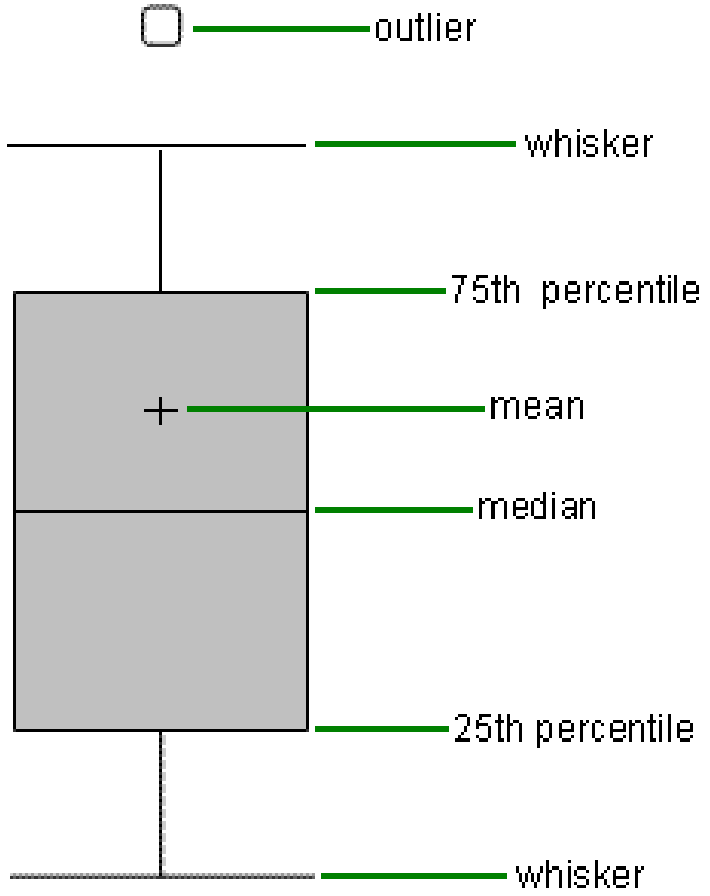
- 어떠한 변수에 대해서 구간별 빈도수를 나타낸 그래프
- 가로축은 구간을 , 세로축은 빈도를 나타내며 도수분포를 그림으로 나타냄
- 막대그래프와 달리 막대 사이의 간격이 없으며 평균과 분포를 아래처럼 비교적 시각적으로 뚜렷하게 볼 수 있음



Unit 02 | 전처리 및 EDA

✓ Boxplot

- 자료로부터 얻어낸 통계량인 5가지 요약 수치(최소값, Q1, Q2, Q3, 최댓값) 이용
- 데이터 집합의 범위와 중앙값을 빠르게 확인 가능
- 통계적으로 이상치(outlier)가 있는지 확인 가능



4) Feature 간 관계 살펴보기

- Categorical – Categorical
  - 교차테이블, 모자이크 플롯을 이용해 각 속성값의 쌍에 해당하는 값 개수를 표시

- Numeric – Categorical
  - 각 카테고리별 통계값(평균, 중간 값 등)을 관찰
  - Side-by-side box plot

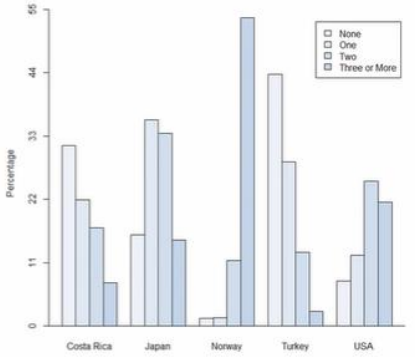
- Numeric – Numeric
  - 상관계수를 통해 두 속성 간의 연관성 확인
  - 산포도를 통해 경향 파악

Unit 02 | 전처리 및 EDA

Two Categorical Variables

Table 5
Number of computers in home by country

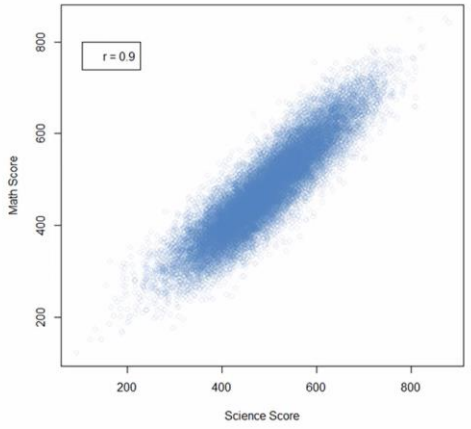
Country	None	One	Two	Three+
Costa Rica	31.36	21.94	17.10	7.50
Japan	15.80	35.79	33.49	14.89
Norway	1.31	1.49	11.37	53.52
Turkey	43.67	28.49	12.84	2.59
USA	7.86	12.29	25.20	21.50



Two Continuous Variables

Table 6
PISA test score

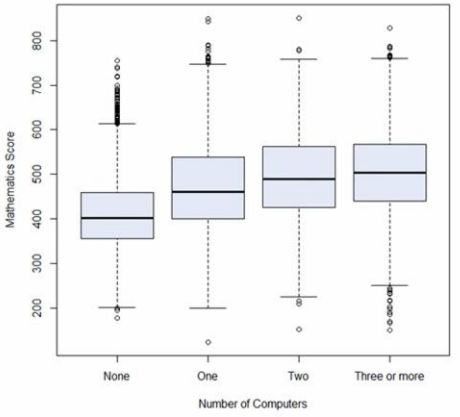
Score	Mean	S.D.
Science	491.47	97.03
Math	477.46	97.87



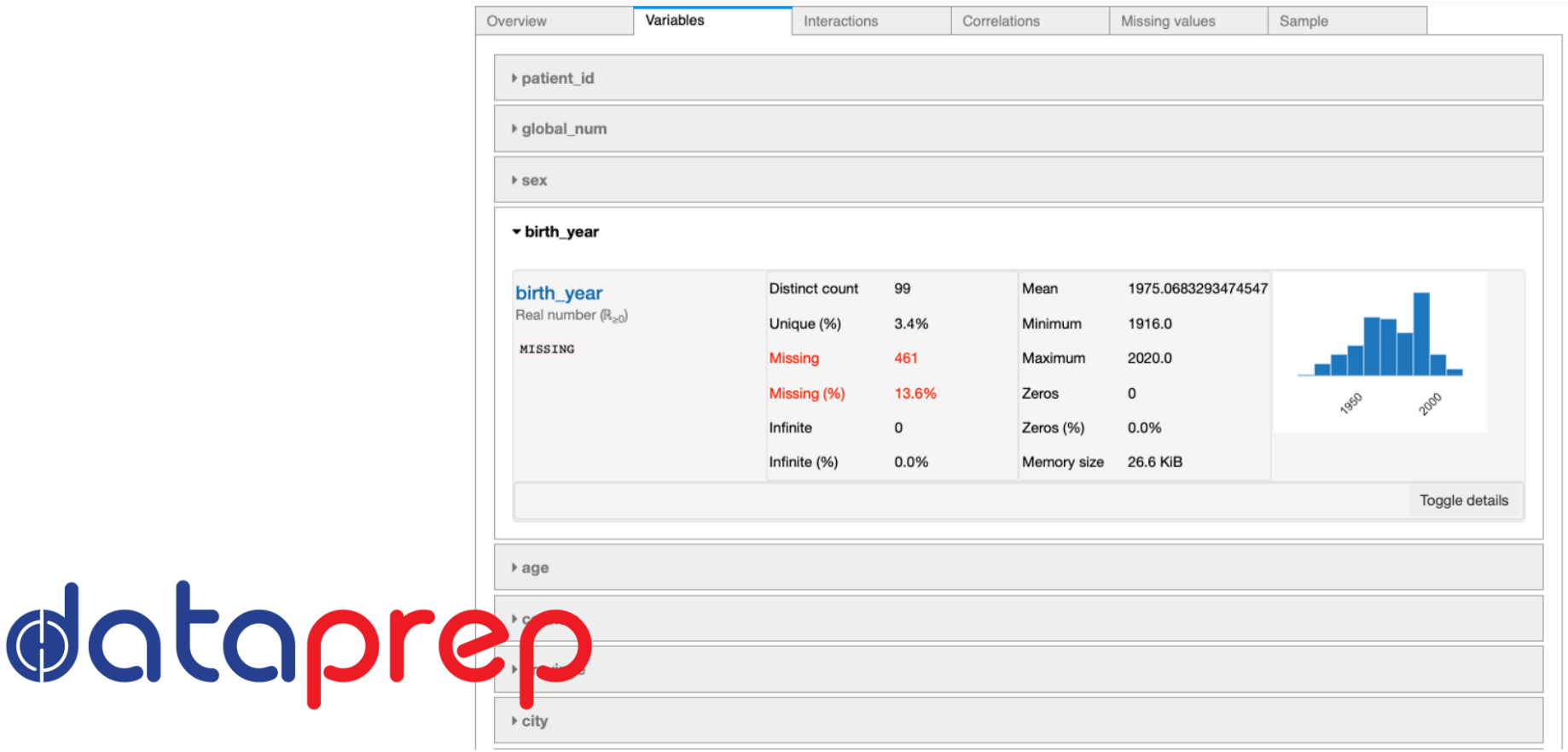
One Continuous and One Categorical

Table 7
Mathematics score by number of computers

Group	Mean	S.D.
None	413.15	83.47
One	471.46	96.87
Two	494.93	94.91
Three+	504.12	91.84

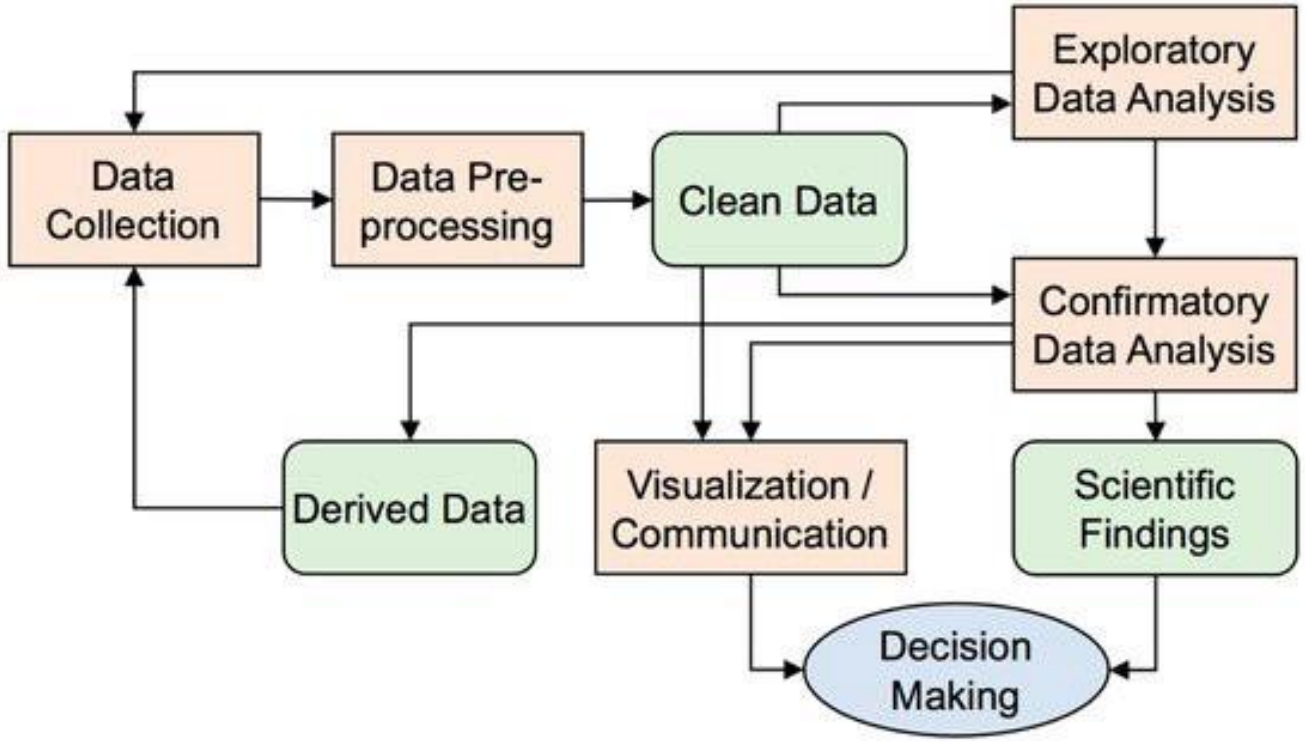


+) dataprep.eda, pandas profiling



Unit 02 | 전처리 및 EDA

EDA 그 이후



- EDA단계에서 얻은 이해를 바탕으로 사용하고자 하는 분석 방법에 최적화된 방향으로 데이터 전처리를 진행
- EDA단계에서 얻은 추후 분석에서 조심해야 할 부분, 주의해야할 점, 변환해야 할 변수에 대한 정보 기록

Unit 02 | 전처리 및 EDA 방법

데이터에 대해 충분한 시간과 많은 노력을 쏟는다면,

분석 프로젝트에서 좋은 결과를 얻을 수 있다.

실습을 통해 데이터를 하나하나 뜯어보자 !

Unit 03 | 실습 및 과제

✓ 과제

- 파이썬을 이용하여 전처리 및 EDA를 진행해주세요
  - 결측치 처리
  - 유의미한 시각화 7개 이상
  - 수치형 변수 간 상관관계 파악
  - 파생변수 생성

Unit 00 | 참고자료

투빅스 11기 유기윤님 1주차 EDA 강의자료

투빅스 13기 김현선님 1주차 EDA 강의자료

투빅스 14기 이혜린님 1주차 EDA 강의자료

투빅스 15기 이수민님 1주차 EDA 강의자료

들어주셔서 감사합니다.